Parte I
Comprensión de los datos

El primer paso para comprender los datos es escuchar lo que dicen, "dejar que los números hablen por sí mismos". Y éstos hablan de forma clara solamente cuando les ayudamos a hablar organizando, representando y haciendo preguntas. Esto es el *análisis de datos*. El grado de confianza en lo que dicen los datos depende de su origen. Por tanto, también estamos interesados en *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 de este libro 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 poder confiar en los resultados de 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 50.000 hogares, aunque el objetivo sea el de sacar conclusiones referidas a la totalidad de los 100 millones de hogares de aquel país. Este es un problema complejo. 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. Las condiciones que exige la inferencia estadística no nos preocupan en los capítulos 1 y 2. Lo que nos preocupa es el examen sistemático de los datos y las herramientas que utilizamos para tal fin.

Por supuesto que a menudo queremos utilizar los datos para alcanzar conclusiones más generales; que esto sea posible depende sobre todo de cómo se obtuvieron. Los buenos datos raramente "caen del cielo"; son producto del esfuerzo humano, al igual que los videojuegos o las medias de nailon. El capítulo 3

nos muestra cómo obtener buenos datos y cómo decidir si podemos confiar en los obtenidos por los demás.

El estudio del análisis y obtención de datos te proporciona ideas y herramientas que te serán de gran utilidad cuando tengas que vértelas con números. La inferencia exige que un 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 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.

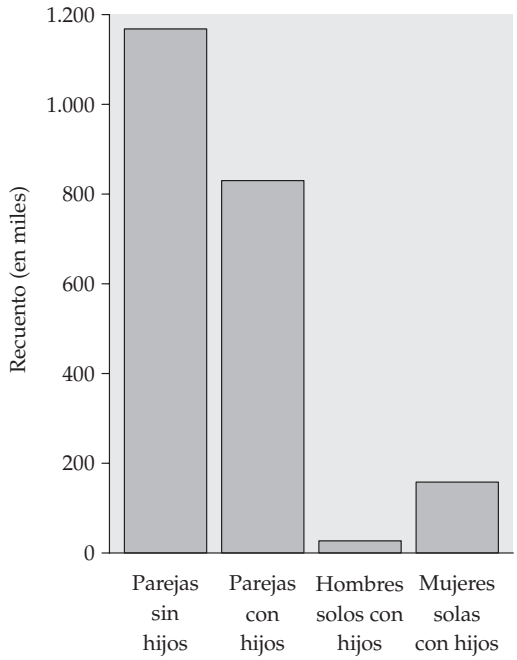


Figura 1.1(a). Diagrama de barras del número de familias por tipos en Suecia.

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. ■

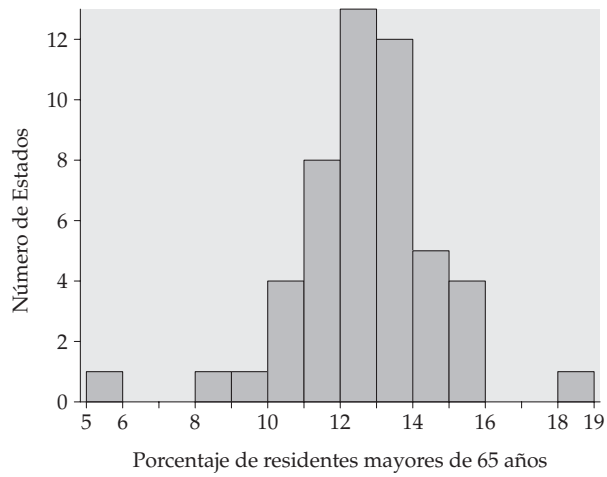


Figura 1.2. Histograma del porcentaje de residentes mayores de 65 años en los 50 Estados de EE UU. Datos de la tabla 1.1.

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. ■

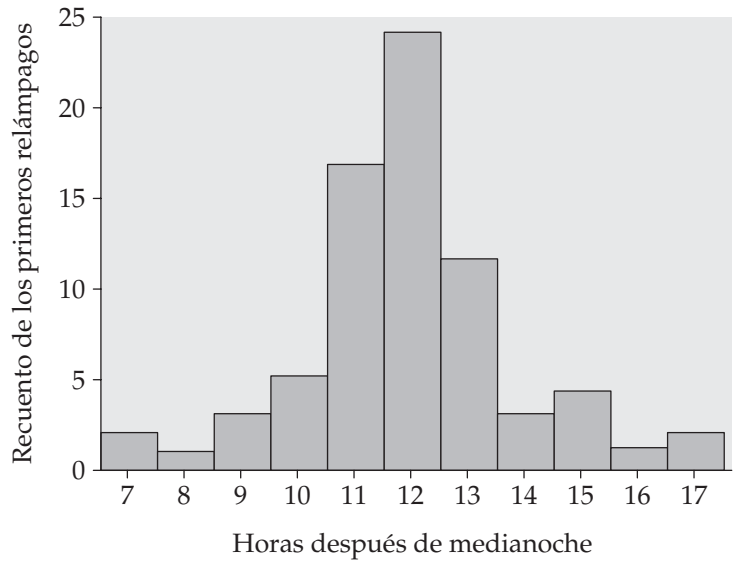


Figura 1.3. Distribución de la hora en la que se produce el primer relámpago del día en una localidad de Colorado, EE UU.

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

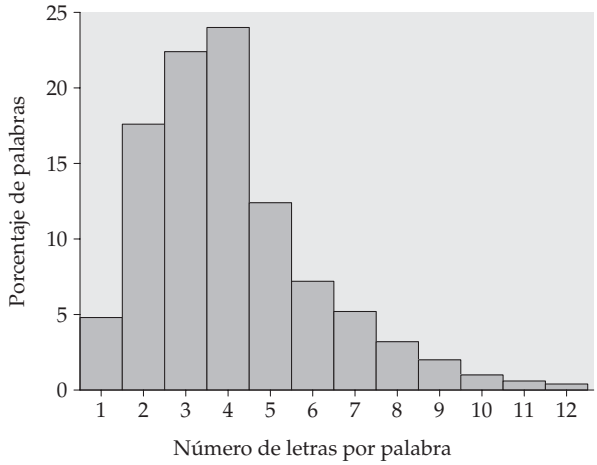


Figura 1.4. Distribución de la longitud de las palabras utilizadas en las obras de Shakespeare.

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.

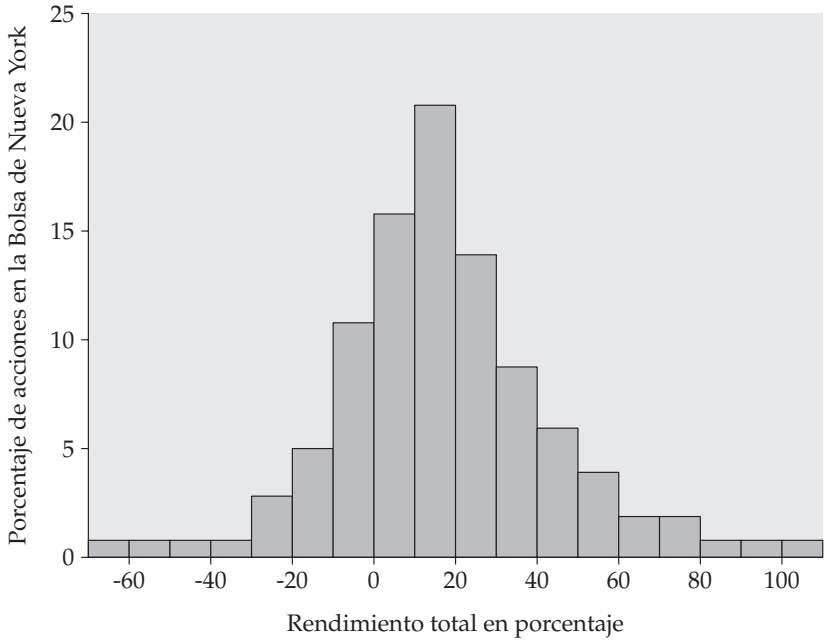


Figura 1.5. Distribución de rendimientos totales de todas las acciones de la Bolsa de Nueva York durante un año. Para el ejercicio 1.8.

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

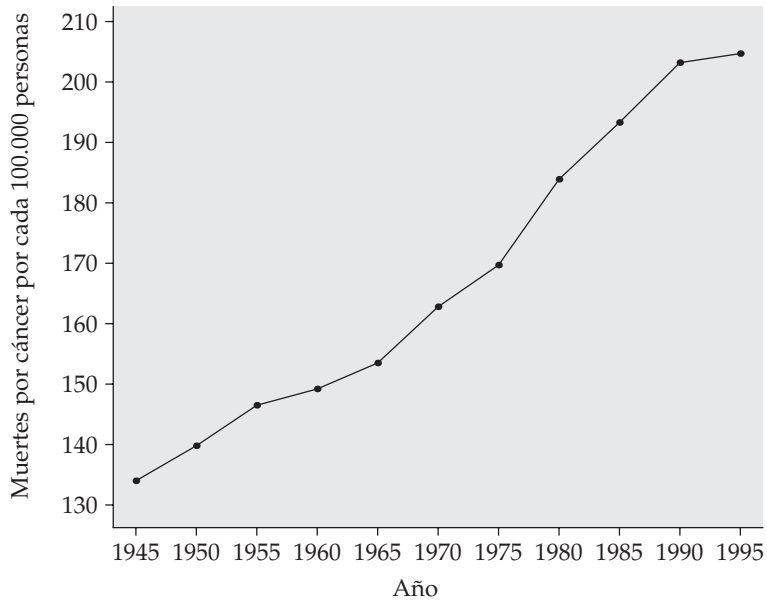


Figura 1.7. Gráfico temporal correspondiente a las tasas de mortalidad por cáncer en EE UU (número de muertes por cada 100.000 personas), desde 1945 hasta 1995.

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.

Longitud	1	2	3	4	5	6	7	8
Porcentaje	3,6	14,8	18,7	16,0	12,5	8,2	8,1	5,9
Longitud	9	10	11	12	13	14	15	
Porcentaje	4,4	3,6	2,1	0,9	0,6	0,4	0,2	

(a) Dibuja un histograma correspondiente a esta distribución. Describe la forma, el centro y la dispersión.

(b) ¿Cómo podemos comparar la distribución de la longitud de las palabras utilizadas en *Popular Science* con la distribución de la longitud de las palabras en las obras de Shakespeare de la figura 1.4? Fíjate especialmente en las palabras cortas (2, 3 y 4 letras) y en las palabras muy largas (más de 10 letras).

1.14. Huracanes. El histograma de la figura 1.8 muestra el número de huracanes que alcanzaron la costa este de EE UU durante un periodo de 70 años.[8] Describe de manera breve la forma de esta distribución. ¿Dónde queda aproximadamente su centro?

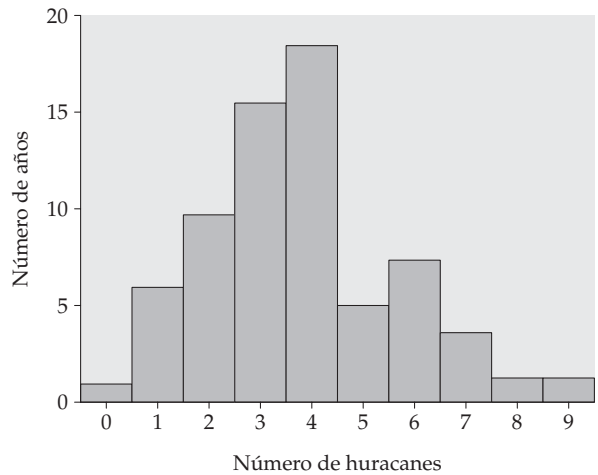


Figura 1.8. Distribución del número de huracanes en la costa este de EE UU durante un periodo de 70 años. Para el ejercicio 1.14.

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

1.15. Número de goles. La figura 1.9 muestra la distribución del número de goles de los jugadores de primera división de la liga española de fútbol que al menos marcaron 5 goles durante la temporada 1999/2000.

(a) La distribución, ¿es aproximadamente simétrica, claramente asimétrica o ninguna de la dos cosas?

(b) ¿Cuál es el número de goles típico de un jugador de la liga española de fútbol de la temporada 1999/2000? ¿Cuáles son el máximo y el mínimo de goles marcados?

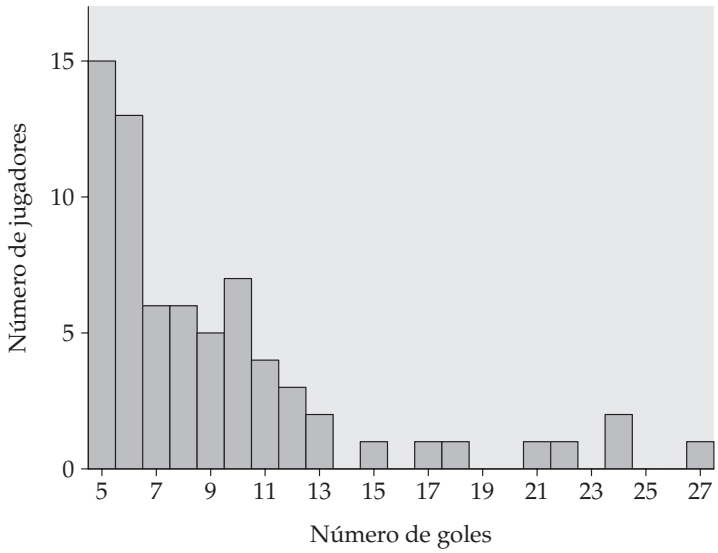


Figura 1.9. Distribución del número de goles de los jugadores de primera división de la liga española de fútbol durante la temporada 1999/2000.

1.16. Supón que tus amigos y tú mismo vaciáis vuestros monederos y vais apuntando la fecha que aparece en cada moneda que sacáis. La distribución de estos datos es asimétrica hacia la izquierda. Explica por qué.

1.17. Pirámide de edad en EE UU. La pirámide de edad de un país tiene una gran influencia sobre sus condiciones sociales y económicas. La tabla 1.3 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.

Tabla 1.3. Pirámides de edad de 1950 y 2075 en los EE UU (en millones de personas).

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

(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 sobre los porcentajes de 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?

1.18. Goles marcados por Paulino Alcántara. He aquí el número de goles que marcó Paulino Alcántara mientras fue jugador del F.C. Barcelona, desde la temporada 1911/12 hasta la temporada 1926/27:

6 15 21 25 33 0 5 42 47 19 42 34 39 6 15 8

Dibuja un diagrama de tallos con estos datos. La distribución, ¿es aproximadamente simétrica, claramente asimétrica o nada de esto?
En un año típico, ¿cuántos goles marcó aproximadamente Paulino Alcántara? ¿Existe alguna observación atípica?

1.19. Goles marcados por Ladislao Kubala. Ladislao Kubala ha sido uno de los mejores jugadores de fútbol de todos los tiempos. He aquí el número de goles que

marcó por temporada mientras fue jugador del F.C. Barcelona desde la temporada 1950/51 hasta la temporada 1960/61:

16 48 18 28 19 22 14 19 17 25 17

Diagrama de tallos doble

Un **diagrama de tallos doble** nos ayuda a comparar dos distribuciones. Sitúa los tallos como es habitual. Sin embargo, traza dos líneas verticales una a cada lado de los tallos. A la derecha, sitúa las hojas correspondientes a los goles de Paulino Alcántara (ejercicio 1.18). A la izquierda, sitúa las hojas correspondientes a los goles de Ladislao Kubala. Sitúa las hojas de cada tallo en orden creciente a partir del tallo. Describe brevemente las diferencias existentes entre ambas distribuciones.

1.20. Mercado en baja. Los inversores hablan de un "mercado en baja" cuando el valor de las acciones cae sustancialmente. La tabla 1.4 proporciona datos de todas las caídas de al menos un 10% del índice Standard & Poor's entre 1940 y 1977. Los datos muestran la bajada de los índices respecto al valor máximo y los meses que se mantuvo dicha bajada.

Tabla 1.4. Valor de las caídas y duración del índice Standard & Poor's.

Año	Descenso (porcentaje)	Duración (meses)	Año	Descenso (porcentaje)	Duración (meses)
1940-1942	42	28	1966	22	8
1946	27	5	1968-1970	36	18
1950	14	1	1973-1974	48	21
1953	15	8	1981-1982	26	19
1955	10	1	1983-1984	14	10
1956-1957	22	15	1987	34	3
1959-1960	14	15	1990	20	3
1962	26	6			

(a) Dibuja un diagrama de tallos con los porcentajes de las bajadas del valor de las acciones durante estos años. Vuelve a dibujar el diagrama de tallos, pero dividiendo los tallos. ¿Qué diagrama prefieres? ¿Por qué?

(b) La forma de esta distribución es irregular, de todas formas la podemos describir como algo asimétrica. La distribución, ¿es asimétrica hacia la derecha o hacia la izquierda?

(c) Describe el centro y la dispersión de los datos. ¿Qué le dirías a un inversor sobre la disminución del valor de las acciones en años con el mercado en baja?

1.21. Maratón de Boston. A partir de 1972, se permitió la participación de mujeres en la maratón de Boston. En la tabla 1.5, se muestran los tiempos (en minutos) de las mujeres que ganaron desde 1972 hasta 1998.

(a) Dibuja un diagrama temporal con los tiempos de las ganadoras.

(b) Describe de forma breve la distribución de los tiempos de las ganadoras de la maratón a lo largo de estos años. En los últimos años, ¿los tiempos han dejado de bajar?

Tabla 1.5. Tiempos de las ganadoras de la maratón de Boston.

Año	Tiempo	Año	Tiempo	Año	Tiempo
1972	190	1981	147	1990	145
1973	186	1982	150	1991	144
1974	167	1983	143	1992	144
1975	162	1984	149	1993	145
1976	167	1985	154	1994	142
1977	168	1986	145	1995	145
1978	165	1987	146	1996	147
1979	155	1988	145	1997	146
1980	154	1989	144	1998	143

1.22. La epidemia de gripe de 1918. Entre 1918 y 1919 una epidemia de gripe mató a más de 25 millones de personas en todo el planeta. He aquí datos sobre el número de nuevos casos de gripe y la cantidad de muertos en San Francisco, semana a semana, desde el 5 de octubre de 1918 hasta el 25 de enero de 1919. La fecha corresponde al último día de la semana.[9]

Fecha	Nuevos casos	Muertos	Fecha	Nuevos casos	Muertos
5-oct	36	0	7-dic	722	50
12-oct	531	0	14-dic	1.517	71
19-oct	4.233	130	21-dic	1.828	137
26-oct	8.682	552	28-dic	1.539	178
2-nov	7.164	738	4-ene	2.416	194
9-nov	2.229	414	11-ene	3.148	290
16-nov	600	198	18-ene	3.465	310
23-nov	164	90	25-ene	1.440	149
30-nov	57	56			

[9] A. W. Crosby, *America's Forgotten Pandemic: The Influenza of 1918*, Cambridge University Press, Nueva York, 1989.

(a) Dibuja un diagrama temporal con los datos de nuevos casos de gripe. Basándote en tu diagrama, describe la progresión de la enfermedad.

(b) Nos gustaría comparar la distribución del número de nuevos casos con la distribución del número de muertos. Para conseguir que las magnitudes de las dos variables sean similares y facilitar la comparación, representa el número de muertos con relación al tiempo desde el 5 de octubre hasta el 25 de enero. Luego, utilizando otro color, representa en el mismo gráfico el número de nuevos casos dividido por 10. ¿Qué ves? Concretamente, ¿cuál es el desfase entre el cambio en el número de nuevos casos y el cambio en el número de muertos?

1.23. ¡Fíjate en las escalas! La impresión que proporciona un gráfico temporal depende de las escalas que utilices en los dos ejes. Si estiras el eje de las ordenadas y comprimes el eje de las abscisas, los cambios aparecen como más rápidos. En cambio, si comprimes el eje de las ordenadas y estiras el eje de las abscisas los cambios aparecen como más lentos. Dibuja dos diagramas temporales más con los datos del ejemplo 1.6, de manera que en un gráfico dé la impresión de que las muertes por cáncer aumentan muy rápidamente y en el otro, en cambio, que dicho incremento parezca muy suave. La moraleja de este ejercicio es: "Fíjate en las escalas cuando mires un diagrama temporal".

1.24. Población de los Estados. La tabla 1.6 presenta algunos datos sobre Estados europeos. La primera columna identifica los Estados. La segunda indica la región socio-política a la que pertenece cada uno de ellos: 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 en kilómetros cuadrados. 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 (la media del número de ejemplares vendidos cada día) y el número de aparatos de televisión por cada 1.000 habitantes. 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*.

Dibuja un diagrama de tallos sobre la población de los Estados. Describe brevemente la forma, el centro y la dispersión de la distribución poblacional. Explica por qué la forma de la distribución no es sorprendente. ¿Hay algún Estado que consideres que es una observación atípica?

Tabla 1.6. Datos sobre los Estados europeos.

Estado	Región	Población (1.000 hab.) 1.993	Superficie (km^2)	PIB *per cápita* 1994 (dólares)	Periódicos*	Televisiones*	% 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	135	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	UE	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
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
Yugoslavia	EE	10.623	87.968	1.000	52	179	–

* Por 1.000 hab.

EE = Estado del Este.

UE = Unión Europea.

OT = Otros Estados.

1.25. Distribución del PIB *per cápita*. Dibuja un diagrama de tallos con la distribución del PIB *per cápita* de los Estados europeos (tabla 1.6). Describe brevemente la forma de la distribución. Halla el punto medio de los datos y señala su valor en el diagrama de tallos.

1.26. Televisores por cada 1.000 habitantes. Representa gráficamente la distribución del número de televisores por cada 1.000 habitantes en los Estados europeos (tabla 1.6). ¿Cuál es la forma de la distribución? ¿Existe alguna observación atípica o desviación notable?

1.3 Descripción de las distribuciones con números

Ladislao Kubala posiblemente sea el mejor jugador que nunca haya tenido el F.C. Barcelona. He aquí el número de goles por temporada que marcó este jugador mientras estuvo en el F.C. Barcelona:

Temporada	Goles	Temporada	Goles
1950/51	16	1956/57	14
1951/52	48	1957/58	19
1952/53	18	1958/59	17
1953/54	28	1959/60	25
1954/55	19	1960/61	17
1955/56	22		

El diagrama de tallos de la figura 1.10 muestra la forma, el centro y la dispersión de estos datos. La distribución es asimétrica hacia la derecha. El centro es aproximadamente 18 y la dispersión va de 14 hasta 48. La forma, el centro y la dispersión proporcionan una buena descripción de la distribución general de cualquier distribución de una variable numérica. A continuación, veremos cómo caracterizar el centro y la dispersión de cualquier distribución.

```
1 | 4 6 7 7 8 9 9
2 | 2 5 8
3 |
4 | 8
```

Figura 1.10. Distribución del número de goles marcados por temporada por Ladislao Kubala mientras fue jugador del F.C. Barcelona.

1.3.1 Una medida de centro: la media

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

LA MEDIA, $\bar{x}$

Para hallar la **media** de un conjunto de observaciones, suma sus valores y divide por el número de observaciones. Si las n observaciones son x_1, x_2, ..., x_n, su media es

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

o de forma más compacta

$$\bar{x} = \frac{1}{n} \sum x_i$$

La $\sum$ (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 distinguir las n observaciones. Los subíndices 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 1.7. Goles marcados por Ladislao Kubala

El número medio de goles que marcó Ladislao Kubala mientras fue jugador del F.C. Barcelona desde la temporada 1950/51 hasta la temporada 1960/61 es

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{16 + 48 + \cdots + 17}{11} \\ &= \frac{195}{11} = 17{,}72 \end{aligned}$$

En la práctica no es necesario que sumes y dividas; puedes entrar los datos en una calculadora y utilizar la función que calcula la media aritmética. No obstante, tienes que saber qué es lo que hace la calculadora. La temporada 1951/52, el número de goles que marcó Kubala fue extraordinario: 48. ¿Cambia mucho la media si excluimos este valor? La distribución, ¿es simétrica? ¿Cuál es la posición de la media? Si eliminamos la temporada 1951/52, la media del número de goles marcados por Kubala es 17,72. ■

El ejemplo 1.7 ilustra un hecho importante sobre la media como medida de centro: 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. Debido a que la media no puede resistir la influencia de observaciones extremas, decimos que la media no es una **medida robusta** de centro.

Medida robusta

APLICA TUS CONOCIMIENTOS

1.27. Actitud de los estudiantes. 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:

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 una calculadora. Comprueba que obtienes el mismo resultado.

(b) El diagrama de tallos del ejercicio 1.9 sugiere que una puntuación de 200 es una observación atípica. Utiliza tu 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

En la sección 1.2, utilizamos el valor central de una distribución como una medida informal de centro. La *mediana* es la formalización de este valor central, mediante una regla específica para su cálculo.

LA MEDIANA *M*

La **mediana *M*** es el punto medio de una distribución, es decir, es el número tal que la mitad de las observaciones son menores y la otra mitad, mayores. 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, entonces la mediana M es la observación central de la lista ordenada. Halla la posición de la mediana contando $\frac{(n+1)}{2}$ observaciones desde el comienzo de la lista.

3. Si el número de observaciones n es par, 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 $\frac{(n+1)}{2}$ observaciones desde el comienzo de la lista.

Fíjate en que la fórmula $\frac{(n+1)}{2}$ no da la mediana, simplemente sitúa la mediana en la lista ordenada. Para el cálculo de la mediana son necesarios pocos pasos, por tanto, es fácil hallarla a mano cuando se tiene un conjunto de datos pequeño. Sin embargo, ordenar incluso un número de datos no demasiado elevado es pesado. En consecuencia, hallar la mediana a mano es molesto. Incluso las calculadoras más sencillas tienen la tecla $\bar{x}$. Sin embargo, para hallar la mediana es necesario utilizar un ordenador o una calculadora con funciones estadísticas avanzadas.

EJEMPLO 1.8. Cálculo de la mediana

Halla la mediana del número de goles de Paulino Alcántara mientras fue jugador del F.C. Barcelona. En primer lugar, ordena los datos en forma creciente:

0 5 6 6 8 15 15 **19** **21** 25 33 34 39 42 42 47

En total tenemos $n = 16$ observaciones, un número par. No hay una única observación central, sino dos. Son los dos valores que se han marcado en negrita.

Estas observaciones tienen 7 observaciones a la izquierda y 7 a la derecha. La mediana se halla a medio camino de estas observaciones. Para localizarla podemos utilizar la fórmula:

$$\text{Localización de } M = \frac{n+1}{2} = \frac{17}{2} = 8{,}5$$

El valor 8,5 indica "a medio camino entre la octava y la novena observación". Este resultado concuerda con lo que hemos visto a simple vista.

¿Cuál es el valor de la mediana? Podemos calcular su valor de la siguiente manera:

$$M = \frac{19+21}{2} = 20$$

Podemos comparar Paulino Alcántara con Ladislao Kubala. Éstos son los goles, ordenados, que marcó Kubala mientras fue jugador del F.C. Barcelona

14 16 17 17 18 **19** 19 22 25 28 48

Tenemos un número impar de observaciones, por tanto, existe una observación central. Esta observación es la mediana. Es el valor 19 que se ha marcado en negrita. Tiene 5 observaciones a la izquierda y 5 observaciones a la derecha. Aunque Kubala ha sido el mejor jugador de todos los tiempos del F.C. Barcelona, la mediana de Alcántara es ligeramente mayor.

Debido a que $n = 11$, nuestra fórmula para localizar la mediana da

$$\text{Localización de } M = \frac{n+1}{2} = \frac{12}{2} = 6$$

Es decir, la mediana es la sexta observación. Es más rápido utilizar esta fórmula que localizar a ojo la mediana. ■

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 una observación atípica tira de la media. La media de los goles de Kubala con todas las observaciones es 22,09; si eliminamos la observación atípica, la media es 17,72. Sin embargo, el valor de la mediana cambia mucho menos, va de 19 a 18,5. A diferencia de la media, la mediana es *robusta*. Si la temporada 1951/52 Kubala hubiera marcado 480 goles, el valor de la

mediana no cambiaría en absoluto. El valor 480 tan sólo es un valor más entre los valores mayores que el valor central, no importa si se halla muy alejado o poco. En cambio, en el cálculo de la media se tienen en cuenta los valores de todas las observaciones, por tanto, un valor muy grande hace que aumente su valor.

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 muy 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 viviendas vendidas en España en 1993 fue de 139.400 €. En cambio, el precio mediano fue de 117.000 €. 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 central") en lugar de la media ("el valor promedio"). De todas formas, si fueras un inspector de Hacienda interesado en el valor total de tu zona, tendrás que utilizar la media. El valor total es la media multiplicada por el número total de casas. El valor total no está relacionado con la mediana. Aunque la media y la mediana miden el centro de maneras diferentes; las dos son útiles.

APLICA TUS CONOCIMIENTOS

1.28. Médicos suizos. Un estudio en Suiza examinó el número de cesáreas llevadas a cabo por 15 médicos (hombres) durante un año. Sus resultados fueron

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

(a) Dibuja un histograma con estos datos. Fíjate en que existen dos observaciones atípicas.

(b) Halla la media y la mediana del número de cesáreas. ¿Cómo se puede explicar, a partir de las observaciones atípicas, la diferencia entre ambas?

(c) Halla la media y la mediana del número de cesáreas sin las dos observaciones atípicas. Los resultados en (b) y en (c), ¿ilustran la robustez de la mediana y la falta de robustez de la media?

1.29. Los más ricos. En EE UU la distribución de los ingresos individuales es muy asimétrica hacia la derecha. En 1997 la media y la mediana de los ingresos del 1% de los estadounidenses más ricos era de 330.000 y 675.000 dólares,

respectivamente. ¿Cuál de estos valores corresponde a la media y cuál a la mediana? Justifica tu respuesta.

1.30. En el ejercicio 1.27 hallaste 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.

1.3.4 Una medida de dispersión: los cuartiles

La media y la mediana proporcionan dos medidas distintas del centro de una distribución. Sin embargo, caracterizar una distribución sólo con una medida de su centro puede ser engañoso. Dos provincias con la misma mediana de ingresos 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, 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.

Una manera de medir la dispersión es dar las observaciones máxima y mínima. Por ejemplo, el número de goles marcados por temporada por Paulino Alcántara va de 0, una temporada que Paulino estuvo mucho tiempo en Filipinas, hasta 47. Estas dos observaciones nos dan la dispersión total de los datos. Sin embargo, la presencia de alguna observación atípica nos puede enmascarar esta medida de dispersión. Podríamos mejorar nuestra descripción de la dispersión fijándonos también en la dispersión del 50% de los valores centrales de los datos. Los *cuartiles* determinan entre qué valores se encuentra la mitad central de las observaciones. Ordenemos las observaciones de menor a mayor. El *primer cuartil* separa el primer 25% de las observaciones. El *tercer cuartil* separa el primer 75% de observaciones. En otras palabras, el primer cuartil es mayor que el 25% de las observaciones. El tercer cuartil es mayor que el 75% de las observaciones. El segundo cuartil es la mediana. El segundo cuartil 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.

LOS 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 global.
3. El **tercer cuartil Q_3** es la mediana de las observaciones situadas a la derecha de la mediana global.

Veamos un ejemplo de cómo utilizar las reglas anteriores para hallar los cuartiles cuando se tiene un número par y un número impar de observaciones.

EJEMPLO 1.9. Halla los cuartiles

El número de goles que marcó por temporada Paulino Alcántara, ordenados, es

0	5	6	6	8	15	15	19	21	25	33	34	39	42	42	47
			↑				↑				↑				
			Q_1				M				Q_3				

Tenemos un número par de observaciones, por tanto, la mediana se halla a medio camino de las dos observaciones centrales: la octava y la novena. El primer cuartil es la mediana de las primeras 8 observaciones, ya que éstas se hallan a la izquierda de la mediana. Comprueba que los cuartiles son $Q_1 = 7$ y $Q_3 = 36{,}5$. Cuando el número de observaciones es par, todas ellas intervienen en el cálculo de los cuartiles.

Fíjate en que los cuartiles son robustos. Por ejemplo, si el récord de Alcátara fuera de 470 goles, en vez de 47, el valor de Q_3 no cambiaría.

Los datos sobre los goles por temporada de Kubala, también ordenados, son

14	16	17	17	18	**19**	19	22	25	28	48
		↑			↑			↑		
		Q_1			M			Q_3		

Tenemos un número impar de observaciones, por tanto, la mediana es el valor central, que se ha marcado en negrita. El primer cuartil es la mediana de las 5 observaciones que quedan a la izquierda de la mediana. Por tanto, $Q_1 = 17$. También puedes utilizar la fórmula que vimos para localizar la mediana con $n = 5$ observaciones:

$$\text{Localización de } Q_1 = \frac{n+1}{2} = \frac{5+1}{2} = 3$$

El tercer cuartil es la mediana de las 5 observaciones que quedan a la derecha de la mediana, $Q_3 = 25$. La mediana global queda fuera del cálculo de los cuartiles cuando hay un número impar de observaciones. ■

Ve con cuidado cuando, tal como ocurre en estos ejemplos, diversas observaciones toman el mismo valor numérico. Escribe todas las observaciones y aplica las reglas como si todos los valores fueran distintos. Algunos programas estadísticos utilizan una regla un poco distinta para hallar los cuartiles, por lo que los resultados del ordenador pueden ser algo distintos de los que calcules a mano. Pero no te preocupes por eso. Las diferencias serán siempre demasiado pequeñas para ser importantes.

1.3.5 Los cinco números resumen y los diagramas de caja

Aunque las observaciones máxima y mínima nos dicen poco sobre la distribución en conjunto, sin embargo nos dan información sobre sus colas. Esta información no queda reflejada si solamente conocemos Q_1, M y Q_3. Para tener un resumen rápido del centro y la dispersión, combina los cinco números.

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

$$\text{Mínima} \quad Q_1 \quad M \quad Q_3 \quad \text{Máxima}$$

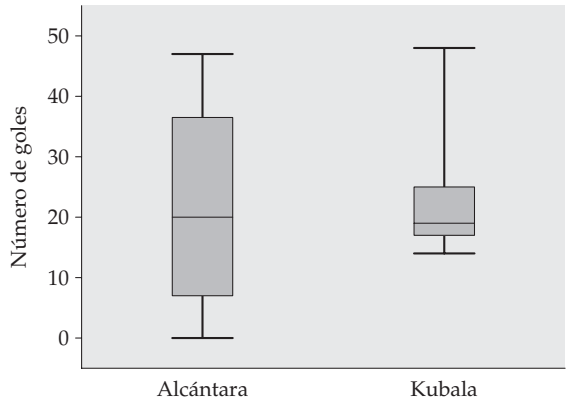


Figura 1.11. Diagramas de caja en un mismo gráfico para comparar el número de goles por temporada de Paulino Alcántara y de Ladislao Kubala.

Estos cinco números proporcionan una descripción razonablemente completa del centro y la dispersión. Los cinco números resumen del ejemplo 1.9 son

0 7 20 36,5 47

para Paulino Alcántara y

14 17 19 25 48

para Ladislao Kubala. Los cinco números resumen de una distribución nos conducen a un nuevo gráfico, el diagrama de caja. La figura 1.11 nos muestra los diagramas de caja correspondientes a estas distribuciones.

DIAGRAMA DE CAJA

Un **diagrama de caja** muestra gráficamente los cinco números resumen.

- Los lados superior e inferior de la caja corresponden a los cuartiles.
- El segmento del interior de la caja corresponde a la mediana.
- Los extremos de los segmentos perpendiculares a los lados superior e inferior de la caja corresponden a los valores máximo y mínimo, respectivamente.

Debido a que los diagramas de caja proporcionan menos detalles que los histogramas o los diagramas de tallos, se utilizan para comparar simultáneamente varias distribuciones, como en la figura 1.11. Los diagramas de caja se pueden dibujar de forma horizontal o vertical. Asegúrate de incluir una escala numérica en el gráfico. Cuando mires un diagrama de caja, localiza primero 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 de los segmentos perpendiculares a los lados superior e inferior de la caja (observaciones máxima y mínima) muestran la dispersión de todos los datos. En la figura 1.11 vemos que Kubala era mucho más regular que Alcántara, ya que la dispersión de su número de goles por temporada es mucho menor. Este hecho queda especialmente claro si nos fijamos en la dispersión del 50% de los datos centrales.

Un diagrama de caja también informa sobre la simetría o 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 global.

APLICA TUS CONOCIMIENTOS

1.31. Médicos suizos. El ejercicio 1.28 proporciona el número de cesáreas realizadas por 15 médicos en Suiza. El mismo estudio también proporciona el número de cesáreas llevadas a cabo por 10 doctoras:

5 7 10 14 18 19 25 29 31 33

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

(b) Dibuja un diagrama de tallos doble para comparar el número de operaciones realizadas por los doctores y las doctoras. ¿Cuáles son tus conclusiones?

1.32. ¿Qué edad tienen los presidentes de EE UU? ¿Qué edad tenían los presidentes de EE UU al inicio de su mandato? Bill Clinton tenía 46 años, ¿era muy joven cuando tomó posesión de su cargo? La tabla 1.7 proporciona las edades de todos los presidentes de EE UU al inicio de su mandato.

(a) Representa mediante un diagrama de tallos la distribución de las edades. A partir de la forma de la distribución, ¿crees que la mediana tiene que ser mucho menor que la media, igual o mucho mayor?

Tabla 1.7. Edades de los presidentes de EE UU al inicio de su mandato.

Presidente	Año	Presidente	Año	Presidente	Año
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. B. 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

(b) Calcula la media y 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 de las edades de los presidentes al inicio de su mandato? ¿Bill Clinton estaba entre el 25% de presidentes más jóvenes?

1.33. PIB *per cápita* de Estados europeos. La tabla 1.6 contiene datos sobre los Estados europeos. Queremos comparar la distribución del PIB *per cápita* de los países de la Unión Europea (UE) con la de los países que formaban parte del bloque soviético (EE). Entramos los datos en el ordenador con los nombres UE para los países de la Unión Europea y EE para los países del extinto bloque soviético. He aquí los resultados del programa estadístico Minitab que proporciona los cinco números resumen junto con otra información (otros programas ofrecen resultados similares).

```
EE
     N      MEAN   MEDIAN    STDEV    MIN     MAX      Q1      Q3
    19    2124.7     2160   1541.3    360    7140    1080    2590
UE
     N      MEAN   MEDIAN    STDEV    MIN     MAX      Q1      Q3
    15   21078.1    22445   7900.7   7480   39850   16020   24290
```

Utiliza estos resultados para dibujar en un mismo gráfico el diagrama de caja de los países de la Unión Europea (UE) y el diagrama de caja de los países que formaban parte del bloque soviético (EE). Describe brevemente la comparación de las dos distribuciones.

1.3.6 Una medida de dispersión: la desviación típica

Los cinco números resumen no son la descripción numérica más común de una distribución. 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.

LA DESVIACIÓN TÍPICA

La **varianza s^2** de un conjunto de observaciones es la suma de los cuadrados de las desviaciones de las observaciones respecto a su media dividido por $n-1$. Algebraicamente la varianza de n observaciones $x_1, x_2, \ldots, x_n$ es

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

o, de 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:

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

En la práctica, utilizaremos una calculadora o un ordenador para calcular la desviación típica. De todas maneras, calcular algunos casos, paso a paso, te ayudará a comprender la varianza y la desviación típica.

EJEMPLO 1.10. Cálculo de la desviación típica

El nivel metabólico de una persona es el ritmo al que el cuerpo consume energía. Este nivel 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 cada 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:

$$\begin{aligned}\bar{x} &= \frac{1.792 + 1.666 + 1.362 + 1.614 + 1.460 + 1.867 + 1.439}{7} \\ &= \frac{11.200}{7} = 1.600 \text{ calorías}\end{aligned}$$

La figura 1.12 muestra los datos como puntos sobre una escala numérica, la media se ha marcado con un asterisco (*). Las flechas señalan la situación de dos desviaciones con relación a la media. Estas desviaciones muestran la dispersión de los datos con relación a dicha media y son el punto de partida para los cálculos de la varianza y de la desviación típica.

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	$1.439 - 1.600 = -161$	$(-161)^2 = 25.921$
	Total $= 0$	Total $= 214.870$

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

$$s^2 = \frac{214.870}{6} = 35.811{,}67$$

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

$$s = \sqrt{35.811{,}67} = 189{,}24 \text{ calorías}$$

■

Fíjate en que para calcular la varianza s^2 dividimos el sumatorio por el número de observaciones menos uno, es decir, por $n - 1$, en vez de n. La razón es que la suma de las desviaciones $x_i - \bar{x}$ es siempre cero, la última desviación se puede hallar cuando se conocen las otras $n - 1$. Solamente $n - 1$ de las desviaciones al cuadrado pueden variar libremente. Esta es la razón por la que dividimos 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 $n - 1$. Asegúrate de dividir por $n - 1$.

Grados de libertad

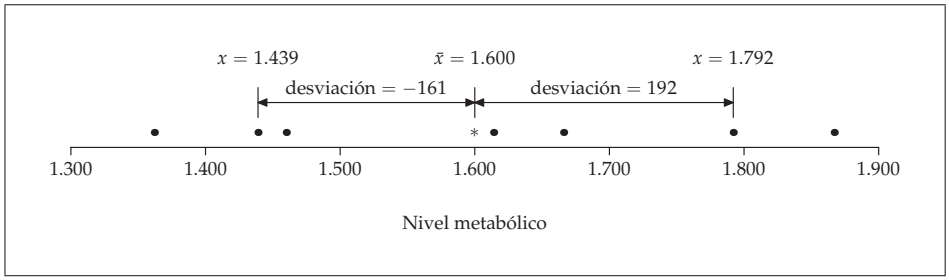


Figura 1.12. Niveles metabólicos de siete hombres, con la media (∗) y las desviaciones de dos observaciones respecto a la media.

Más importante que los detalles del cálculo a mano son las propiedades que determinan la utilidad de la desviación típica:

- s mide la dispersión con relación a la media y se debe utilizar solamente cuando se elige la media como medida de centro.
- $s = 0$ solamente cuando *no hay dispersión*. Esto ocurre únicamente cuando todas las observaciones toman el mismo valor. En caso contrario $s > 0$. A medida que las observaciones se separan más de la media, s se hace mayor.
- s tiene las mismas unidades de medida que las observaciones originales. Por ejemplo, si mides los niveles metabólicos en calorías, s también se expresa en calorías. Este es un motivo para preferir s a la varianza s^2, que se expresa en calorías al cuadrado.
- Igual que ocurre con la media $\bar{x}$, s no es robusta: fuertes asimetrías o unas pocas observaciones atípicas pueden hacer que aumente mucho s. Por ejemplo, la desviación típica del número de goles que marcó Kubala por temporada es 9,51 (puedes comprobarlo con tu calculadora). Si eliminamos la temporada 1951/52, una temporada con un número extraordinario de goles, la desviación típica de las restantes temporadas es 4,30.

Puede ser que creas que la importancia de la desviación típica no está suficientemente justificada. En la próxima sección veremos que la desviación típica es la medida natural de dispersión para una importante 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.

1.3.7 Elección de medidas de centro y de dispersión

Hemos visto dos maneras de describir el centro y la dispersión de una distribución: los cinco números resumen, por un lado, y $\bar{x}$ y s, por otro. ¿Cuál de estas dos maneras tenemos que escoger? Debido a que los dos lados de una distribución muy asimétrica tienen dispersiones distintas, no existe la posibilidad de que con un solo número, como por ejemplo s, podamos describir bien la dispersión. En tal caso, es mejor utilizar los cinco números resumen, con los dos cuartiles y los dos valores extremos.

ELECCIÓN DE UN RESUMEN NUMÉRICO

Para describir una distribución asimétrica o una distribución con observaciones atípicas muy claras, es mejor utilizar los cinco números resumen. Utiliza $\bar{x}$ y s sólo en el caso de distribuciones razonablemente simétricas que no presenten observaciones atípicas.

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 1.36 proporciona un ejemplo de distribución para la cual los resúmenes numéricos son engañosos. **Representa siempre tus datos gráficamente.**

APLICA TUS CONOCIMIENTOS

1.34. La concentración de determinadas sustancias en la sangre influye en la salud de las personas. He aquí las mediciones del nivel de fosfatos 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 1.10 ilustra este método.

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

1.35. Ferenc Puskas. El gran jugador de fútbol Ferenc Puskas, conocido popularmente como Cañoncito Pum, jugó de la temporada 1948/49 a la 1956/57 en el Kispest de Budapest. En 1956 huyó de Hungría cuando estalló la Revolución húngara y estuvo dos temporadas sin jugar. En la temporada 1958/59 fichó por el Real Madrid y estuvo en activo en este equipo como jugador de la liga española hasta la temporada 1965/66. He aquí el número de goles que marcó por temporada:

Temporada	Goles	Temporada	Goles
1948/49	31	1957/58	0
1949/50	25	1958/59	21
1950/51	21	1959/60	25
1951/52	22	1960/61	27
1952/53	27	1961/62	20
1953/54	21	1962/63	26
1954/55	18	1963/64	20
1955/56	5	1964/65	11
1956/57	0	1965/66	4

(a) Utiliza tu calculadora para hallar la media $\bar{x}$ y la desviación típica s del número de goles en la liga desde la temporada 1948/49 hasta la temporada 1965/66.

(b) Utiliza tu calculadora para hallar $\bar{x}$ y s una vez eliminadas las temporadas 1956/57 y 1957/58. ¿Cómo afecta la eliminación de estas dos temporadas a los valores de $\bar{x}$ y s?

1.36. PIB *per cápita* de los Estados europeos. El ejercicio 1.33 proporciona resúmenes numéricos sobre el PIB *per cápita* de los Estados europeos pertenecientes a la Unión Europea y sobre los Estados que formaban parte del antiguo bloque soviético. Ahora considera todos los Estados europeos conjuntamente. Calcula

los cinco números resumen y dibuja el diagrama de caja correspondiente. Estos resúmenes numéricos (y el diagrama de caja derivado del mismo) no muestran una de las características más importantes de esta distribución. Dibuja un diagrama de tallos con todos los datos sobre el PIB *per cápita* de los Estados europeos de la tabla 1.6. ¿Cuál es la forma de la distribución? Recuerda que debes empezar siempre representando gráficamente tus datos —los resúmenes numéricos no son una descripción completa—.

RESUMEN DE LA SECCIÓN 1.3

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 la media aritmética 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 también a su izquierda.

Los **cinco números resumen** son la mediana, los cuartiles y las observaciones extremas máxima y mínima. Los cinco números resumen proporcionan una descripción rápida de una distribución. La mediana describe el centro; 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 dan la dispersión 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 positiva, 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.

Una **medida robusta** de cualquier aspecto de una distribución se ve relativamente poco afectada por cambios en los valores numéricos de una pequeña parte del número total de observaciones, sin importar la magnitud de estos cambios. La mediana y los cuartiles son robustas, en cambio, la media y la desviación típica no lo son.

La media y la desviación típica son buenas descripciones de las distribuciones simétricas sin observaciones atípicas. Son especialmente útiles en el caso de distribuciones normales que veremos en la siguiente sección. Los cinco números resumen son la mejor síntesis de las distribuciones asimétricas.

EJERCICIOS DE LA SECCIÓN 1.3

1.37. El año pasado una pequeña empresa de consultoría pagó a cada uno de sus cinco administrativos 22.000 € y a los dos titulados universitarios, 50.000. Finalmente, el propietario de la empresa cobró 270.000 €. ¿Cuál es el salario medio pagado en esta empresa? ¿Cuántos empleados ganan menos de la media? ¿Cuál es el salario mediano?

1.38. Elecciones presidenciales. El porcentaje de votos que obtuvo cada uno de los candidatos a la presidencia de EE UU que ganó las elecciones desde 1948 hasta 1996 es el siguiente:

Año	Porcentaje	Año	Porcentaje
1948	49,6	1976	50,1
1952	55,1	1980	50,7
1956	57,4	1984	58,8
1960	49,7	1988	53,9
1964	61,1	1992	43,2
1968	43,4	1996	49,2
1972	60,7		

(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 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. Hállalo. ¿En qué años se obtuvieron victorias aplastantes?

1.39. ¿Cuántas calorías contienen las salchichas? 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 cerdo, y de 17 marcas de salchichas hechas con carne de pollo.[10] Éstos son 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 cerdo:

```
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 pollo:

```
Mean = 122.5  Standard deviation = 25.48  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 pollo, ¿significa ingerir menos calorías que comer las hechas con carne de ternera o de cerdo?

1.40. Porcentaje del PIB destinado a educación. La columna "% PIB en educación pública" de la tabla 1.6 proporciona el porcentaje del PIB de los Estados europeos dedicado a educación pública. Queremos comparar el porcentaje dedicado por los Estados de la Unión Europea con el dedicado por los países del Este que formaban parte del bloque soviético.

(a) Haz una lista (con los valores ordenados) de los datos del porcentaje del PIB destinado a educación pública de los Estados de la Unión Europea y otra lista con los datos 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 ambas distribuciones. Describe brevemente lo que observas.

[10] *Consumer Reports*, junio 1986, págs. 366-367. Un estudio más reciente aparece en *Consumer Reports*, julio 1993, págs. 415-419.

1.41. Densidad de la Tierra. 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):[11]

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?

1.42. $\bar{x}$ y s no son suficientes. La media $\bar{x}$ y la desviación típica s como medidas de centro y de dispersión no son una descripción completa de una distribución. Conjuntos de datos de distintas formas pueden tener la misma media y desviación típica. Para demostrar este hecho, utiliza tu calculadora y halla $\bar{x}$ y s de los siguientes conjuntos de datos. A continuación dibuja un diagrama de tallos de cada uno de ellos. Comenta la forma de cada distribución.

Datos A	9,14	8,14	8,74	8,77	9,26	8,10	6,13	3,10	9,13	7,26	4,74
Datos B	6,58	5,76	7,71	8,84	8,47	7,04	5,25	5,56	7,91	6,89	12,50

1.43. La tabla 1.1 facilita datos sobre el porcentaje de gente con al menos 65 años en cada uno de los Estados de EE UU. La figura 1.2 es un histograma correspondiente a estos datos. Como descripción numérica breve, ¿qué prefieres, los cinco números resumen o $\bar{x}$ y s? ¿Por qué? Calcula la descripción que prefieras.

1.44. ¿Acciones calientes? La tabla 1.8 proporciona los rendimientos, expresados como porcentajes mensuales, de las acciones de Philip Morris en un periodo comprendido entre el mes de julio de 1990 y el mes de mayo de 1997. (El rendimiento

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

de una acción deriva de la variación de su precio y de los dividendos pagados, aquí se expresa como un porcentaje de su valor al inicio de cada mes.)

(a) Dibuja un diagrama de tallos o un histograma con estos datos. ¿Cómo has decidido qué representación gráfica utilizar?

(b) Existe una clara observación atípica. ¿Cuál es el valor de esta observación? (La bajada de la cotización de estas acciones, se puede explicar por las nuevas acciones emprendidas en contra de las tabacaleras.) Después de eliminar la observación atípica, describe la forma, el centro y la dispersión de los datos.

(c) En los estudios sobre inversiones, es frecuente utilizar la media y la desviación típica para resumir y comparar los rendimientos de las acciones. Halla la media y la desviación típica de los rendimientos de las acciones de la tabla 1.8. Si invirtieras 100 € en estas acciones al comienzo de un mes y obtuvieras el rendimiento medio, ¿cuánto tendrías al final del mes?

(d) Si invirtieras 100 € en estas acciones al comienzo del peor mes (la observación atípica), ¿cuánto tendrías al final del mes? Halla otra vez la media y la desviación típica, pero dejando fuera la observación atípica. ¿En qué medida afecta esta observación atípica los valores de la media y de la desviación típica? La eliminación de esta observación atípica, ¿cambiaría el valor de la mediana? ¿Y los cuartiles? Sin hacer los cálculos, ¿cómo lo puedes saber?

Tabla 1.8. Rendimientos mensuales, expresados como porcentaje, de las acciones de Philip Morris, desde julio de 1990 hasta mayo de 1997.

−5,7	1,2	4,1	3,2	7,3	7,5	18,6	3,7	−1,8	2,4
−6,5	6,7	9,4	−2	−2,8	−3,4	19,2	−4,8	0,5	−0,6
2,8	−0,5	−4,5	8,7	2,7	4,1	−10,3	4,8	−2,3	−3,1
−10,2	−3,7	−26,6	7,2	−2,9	−2,3	3,5	−4,6	17,2	4,2
0,5	8,3	−7,1	−8,4	7,7	−9,6	6	6,8	10,9	1,6
0,2	−2,4	−2,4	3,9	1,7	9	3,6	7,6	3,2	−3,7
4,2	13,2	0,9	4,2	4	2,8	6,7	−10,4	2,7	10,3
5,7	0,6	−14,2	1,3	2,9	11,8	10,6	5,2	13,8	−14,7
3,5	11,7	1,3							

1.45. Salarios de atletas. Los jugadores del equipo de béisbol de los Orioles de Baltimore en EE UU fueron los mejor pagados durante la liga estadounidense de 1998. He aquí sus salarios en miles de dólares. (Por ejemplo, 6.495 significa 6.495.000 dólares.)

6.495	6.486	6.300	6.269	5.442	5.391	3.600	3.600	3.583
3.089	2.850	2.500	1.950	1.663	1.367	1.333	1.150	900
856	800	800	665	650	450	450	170	170

1.46. Valor neto de un patrimonio. El valor neto de un patrimonio es el valor total de las posesiones e inversiones menos las deudas totales. En 1997 el valor medio y la mediana de los patrimonios europeos eran de 51.000 y 212.000 €, respectivamente. ¿Cuál de estos valores corresponde a la media? ¿Y a la mediana? Justifica tus respuestas.

1.47. Salarios millonarios de los jugadores de la NBA. En un artículo se comenta que de los 411 jugadores de la NBA (National Basketball Association) sólo 139 ganaban más de 2,36 millones de dólares. En el artículo no queda claro si 2,36 es la media o la mediana de las ganancias de los jugadores de baloncesto de la NBA. ¿Tú que crees, que es la media o la mediana? ¿Por qué?

1.48. ¿Media o mediana? En cada una de las situaciones siguientes, ¿qué medida de centro deberías utilizar, la media o la mediana?

(a) El Ayuntamiento de Barcelona está considerando la posibilidad de aplicar un nuevo impuesto sobre los ingresos de los hogares de la ciudad. Para ello, quiere conocer los ingresos totales de los hogares.

(b) En un estudio sobre el nivel de vida de los barceloneses, un sociólogo quiere conocer la renta típica de un hogar de la ciudad.

1.49. 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)? Justifica tu respuesta.

1.4 Distribuciones normales

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

1. Siempre representa gráficamente tus datos, habitualmente con un diagrama de tallos o con un histograma.
2. Identifica su aspecto general (forma, centro y dispersión) y las desviaciones sorprendentes, como son las observaciones atípicas.
3. Calcula un resumen numérico para describir de forma breve el centro y la dispersión de la distribución.

4. He aquí un elemento más que hay que añadir a esta estrategia: algunas veces la forma de la distribución de un gran número de observaciones es tan regular que la podemos describir mediante una curva lisa.

1.4.1 Curvas de densidad

La figura 1.13 es un histograma sobre las notas de Lengua de los 947 estudiantes de séptimo curso de la ciudad de Gary, Indiana, EE UU, en un examen a nivel nacional.[12] Las notas de muchos estudiantes tienen una distribución bastante regular. El histograma es simétrico, ya que ambos lados disminuyen de forma suave desde el pico central. No hay espacios vacíos destacables ni observaciones atípicas. La curva dibujada a través de la parte alta de las barras del histograma de la figura 1.13 es una buena descripción del aspecto general de los datos. Esta curva es un **modelo matemático** de la distribución, es decir, es una descripción idealizada. La curva de densidad describe de forma compacta el aspecto general de los datos, ignora las pequeñas irregularidades así como las observaciones atípicas.

Modelo matemático

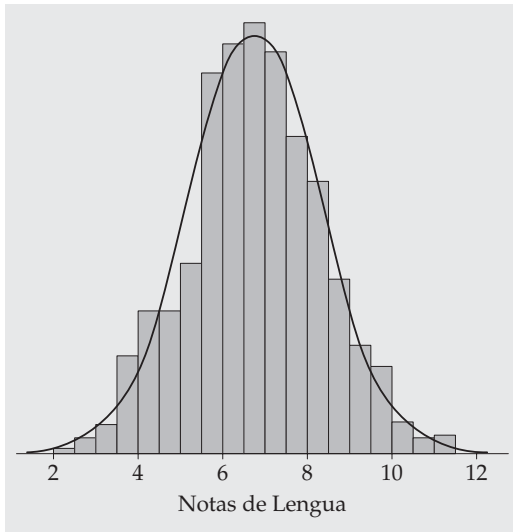


Figura 1.13. Histograma de las notas de Lengua de todos los alumnos de séptimo curso de la ciudad de Gary, Indiana, EE UU. La curva muestra la forma de la distribución.

[12]Datos proporcionados por Celeste Foster, Departamento de Educación, Purdue University.

Veremos que es más fácil trabajar con la curva de la figura 1.13 que con el histograma. La razón es que el histograma depende de la elección que hacemos de las clases. Con un poco de cuidado podemos utilizar una curva que no dependa de ninguna decisión nuestra. Veamos cómo hacerlo.

EJEMPLO 1.11. De un histograma a una curva de densidad

Nuestra vista observa las *áreas* de las barras de un histograma. Estas áreas representan proporciones de observaciones. La figura 1.14(a) es una copia de la figura 1.13 en la que se han sombreado las barras de la izquierda. Estas barras sombreadas representan a los estudiantes con notas de Lengua menores o iguales a 6,0. Hay 287 estudiantes de este tipo, que representan en total la proporción $\frac{287}{947} = 0{,}303$ de todos los estudiantes de séptimo curso de Gary.

Ahora fíjate en la curva trazada sobre las barras. En la figura 1.14(b), se ha sombreado el área por debajo de la curva situada a la izquierda de 6,0. Ajusta la escala del gráfico de manera que *el área total por debajo de la curva sea exactamente 1*. Este área representa la proporción 1, es decir, la totalidad de las observaciones. Por tanto, las áreas por debajo de la curva representan proporciones de observaciones. Ahora, la curva es una *curva de densidad*. El área sombreada por debajo de la curva de la figura 1.14(b) representa la proporción de estudiantes con notas menores o iguales a 6,0. Este área es 0,293, la diferencia con el área del histograma es solamente 0,010. Puedes ver que las áreas por debajo de la curva de densidad dan aproximaciones bastante buenas de las áreas dadas por el histograma. ■

CURVA DE DENSIDAD

Una **curva de densidad** es una curva que:

- se halla siempre en el eje de las abscisas o por encima de él, y
- define por debajo un área exactamente igual a 1.

Una curva de densidad describe el aspecto general de una distribución. El área por debajo de la curva, y entre cualquier intervalo de valores, es la proporción de todas las observaciones que están situadas en dicho intervalo.

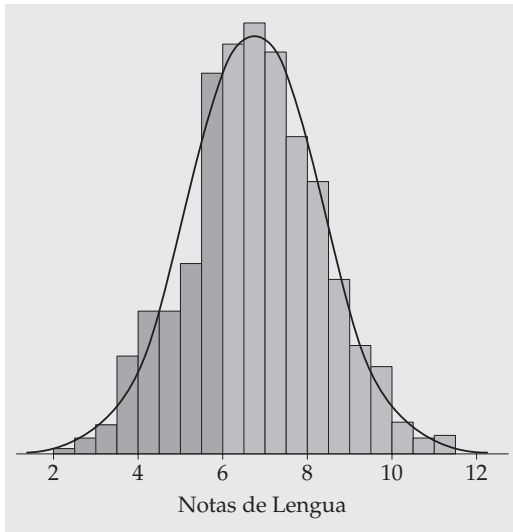


Figura 1.14(a). A partir del histograma, la proporción de notas menores o iguales que 6,0 es 0,303.

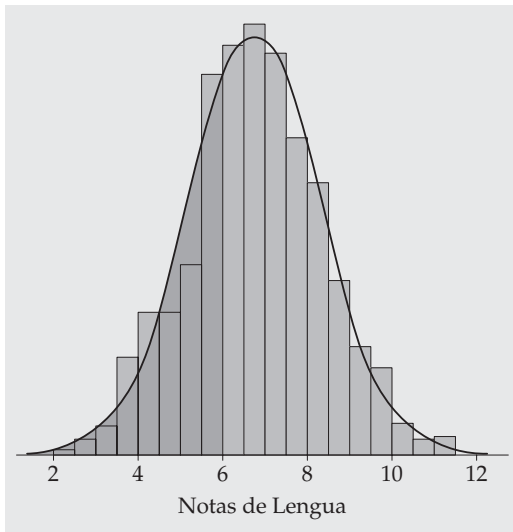


Figura 1.14(b). A partir de la curva de densidad, la proporción de notas menores o iguales que 6,0 es 0,293.

Curva normal

La curva de densidad de las figuras 1.13 y 1.14 son **curvas normales**. Las curvas de densidad, al igual que las distribuciones, pueden tener muchas formas. La figura 1.15 muestra dos de estas curvas: una simétrica y otra asimétrica hacia la derecha. Una curva de densidad es, a menudo, una descripción adecuada del aspecto general de una distribución. La curva no describe las observaciones atípicas, que son desviaciones del aspecto general. Por supuesto que ningún conjunto de datos reales es descrito exactamente por una curva de densidad. La curva es una aproximación fácil de utilizar y lo suficientemente precisa para ser utilizada en la práctica.

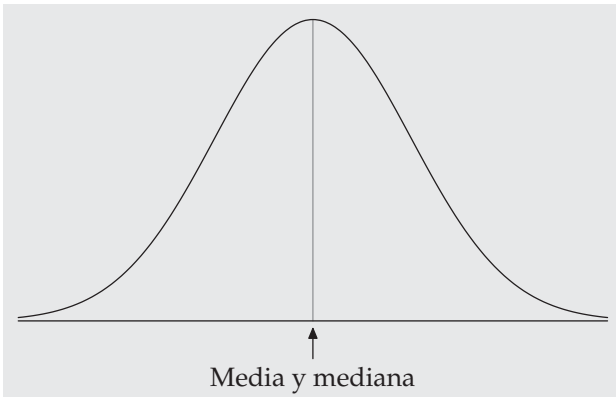


Figura 1.15(a). La mediana y la media de una curva de densidad simétrica.

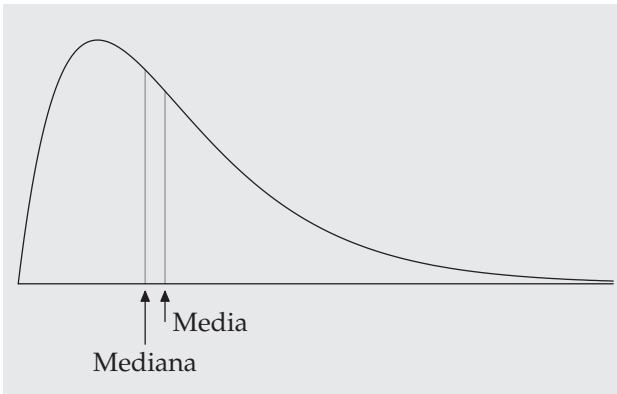


Figura 1.15(b). La mediana y la media de una curva de densidad asimétrica hacia la derecha.

1.4.2 Mediana y media de una curva de densidad

Nuestras medidas de centro y de dispersión se aplican tanto a las curvas de densidad como a los datos reales. La mediana y los cuartiles son fáciles de calcular.

Las áreas por debajo de una curva de densidad representan proporciones del número total de observaciones. La mediana es el punto del eje de las abscisas que tiene el mismo número de observaciones a ambos lados. En consecuencia, **la mediana de una curva de densidad es el punto del eje de abscisas que divide la curva en dos áreas iguales**, es decir, el punto que tiene la mitad del área por debajo de la curva de densidad a su izquierda y la otra mitad a su derecha. Los cuartiles dividen el área por debajo de la curva en cuatro partes iguales. Una cuarta parte del área por debajo de la curva queda a la izquierda del primer cuartil, y tres cuartas partes del área quedan a la izquierda del tercer cuartil. Puedes localizar, aproximadamente, la mediana y los cuartiles de cualquier curva de densidad a simple vista, dividiendo el área por debajo de la curva en cuatro partes iguales.

Debido a que las curvas de densidad son distribuciones idealizadas, una curva de densidad simétrica es exactamente simétrica. La mediana de una curva de densidad simétrica se encuentra, por tanto, en su centro. La figura 1.15(a) muestra la mediana de una curva simétrica. No es fácil situar el punto del eje de las abscisas que divide una curva de densidad asimétrica en dos áreas iguales, pero existen métodos matemáticos para encontrar la mediana de cualquier curva de densidad. Hemos utilizado uno de estos métodos para señalar la mediana en la curva de densidad de la figura 1.15(b).

¿Qué se puede decir acerca de la media? La media de un conjunto de observaciones es su media aritmética. Si pensamos en las observaciones como pesos distribuidos a lo largo de una vara, la media es el punto de equilibrio de la vara. Este hecho también es cierto para las curvas de densidad. **La media es el punto en el que se equilibraría el área por debajo de la curva si estuviera constituida por un material sólido**. La figura 1.16 ilustra este hecho. Una curva simétrica se equilibra en su centro ya que los dos lados son idénticos. Por tanto, **la media y la mediana de una curva de densidad simétrica son iguales**, tal como ilustra la figura Figura 1.15(a). Sabemos que la media de una distribución asimétrica está desplazada hacia la cola larga. La figura 1.15(b) muestra cómo la media de una curva de densidad asimétrica está más desplazada hacia la cola larga que la mediana. Es difícil localizar a simple vista el punto de equilibrio de una curva asimétrica. Existen procedimientos matemáticos para calcular la media de cualquier curva de densidad; utilizándolos fuimos capaces tanto de localizarla en la figura 1.15(b) como de situar en ella la mediana.

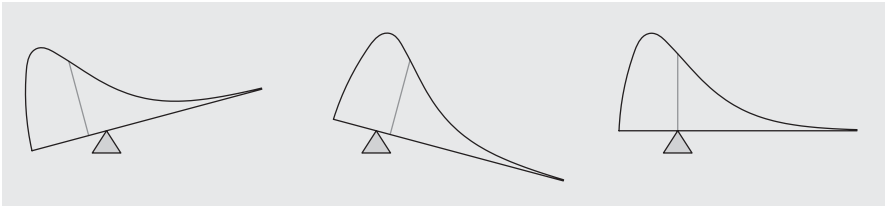


Figura 1.16. La media sería el punto de equilibrio del área por debajo de la curva en caso de que estuviera constituida por un material sólido.

MEDIANA Y MEDIA DE UNA CURVA DE DENSIDAD

La **mediana** de una curva de densidad es el punto que divide el área por debajo de la curva en dos mitades iguales.

La **media** de una curva de densidad es el punto de equilibrio, aquel en el que la curva se equilibraría si ésta estuviera hecha de un material sólido.

La mediana y la media son iguales en el caso de curvas de densidad simétricas. Las dos se encuentran en el centro de la curva. La media de una curva asimétrica se desplaza de la mediana, hacia la cola larga.

A simple vista podemos situar de una manera aproximada la media, la mediana y los cuartiles de una curva de densidad. Esto no ocurre con la desviación típica. Cuando sea necesario podemos acudir a métodos matemáticos más avanzados para conocer el valor de la desviación típica. El estudio de estos métodos matemáticos forma parte de la estadística teórica. Aunque nos centremos en la estadística aplicada, a menudo utilizaremos los resultados de los estudios matemáticos.

Como una curva de densidad es una descripción idealizada de la distribución de datos, necesitamos distinguir entre la media y la desviación típica de una curva de densidad, y la media $\bar{x}$ y la desviación típica s calculada a partir de observaciones reales. La notación habitual de la media de una distribución idealizada es $\boldsymbol{\mu}$ (la letra griega my). La desviación típica de una curva de densidad se representa por $\boldsymbol{\sigma}$ (la letra griega sigma minúscula).

Media μ

Desviación típica σ

APLICA TUS CONOCIMIENTOS

1.50.(a) Dibuja una curva de densidad que sea simétrica pero que tenga una forma distinta a la curva de la figura 1.15(a).

(b) Dibuja una curva de densidad que sea muy asimétrica hacia la izquierda.

1.51. La figura 1.17 muestra la curva de densidad de una *distribución uniforme*. La curva toma el valor constante 1 para todos los valores situados en el intervalo definido entre 0 y 1, y el valor 0 para los restantes valores. Esto significa que los datos descritos por esta distribución toman valores con una dispersión uniforme entre 0 y 1. Utiliza las áreas por debajo de la curva de densidad para responder a las siguientes preguntas:

(a) ¿Por qué el área total por debajo de la curva es igual a 1?

(b) ¿Qué porcentaje de las observaciones es mayor que 0,8?

(c) ¿Qué porcentaje de las observaciones es menor que 0,6?

(d) ¿Qué porcentaje de las observaciones queda entre 0,25 y 0,75?

(e) ¿Cuál es la media μ de esta distribución?

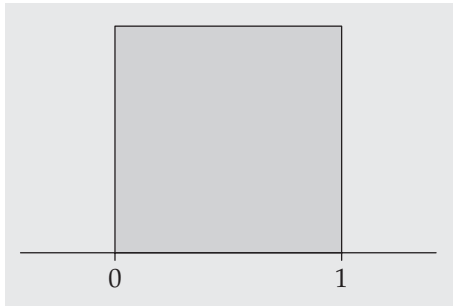


Figura 1.17. La curva de densidad de la distribución uniforme del ejercicio 1.51.

1.52. La figura 1.18 muestra tres curvas de densidad. En cada una de ellas se han señalado tres puntos. ¿Cuáles corresponden a la media y cuáles a la mediana?

1.4.3 Distribuciones normales

Una clase particularmente importante de curvas de densidad se ha visto ya en las figuras 1.13 y 1.15(a). Estas curvas son simétricas, con un solo pico y tienen forma

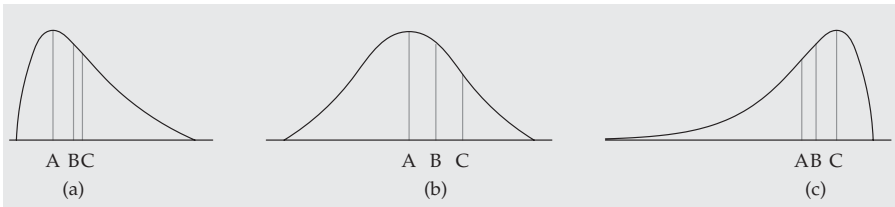


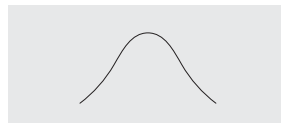
Figura 1.18. Las tres curvas de densidad del ejercicio 1.52.

Distribuciones normales

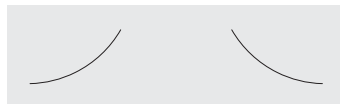
de campana. Se les llama *curvas normales* y describen las **distribuciones normales**. Todas las distribuciones normales tienen el mismo aspecto. La curva de densidad exacta de una distribución normal concreta se describe dando su media μ y su desviación típica σ. La media se sitúa en el centro de la curva simétrica, en el mismo lugar que la mediana. Si se cambia μ sin cambiar σ se provoca un desplazamiento de la curva de densidad a lo largo del eje de las abscisas sin que cambie su dispersión. La desviación típica σ controla la dispersión de la curva normal. La figura 1.19 muestra dos curvas normales con diferentes valores de σ. La curva con una mayor desviación típica presenta una mayor dispersión.

La desviación típica σ es la medida natural de la dispersión de las distribuciones normales. La forma de una curva normal no sólo queda completamente determinada por μ y σ, sino que además es posible situar σ a simple vista en la curva. Vamos a ver cómo.

Imagínate que bajas esquiando una montaña que tiene la forma de una curva normal. Al principio, cuando dejas el pico, la pendiente es muy fuerte.



Afortunadamente, antes de encontrarte al final de la pendiente, ésta se hace más suave:



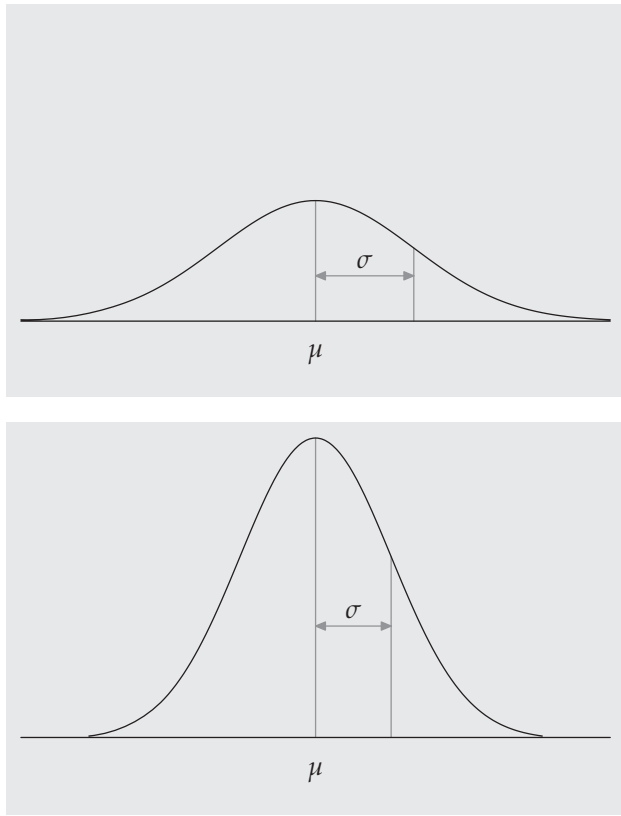


Figura 1.19. Dos curvas normales que muestran la media μ y la desviación típica σ.

Los puntos en los que tiene lugar este cambio de curvatura se hallan a una distancia σ, a ambos lados de la media μ. Puedes percibir este cambio si repasas la curva con un lápiz y, por tanto, puedes hallar la desviación típica. Recuerda que en general μ y σ no describen de manera completa la forma de la mayor parte de las distribuciones, y que la forma de la curva de densidad no nos permite conocer σ. Esto solamente ocurre para las distribuciones normales.

¿Por qué son tan importantes las distribuciones normales en estadística? Hay tres razones. La primera, porque las distribuciones normales son buenas descripciones de algunas distribuciones de *datos reales*. Distribuciones que se aproximan a la normal son, por ejemplo, las de los resultados de un examen que hace mucha

gente (como, por ejemplo, las notas de las pruebas de acceso a la universidad o los resultados de muchas pruebas psicológicas), las medidas repetidas de una misma cantidad, y las de algunas características morfométricas de poblaciones biológicas (como la longitud de las cucarachas o la producción de maíz). La segunda, porque las distribuciones normales son buenas aproximaciones a los resultados de muchos tipos de *fenómenos aleatorios*, tales como lanzar una moneda al aire muchas veces. La tercera, y más importante, porque, como veremos, muchos procedimientos de *inferencia estadística* basados en distribuciones normales, dan buenos resultados cuando se aplican a distribuciones aproximadamente simétricas. De todas formas, a pesar de que muchos conjuntos de datos tienen distribuciones normales, muchos otros carecen de ellas. Por ejemplo, la mayoría de las distribuciones de ingresos son asimétricas hacia la derecha y, por tanto, no son normales. Los datos no normales, al igual que la gente no normal, no sólo son frecuentes, sino que a veces son más interesantes que sus homólogos normales.

1.4.4 Regla del 68-95-99,7

Aunque existen muchas curvas normales, todas ellas tienen propiedades comunes. En particular, todas las distribuciones normales cumplen las propiedades descritas por la siguiente regla.

LA REGLA DEL 68-95-99,7

En una distribución normal de media μ y desviación típica σ:

- El **68%** de todas las observaciones se encuentran dentro del intervalo $\mu \pm \sigma$.
- El **95%** de todas las observaciones se encuentran dentro del intervalo $\mu \pm 2\sigma$.
- El **99,7%** de todas las observaciones se encuentran dentro del intervalo $\mu \pm 3\sigma$.

La figura 1.20 ilustra la regla del 68-95-99,7. Recordando estos tres números, puedes caracterizar una distribución normal sin realizar cálculos muy detallados.

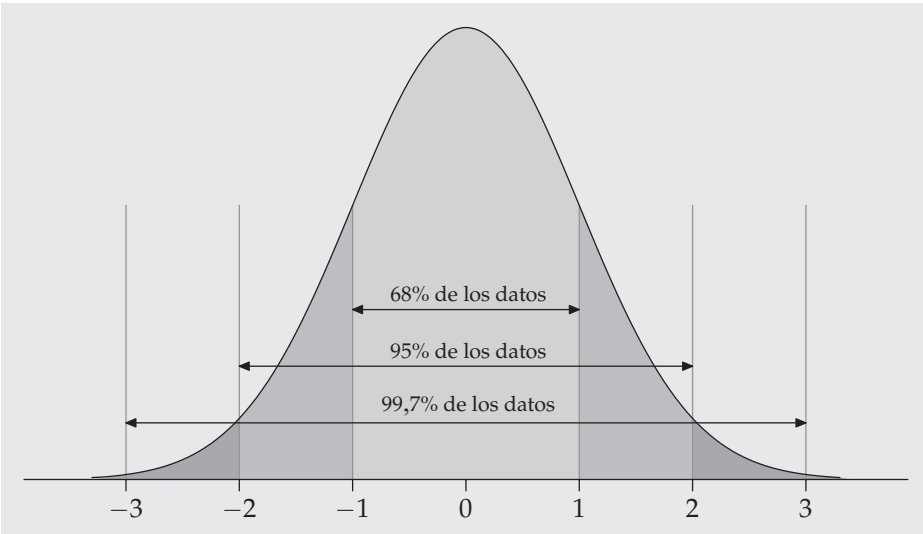


Figura 1.20. La regla del 68-95-99,7 para distribuciones normales.

EJEMPLO 1.12. Utilización de la regla del 68-95-99,7

La distribución de las alturas de las chicas entre 18 y 24 años es aproximadamente normal con media $\mu = 1{,}64$ m y desviación típica $\sigma = 0{,}06$ m. La figura 1.21 muestra la aplicación de la regla 68-95-99,7 a este ejemplo.

Para esta distribución, dos veces la desviación típica es igual a 0,12. El 95 de la regla del 68-95-99,7 indica que el 95% central de las chicas miden entre $1{,}64 - 0{,}12$ y $1{,}64 + 0{,}12$ m de altura, es decir, entre 1,52 y 1,76 m. Esto es exactamente así para una distribución exactamente normal. Es aproximadamente cierto para las alturas de las chicas, ya que la distribución de alturas es aproximadamente normal.

El 5% restante de las chicas tienen alturas situadas fuera del intervalo que va de 1,52 a 1,76 m. Debido a que las distribuciones normales son simétricas, la mitad de estas chicas estarán situadas en la parte alta. Es decir, el 2,5% de las chicas más altas miden más de 1,76 m.

El 99,7 de la regla del 68-95-99,7 indica que prácticamente todas las chicas (el 99,7%) tienen alturas entre $\mu - 3\sigma$ y $\mu + 3\sigma$. Este intervalo de alturas va desde 1,46 hasta 1,82 m. ■

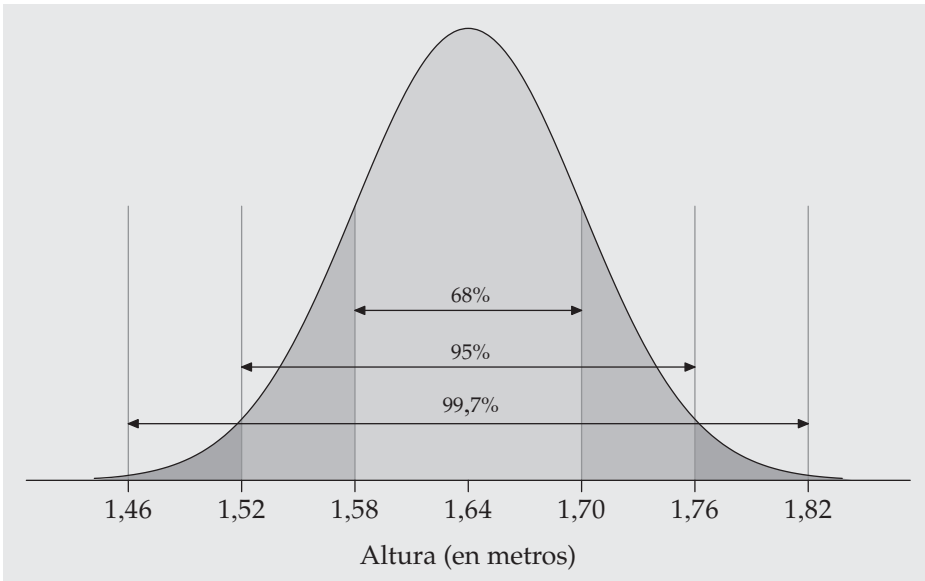


Figura 1.21. La regla del 68-95-99,7 aplicada a la distribución de las alturas de chicas. Aquí, $\mu = 1{,}64$ m y $\sigma = 0{,}06$ m.

Como utilizaremos las distribuciones normales con frecuencia, es útil introducir una notación breve. Nos referimos a una distribución normal de media μ y de desviación típica σ, como a una $N(\mu, \sigma)$. Por ejemplo, la distribución de las alturas de las chicas es una $N(1{,}64, 0{,}06)$.

APLICA TUS CONOCIMIENTOS

1.53. Alturas de hombres. La distribución de alturas de los hombres adultos es aproximadamente normal, con una media de 1,75 m y una desviación típica de 0,06 m. Dibuja una curva normal en la que sitúes correctamente su media y su desviación típica. (Sugerencia: primero dibuja la curva, localiza sobre ella los puntos de inflexión y proyecta estos puntos sobre el eje de las abscisas.)

1.54. Más sobre alturas de hombres. La distribución de las alturas de los hombres adultos es aproximadamente normal con una media de 1,75 m y desviación típica de 0,06 m. Usa la regla del 68-95-99,7 para responder a las siguientes preguntas:

(a) ¿Qué porcentaje de hombres son más altos que 1,87 m?
(b) ¿Entre qué alturas se encuentra el 95% central de la población de hombres?
(c) ¿Qué porcentaje de hombres tiene una altura inferior a 1,69 m?

1.55. Coeficientes de inteligencia. La distribución de los coeficientes de inteligencia de hombres entre 20 y 34 años tiene aproximadamente una distribución normal de media $\mu = 110$ y desviación típica $\sigma = 25$. Utiliza la regla del 68-95-99,7 para responder a las siguientes preguntas:

(a) De los hombres entre 20 y 34 años, ¿qué porcentaje tiene un coeficiente intelectual superior a 110?
(b) ¿Qué porcentaje tiene un coeficiente intelectual superior a 160?
(c) ¿En qué intervalo se encuentra el 95% central de la población?

1.4.5 Distribución normal estandarizada

Tal como sugiere la regla del 68-95-99,7, todas las distribuciones normales comparten muchas propiedades comunes. De hecho, todas ellas son iguales si tomamos como unidad de medida σ a partir de un centro que es la media μ. Pasar a estas unidades se llama *estandarizar*. Para estandarizar un valor, réstale la media de la distribución y luego divídelo por la desviación típica.

ESTANDARIZACIÓN Y VALORES *Z*

Si x es una observación de una distribución de media μ y desviación típica σ, el **valor estandarizado** de x es

$$z = \frac{x - \mu}{\sigma}$$

Los valores estandarizados se llaman a menudo **valores *z*.**

Un valor z nos dice a cuántas desviaciones típicas se encuentra la observación original de la media y en qué dirección. Las observaciones mayores que la media son positivas y las menores, negativas.

EJEMPLO 1.13. Estandarización de las alturas de las chicas

La distribución de las alturas de las chicas es aproximadamente normal con $\mu = 1{,}64$ m y $\sigma = 0{,}06$ m. La altura estandarizada es

$$z = \frac{\text{altura} - 1{,}64}{0{,}06}$$

La altura estandarizada de una chica es el número de desviaciones típicas que su altura difiere de la media de las alturas de todas las chicas. Por ejemplo, una chica de 1,72 m de altura tiene una altura estandarizada de

$$z = \frac{1{,}72 - 1{,}64}{0{,}06} = 1{,}33$$

o, lo que es lo mismo, su altura es 1,33 desviaciones típicas mayor que la media. De manera similar, una chica de 1,56 m tiene una altura estandarizada de

$$z = \frac{1{,}56 - 1{,}64}{0{,}06} = -1{,}33$$

es decir, 1,33 desviaciones típicas menos que la altura media. ■

Si la variable que estandarizamos tiene una distribución normal, la estandarización no hace más que dar una escala común. La estandarización transforma todas las distribuciones normales en una sola distribución, y ésta sigue siendo normal. La estandarización de una variable que tenga una distribución normal genera una nueva variable que tiene la *distribución normal estandarizada*.

DISTRIBUCIÓN NORMAL ESTANDARIZADA

La **distribución normal estandarizada** es la distribución normal $N(0, 1)$ de media 0 y desviación típica 1.

Si una variable x tienen una distribución normal $N(\mu, \sigma)$ de media μ y desviación típica σ, entonces la variable estandarizada

$$z = \frac{x - \mu}{\sigma}$$

tiene una distribución normal estandarizada.

APLICA TUS CONOCIMIENTOS

1.56. SAT versus ACT. Meritxell obtuvo 680 puntos en el examen de Matemáticas de la prueba SAT (*Scholastic Assessment Test*) de acceso a la universidad en EE UU. La distribución de las notas de Matemáticas en la prueba SAT es normal con media igual a 500 y desviación típica igual a 100. Clara obtuvo 27 puntos en el examen de Matemáticas de otra prueba de acceso a la universidad también en EE UU, la prueba ACT (*American College Testing*). Las notas de Matemáticas en la prueba ACT tienen también una distribución normal pero de media 18 y desviación típica 6. Halla las notas estandarizadas de ambas estudiantes. Suponiendo que los dos exámenes sean similares, ¿qué estudiante obtuvo mayor puntuación?

1.4.6 Cálculos con distribuciones normales

El área por debajo de una curva de densidad es una proporción de observaciones de una distribución. Cualquier pregunta sobre qué proporción de observaciones se encuentra en algún intervalo de valores se puede responder hallando el área por debajo de la curva en ese intervalo. Como todas las distribuciones normales son la misma cuando las estandarizamos, podemos hallar las áreas por debajo de cualquier curva normal utilizando una sola tabla, una tabla que da las áreas por debajo de la curva normal estandarizada.

EJEMPLO 1.14. Utilización de la distribución normal estandarizada

¿Qué proporción de todas las chicas miden menos de 1,72 m? Esta proporción es el área por debajo de la $N(1{,}64, 0{,}06)$ situada a la izquierda de 1,72. Como la altura estandarizada correspondiente a 1,72 es

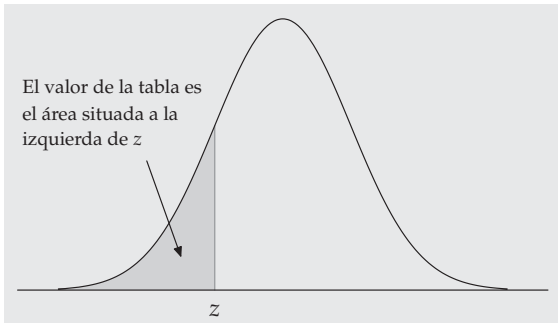
$$z = \frac{x - \mu}{\sigma} = \frac{1{,}72 - 1{,}64}{0{,}06} = 1{,}33$$

este área es la misma que el área por debajo de la curva normal estandarizada situada a la izquierda de $z = 1{,}33$. La figura 1.22(a) muestra este área. ■

Muchas calculadoras dan las áreas por debajo de una curva normal estandarizada. Si tu calculadora no lo hace, la tabla A, que se encuentra al final del libro, da algunas de estas áreas.

LA TABLA NORMAL ESTANDARIZADA

La **tabla A** es la tabla de las áreas por debajo de la curva normal estandarizada. El valor de la tabla correspondiente a cada valor de z es el área por debajo de la curva situada a la izquierda de z.



EJEMPLO 1.15. Utilización de la tabla normal estandarizada

Problema: Halla la proporción de observaciones de la distribución normal estandarizada que son menores que 1,33.

Solución: Para hallar el área situada a la izquierda de 1,33, localiza 1,3 en la columna de la izquierda de la tabla A, luego localiza el dígito restante 3 como 0,03 en la fila superior. El valor del área viene dado, en el cuerpo central de la tabla, en el lugar donde se cruzan la fila donde se halla 1,3 y la columna donde está 0,03. Este valor es 0,9082. La figura 1.22(a) ilustra la relación entre el valor de $z = 1{,}33$ y el área 0,9082. Debido a que $z = 1{,}33$ es el valor estandarizado de 1,72 m, la proporción de chicas que miden menos de 1,72 m es 0,9082 (cerca del 91%).

Problema: Halla la proporción de observaciones de la distribución normal estandarizada que son mayores que $-2{,}15$.

Solución: Entra en la tabla A con el valor $z = -2{,}15$. Es decir, halla $-2{,}1$ en la columna de la izquierda y 0,05 en la fila superior. El valor de la tabla es 0,0158. Este valor es el área situada a la izquierda de $-2{,}15$. Como el área total por debajo de la curva es 1, el área situada a la derecha de $-2{,}15$ es $1 - 0{,}0158 = 0{,}9842$. La figura 1.22(b) ilustra estas áreas. ■

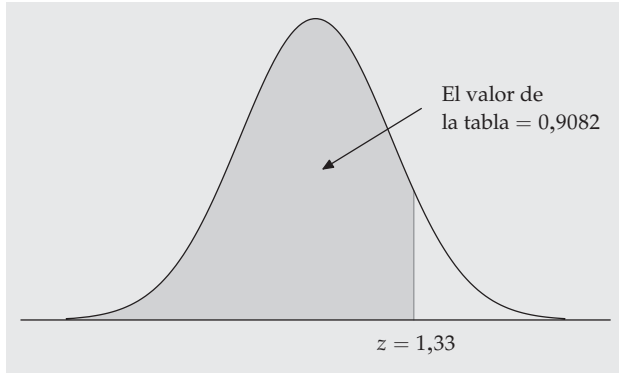


Figura 1.22(a). Área por debajo de una curva normal estandarizada situada a la izquierda del punto $z = 1{,}33$ es 0,9082. La tabla A proporciona las áreas por debajo de la curva normal estandarizada.

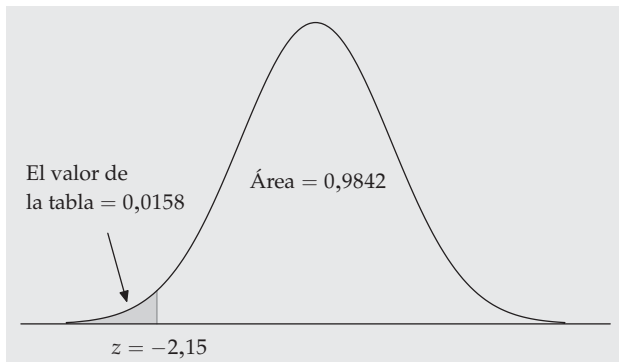


Figura 1.22(b). Áreas por debajo de la curva normal estandarizada situadas a la derecha y a la izquierda de $z = -2{,}15$. La tabla A sólo da las áreas situadas a la izquierda de z.

Podemos responder a cualquier pregunta sobre proporciones de observaciones de una distribución normal, estandarizando y luego utilizando la tabla normal estandarizada. He aquí un resumen del método para hallar la proporción de la distribución en cualquier región.

CÓMO HALLAR PROPORCIONES NORMALES

1. Plantea el problema en términos de la variable observada x.
2. Estandariza x para replantear el problema en términos de la variable normal estandarizada z. Sitúa el área de interés en la curva normal estandarizada.
3. Halla el área buscada por debajo de la curva normal estandarizada, utilizando la tabla A y el hecho de que el área por debajo de la curva es 1.

EJEMPLO 1.16. Cálculos con distribuciones normales

El nivel de colesterol en la sangre es importante, ya que un nivel alto puede aumentar el riesgo de enfermedades coronarias. En una gran población de gente de la misma edad y sexo, la distribución del nivel de colesterol es aproximadamente normal. Para chicos de 14 años, la media es $\mu = 170$ miligramos de colesterol por decilitro de sangre (mg/dl) y la desviación típica es $\sigma = 30$ mg/dl.[13] Los niveles de colesterol superiores a 240 mg/dl pueden exigir atención médica. ¿Qué porcentaje de los chicos de 14 años tienen más de 240 mg/dl de colesterol?

1. *Plantea el problema*. Llama x al nivel de colesterol en la sangre x. La variable x tiene una distribución $N(170, 30)$. Queremos la proporción de chicos con $x > 240$.
2. *Estandariza*. Resta la media, luego divide por la desviación típica, para convertir x en una z normal estandarizada:

$$\begin{aligned} x &> 240 \\ \frac{x-170}{30} &> \frac{240-170}{30} \\ z &> 2{,}33 \end{aligned}$$

[13] P. S. Levy *et al*., "Total serum cholesterol values for youths 12-17 years", *Vital and Health Statistics Series 11*, nº 150, 1975, U.S. National Center for Health Statistics.

La figura 1.23 muestra la curva normal estandarizada. Se ha sombreado el área de interés.

3. *Utiliza la tabla*. En la tabla A vemos que la proporción de observaciones menores que 2,33 es 0,9901. Cerca del 99% de los chicos tienen niveles de colesterol menores que 240. El área situada a la derecha de 2,33 es, por tanto, $1 - 0{,}9901 = 0{,}0099$. Este área es aproximadamente 0,01, o un 1%. Sólo un 1% de los chicos tienen niveles de colesterol tan altos. ■

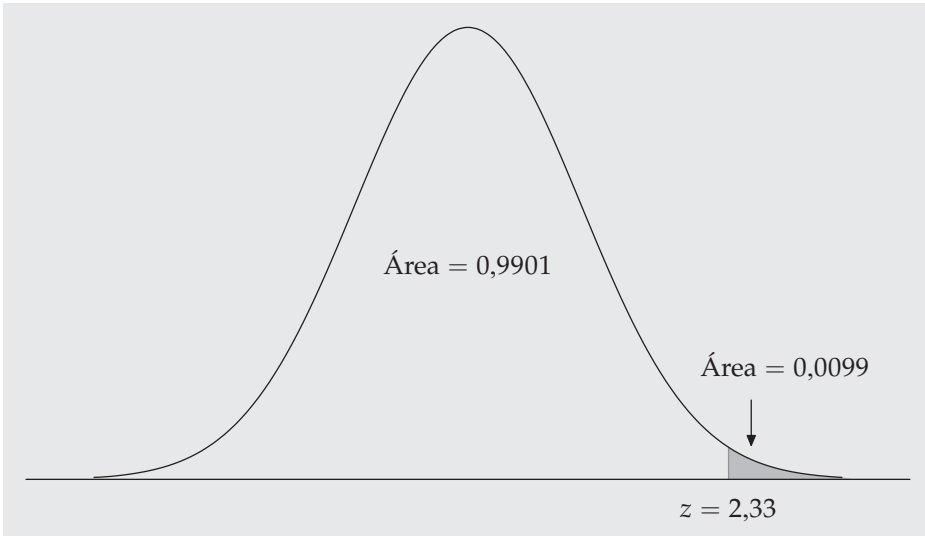


Figura 1.23. Las áreas por debajo de la curva normal estandarizada del ejemplo 1.16.

En una distribución normal, la proporción de observaciones con $x > 240$ es igual que la proporción de observaciones con $x \geq 240$. El área por debajo de la curva y exactamente encima de 240 es cero, por consiguiente, las áreas por debajo de la curva con $x > 240$ y $x \geq 240$ son iguales. Esto no es cierto en el caso de datos reales. Puede haber un chico con exactamente 240 mg/dl de colesterol en la sangre. La distribución normal es sólo una aproximación fácil de utilizar, no es una descripción de cada uno de los detalles de los datos reales.

Para que no te equivoques cuando utilices un programa estadístico o cuando uses la tabla A para hacer cálculos con distribuciones normales, te aconsejamos que hagas un dibujo de una normal y señales el área que quieres; luego, fíjate en el área que te da la tabla o el programa estadístico. He aquí un ejemplo.

EJEMPLO 1.17. Más sobre cálculos con distribuciones normales

¿Qué porcentaje de chicos de 14 años tienen un nivel de colesterol en la sangre entre 170 y 240 mg/dl?

1. *Plantea el problema*. Estamos interesados en conocer la proporción de chicos con $170 \leq x \leq 240$.
2. *Estandariza*:

$$\begin{array}{rcccl} 170 & \leq & x & \leq & 240 \\ \dfrac{170-170}{30} & \leq & \dfrac{x-170}{30} & \leq & \dfrac{240-170}{30} \\ 0 & \leq & z & \leq & 2{,}33 \end{array}$$

   La figura 1.24 nos muestra el área por debajo de la curva normal estandarizada.
3. *Utiliza la tabla*. El área entre 2,33 y 0 es el área a la izquierda de 2,33 *menos* el área situada a la izquierda de 0. Mira la figura 1.24 para comprobarlo. De la tabla A tenemos,

$$\begin{array}{rcl} \text{área entre 0 y 2,33} & = & \text{área a la izquierda de 2,33} - \text{área a la izquierda de 0,00} \\ & = & 0{,}9901 - 0{,}5000 = 0{,}4901 \end{array}$$

   Un 49% de los chicos tiene niveles de colesterol entre 170 y 240 mg/dl. ■

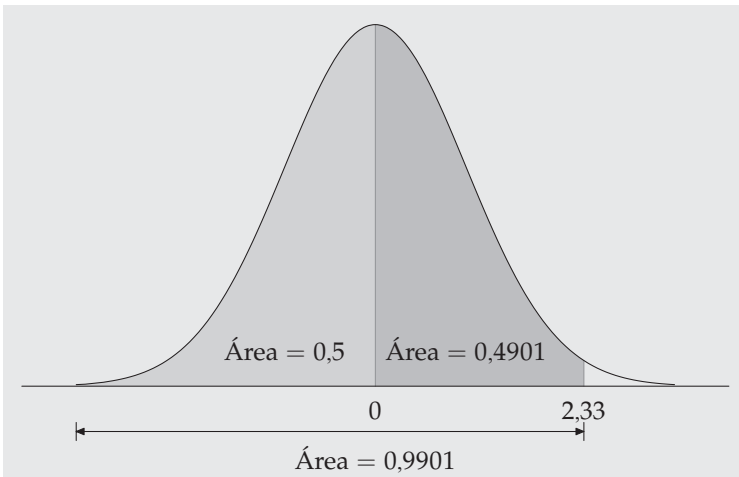


Figura 1.24. Las áreas por debajo de la curva normal estandarizada del ejemplo 1.17.

¿Qué habría ocurrido si hubiéramos encontrado una z que quedara fuera del intervalo de valores cubierto por la tabla A? Los valores z de la tabla A sólo dejan un área de 0,0002 en cada una de las colas. A efectos prácticos podemos considerar que el área por debajo de la curva normal estandarizada fuera del intervalo de valores z cubierto por la tabla es 0.

APLICA TUS CONOCIMIENTOS

1.57. Utiliza la tabla A para hallar proporciones de observaciones a partir de la distribución normal estandarizada que satisfagan cada una de las afirmaciones siguientes. En cada caso, dibuja la curva normal estandarizada y sombrea el área por debajo de la curva que corresponda.

(a) $z < 2{,}85$

(b) $z > 2{,}85$

(c) $z > -1{,}66$

(d) $-1{,}66 < z < 2{,}85$

1.58. Fuerza de tiro de una locomotora. Una importante medida del comportamiento de una locomotora es su adherencia, que es su fuerza de tiro expresada como un múltiplo de su peso. La adherencia de un determinado modelo de locomotora Diesel de 4.400 caballos de potencia varía según una distribución normal de media $\mu = 0{,}37$ y desviación típica $\sigma = 0{,}04$.

(a) ¿Qué proporción de adhesiones, expresadas como se ha comentado anteriormente, son mayores de 0,40?

(b) ¿Qué proporción de adhesiones se hallan entre 0,40 y 0,50?

(c) Las mejoras en el control informático de las locomotoras cambian la distribución normal de manera que $\mu = 0{,}41$ y $\sigma = 0{,}02$. Teniendo en cuenta estas mejoras, halla las proporciones en (a) y (b).

1.4.7 Cómo hallar un valor dada una proporción

Los ejemplos 1.16 y 1.17 ilustran la utilización de la tabla A para hallar la proporción de observaciones que satisfacen una determinada condición como, por ejemplo, que "el contenido de colesterol en la sangre esté entre 170 y 240 mg/dl". En cambio, nos podría interesar hallar el valor observado con una proporción dada de observaciones menores o mayores. Para ello, utiliza la tabla A en dirección

contraria. Halla la proporción dada en el cuerpo central de la tabla, lee la z correspondiente a partir de la columna de la izquierda y de la fila superior, luego efectúa la operación contraria a la realizada al estandarizar para obtener el valor observado. He aquí un ejemplo.

EJEMPLO 1.18. Cálculos normales 'hacia atrás'

Las notas de Lengua en la prueba SAT (*Scholastic Assessment Test*) de acceso a la universidad de los estudiantes de secundaria estadounidenses tienen aproximadamente una distribución $N(505, 110)$. ¿Cuál debe ser la nota de un alumno para pertenecer al 10% de estudiantes que tienen mejores notas?

1. *Plantea el problema*. Queremos hallar la nota x con un área a su *derecha* de 0,1 por debajo de una curva normal de media $\mu = 505$ y desviación típica $\sigma = 110$. Es lo mismo que hallar la nota SAT x con un área de 0,9 a su *izquierda*. La figura 1.25 plantea el problema de forma gráfica. Como la tabla A sólo da las áreas situadas a la izquierda de los valores z, plantea siempre el problema en términos del área situada a la izquierda de x.
2. *Utiliza la tabla*. Busca en el cuerpo central de la tabla A el valor más cercano a 0,9. Es 0,8997. Este es el valor correspondiente a $z = 1{,}28$. Por tanto, $z = 1{,}28$ es el valor estandarizado con un área de 0,9 a su izquierda.
3. *Desestandariza* para expresar el valor z como un valor x de la distribución normal correspondiente. Sabemos que el valor estandarizado de la x desconocida es $z = 1{,}28$. Por tanto, x satisface

$$\frac{x - 505}{110} = 1{,}28$$

Despejando x en la ecuación, tenemos:

$$x = 505 + (1{,}28)(110) = 645{,}8$$

Esta ecuación tienen sentido: expresa que x se halla a 1,28 desviaciones típicas a la derecha de la media de esta distribución normal. Éste es el significado del valor de $z = 1{,}28$ "desestandarizado". Vemos que el estudiante debe tener una puntuación de al menos 646 para estar entre el 10% de los estudiantes mejores. ■

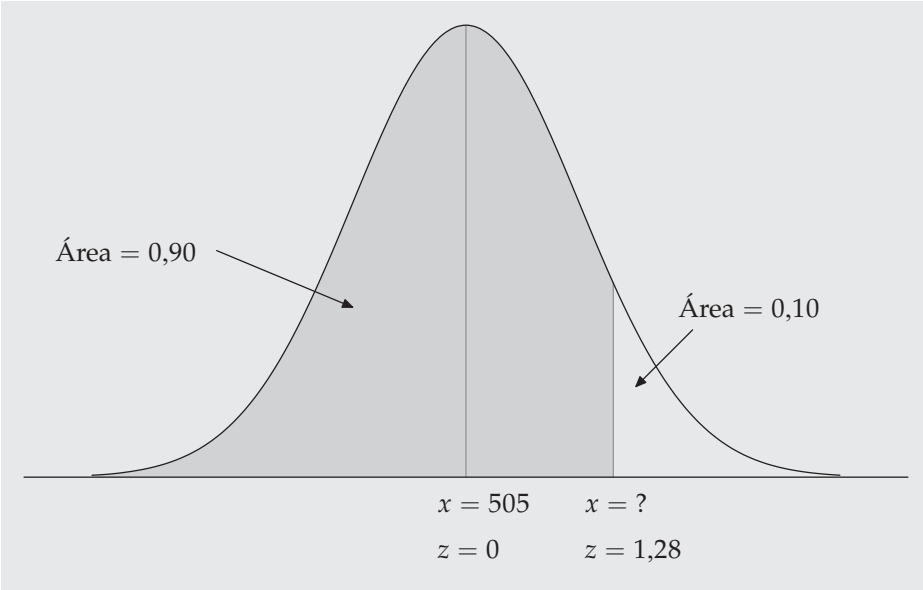


Figura 1.25. Localización del punto de una curva normal estandarizada con un área de 0,10 a su derecha.

He aquí la fórmula general para desestandarizar un valor z. Para hallar el valor x de la distribución normal con media μ y desviación típica σ correspondiente a un valor normal estandarizado z, utiliza

$$x = \mu + z\sigma$$

APLICA TUS CONOCIMIENTOS

1.59. Utiliza la tabla A para hallar el valor de z de una distribución normal estandarizada que cumpla cada una de las siguientes condiciones. (Utiliza el valor z de la tabla A que satisfaga de forma más aproximada la condición.) En cada caso, dibuja una normal y sitúa tu valor z en el eje de las abscisas.

(a) El valor z tal que el 25% de las observaciones sean menores.

(b) El valor z tal que el 40% de las observaciones sean mayores.

1.60. Coeficientes de inteligencia. Los coeficientes de una prueba de inteligencia (*Wechsler Adult Intelligence Scale*) para un grupo de adultos entre 20 y 34 años tienen una distribución aproximadamente normal de media $\mu = 110$ y desviación típica $\sigma = 25$.

(a) ¿Qué porcentaje de personas entre 20 y 34 años tiene un coeficiente de inteligencia mayor que 100?

(b) ¿Qué valor del coeficiente de inteligencia es necesario para estar entre el 25% que obtiene peores resultados?

(c) ¿Qué valor del coeficiente de inteligencia es necesario para estar entre el 5% que obtienen mejores resultados?

RESUMEN DE LA SECCIÓN 1.4

Algunas veces podemos describir el aspecto general de una distribución mediante una **curva de densidad**. Ésta siempre tiene un área total por debajo igual a 1. El área por debajo de una curva de densidad da la proporción de observaciones situadas en el intervalo seleccionado.

Una curva de densidad es una descripción idealizada de la forma general de una distribución que suaviza las irregularidades de los datos reales. Escribe la media de una curva de densidad como μ y la desviación típica de una curva de densidad como σ, para distinguirlas de la media $\bar{x}$ y de la desviación típica s de los datos reales.

La media, la mediana y los cuartiles de una curva de densidad se pueden localizar a simple vista. La **media μ** es el punto de equilibrio de la curva. La **mediana** divide el área por debajo de la curva en dos mitades iguales. Los **cuartiles**, juntamente con la mediana, dividen el área por debajo de la curva en cuartos. Para la mayoría de las curvas de densidad, la **desviación típica σ** no se puede localizar a simple vista.

En una curva de densidad simétrica, la media y la mediana coinciden. En una curva asimétrica la media está situada más hacia la cola larga que la mediana.

Las **distribuciones normales** se describen mediante una familia especial de curvas de densidad simétricas, en forma de campana, llamadas **curvas normales**. La media μ y la desviación típica σ caracterizan completamente una distribución normal $N(\mu, \sigma)$. La media es el centro de la curva y σ es la distancia a ambos lados de μ en la que la curva presenta una inflexión.

Para **estandarizar** cualquier observación x, réstale la media de la distribución y divide el resultado por su desviación típica. El **valor z** resultante

$$z = \frac{x - \mu}{\sigma}$$

nos dice a cuántas desviaciones típicas se halla x de la media.

Todas las distribuciones normales son la misma cuando las mediciones se expresan en unidades estandarizadas. En particular, todas las distribuciones normales satisfacen la **regla del 68-95-99,7**, que describe qué porcentajes de observaciones se encuentran a menos de una, dos o tres desviaciones típicas de la media.

Si x tiene la distribución $N(\mu, \sigma)$, entonces **la variable estandarizada** $z = \frac{(x-\mu)}{\sigma}$ tiene la **distribución normal estandarizada** $N(0, 1)$ de media 0 y desviación típica 1. La tabla A da la proporción de observaciones normales estandarizadas que son menores que z, para muchos valores de z. Estandarizando, podemos utilizar la tabla A para cualquier distribución normal.

EJERCICIOS DE LA SECCIÓN 1.4

1.61. La figura 1.26 muestra dos curvas normales, ambas con media 0. ¿Podrías decir cuánto valen aproximadamente las desviaciones típicas de estas curvas?

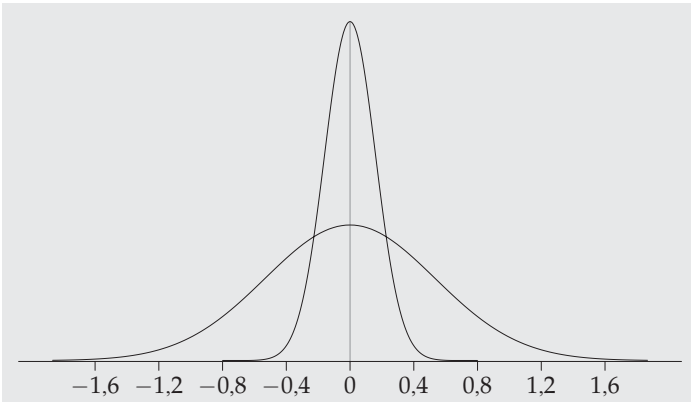


Figura 1.26. Dos curvas normales con la misma media pero con desviaciones típicas distintas. Para el ejercicio 1.61.

1.62. Perímetros craneales de los soldados. Según fuentes del ejército de EE UU los perímetros craneales de los soldados tienen una distribución normal con media 57,9 cm y desviación típica 2,8 cm. Utiliza la regla del 68-95-99,7 para responder a las siguientes preguntas:

(a) ¿Qué porcentaje de soldados tiene un perímetro craneal mayor de 60,7 cm?

(b) ¿Qué porcentaje de soldados tiene un perímetro craneal situado entre 55,1 y 60,7 cm?

1.63. Duración del embarazo. La duración del embarazo humano desde la fecundación del óvulo hasta el parto varía de acuerdo con una distribución aproximadamente normal, con una media de 266 días y una desviación típica de 16. Utiliza la regla del 68-95-99,7 para responder a las siguientes preguntas.

(a) ¿Entre qué valores se encuentra la duración del embarazo del 95% central de la población?

(b) ¿Qué duración tiene el 2,5% de los embarazos más cortos?

1.64. Tres grandes récords. Tres célebres récords del mundo del béisbol son las medias de bateos de 0,420 de Ty Cobb en 1911, de 0,406 de Ted Williams en 1941 y la de George Brett de 0,390 en 1980. Estas medias de bateos no se pueden comparar directamente porque la forma de las distribuciones ha ido cambiando a lo largo de los años. Las distribuciones de las medias de bateos en los distintos años son bastante simétricas y razonablemente normales. Mientras la media de las distribuciones se ha mantenido relativamente constante en los últimos años, las desviaciones típicas de estas distribuciones ha ido disminuyendo. He aquí los resultados:

Década	Media	Desviación típica
Años diez	0,266	0,0371
Años cuarenta	0,267	0,0326
Años setenta	0,261	0,0317

Calcula los valores estandarizados de las medias de bateos de Cobb, de Williams y de Brett para determinar cómo se encuentran situados entre sí.[14]

1.65. Utiliza la tabla A para hallar la proporción de observaciones de una distribución normal estandarizada que se sitúan en cada una de las siguientes regiones. En cada caso, dibuja la distribución normal y sombrea el área correspondiente a cada región.

(a) $z \leq -2{,}25$

(b) $z \geq 2{,}25$

(c) $z > 1{,}77$

(d) $-2{,}25 < z < 1{,}77$

[14]Stephen Jay Gould, "Entropic homogeneity isn't why no one hits 400 any more", *Discover*, agosto 1986, págs. 60-66.

1.66. (a) Halla el número z tal que la proporción de observaciones menores que z en una distribución normal estandarizada sea 0,8.

(b) Halla el número z tal que un 35% de las observaciones de una distribución normal estandarizada sea mayor que z.

1.67. Mercado bursátil. La tasa de rendimiento anual de las acciones tiene una distribución aproximadamente normal. Desde 1945, esas tasas de rendimiento de los 500 valores que componen el índice Standard & Poor's tienen una media anual del 12% y una desviación típica del 16,5%. Tomando los datos anteriores como referencia para un periodo plurianual bastante largo, responde a las siguientes preguntas:

(a) ¿En qué intervalo hallamos el 95% central de las tasas de rendimiento anuales?

(b) Se considera que la Bolsa está en crisis si la tasa de rendimiento es menor que 0. ¿En qué porcentaje de años la Bolsa está en crisis?

(c) ¿En qué porcentaje de años el índice gana al menos el 25%?

1.68. Duración del embarazo. La duración del embarazo humano desde la fecundación del óvulo hasta el parto tiene una distribución aproximadamente normal, con una media de 266 días y una desviación típica de 16 días.

(a) ¿Qué porcentaje de embarazos dura menos de 240 días (aproximadamente 8 meses)?

(b) ¿Qué porcentaje de embarazos tiene una duración comprendida entre 240 y 270 días (de una manera aproximada entre 8 y 9 meses)?

(c) ¿Qué duración tiene el 20% de los embarazos más largos?

1.69. Los niños de ahora, ¿son más inteligentes? Cuando la prueba de inteligencia de Stanford-Binet (prueba IQ) se empezó a utilizar en 1932, la prueba se ajustó de manera que los resultados de cada grupo de edad de niños tuviera aproximadamente una distribución normal de media $\mu = 100$ y desviación típica $\sigma = 15$. La prueba se reajusta de regularmente para que la media se mantenga en 100. Si hoy los niños de EE UU pasaran la prueba de 1932, su media sería de 120. La explicación de este aumento en el coeficiente de inteligencia a lo largo del tiempo desconoce; de todas formas podría ser debido a una mejor alimentación en la infancia y a más experiencia de los niños a la hora de pasar este tipo de pruebas.[15]

[15]Ulric Neisser, "Rising scores on intelligence tests", *American Scientist*, septiembre-octubre 1997.

(a) A menudo, los coeficientes de inteligencia superiores a 130 se califican como "muy superiores". ¿Qué porcentaje de niños tienen coeficientes muy superiores?

(b) Si los niños de hoy pasaran la prueba de 1932, ¿qué porcentaje de niños tendría coeficientes muy superiores? (Supón que la desviación típica $\sigma = 15$ se mantiene constante.)

1.70. La mediana de cualquier distribución normal es igual a su media. Podemos utilizar los cálculos normales para hallar los cuartiles de una distribución normal.

(a) ¿Cuál es el área por debajo de la curva normal estandarizada situada a la izquierda del primer cuartil? Utiliza este resultado para hallar el valor del primer cuartil de una distribución normal estandarizada. De forma similar, halla el tercer cuartil.

(b) El resultado que obtuviste en (a) te proporciona el valor z de los cuartiles de cualquier distribución normal. ¿Cuáles son los cuartiles de la duración del embarazo humano? (Utiliza la distribución del ejercicio 1.68.)

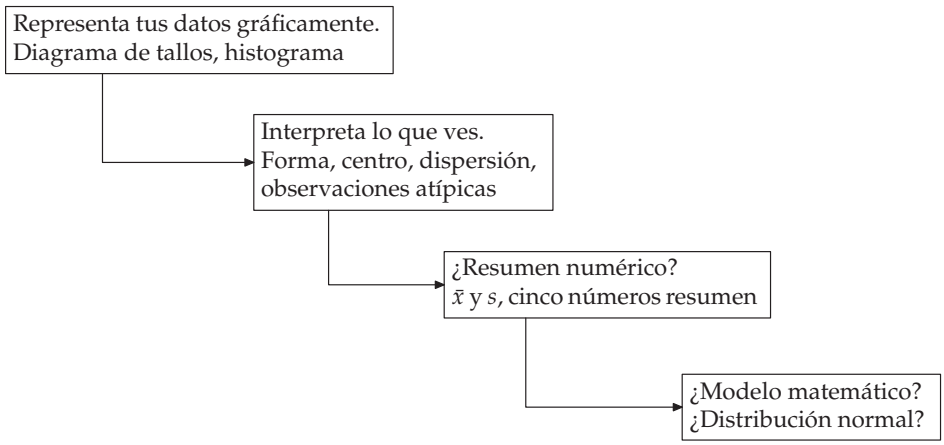
1.71. Los *deciles* de cualquier distribución son los puntos que señalan el 10% de las observaciones menores y el 10% de las mayores. Los deciles de una curva de densidad son, por tanto, los puntos a la izquierda de los cuales hay un área de 0,1 y de 0,9 por debajo de la curva.

(a) ¿Cuáles son los deciles de una distribución normal estandarizada?

(b) La altura de las mujeres tiene aproximadamente una distribución normal con media 1,64 m y desviación típica 0,06 m. ¿Cuáles son los deciles de esta distribución?

REPASO DEL CAPÍTULO 1

El análisis de datos es el arte de describir los datos utilizando gráficos y resúmenes numéricos. El propósito del análisis de datos es describir las características más importantes de un conjunto de datos. Este capítulo introduce el análisis de datos presentando las ideas y las herramientas estadísticas necesarias para describir la distribución de una sola variable. La siguiente figura te ayudará a organizar estas ideas tan importantes. Las preguntas que aparecen en los dos últimos cuadros de la figura nos recuerdan que la utilidad de los resúmenes numéricos y de los modelos tales como las distribuciones normales depende de lo que hallemos cuando analizamos los datos usando gráficos. He aquí una lista de repaso de



las habilidades más importantes que debes haber adquirido durante el estudio de este capítulo.

A. DATOS

1. Identificar los individuos y las variables de un conjunto de datos.
2. Identificar si las variables son categóricas o cuantitativas. Identificar las unidades de medida de las variables cuantitativas.

B. REPRESENTACIONES GRÁFICAS

1. Dibujar un diagrama de barras correspondiente a una variable categórica. Interpretar los diagramas de barras y de sectores.
2. Dibujar un histograma de la distribución de una variable cuantitativa.
3. Dibujar un diagrama de tallos de la distribución de un conjunto pequeño de observaciones. Cuando sea necesario, redondea las hojas o divide los tallos para mejorar el gráfico.

C. ANÁLISIS DE UNA DISTRIBUCIÓN (VARIABLES CUANTITATIVAS)

1. Identificar la forma de una distribución y las desviaciones más importantes.
2. Valorar a partir de un diagrama de tallos, o de un histograma, si la forma

de una distribución es aproximadamente simétrica, claramente asimétrica o ninguna de las dos cosas. Identificar si la distribución tiene más de un pico.
3. Describir el aspecto general dando medidas numéricas de centro y dispersión, además de la descripción verbal de su forma.
4. Decidir qué medidas de centro o de dispersión son las más apropiadas: la media y la desviación típica (especialmente para distribuciones simétricas) o los cinco números resumen (especialmente para distribuciones asimétricas).
5. Identificar las observaciones atípicas.

D. GRÁFICOS TEMPORALES

1. Dibujar un gráfico temporal, situando el tiempo en el eje de las abscisas y los valores observados en el eje de las ordenadas.
2. Identificar tendencias u otros rasgos generales de los gráficos temporales.

E. MEDIDA DE CENTRO

1. Calcular la media $\bar{x}$ de un conjunto de datos.
2. Calcular la mediana M de un conjunto de observaciones.
3. Comprender que la mediana es más robusta (se ve menos afectada por las observaciones extremas) que la media. Saber que la asimetría de una distribución desplaza la media hacia la cola más larga.

F. MEDIDA DE DISPERSIÓN

1. Calcular los cuartiles Q_1 y Q_3 de un conjunto de datos.
2. Calcular los cinco números resumen y dibujar un diagrama de caja; valorar el centro, la dispersión, la simetría o la asimetría a partir de un diagrama de caja.
3. Calcular la desviación típica s de un conjunto de observaciones utilizando una calculadora.
4. Conocer las propiedades básicas de s: $s \geq 0$ siempre; $s = 0$ sólo cuando todas las observaciones son iguales; s aumenta a medida que aumenta la dispersión de los datos; s se mide en las mismas unidades que los datos originales; las observaciones atípicas y las asimetrías tiran fuertemente de s.

G. CURVAS DE DENSIDAD

1. Saber que las áreas por debajo de una curva de densidad representan proporciones de todas las observaciones y que el área total por debajo de una curva de densidad es 1.
2. Localizar de forma aproximada la mediana (punto de igualdad de áreas) y la media (punto de equilibrio) de una curva de densidad.
3. Saber que la media y la mediana se encuentran en el centro de una curva de densidad simétrica, y que la media se desplaza hacia la cola larga de una curva asimétrica.

H. DISTRIBUCIONES NORMALES

1. Reconocer la forma de una curva normal. Ser capaz de estimar el valor de la media y de la desviación típica en este tipo de curvas.
2. Utilizar la regla del 68-95-99,7 y la simetría para establecer qué porcentaje de las observaciones de una distribución se encuentra a menos de una, dos o tres desviaciones típicas de la media.
3. Hallar el valor estandarizado (valor z) de una observación. Interpretar los valores z. Saber que cualquier distribución normal se transforma en una normal estandarizada $N(0,1)$ cuando se estandariza.
4. Dada una variable normal de media μ y desviación típica σ, calcular la proporción de valores que son mayores o menores que un número determinado, o que se encuentran entre dos números.
5. Dada una variable con una distribución normal de media μ y desviación típica σ, calcular el punto tal que una determinada proporción de todos los valores sea mayor que dicho punto. Calcular también el punto tal que una determinada proporción de valores sea menor que dicho punto.

EJERCICIOS DE REPASO DEL CAPÍTULO 1

1.72. ¿Preferencias en la votación? Las preferencias políticas de los españoles dependen de la edad, de los ingresos y del sexo de los votantes. Una investigadora selecciona una amplia muestra de votantes. De cada uno de ellos, la investigadora registra el sexo, la edad, los ingresos y si votó al Partido Popular, al Partido Socialista o a otro partido en las últimas elecciones. De estas variables, ¿cuáles son categóricas y cuáles son cuantitativas?

1.73. Armas asesinas. El Anuario Estadístico de 1997 de los Estados Unidos, proporciona datos del FBI sobre asesinatos en 1995. En ese año, el 55,8% de todos los asesinatos se cometieron con pistolas, el 12,4% con otras armas de fuego, el 12,6% con armas blancas, el 5,9% con alguna parte del cuerpo (en general las manos y los pies) y el 4,5% con algún objeto contundente. Representa gráficamente estos datos. ¿Necesitas la categoría de "otros métodos"?

1.74. Nunca en domingo. En la provincia canadiense de Ontario se ha realizado un estudio estadístico sobre el funcionamiento del sistema de sanidad pública. Los diagramas de barras de la figura 1.27 proceden del estudio de los ingresos y las altas de los hospitales de Ontario.[16] Estos diagramas muestran el número de pacientes con problemas de corazón que fueron ingresados y dados de alta cada día de la semana durante un periodo de 2 años.

(a) Explica por qué no cabe esperar diferencias en el número de ingresos de pacientes con cardiopatías en los distintos días de la semana. ¿Es ésta una deducción correcta a partir de los datos que se aportan?

(b) Describe la distribución de las altas. ¿Existe alguna diferencia con la distribución de ingresos? ¿Cómo se puede explicar esta diferencia?

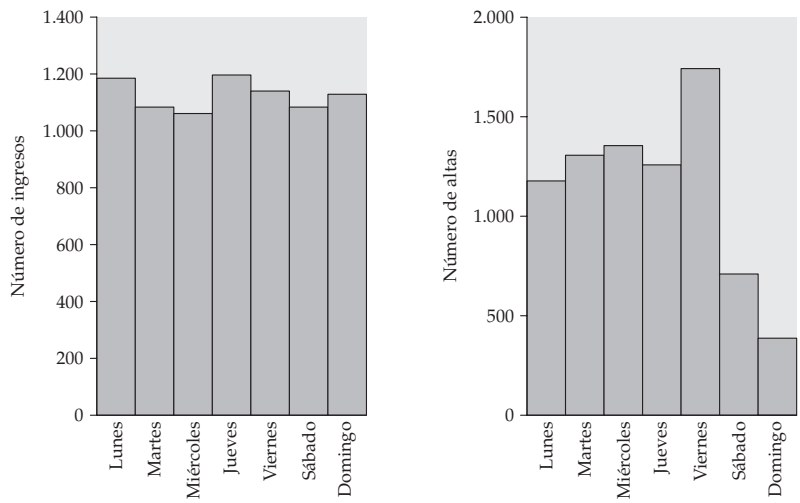


Figura 1.27. Diagramas de barras del número de ingresos y altas de pacientes con cardiopatías, cada día de la semana en los hospitales de Ontario, Canadá.

[16] Antoni Basinski, "Almost never on Sunday: implications of the patterns of admission and discharge for common conditions", Institute for Clinical Evaluative Sciences in Ontario, 18 octubre 1993.

1.75. De casa a la universidad. El profesor Moore, autor de este libro, que vive a unos kilómetros del campus universitario de la Universidad de Purdue, ha registrado durante 42 días el tiempo que tarda conduciendo desde su casa hasta la universidad. He aquí los tiempos (en minutos) correspondientes a 42 días consecutivos:

8,25	7,83	8,30	8,42	8,50	8,67	8,17	9,00	9,00	8,17	7,92
9,00	8,50	9,00	7,75	7,92	8,00	8,08	8,42	8,75	8,08	9,75
8,33	7,83	7,92	8,58	7,83	8,42	7,75	7,42	6,75	7,42	8,50
8,67	10,17	8,75	8,58	8,67	9,17	9,08	8,83	8,67		

(a) Dibuja un diagrama de tallos correspondiente a estos datos. (Redondea a décimas de minuto y divide los tallos.) La distribución de datos, ¿es aproximadamente simétrica, claramente asimétrica o ninguna de las dos cosas? ¿Existen observaciones atípicas?

(b) Dibuja un diagrama temporal con estos datos. (Marca en el eje de abscisas los días consecutivos desde el 1 hasta el 42.) El gráfico no deja entrever ninguna tendencia. De todas formas, se observa que hay un día en el que la duración del trayecto fue muy corta y dos días en que fue muy larga. Señala estos valores en tu gráfico.

(c) Las tres observaciones que se salen un poco de la distribución general se pueden explicar. El día que el profesor Moore tardó muy poco tiempo en llegar a la universidad corresponde al Día de Acción de Gracias (*Thanksgiving Day*), que en EE UU es festivo. Los dos días que tardó mucho más de lo normal corresponden, por un lado, a un día en el que ocurrió un accidente con las inevitables retenciones, y, por otro, a un día en el que se produjeron unas fuertes nevadas que dificultaron la conducción. Elimina estas tres observaciones y calcula la media y la desviación típica de las restantes 39 observaciones.

(d) Haz un recuento del número de observaciones entre $\bar{x} - s$ y $\bar{x} + s$, entre $\bar{x} - 2s$ y $\bar{x} + 2s$, y finalmente entre $\bar{x} - 3s$ y $\bar{x} + 3s$. Halla el porcentaje de observaciones en cada uno de los intervalos anteriores. Compara estos porcentajes con los que les correspondería de acuerdo con la regla del 68-95-99,7.

1.76. Rendimiento de las acciones. La tabla 1.8 proporciona los rendimientos mensuales de las acciones de Philip Morris para el periodo que va de julio de 1990 a mayo de 1997. Los datos se presentan ordenados cronológicamente empezando con $-5{,}7\%$, el rendimiento de julio de 1990. Dibuja un diagrama temporal con estos datos. Este periodo corresponde a una época de movilizaciones crecientes en contra del tabaco. Por tanto, cabe esperar una tendencia decreciente en los rendimientos de las acciones. Sin embargo, también aparece un periodo en el cual

el valor de las acciones crece de forma rápida. ¿Qué puede haber provocado esta tendencia creciente? ¿Qué muestra tu diagrama temporal?

1.77. Nueva variedad de maíz. El maíz es un alimento importante para los animales. De todas formas, este alimento carece de algunos aminoácidos que son esenciales. Un grupo de científicos desarrolló una nueva variedad que sí contenía dichos aminoácidos a niveles apreciables. Para comprobar el valor de esta nueva variedad para la alimentación animal se llevó a cabo el siguiente experimento: a un grupo de 20 pollos de un día se les suministró un pienso que contenía harina de maíz de la nueva variedad. A otro grupo de 20 pollos (grupo de control) se le alimentó con un pienso idéntico al anterior, aunque no contenía harina de la variedad mejorada de maíz. Los resultados que se obtuvieron sobre las ganancias de peso de los pollos (en gramos), al cabo de 21 días de alimentación, fueron los siguientes:[17]

Variedad normal				Variedad mejorada			
380	321	366	356	361	447	401	375
283	349	402	462	434	403	393	426
356	410	329	399	406	318	467	407
350	384	316	272	427	420	477	392
345	455	360	431	430	339	410	326

(a) Calcula los cinco números resumen correspondientes a la ganancia de peso de los dos grupos de pollos. Para comparar las dos distribuciones, representa los dos diagramas de caja en un mismo gráfico. ¿Qué se puede deducir de estos diagramas de caja?

(b) En el trabajo original donde aparecieron los datos, los autores calcularon las medias y las desviaciones típicas de cada grupo de pollos. ¿Cuáles son sus valores? ¿Qué diferencia hay entre las medias de cada grupo?

1.78. Alfredo Di Stefano. Antes de ir a España y fichar por el Real Madrid en la temporada 1952/53 y posteriormente por el Real Club Deportivo Español de Barcelona la temporada 1964/65, Alfredo Di Stefano jugó en varios equipos suramericanos: River Plate de Buenos Aires, Huracán de Buenos Aires y Millonarios de Bogotá.

[17] G. L. Cromwell *et al*., "A comparison of the nutritive value of *opaque-2*, *floury-2* and normal corn for the chick", *Poultry Science*, 57, 1968, págs. 840-847.

Mientras jugó en Suramérica el número de goles por temporada en la liga fue

Temporada	Goles	Temporada	Goles
1944/45	0	1948/49	24
1945/46	11	1949/50	23
1946/47	27	1950/51	32
1947/48	14	1951/52	19

Mientras jugó en España el número de goles por temporada en la liga fue

Temporada	Goles	Temporada	Goles
1953/54	28	1960/61	21
1954/55	25	1961/62	10
1955/56	24	1962/63	12
1956/57	31	1963/64	11
1957/58	19	1964/65	7
1958/59	23	1965/66	4
1959/60	12		

Calcula los cinco números resumen correspondientes al tiempo que Di Stefano jugó en Suramérica y al tiempo que jugó en España. Sitúa los dos diagramas de caja en un mismo gráfico y compara las dos distribuciones.

1.79. Los todoterrenos, ¿desperdician combustible? La tabla 1.2 da los consumos, en litros a los cien kilómetros, de 26 modelos de coches de tamaño medio de 1998. Aquí presentamos los consumos de 19 modelos de todoterreno de ese año.[18]

Modelo	Consumo (litros/100 km)	Modelo	Consumo (litros/100 km)
Acura SLX	12,5	Jeep Wrangler	12,5
Chevrolet Blazer	11,8	Land Rover	14,8
Chevrolet Tahoe	12,5	Mazda MPV	12,5
Dodge Durango	13,9	Mercedes-Benz ML320	11,3
Ford Expedition	13,1	Mitsubishi Montero	11,8
Ford Explorer	12,5	Nissan Pathfinder	12,5
Honda Passport	11,8	Suzuki Sidekick	9,1
Infiniti QX4	12,5	Toyota RAV4	9,1
Isuzu Trooper	12,5	Toyota 4Runner	10,8
Jeep Grand Cherokee	13,1		

[18] Véase nota 2.

(a) Describe gráfica y numéricamente los consumos en carretera de los 4×4. ¿Cuáles son las principales características de esta distribución?

(b) Dibuja diagramas de caja para comparar la distribución de los automóviles medianos con la de los 4×4. ¿Cuáles son las principales diferencias entre estas dos distribuciones?

1.80. Supervivencia de conejillos de Indias. En la tabla 1.9 se presentan los tiempos de supervivencia, en días, de 72 conejillos de Indias después de que se les inyectara el bacilo de la tuberculosis en un experimento médico.[19] La distribución de los tiempos de supervivencia, ya sea de máquinas (sobrecargadas), ya sea de personas enfermas (por ejemplo, personas que están bajo tratamiento oncológico), se suele caracterizar por ser asimétricas hacia la derecha.

Tabla 1.9. Tiempos de supervivencia (en días) de conejillos de Indias en un experimento médico.

43	45	53	56	56	57	58	66	67	73
74	79	80	80	81	81	81	82	83	83
84	88	89	91	91	92	92	97	99	99
100	100	101	102	102	102	103	104	107	108
109	113	114	118	121	123	126	128	137	138
139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511
522	598								

(a) Representa gráficamente estos datos y describe sus características más destacables. La distribución, ¿es asimétrica hacia la derecha?

(b) He aquí los resultados del programa estadístico Data Desk correspondientes a estos datos:

```
Summary statistics for dias

Mean 141.84722
Median 102.50000
Cases 72
StdDev 109.20863
Min 43
Max 598
25th%ile 82.250000
75th%ile 153.75000
```

[19] T. Bjerkedal, "Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli", *American Journal of Hygiene*, 72, 1960, págs. 130-148.

Explica cómo la relación entre la media y la mediana refleja la asimetría de los datos. ("Cases" significa número de observaciones, "25th%ile" significa primer cuartil, también se llama "25 th percentil", ya que el 25% de los datos quedan a la derecha de este valor. De forma similar "75 th%ile" es el tercer cuartil.)

(c) Calcula los cinco números resumen y explica brevemente cómo se puede detectar la asimetría de los datos a partir de ellos.

1.81. Acciones calientes. La tasa de rendimiento de una acción se deriva de la variación de su precio y de los dividendos pagados, y normalmente se expresa como un porcentaje respecto a su valor inicial. A continuación se presentan datos sobre las tasas de rendimiento mensuales de las acciones de los almacenes Wal-Mart desde el año 1973 hasta el año 1991. Tenemos un total de 228 observaciones. La figura 1.28 muestra los resultados de un programa estadístico que describe la distribución de estos datos. Fíjate en que el tallo está constituido por las decenas de los porcentajes. Las hojas están constituidas por las unidades. El diagrama de tallos divide los tallos para que la representación sea mejor. El programa proporciona las observaciones atípicas mayores y las menores de forma separada. No las incluye en el diagrama de tallos.

(a) Calcula los cinco números resumen de estos datos.

(b) Describe las principales características de la distribución.

(c) Si tuvieras 1.000 dólares en acciones de Wal-Mart al inicio del mejor mes de los 19 años considerados, ¿cuánto dinero habrías ganado al final del mes? Si tuvieras 1.000 dólares en acciones al comienzo del peor mes, ¿cuánto valdría tu dinero al final de dicho mes?

1.82. El criterio 1,5 × RI. Un criterio que puedes utilizar para detectar observaciones atípicas de un conjunto de datos es el siguiente:

1. Halla los cuartiles Q_1 y Q_3 y el **recorrido intercuartílico** $\text{RI} = Q_3 - Q_1$. El recorrido intercuartílico es la dispersión del 50% de los datos centrales.
2. Califica como atípica una observación si se sitúa más a la izquierda de $1{,}5 \times \text{RI}$ desde el primer cuartil o más a la derecha de $1{,}5 \times \text{RI}$ a partir del tercer cuartil.

Recorrido intercuartílico

Halla el recorrido intercuartílico RI correspondiente a los datos del ejercicio anterior. De acuerdo con el criterio que acabamos de ver, ¿existe alguna observación atípica? ¿Crees que este criterio es el mismo que utiliza el programa estadístico para seleccionar las observaciones atípicas?

```
Mean = 3.064
Standard deviation = 11.49

N = 228  Median = 3.4691
Quartiles = -2.950258, 8.4511

Decimal point is 1 place to the right of the colon

Low: -34.04255  -31.25000  -27.06271  -26.61290

 -1   :   985
 -1   :   444443322222110000
 -0   :   999988777666666665555
 -0   :   44444444333333322222222222221111111100
  0   :   00000111111111111222222333333344444444
  0   :   55555555555555555555566666666667777777888888888899999
  1   :   00000000011111111122233334444
  1   :   55566667889
  2   :   011334

High:  32.01923  41.80531  42.05607  57.89474  58.67769
```

Figura 1.28. Resultados de un programa estadístico que describe los rendimientos mensuales de las acciones de Walt-Mart. Para el ejercicio 1.81.

1.83. Rendimiento de acciones. ¿Crees que ha cambiado el rendimiento de las acciones de Wal-Mart en los 19 años que van desde 1973 hasta 1991? En el ejercicio 1.81 vimos la distribución de los 228 rendimientos mensuales. Este tipo de descripción no puede responder a preguntas sobre los cambios acaecidos a lo largo del tiempo. La figura 1.29 es un tipo de gráfico temporal. En lugar de representar todas las observaciones, éstas se presentan agrupadas por años en forma de diagramas de caja. Cada año tenemos 12 rendimientos mensuales.

(a) ¿Se observa alguna tendencia en los rendimientos mensuales típicos a lo largo de estos años?

(b) ¿Se observa alguna tendencia en la dispersión anual de los datos?

(c) El diagrama de tallos de la figura 1.28 señala algunas observaciones atípicas. ¿Cuáles de éstas se pueden detectar en los diagramas de caja? ¿En qué años

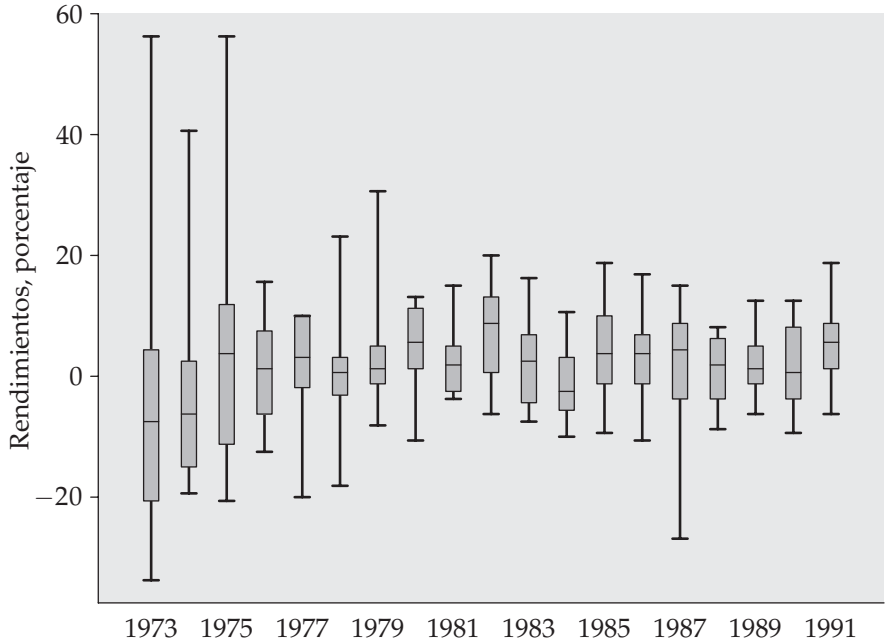


Figura 1.29. Diagramas de caja que permiten la comparación de las distribuciones mensuales de los rendimientos de las acciones de Wal-Mart durante 19 años.

ocurren? ¿Refuerza esto las conclusiones que has obtenido en el apartado (b)? ¿Hay alguna observación atípica especialmente sorprendente después de tener en cuenta tu respuesta en (b)?

1.84. Julia dice: "La gente ahora vive más años que antes, por tanto, es probable que los nuevos presidentes de EE UU sean mayores que los anteriores cuando acceden a la Casa Blanca". Juan responde: "No, a los votantes de ahora les gusta la juventud y no respetan a la gente mayor, por tanto, es probable que los presidentes de EE UU sean más jóvenes que hace unos años".

Dibuja un gráfico temporal con la edad de los presidentes de EE UU que tienes en la tabla 1.7. En el eje de las abscisas sitúalos desde el primero, que corresponde a Washington, hasta el que ocupa el lugar número 42, que corresponde a Clinton. ¿Se observa alguna tendencia a lo largo del tiempo? ¿A quién da la razón los datos, a Julia o a Juan?

1.85. Coste de la capacidad de los ordenadores. Los usuarios de informática saben que el coste de la potencia de los ordenadores ha ido disminuyendo de forma muy rápida. Por ejemplo, el coste de la capacidad, en megabytes, de los mayores discos duros del mercado de los ordenadores personales para Macintosh es[20]

Año	1992	1993	1994	1995	1996
Coste (€)	5,07	2,40	1,14	0,53	0,36

Estos costes se han ajustado de acuerdo con la inflación de cada año para facilitar su comparación. Dibuja un diagrama temporal con estos datos. Señala si observas alguna tendencia.

1.86. Grandes robles y pequeñas bellotas. De las 50 especies de roble de los EE UU, 28 crecen en la costa atlántica y 11 en la costa de California. Estamos interesados en la distribución del tamaño de las bellotas de los robles. He aquí datos sobre el volumen de bellotas (en centímetros cúbicos) de estas 39 especies de roble:

Atlántico							**California**		
1,4	3,4	9,1	1,6	10,5	2,5	0,9	4,1	5,9	17,1
6,8	1,8	0,3	0,9	0,8	2,0	1,1	1,6	2,6	0,4
0,6	1,8	4,8	1,1	3,0	1,1	1,1	2,0	6,0	7,1
3,6	8,1	3,6	1,8	0,4	1,1	1,2	5,5	1,0	

(a) Dibuja un histograma con los 39 volúmenes de bellota. Describe la distribución. Incluye un resumen numérico adecuado.

(b) Compara las distribuciones de las regiones atlántica y californiana con un gráfico y con resúmenes numéricos. ¿Qué has hallado?

1.87. La tabla 1.6 hace referencia a los Estados europeos. Existe mucha más información. Entra en la página web de la Comisión Europea o acércate a la biblioteca de la Comisión Europea que tengas más próxima y busca más datos estadísticos sobre los Estados europeos.

[20] Apareció en *MacWorld*, septiembre 1996, pág. 145.

(a) ¿Qué porcentaje representa la población ocupada en agricultura en cada Estado?

(b) Compara la inflación de los distintos Estados europeos. Representa gráficamente la información que hayas encontrado. Calcula los resúmenes numéricos más adecuados. ¿Cuáles son tus conclusiones?

1.88. Adopción de la cultura anglosajona. La prueba ARSMA (*Acculturation Rating Scale for Mexican Americans*) es una prueba psicológica que se utiliza para determinar el nivel de integración cultural de los estadounidenses de origen mexicano. Las puntuaciones posibles van de 1,0 a 5,0. Los valores más altos de esta escala corresponden a niveles más elevados de adaptación a la cultura anglosajona. Cuando se efectuó esta prueba con una población experimental, se observó que la distribución de las puntuaciones era aproximadamente normal con una media de 3,0 y una desviación típica de 0,8. Un investigador cree que los mexicanos recién llegados a EE UU tienen una puntuación media próxima a 1,7 y que la de la siguiente generación está próxima a 2,1. ¿Qué proporción de la población experimental tiene puntuaciones menores de 1,7? ¿Y entre 1,7 y 2,1?

1.89. Perímetro craneal de soldados. Según datos del ejército de EE UU, la distribución del perímetro craneal entre sus soldados es aproximadamente normal con una media de 57,9 cm y una desviación típica de 2,8 cm. Los cascos militares se producen de forma industrial excepto para los soldados con perímetros craneales situados en el 5% superior o bien en el 5% inferior, para los cuales se hacen a medida. ¿Para qué perímetros craneales se hacen estos cascos a medida?

1.90. Adopción de la cultura anglosajona. El ejercicio 1.88 describió la prueba ARSMA. ¿Cuál debe ser el resultado de un estadounidense de origen mexicano para pertenecer al 30% de la población experimental que obtuvo mejores resultados en la prueba? ¿Qué resultados definen el 30% para los cuales la cultura mexicano-española tiene un mayor peso?