

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.

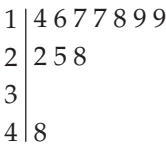


Figura 1.10. Distribución del número de goles marcados por temporada por Ladislao Kubala mientras fue jugador del F.C. Barcelona.

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 $x_1, x_2, \dots, x_n$, su media es

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

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

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

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 universitarios 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 Alcántara 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

Mínima Q_1 M Q_3 Máxima

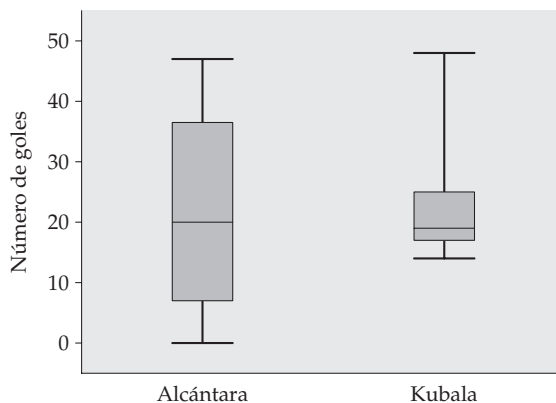


Figura 1.11. Diagramas de caja en un mismo gráfico para comparar el número de goles por temporada de Paulino Alcántara y de Ladislao Kubala.

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 $x_1, x_2, \dots, x_n$ es

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

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 $n - 1$.

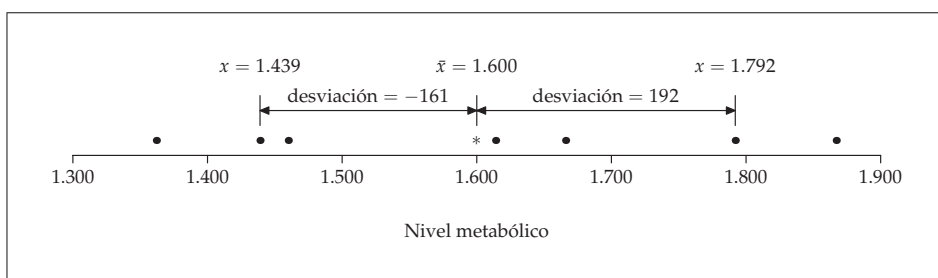


Figura 1.12. Niveles metabólicos de siete hombres, con la media (*) y las desviaciones de dos observaciones respecto a la media.

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 Kuba 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

¹¹S. 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.