Curso Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

Licencia creative commons

Estadísticos muestrales

La tabla de frecuencias sintetiza la información de la variable estudiada en la muestra, pero en muchas ocasiones es insuficiente para describir determinados aspectos de la distribución.

Para describir adecuadamente el comportamiento de la variable se calculan unas medidas llamadas **estadísticos muestrales** que son indicadores de distintos aspectos de la distribución muestral.

Los estadísticos se clasifican en tres grupos:

Estadísticos de Posición: Miden en torno a qué valores se agrupan los datos y cómo se reparten en la distribución.

Estadísticos de Dispersión: Miden la heterogeneidad de los datos.

Estadísticos de Forma: Miden aspectos de la forma que tiene la distribución de los datos, como la simetría o el apuntamiento.

Estadísticos de posición

Pueden ser de dos tipos:

Estadísticos de Tendencia Central: Determinan valores alrededor de los cuales se agrupa la distribución. Estas medidas suelen utilizarse como valores representativos de la muestra. Las más importantes son:

- Media aritmética
- Mediana
- Moda

Otros estadísticos de Posición: Dividen la distribución en partes con el mismo número de observaciones. Las más importantes son:

- Cuantiles: Cuartiles, Deciles, Percentiles.

Media aritmética

Definición (Media aritmética muestral $\bar{x}$)

La *media aritmética muestral* de una variable X es la suma de los valores observados en la muestra dividida por el tamaño muestral

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

A partir de la tabla de frecuencias puede calcularse como:

$$\bar{x} = \frac{\sum x_i n_i}{n} = \sum x_i f_i$$

En la mayoría de los casos, la media aritmética es la medida que mejor representa a la muestra.

¡Ojo! No puede calcularse para atributos.

En el ejemplo anterior del número de hijos tenemos

$$\bar{x} = \frac{1+2+4+2+2+2+3+2+1+1+0+2+2}{25} + $$
$$+ \frac{0+2+2+1+2+2+3+1+2+2+1+2}{25} = \frac{44}{25} = 1{,}76 \text{ hijos.}$$

o bien, desde la tabla de frecuencias

x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
0	2	0,08	0	0
1	6	0,24	6	0,24
2	14	0,56	28	1,12
3	2	0,08	6	0,24
4	1	0,04	4	0,16
$\sum$	25	1	44	1,76

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{44}{25} = 1{,}76 \qquad \bar{x} = \sum x_i f_i = 1{,}76.$$

Es decir, el número de hijos que mejor representa a la muestra es $1{,}76$ hijos.

Cálculo de la media aritmética

Ejemplo con datos agrupados

En el ejemplo anterior de las estaturas se tiene

$$\bar{x} = \frac{179 + 173 + \cdots + 187}{30} = 175{,}07 \text{ cm.}$$

o bien, desde la tabla de frecuencias utilizando las marcas de clase:

X	x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
$(150, 160]$	155	2	0,07	310	10,33
$(160, 170]$	165	8	0,27	1320	44,00
$(170, 180]$	175	11	0,36	1925	64,17
$(180, 190]$	185	7	0,23	1295	43,17
$(190, 200]$	195	2	0,07	390	13
$\sum$		30	1	5240	174,67

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{5240}{30} = 174{,}67 \qquad \bar{x} = \sum x_i f_i = 174{,}67.$$

Al agrupar datos el cálculo de estadísticos desde la tabla puede diferir ligeramente del valor real obtenido directamente desde la muestra, ya que no se trabaja con los datos reales sino con los representantes de las clases.

Media ponderada

En algunos casos, los valores de la muestra no tienen la misma importancia. En este caso la media aritmética no es una buena medida de representatividad ya que en ella todos los valores de la muestra tienen el mismo peso. En este caso es mucho mejor utilizar otra medida de tendencia central conocida como media ponderada.

Definición (Media ponderada muestral $\bar{x}_p$)

Dada una muestra de n valores en la que cada valor x_i tiene asociado un peso p_i, la *media ponderada muestral* de la variable X es la suma de los productos de cada valor observado en la muestra por su peso, dividida por la suma de todos los pesos

$$\bar{x}_p = \frac{\sum x_i p_i}{\sum p_i}$$

A partir de la tabla de frecuencias puede calcularse como:

$$\bar{x}_p = \frac{\sum x_i p_i n_i}{\sum p_i}$$

Cálculo de la media ponderada

Supongase que un alumno quiere calcular la nota media de las asignaturas de un curso.

Asignatura	Créditos	Nota
Matemáticas	6	5
Lengua	4	3
Química	8	6

La media aritmética vale

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5+3+6}{3} = 4{,}67 \text{ puntos},$$

Sin embargo, esta nota no representa bien el rendimiento académico del alumno ya que en ella han tenido igual peso todas las asignaturas, cuando la química debería tener más peso que la lengua al tener más créditos.

Es más lógico calcular la media ponderada, tomando como pesos los créditos de cada asignatura:

$$\bar{x}_p = \frac{\sum x_i p_i}{\sum p_i} = \frac{5\cdot 6 + 3\cdot 4 + 6\cdot 8}{6+4+8} = \frac{90}{18} = 5 \text{ puntos}.$$

Mediana

Definición (Mediana muestral Me)

La *mediana muestral* de una variable X es el valor de la variable que, una vez ordenados los valores de la muestra de menor a mayor, deja el mismo número de valores por debajo y por encima de él.

La mediana cumple $N_{Me} = n/2$ y $F_{Me} = 0{,}5$.

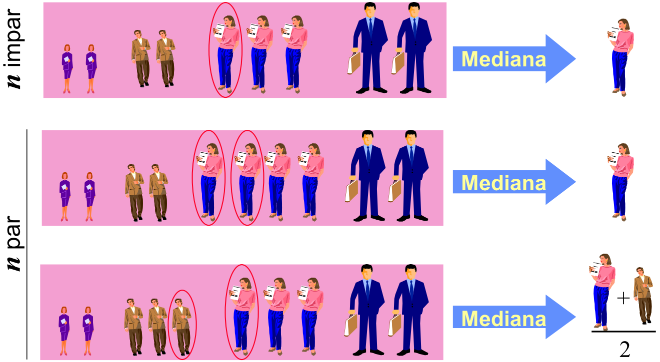
El cálculo de la mediana se realiza de forma distinta según se hayan agrupado los datos o no.

¡Ojo! No puede calcularse para atributos nominales.

Cálculo de la mediana con datos no agrupados

Con datos no agrupados pueden darse varios casos:

- Tamaño muestral impar: La mediana es el valor que ocupa la posición $\frac{n+1}{2}$.
- Tamaño muestral par: La mediana es la media de los valores que ocupan las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$.



Cálculo de la mediana

Ejemplo con datos no agrupados

En el ejemplo anterior del número de hijos, el tamaño muestral es 25, de manera que al ser impar se deben ordenar los datos de menor a mayor y buscar el que ocupa la posición $\frac{25+1}{2} = 13$.

$$0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, \boxed{2}, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4$$

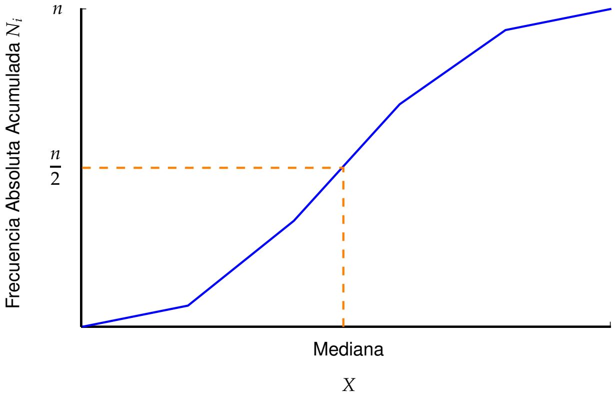
y la mediana es 2 hijos.

Si se trabaja con la tabla de frecuencias, se debe buscar el primer valor cuya frecuencia absoluta acumulada iguala o supera a 13, que es la posición que le corresponde a la mediana, o bien el primer valor cuya frecuencia relativa acumulada iguala o supera a $0{,}5$:

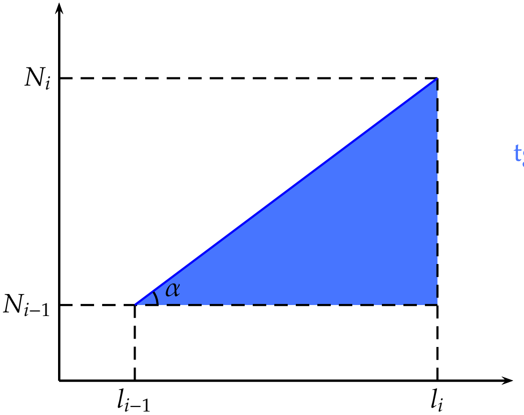
x_i	n_i	f_i	N_i	F_i
0	2	0,08	2	0,08
1	6	0,24	8	0,32
$\boxed{2}$	14	0,56	22	0,88
3	2	0,08	24	0,96
4	1	0,04	25	1
$\sum$	25	1		

Cálculo de la mediana con datos agrupados

Con datos agrupados la mediana se calcula interpolando en el polígono de frecuencias absolutas acumuladas para el valor $n/2$.

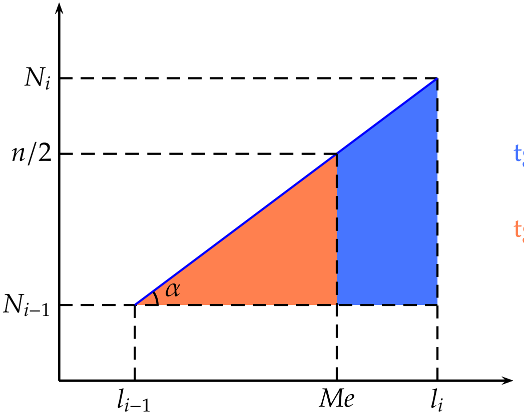


Interpolación en el polígono de frecuencias absolutas acumuladas



$$\mathrm{tg}(\alpha) = \frac{N_i - N_{i-1}}{l_i - l_{i-1}}$$

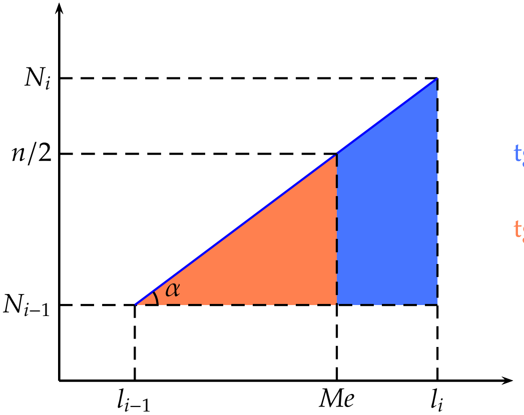
Interpolación en el polígono de frecuencias absolutas acumuladas



$$\mathrm{tg}(\alpha) = \frac{N_i - N_{i-1}}{l_i - l_{i-1}}$$

$$\mathrm{tg}(\alpha) = \frac{n/2 - N_{i-1}}{Me - l_{i-1}}$$

Interpolación en el polígono de frecuencias absolutas acumuladas



$$\operatorname{tg}(\alpha) = \frac{N_i - N_{i-1}}{l_i - l_{i-1}}$$

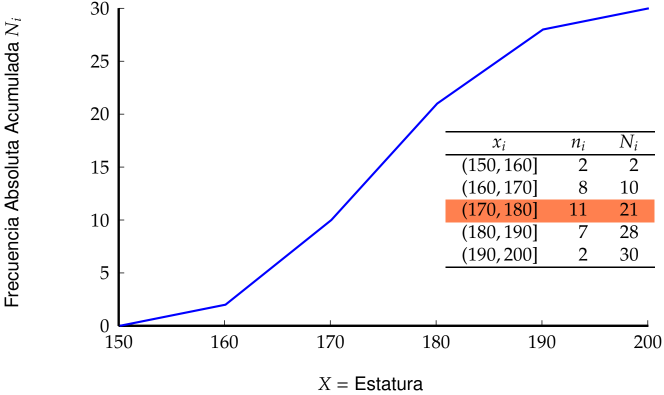
$$\operatorname{tg}(\alpha) = \frac{n/2 - N_{i-1}}{Me - l_{i-1}}$$

$$Me = l_{i-1} + \frac{n/2 - N_{i-1}}{N_i - N_{i-1}}(l_i - l_{i-1}) = l_{i-1} + \frac{n/2 - N_{i-1}}{n_i} a_i$$

Cálculo de la mediana

Ejemplo con datos agrupados

En el ejemplo de las estaturas $n/2 = 30/2 = 15$. Si miramos en el polígono de frecuencias acumuladas comprobamos que la mediana caerá en el intervalo $(170, 180]$.

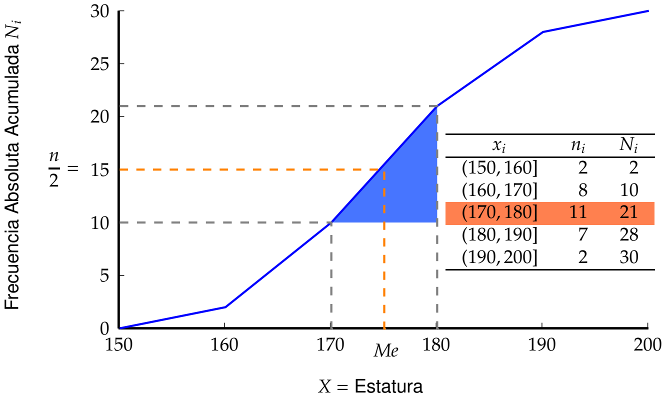


x_i	n_i	N_i
$(150, 160]$	2	2
$(160, 170]$	8	10
$(170, 180]$	11	21
$(180, 190]$	7	28
$(190, 200]$	2	30

Cálculo de la mediana

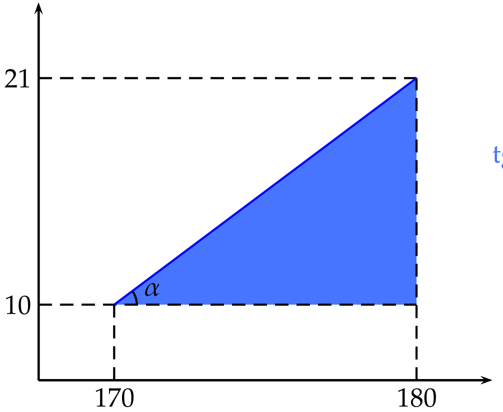
Ejemplo con datos agrupados

En el ejemplo de las estaturas $n/2 = 30/2 = 15$. Si miramos en el polígono de frecuencias acumuladas comprobamos que la mediana caerá en el intervalo $(170, 180]$.



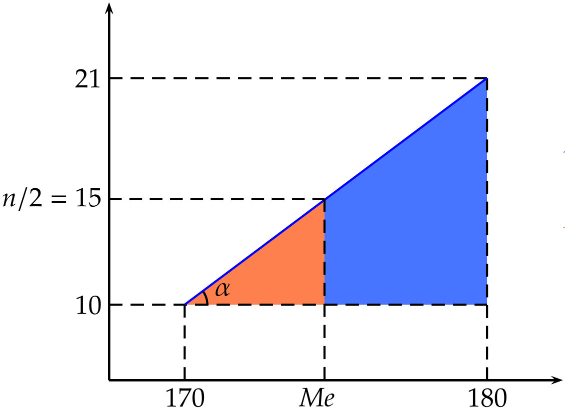
x_i	n_i	N_i
$(150, 160]$	2	2
$(160, 170]$	8	10
$(170, 180]$	11	21
$(180, 190]$	7	28
$(190, 200]$	2	30

Interpolación en el polígono de frecuencias absolutas acumuladas



$$\mathrm{tg}(\alpha) = \frac{21 - 10}{180 - 170}$$

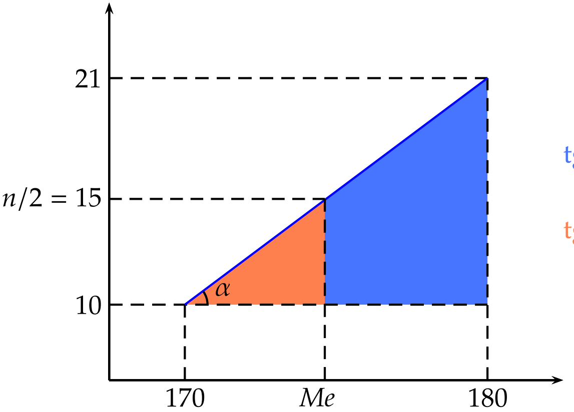
Interpolación en el polígono de frecuencias absolutas acumuladas



$$\operatorname{tg}(\alpha) = \frac{21 - 10}{180 - 170}$$

$$\operatorname{tg}(\alpha) = \frac{15 - 10}{Me - 170}$$

Interpolación en el polígono de frecuencias absolutas acumuladas



$$\operatorname{tg}(\alpha) = \frac{21 - 10}{180 - 170}$$

$$\operatorname{tg}(\alpha) = \frac{15 - 10}{Me - 170}$$

$$Med = 170 + \frac{15 - 10}{21 - 10}(180 - 170) = 170 + \frac{5}{11}10 = 174{,}54$$

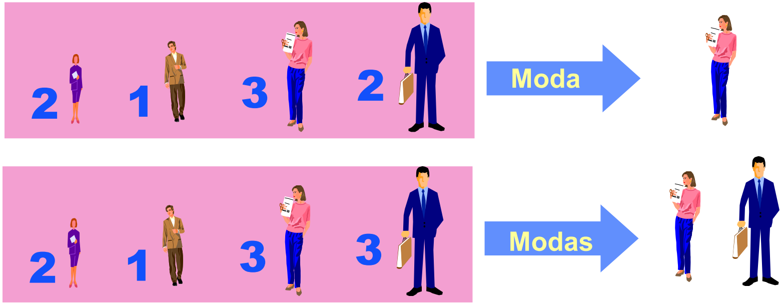
Moda

Definición (Moda muestral Mo)

La *moda muestral* de una variable X es el valor de la variable más frecuente en la muestra.

Con datos agrupados se toma como clase modal la clase con mayor frecuencia en la muestra.

En ocasiones puede haber más de una moda.



Cálculo de la moda

En el ejemplo del número de hijos puede verse fácilmente en la tabla de frecuencias que la moda es $Mo = 2$ hijos.

x_i	n_i
0	2
1	6
$\boxed{2}$	14
3	2
4	1

Y en el ejemplo de las estaturas también puede verse en la tabla de frecuencias que la clase modal es $Mo = (170, 180]$.

x_i	n_i
$(150, 160]$	2
$(160, 170]$	8
$\boxed{(170,180]}$	11
$(180, 190]$	7
$(190, 200]$	2

¿Qué estadístico de tendencia central usar?

En general, siempre que puedan calcularse conviene tomarlas en el siguiente orden:

1. Media. La media utiliza más información que el resto ya que para calcularla se tiene en cuenta la magnitud de los datos.
2. Mediana. La mediana utiliza menos información que la media, pero más que la moda, ya que para calcularla se tiene en cuenta el orden de los datos.
3. Moda. La moda es la que menos información utiliza ya que para calcularla sólo se tienen en cuenta las frecuencias absolutas.

Pero, ¡ojo! la media también es muy sensible a los datos atípicos, así que, tampoco debemos perder de vista la mediana.

Por ejemplo, consideremos la siguiente muestra del número de hijos de 7 matrimonios:

$$0, 0, 1, 1, 2, 2, 15$$

$$\bar{x} = 3 \text{ hijos} \quad \text{y} \quad Me = 1 \text{ hijos}$$

¿Qué representante de la muestra tomarías?

Cuantiles

Son valores de la variable que dividen la distribución, supuesta ordenada de menor a mayor, en partes que contienen el mismo número de datos.

Los más utilizados son:

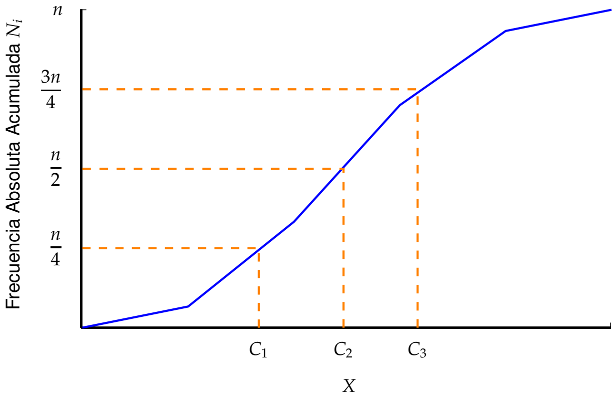
Cuartiles: Dividen la distribución en 4 partes iguales.
Hay tres cuartiles: C_1 (25 % acumulado) , C_2 (50 % acumulado), C_3 (75 % acumulado).

Deciles: Dividen la distribución en 10 partes iguales.
Hay 9 deciles: D_1 (10 % acumulado) ,. . . , D_9 (90 % acumulado).

Percentiles: Dividen la distribución en 100 partes iguales.
Hay 99 percentiles: P_1 (1 % acumulado),. . . , P_{99} (99 % acumulado).

Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. Por ejemplo, en el caso de los cuartiles se buscan los valores que tienen frecuencias absolutas acumuladas $n/4$ (primer cuartil), $n/2$ (segundo cuartil) y $3n/4$ (tercer cuartil) y si se trata de datos agrupados se interpola sobre el polígono de frecuencias acumuladas.



Cálculo de los cuantiles

Ejemplo con datos no agrupados

En el ejemplo anterior del número de hijos se tenían la siguientes frecuencias relativas acumuladas

x_i	F_i
0	0,08
1	0,32
2	0,88
3	0,96
4	1

$$
\begin{aligned}
F_{C_1} &= 0{,}25 \Rightarrow C_1 = 1 \text{ hijos},\\
F_{C_2} &= 0{,}5 \Rightarrow C_2 = 2 \text{ hijos},\\
F_{C_3} &= 0{,}75 \Rightarrow C_3 = 2 \text{ hijos},\\
F_{D_3} &= 0{,}3 \Rightarrow D_3 = 1 \text{ hijos},\\
F_{P_{92}} &= 0{,}92 \Rightarrow P_{92} = 3 \text{ hijos}.
\end{aligned}
$$

Estadísticos de dispersión

Recogen información respecto a la heterogeneidad de la variable y a la concentración de sus valores en torno a algún valor central.

Para las variables cuantitativas, las más empleadas son:

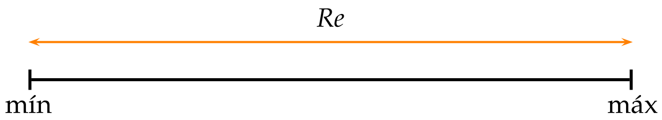
- Recorrido.
- Rango Intercuartílico.
- Varianza.
- Desviación Típica.
- Coeficiente de Variación.

Definición (Recorrido muestral Re)

El *recorrido muestral* de una variable X se define como la diferencia entre el máximo y el mínimo de los valores en la muestra.

$$Re = \max_{x_i} - \min_{x_i}$$

El recorrido da una idea de la máxima variación que hay entre los datos muestrales. No obstante, es muy sensible a datos atípicos ya que suelen aparecer justo en los extremos de la distribución, por lo que no se suele utilizar mucho.



Rango intercuartílico

Para evitar el problema de los datos atípicos en el recorrido, se puede utilizar el primer y tercer cuartil en lugar del mínimo y el máximo.

Definición (Rango intercuartílico muestral RI)

El *rango intercuartílico muestral* de una variable X se define como la diferencia entre el tercer y el primer cuartil de la muestra.

$$RI = C_3 - C_1$$

El rango intercuartílico da una idea de la variación que hay en el 50 % de los datos centrales.

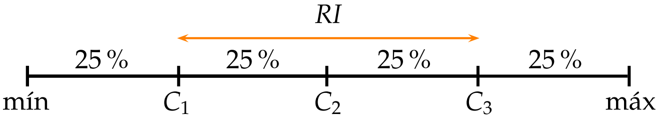


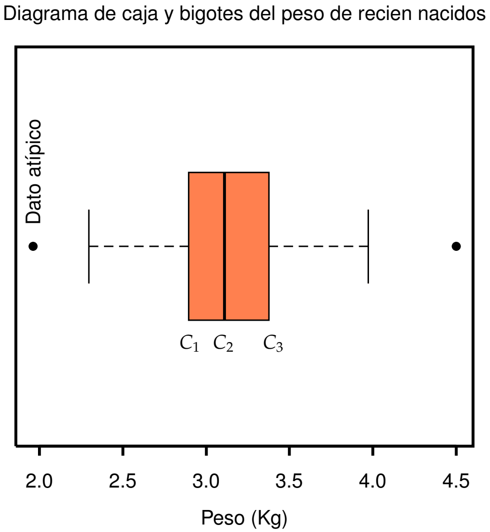
Diagrama de caja y bigotes

La dispersión de una variable suele representarse gráficamente mediante un **diagrama de caja y bigotes**, que consiste en una caja sobre un eje X donde el borde inferior de la caja es el primer cuartil, y el borde superior el tercer cuartil, y por tanto, la anchura de la caja es el rango intercuartílico. En ocasiones también se representa el segundo cuartil con una línea que divide la caja.

También se utiliza para detectar los valores atípicos mediante unos segmentos (bigotes) que salen de los extremos de la caja y que marcan el intervalo de normalidad de los datos.

Diagrama de caja y bigotes

Ejemplo con pesos de recién nacidos



Construcción del diagrama de caja y bigotes

1. Calcular los cuartiles.
2. Dibujar una caja de manera que el extremo inferior caiga sobre el primer cuartil y el extremo superior sobre el tercer cuartil.
3. Dividir la caja con una línea que caiga sobre el segundo cuartil.
4. Para los bigotes inicialmente se determina la posición de los puntos denominados *vallas* v_1 y v_2 restando y sumando respectivamente a primer y tercer cuartil $1{,}5$ veces el rango intercuartílico RI:

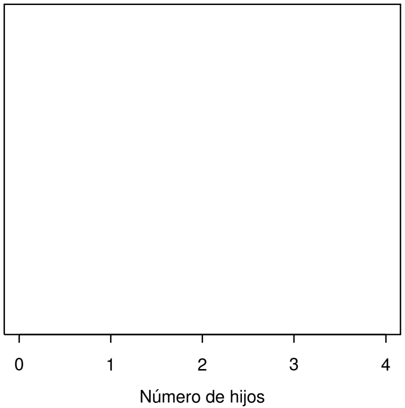
   $$v_1 = C_1 - 1{,}5RI$$
   $$v_2 = C_3 + 1{,}5RI$$

   A partir de las vallas se buscan los valores b_1, que es el mínimo valor de la muestra mayor o igual que v_1, y b_2, que es máximo valor de la muestra menor o igual que v_2. Para el bigote inferior se dibuja un segmento desde el borde inferior de la caja hasta b_1 y para el superior se dibuja un segmento desde el borde superior de la caja hasta b_2.
5. Finalmente, si en la muestra hay algún dato por debajo de v_1 o por encima de v_2 se dibuja un punto sobre dicho valor.

Construcción del diagrama de caja y bigotes

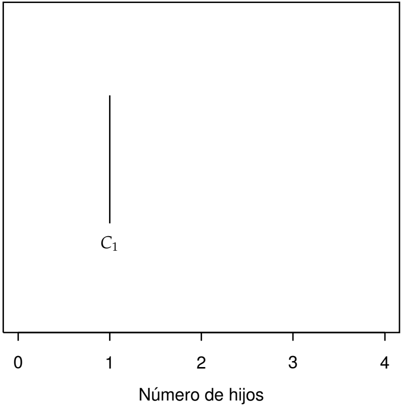
Ejemplo del número de hijos

Diagrama de caja y bigotes del número de hijos



1. Calcular los cuartiles: $C_1 = 1$ hijos

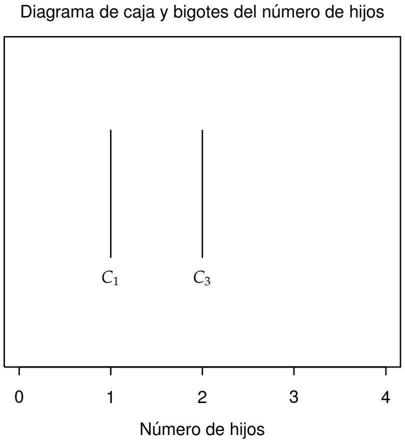
Diagrama de caja y bigotes del número de hijos



Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

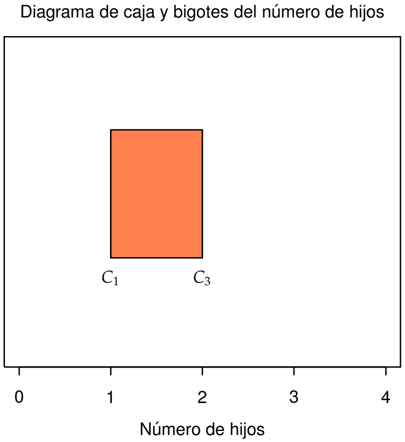
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.



Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

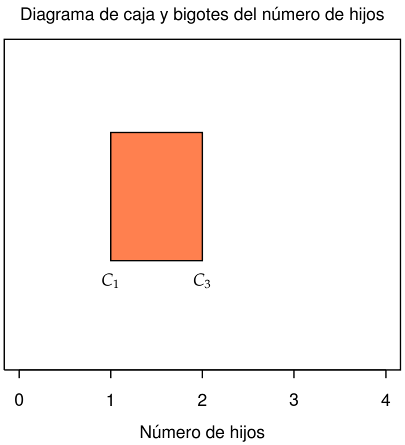
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.



Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

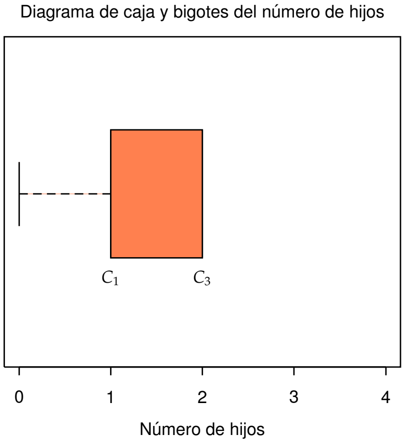
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.
3. Calcular las vallas: $v_1 = 1 - 1{,}5 * 1 = -0{,}5$ y $v_2 = 2 + 1{,}5 * 1 = 3{,}5$.



Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

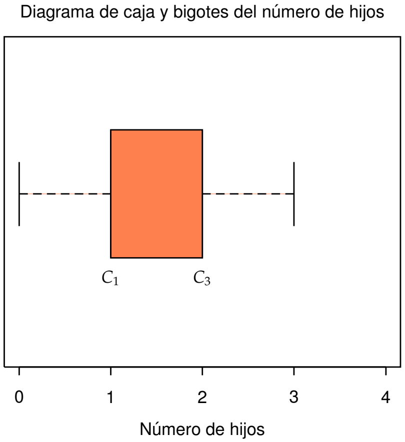
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.
3. Calcular las vallas: $v_1 = 1 - 1{,}5 * 1 = -0{,}5$ y $v_2 = 2 + 1{,}5 * 1 = 3{,}5$.
4. Dibujar los bigotes: $b_1 = 0$ hijos



Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

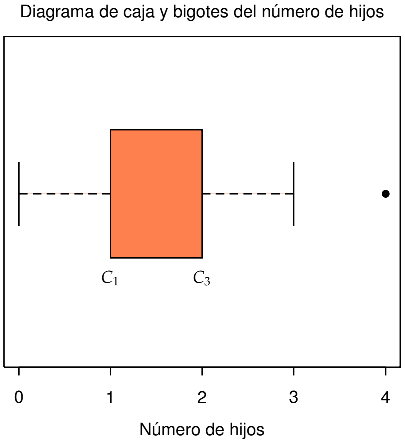
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.
3. Calcular las vallas: $v_1 = 1 - 1{,}5 * 1 = -0{,}5$ y $v_2 = 2 + 1{,}5 * 1 = 3{,}5$.
4. Dibujar los bigotes: $b_1 = 0$ hijos y $b_1 = 3$ hijos.



Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

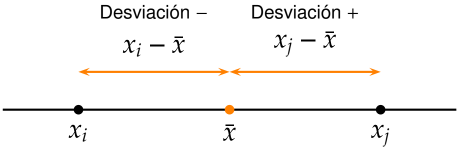
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.
3. Calcular las vallas: $v_1 = 1 - 1{,}5 * 1 = -0{,}5$ y $v_2 = 2 + 1{,}5 * 1 = 3{,}5$.
4. Dibujar los bigotes: $b_1 = 0$ hijos y $b_1 = 3$ hijos.
5. Dibujar los datos atípicos: 4 hijos.



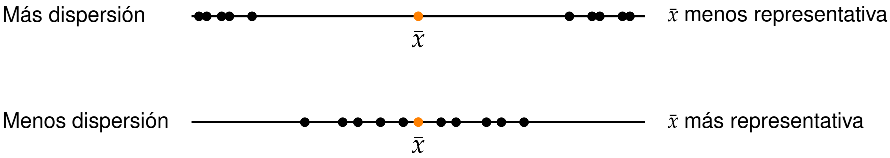
Desviaciones respecto de la media

Otra forma de medir la variabilidad de una variable es estudiar la concentración de los valores en torno a algún estadístico de tendencia central como por ejemplo la media.

Para ello se suele medir la distancia de cada valor a la media. A ese valor se le llama **desviación respecto de la media.**



Si las desviaciones son grandes la media no será tan representativa como cuando la desviaciones sean pequeñas.



¿En qué muestra es más representativa la media?

Varianza y desviación típica

Definición (Varianza s^2)

La *varianza muestral* de una variable X se define como el promedio del cuadrado de las desviaciones de los valores de la muestra respecto de la media muestral.

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

También puede calcularse de manera más sencilla mediante la fórmula

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \sum x_i^2 f_i - \bar{x}^2$$

La varianza tiene las unidades de la variable al cuadrado, por lo que para facilitar su interpretación se suele utilizar su raíz cuadrada:

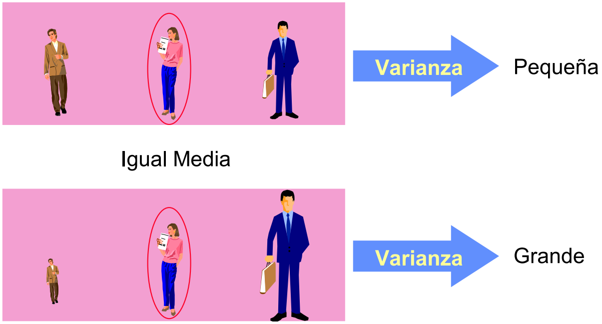
Definición (Desviación típica s)

La *desviación típica muestral* de una variable X se define como la raíz cuadrada positiva de su varianza muestral.

$$s = +\sqrt{s^2}$$

Interpretación de la varianza y la desviación típica

Tanto la varianza como la desviación típica sirven para cuantificar la dispersión de los datos en torno a la media.



Para el número de hijos se puede calcular la varianza a partir de la tabla de frecuencias añadiendo una columna con los cuadrados de los valores:

x_i	n_i	$x_i^2 n_i$
0	2	0
1	6	6
2	14	56
3	2	18
4	1	16
$\sum$	25	96

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{96}{25} - 1{,}76^2 = 0{,}7424 \text{ hijos}^2.$$

Y la desviación típica es $s = \sqrt{0{,}7424} = 0{,}8616$ hijos.

Comparado este valor con el recorrido, que va de 0 a 4 hijos se observa que no es demasiado grande por lo que se puede concluir que no hay mucha dispersión y en consecuencia la media de $1{,}76$ hijos representa bien a los matrimonios de la muestra.

En el ejemplo de las estaturas, al ser datos agrupados, el cálculo se realiza igual que antes pero tomando como valores de la variable las marcas de clase.

X	x_i	n_i	$x_i^2 n_i$
$(150, 160]$	155	2	48050
$(160, 170]$	165	8	217800
$(170, 180]$	175	11	336875
$(180, 190]$	185	7	239575
$(190, 200]$	195	2	76050
$\sum$		30	918350

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{918350}{30} - 174{,}67^2 = 102{,}06 \text{ cm}^2.$$

Y la desviación típica es $s = \sqrt{102{,}06} = 10{,}1$ cm.

Este valor es bastante pequeño, comparado con el recorrido de la variable, que va de 150 a 200 cm, por lo que la variable tiene poca dispersión y en consecuencia su media es muy representativa.

Coeficiente de variación

Tanto la varianza como la desviación típica tienen unidades y eso dificulta a veces su interpretación y su comparación.

Afortunadamente es fácil definir a partir de ellas una medida de dispersión adimensional que es más fácil de interpretar.

Definición (Coeficiente de variación muestral cv)

El *coeficiente de variación muestral* de una variable X se define como el cociente entre su desviación típica muestral y el valor absoluto de su media muestral.

$$cv = \frac{s}{|\bar{x}|}$$

El coeficiente de variación muestral mide la dispersión relativa de los valores de la muestra en torno a la media muestral.

Como no tiene unidades, es muy sencillo de interpretar: Cuanto mayor sea, mayor será la dispersión y menos representativa será la media.

También se utiliza para comparar la dispersión entre muestras distintas incluso si las variables tienen unidades diferentes.

¡Ojo! No tiene sentido cuando la media muestral vale 0 o valores próximos.

En el caso del número de hijos, como $\bar{x} = 1{,}76$ hijos y $s = 0{,}8616$ hijos, se tiene que el coefiente de variación vale

$$cv = \frac{s}{|\bar{x}|} = \frac{0{,}8616}{|1{,}76|} = 0{,}49.$$

En el caso de las estaturas, como $\bar{x} = 174{,}67$ cm y $s = 10{,}1$ cm, se tiene que el coeficiente de variación vale

$$cv = \frac{s}{|\bar{x}|} = \frac{10{,}1}{|174{,}67|} = 0{,}06.$$

Como se puede observar la dispersión relativa en la muestra de estaturas es mucho menor que en la del número de hijos, por lo que la media de las estaturas será más representativa que la media del número de hijos.

Estadísticos de forma

Son medidas que tratan de caracterizar aspectos de la forma de la distribución de una muestra.

Los aspectos más relevantes son:

Simetría: Miden la simetría de la distribución de frecuencias en torno a la media.
El estadístico más utilizado es el *Coeficiente de Asimetría de Fisher*.

Apuntamiento: Miden el apuntamiento de la distribución de frecuencias.
El estadístico más utilizado es el *Coeficiente de Apuntamiento o Curtosis*.

Coeficiente de asimetría

Definición (Coeficiente de asimetría muestral g_1)

El *coeficiente de asimetría muestral* de una variable X se define como el promedio de las desviaciones de los valores de la muestra respecto de la media muestral, elevadas al cubo, dividido por la desviación típica al cubo.

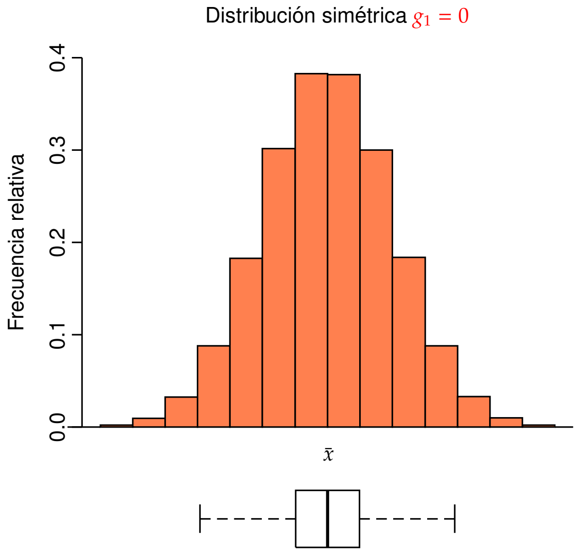
$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n}{s^3} = \frac{\sum (x_i - \bar{x})^3 f_i}{s^3}$$

El coeficiente de asimetría muestral mide el grado de simetría de los valores de la muestra con respecto a la media muestral, de manera que:

- $g_1 = 0$ indica que hay el mismo número de valores a la derecha y a la izquierda de la media (simétrica).
- $g_1 < 0$ indica que la mayoría de los valores son mayores que la media (asimétrica a la izquierda).
- $g_1 > 0$ indica que la mayoría de los valores son menores que la media (asimétrica a la derecha).

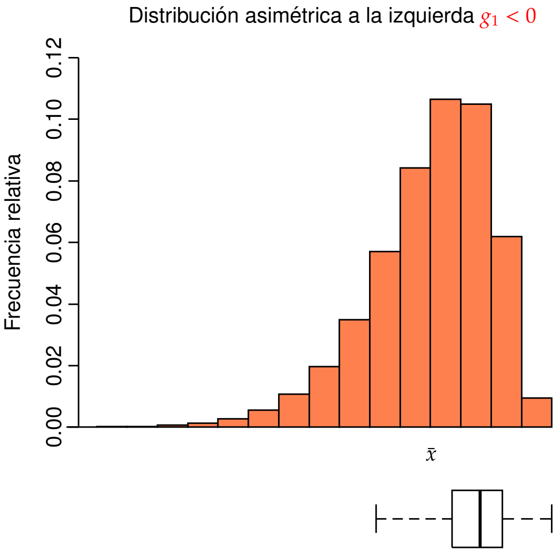
Coeficiente de asimetría

Ejemplo de distribución simétrica



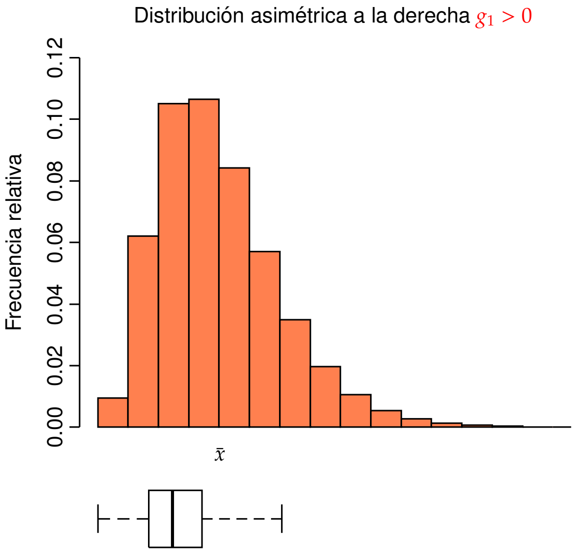
Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la izquierda



Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la derecha



Siguiendo con el ejemplo de las estaturas, podemos calcular el coeficiente de asimetría a partir de la tabla de frecuencias añadiendo una nueva columna con los cubos de las desviaciones a la media $\bar{x} = 174{,}67$ cm:

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
$(150, 160]$	155	2	−19,67	−15221,00
$(160, 170]$	165	8	−9,67	−7233,85
$(170, 180]$	175	11	0,33	0,40
$(180, 190]$	185	7	10,33	7716,12
$(190, 200]$	195	2	20,33	16805,14
$\sum$		30		2066,81

$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n}{s^3} = \frac{2066{,}81/30}{10{,}1^3} = 0{,}07.$$

Al estar tan próximo a 0, este valor indica que la distribución es prácticamente simétrica con respecto a la media.

Coeficiente de apuntamiento o curtosis

Definición (Coeficiente de apuntamiento muestral g_2)

El *coeficiente de apuntamiento muestral* de una variable X se define como el promedio de las desviaciones de los valores de la muestra respecto de la media muestral, elevadas a la cuarta, dividido por la desviación típica a la cuarta y al resultado se le resta 3.

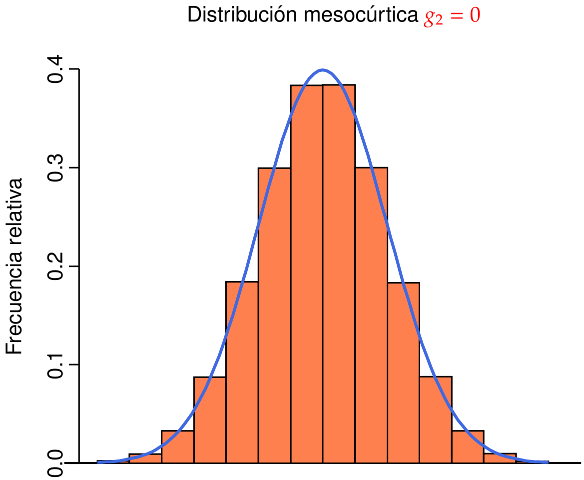
$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{\sum (x_i - \bar{x})^4 f_i}{s^4} - 3$$

El coeficiente de apuntamiento muestral mide el grado de apuntamiento de los valores de la muestra con respecto a una distribución normal de referencia, de manera que:

- $g_2 = 0$ indica que la distribución tienen un apuntamiento normal (*mesocúrtica*).
- $g_2 < 0$ indica que la distribución tiene menos apuntamiento de lo normal (*platicúrtica*).
- $g_2 > 0$ indica que la distribución tiene más apuntamiento de lo normal (*leptocúrtica*).

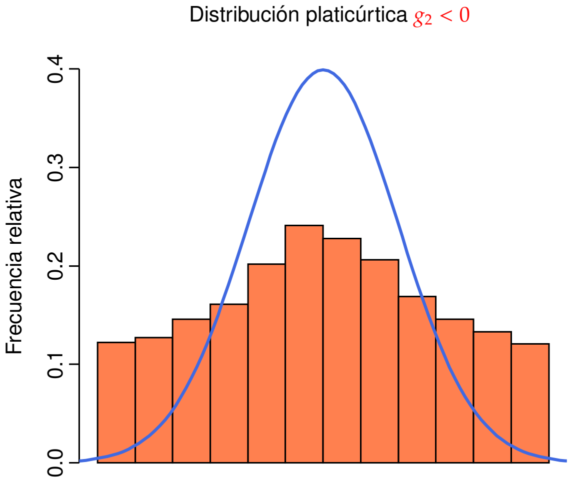
Coeficiente de apuntamiento o curtosis

Ejemplo de distribución mesocúrtica



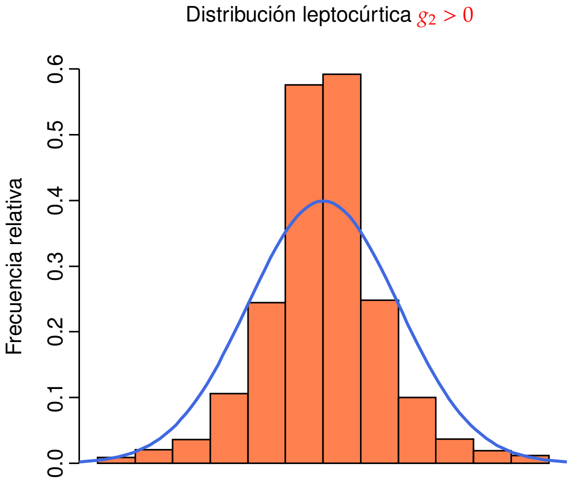
Coeficiente de apuntamiento o curtosis

Ejemplo de distribución platicúrtica



Coeficiente de apuntamiento o curtosis

Ejemplo de distribución leptocúrtica



Cálculo del coeficiente de apuntamiento

Ejemplo con datos agrupados

De nuevo para el ejemplo de las estaturas podemos calcular el coeficiente de asimetría a partir de la tabla de frecuencias añadiendo una nueva columna con las desviaciones a la media $\bar{x} = 174{,}67$ cm elevadas a la cuarta:

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^4 n_i$
$(150, 160]$	155	2	$-19{,}67$	299396,99
$(160, 170]$	165	8	$-9{,}67$	69951,31
$(170, 180]$	175	11	0,33	0,13
$(180, 190]$	185	7	10,33	79707,53
$(190, 200]$	195	2	20,33	341648,49
$\sum$		30		790704,45

$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{790704{,}45/30}{10{,}1^4} - 3 = -0{,}47.$$

Como se trata de un valor negativo, aunque pequeño, podemos decir que la distribución es ligeramente platicúrtica.

Interpretación de los coeficientes de asimetría y apuntamiento

Como se verá más adelante en la parte de inferencia, muchas de las pruebas estadísticas solo pueden aplicarse a poblaciones normales.

Las poblaciones normales se caracterizan por ser simétricas y mesocúrticas, de manera que, tanto el coeficiente de asimetría como el de apuntamiento pueden utilizarse para contrastar si los datos de la muestra provienen de una población normal.

En general, se suele rechazar la hipótesis de normalidad de la población cuando g_1 o g_2 estén fuera del intervalo $[-2, 2]$.

En tal caso, lo habitual es aplicar alguna transformación a la variable para corregir la anormalidad.

Transformaciones de variables

En muchas ocasiones se suelen transformar los datos brutos para trabajar con unas unidades más cómodas, o bien para corregir alguna anormalidad de la distribución.

Por ejemplo, si estamos trabajando con estaturas medidas en metros y tenemos los siguientes valores:

$$1{,}75\text{m}, 1{,}65\text{m}, 1{,}80\text{m},$$

podemos evitar los decimales multiplicando por 100, es decir, pasando de metros a centímetros:

$$175\text{cm}, 165\text{cm}, 180\text{cm},$$

Y si queremos reducir la magnitud de los datos podemos restarles a todos el menor de ellos, en este caso, 165cm:

$$10\text{cm}, 0\text{cm}, 15\text{cm},$$

Está claro que este conjunto de datos es mucho más sencillo que el original. En el fondo lo que se ha hecho es aplicar a los datos la transformación:

$$Y = 100X - 165$$

Transformaciones lineales

Una de las transformaciones más habituales es la *transformación lineal*:

$$Y = a + bX.$$

Se puede comprobar fácilmente que la media y la desviación típica de la variable resultante cumplen:

$$\begin{aligned} \bar{y} &= a + b\bar{x}, \\ s_y &= |b| s_x \end{aligned}$$

Además, el coeficiente de curtosis no se altera y el de asimetría sólo cambia de signo si b es negativo.

Transformación de tipificación y puntuaciones típicas

Una de las transformaciones lineales más habituales es la *tipificación*:

Definición (Variable tipificada)

La *variable tipificada* de una variable estadística X es la variable que resulta de restarle su media y dividir por su desviación típica.

$$Z = \frac{X - \bar{x}}{s_x}$$

La tipificación es muy útil para eliminar la dependencia de una variable respecto de las unidades de medida empleadas.

Los valores tipificados se conocen como **puntuaciones típicas** y miden el número de desviaciones típicas que dista de la media cada observación, lo cual es útil para comparar variables con distintas unidades.

Otra propiedad de la variable tipificada es que tiene media 0 y desviación típica 1:

$$\bar{z} = 0 \qquad s_z = 1$$

Transformación de tipificación y puntuaciones típicas

Ejemplo

Las notas de 5 alumnos en dos asignaturas X e Y son:

Alumno:	1	2	3	4	5		
X :	2	5	4	8	6	$\bar{x} = 5$	$s_x = 2$
Y :	1	9	8	5	2	$\bar{y} = 5$	$s_y = 3{,}16$

¿Han tenido el mismo rendimiento los alumnos que han sacado un 8?

Podría parecer que ambos alumnos han tenido el mismo rendimiento puesto que tienen la misma nota, pero si queremos ver el rendimiento relativo al resto del grupo, tendríamos que tener en cuenta la dispersión de cada muestra y medir sus puntuaciones típicas:

$$\begin{array}{cccccc} X: & -1{,}5 & 0 & -0{,}5 & 1{,}5 & 0{,}5 \\ Y: & -1{,}26 & 1{,}26 & 0{,}95 & 0 & -0{,}95 \end{array}$$

Es decir, el alumno que tiene un 8 en X está $1{,}5$ veces la desviación típica por encima de la media de su grupo, mientras que el alumno que tiene un 8 en Y sólo está $0{,}95$ desviaciones típicas por encima de su media. Así pues, el primer alumno tuvo un rendimiento superior al segundo.

Siguiendo con el ejemplo anterior

¿Cuál es el mejor alumno?

Si simplemente se suman las puntuaciones de cada asignatura se tiene:

Alumno:	1	2	3	4	5
$X:$	2	5	4	8	6
$Y:$	1	9	8	5	2
$\sum$	3	14	12	13	8

El mejor alumno sería el segundo.

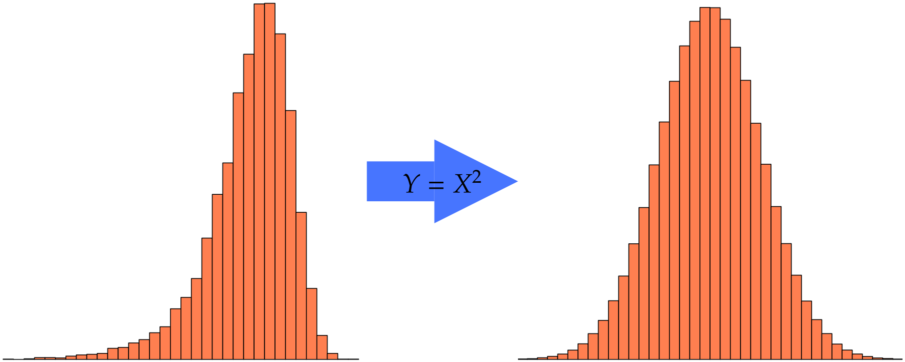
Pero si se considera el rendimiento relativo tomando las puntuaciones típicas se tiene:

Alumno:	1	2	3	4	5
$X:$	$-1{,}5$	0	$-0{,}5$	1,5	0,5
$Y:$	$-1{,}26$	1,26	0,95	0	$-0{,}95$
$\sum$	$-2{,}76$	1,26	0,45	1,5	$-0{,}45$

Y el mejor alumno sería el cuarto.

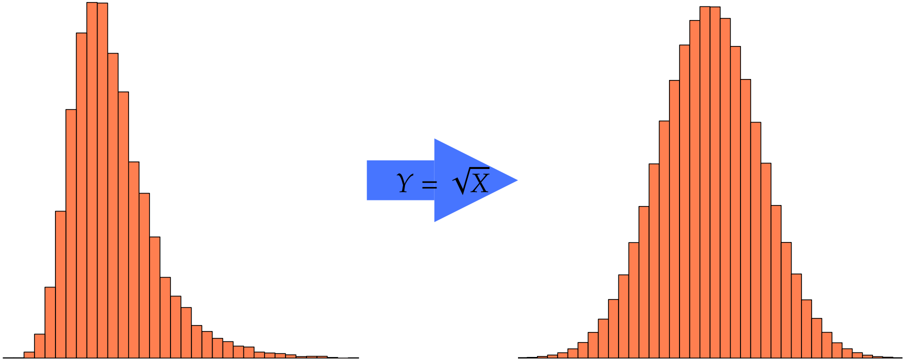
Transformaciones no lineales

La transformación $Y = X^2$ comprime la escala para valores pequeños y la expande para valores altos, de manera que es muy útil para corregir asimetrías hacia la izquierda.



Transformaciones no lineales

Las transformaciones $Y = \sqrt{x}$, $Y = \log X$ y $Y = 1/X$ comprimen la escala para valores altos y la expanden para valores pequeños, de manera que son útiles para corregir asimetrías hacia la derecha.



Variables clasificadoras o factores

En ocasiones interesa describir el comportamiento de una variable, no para toda la muestra, sino para distintos grupos de individuos, como por ejemplo, estudiar las estaturas en hombres y mujeres por separado.

En tal caso se utiliza una nueva variable, llamada **variable clasificadora** o **factor discriminante**, para dividir la muestra en grupos y posteriormente se realiza el estudio descriptivo de la variable principal en cada grupo.

Variables clasificadoras

Usando la misma muestra de estaturas, pero teniendo en cuenta el sexo, tenemos:

Mujeres	173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168.
Hombres	179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187.