Manual Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

Feb 2017

Departamento de Matemática Aplicada y Estadística
CEU San Pablo

Términos de la licencia

Índice general

1 Distribución de frecuencias: Tabulación y gráficos

Estadística descriptiva

La estadística descriptiva es la parte de la estadística encargada de representar, analizar y resumir la información contenida en la muestra.

Tras el proceso de muestreo, es la siguiente etapa de todo estudio estadístico y suele consistir en:

1. Clasificar, agrupar y ordenar los datos de la muestra.
2. Tabular y representar gráficamente los datos de acuerdo a sus frecuencias.
3. Calcular medidas que resuman la información que contiene la muestra (*estadísticos muestrales*).

No tiene poder inferencial ⇒ *No utilizar para sacar conclusiones sobre la población!*

Clasificación de la muestra

El estudio de una variable estadística comienza por medir la variable en los individuos de la muestra y clasificar los valores obtenidos.

Existen dos formas de clasificar estos valores:

Sin agrupar : Ordenar todos los valores obtenidos en la muestra de menor a mayor (si existe orden). Se utiliza con atributos y variables discretas con pocos valores diferentes.

Agrupados : Agrupar los valores en clases (intervalos) y ordenar dichas clases de menor a mayor. Se utiliza con variables continuas y con variables discretas con muchos valores diferentes.

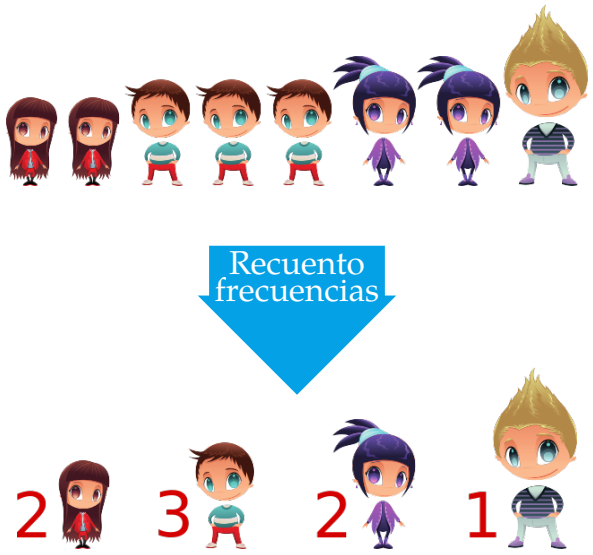
1.1 Distribución de frecuencias

Clasificación de la muestra

X =Estatura

Recuento de frecuencias

X =Estatura



Frecuencias muestrales

Definición 1 (Frecuencias muestrales)**.** Dada una muestra de tamaño n de una variable X, para cada valor x_i de la variable observado en la muestra, se define

- **Frecuencia absoluta** n_i: Es el número de veces que el valor x_i aparece en la muestra.
- **Frecuencia relativa** f_i: Es la proporción de veces que el valor x_i aparece en la muestra.

$$f_i = \frac{n_i}{n}$$

- **Frecuencia absoluta acumulada** N_i: Es el número de valores en la muestra menores o iguales que x_i.

$$N_i = n_1 + \cdots + n_i$$

- **Frecuencia relativa acumulada** F_i: Es la proporción de valores en la muestra menores o iguales que x_i.

$$F_i = \frac{N_i}{n}$$

Tabla de frecuencias

Al conjunto de valores observados en la muestra junto a sus respectivas frecuencias se le denomina **distribución muestral de frecuencias** y suele representarse mediante una **tabla de frecuencias**.

Valores de X	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
x_1	n_1	f_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	N_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k

Tabla de frecuencias

Ejemplo de datos sin agrupar

El número de hijos en 25 familias es

$$1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, \ 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.$$

La tabla de frecuencias asociada a esta muestra es

x_i	n_i	f_i	N_i	F_i
0	2	0.08	2	0.08
1	6	0.24	8	0.32
2	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1
$\sum$	25	1		

Tabla de frecuencias

Ejemplo de datos agrupados

Las estaturas (en cm) de 30 estudiantes es

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

La tabla de frecuencias asociada a esta muestra es

x_i	n_i	f_i	N_i	F_i
$(150, 160]$	2	0.07	2	0.07
$(160, 170]$	8	0.27	10	0.34
$(170, 180]$	11	0.36	21	0.70
$(180, 190]$	7	0.23	28	0.93
$(190, 200]$	2	0.07	30	1
$\sum$	30	1		

Construcción de clases

Cada intervalo de agrupación de datos se denomina **clase** y el centro del intervalo se llama **marca de clase**.

A la hora de agrupar los datos en clases hay que tener en cuenta lo siguiente:

- El número de intervalos no debe ser muy grande ni muy pequeño. Una regla orientativa es tomar un número de intervalos próximo $\sqrt{n}$ o $\log_2(n)$.
- Los intervalos no deben solaparse y deben cubrir todo el rango de valores. Es indiferente si se abren por la izquierda y se cierran por la derecha o al revés.
- El valor más pequeño debe caer dentro del primer intervalo y el más grande dentro del último.

Tabla de frecuencias

Ejemplo con un atributo

Los grupos sanguíneos de 30 personas son

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, A, AB, A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

La tabla de frecuencias asociada a esta muestra es

x_i	n_i	f_i
0	5	0.16
A	14	0.47
B	8	0.27
AB	3	0.10
$\sum$	30	1

¿Por qué en este caso no se construyen las columnas de frecuencias acumuladas?

1.2 Representaciones gráficas

Representaciones gráficas

Es habitual representar la distribución muestral de frecuencias de forma gráfica.

Dependiendo del tipo de variable y de si se han agrupado o no los datos, se utilizan distintos tipos de gráficos:

- Diagrama de barras
- Histograma
- Diagrama de líneas
- Digrama de sectores

Diagrama de barras

Un **diagrama de barras** consiste en un conjunto de barras, una para cada valor o categoría de la variabe, dibujadas en unos ejes cartesianos.

Habitualmente os valores o categorías de la variable se representan en el eje X, y las frecuencias en el eje Y. Para cada valor o categoría de la variabe se dibuja una barra de altura la correspondiente frecuencia. La anchura de la barra es indiferente pero debe haber una separación clara entre las barras.

Dependiendo de la frecuencia representada en el eje Y se tienen distintos tipos de diagramas de barras.

A veces se dibuja un polígono, conocido como **polígono de frecuencias**, uniendo los puntos más altos de cada barra con segmentos.

Diagrama de barras de frecuencias absolutas

Datos sin agrupar

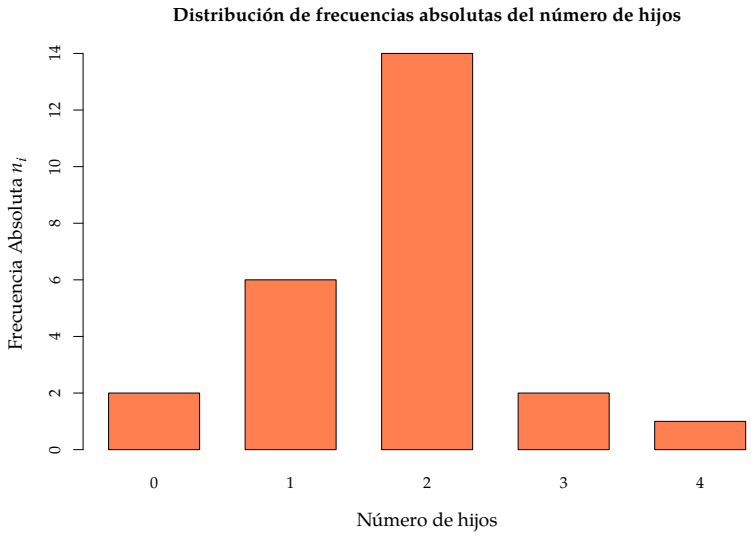


Diagrama de líneas o Polígono de frecuencias absolutas

Datos sin agrupar

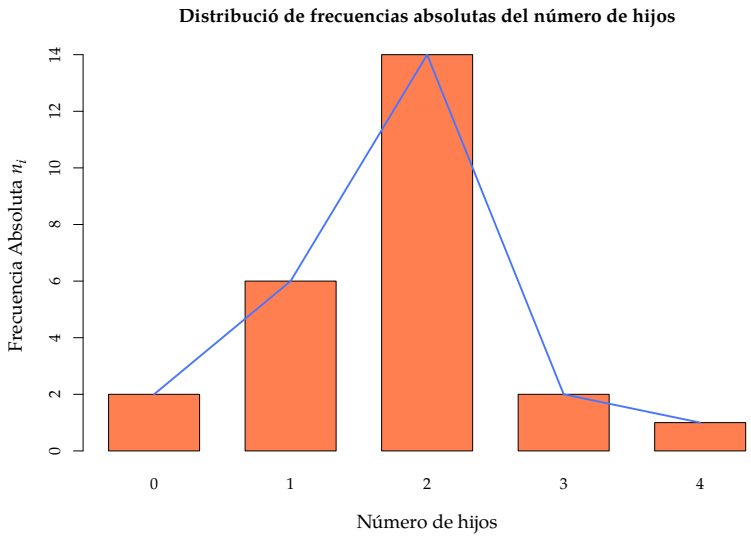


Diagrama de barras de frecuencias acumuladas

Datos sin agrupar

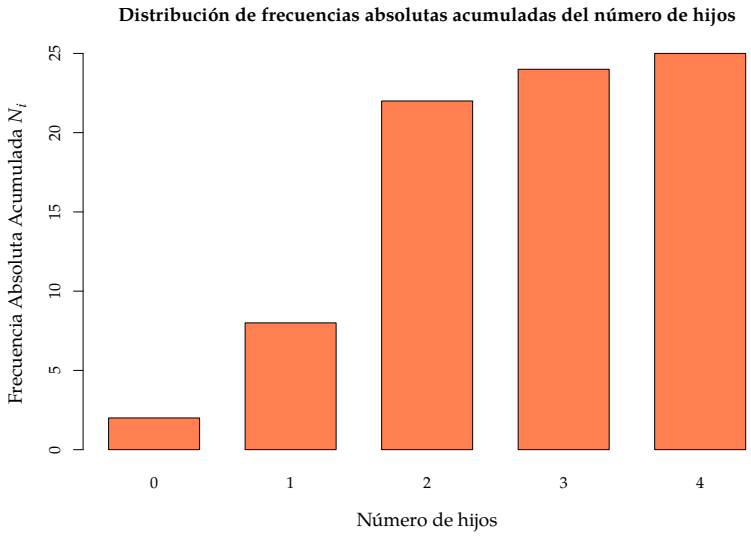
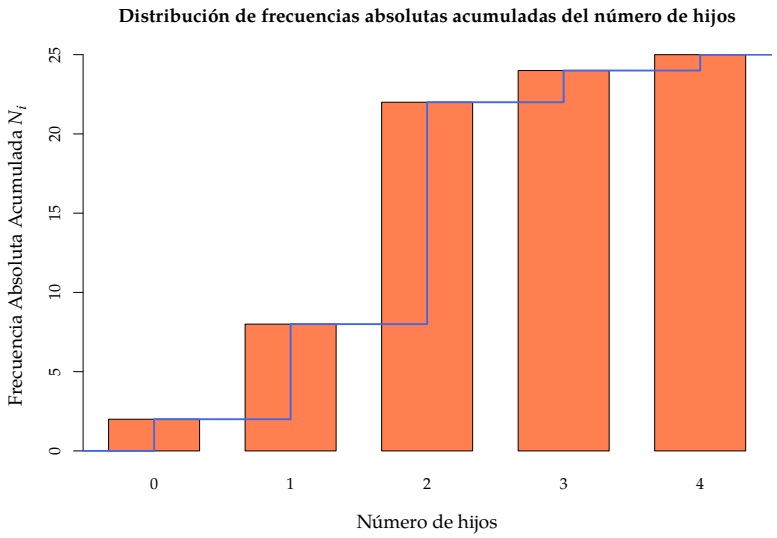


Diagrama de línea o polígono de frecuencias absolutas acumuladas

Datos sin agrupar



Histograma

Un **histograma** es similar a un diagrama de barras pero para datos agrupados.

Habitualmente las clases o intervalos de agrupación se representan en el eje X, y las frecuencias en el eje Y.

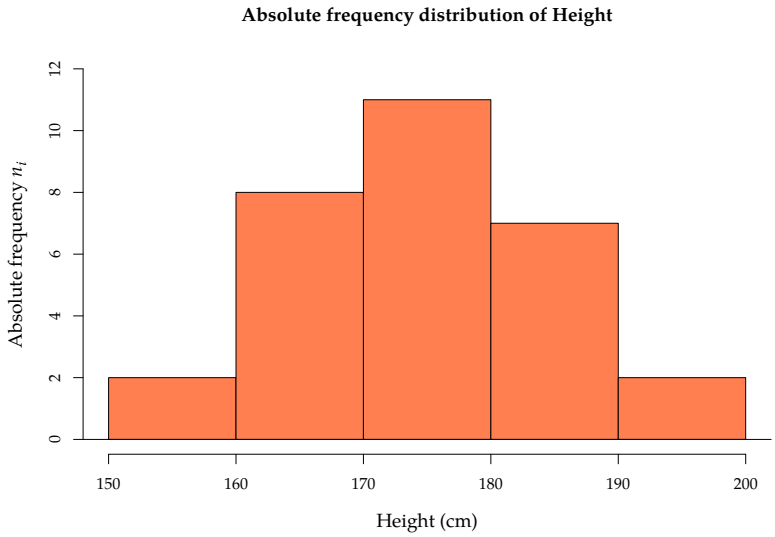
Para cada clase se dibuja una barra de altura la correspondiente frecuencia. A diferencia del diagrama de barras, la anchura del la barra coincide con la anchura de las clases y no hay separación entre dos barras consecutivas.

Dependiendo del tipo de frecuencia representada en el eje Y existen distintos tipos de histogramas.

A veces se dibuja un polígono, conocido como **polígono de frecuencias**, uniendo los puntos más altos de cada barra con segmentos.

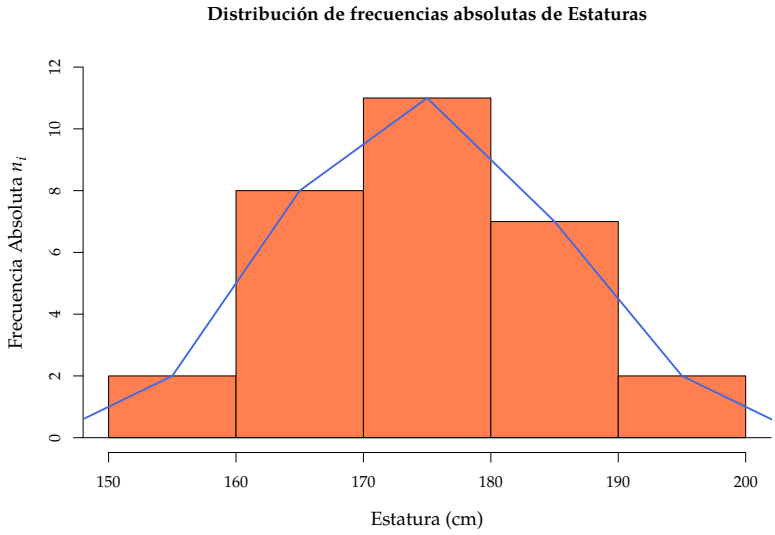
Histograma de frecuencias absolutas

Datos agrupados



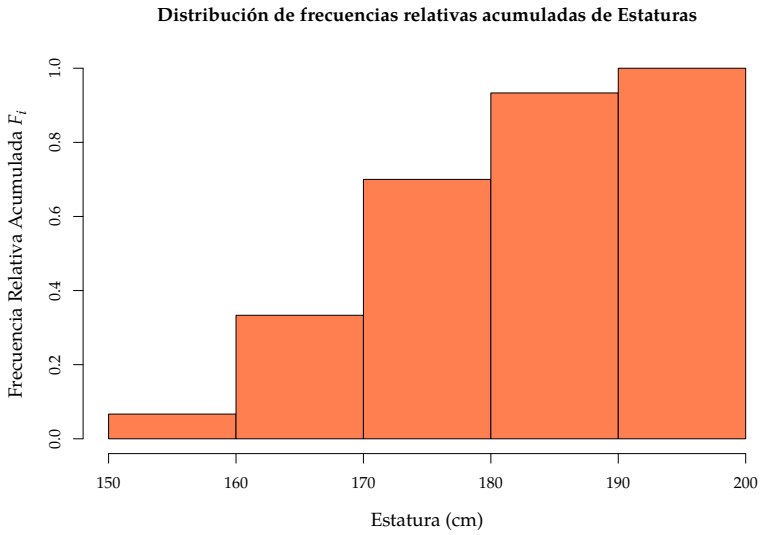
Polígono de frecuencias absolutas

Datos agrupados



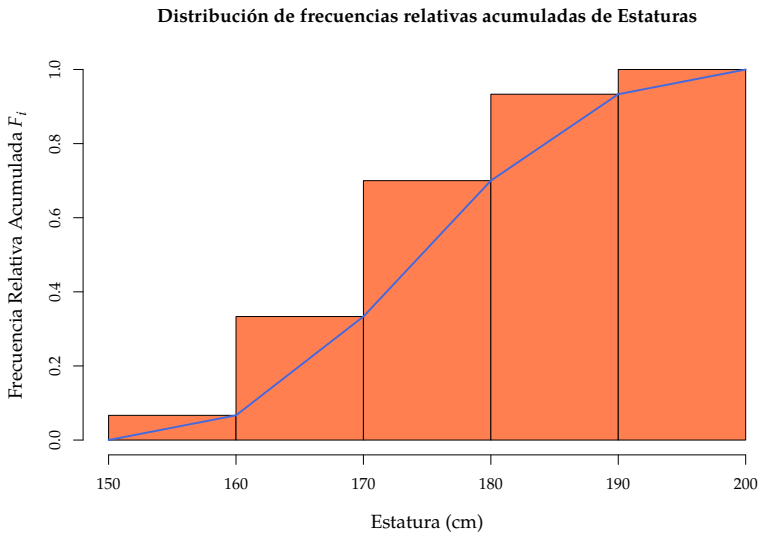
Histograma de frecuencias relativas acumuladas

Datos agrupados



Polígono de frecuencias relativas acumuladas

Datos agrupados



El polígono de frecuencias acumuladas (absolutas o relativas) se conoce como **ojiva**.

Obsérvese que en la ojiva se unen con segmentos los vértices superiores derechos de cada barra, en lugar de los centros, ya que no se consigue acumalar la correspondiente frecuencia hasta el final del intervalo.

Diagrama de sectores

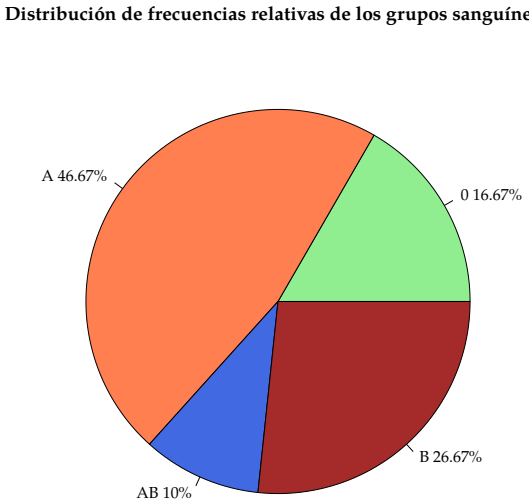
Un **diagrama de sectores** consiste en un círculo divido en porciones, uno por cada valor o categoría de la variable.

Cada porción se conoce como **sector** y su ángulo o área es proporcional a la correspondiente frecuencia del valor o categoría.

Los diagramas de sectores pueden representar frecuencias absolutas o relativas, pero no pueden representar frecuencias acumuladas, y se utilizan sobre todo con atributos nominales. Para atributos ordinales o variables cuantitativas es mejor utilizar diagramas de barras o histogramas, ya es más fácil percibir las diferencias en una dimensión (altura de las barras) que en dos dimensiones (áreas de los sectores).

Diagrama de sectores

Atributos



Datos atípicos

Uno de los principales problemas de las muestras son los **datos atípicos**, que son valores muy distintos de los demás valores en la muestra.

Es muy importante detectar los datos atípicos antes de realizar cualquier análisis de los datos, pues *suelen distorsionar los resultados*.

Aparecen siempre en los extremos de la distribución, y pueden detectarse fácilmente con un diagrama de caja y bigotes (como se verá después).

Tratamiento de los datos atípicos

Cuando trabajemos con muestras grandes, los datos atípicos tienen menor influencia y pueden dejarse en la muestra.

Cuando trabajemos con muestras pequeñas tenemos varias opciones:

- Eliminar el dato atípico si es un error.
- Sustituir el dato atípico por el mayor o menor valor de la distribución que no sea atípico, si no es un error pero que no concuerda con el modelo de distribución teórico de la población.
- Dejarl el dato atípico si no es un error y cambiar el modelo de distribución teórico para ajustarse a los datos atípicos.

1.3 Estadísticos muestrales

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.

1.4 Estadísticos de posición

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 2 (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.

Cálculo de la media aritmética

Ejemplo con datos no agrupados

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 3 (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 4 (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 $F_{Me} = 0.5$.

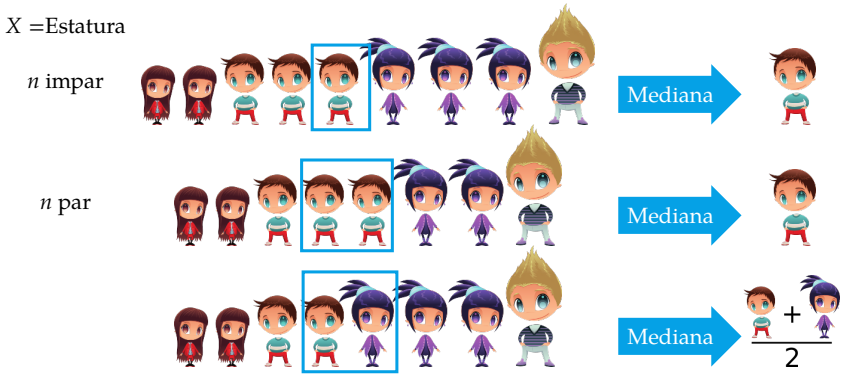
El cálculo de la mediana se realiza de forma distinta según se hayan agrupado los datos o no.

¡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 $\frac{25+1}{2} = 13$.

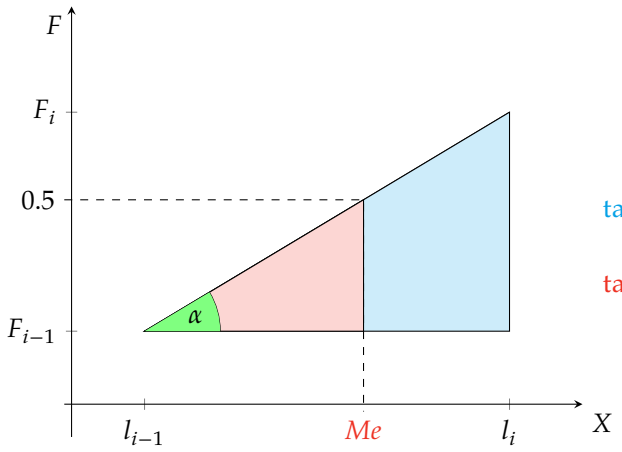
$$0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, \boxed{2}, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4$$

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
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



$$\tan(\alpha) = \frac{F_i - F_{i-1}}{l_i - l_{i-1}}$$

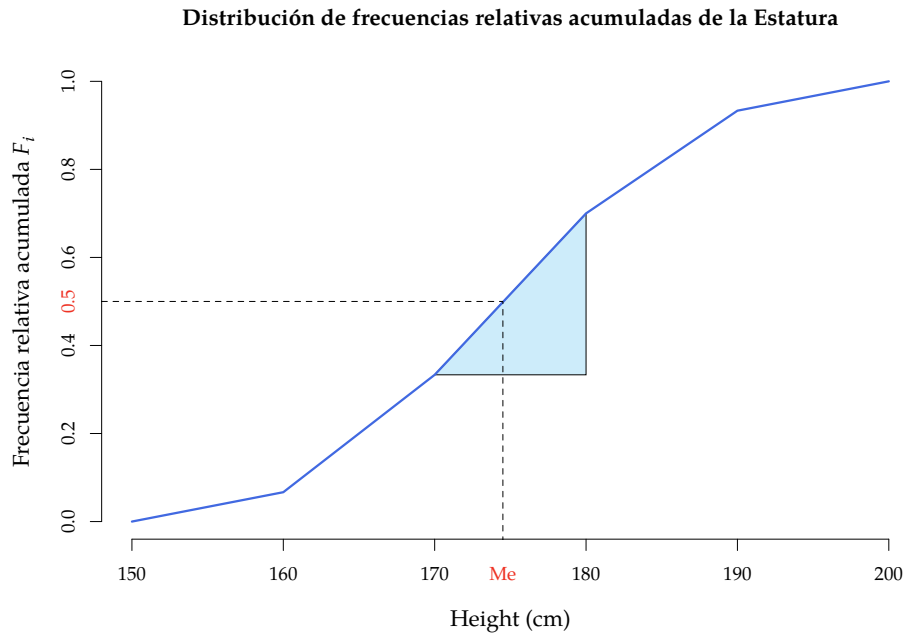
$$\tan(\alpha) = \frac{0.5 - F_{i-1}}{Me - l_{i-1}}$$

$$Me = l_i + \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}(l_i - l_{i-1}) = l_i + \frac{0.5 - F_{i-1}}{f_i}a_i$$

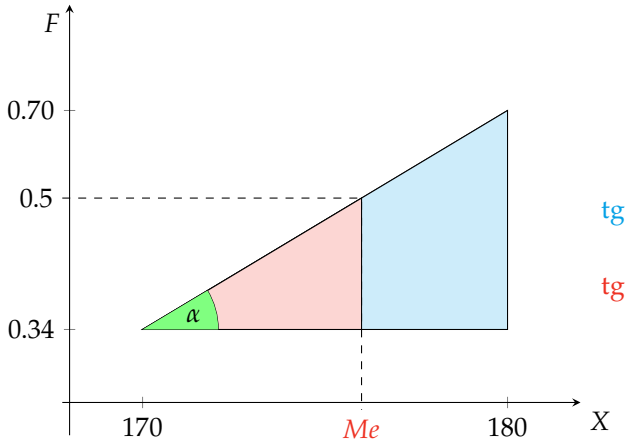
Cálculo de la mediana

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 $(170, 180]$.



Interpolación en el polígono de frecuencias absolutas acumuladas



$$\operatorname{tg}(\alpha) = \frac{0.7 - 0.34}{180 - 170}$$

$$\operatorname{tg}(\alpha) = \frac{0.5 - 0.34}{Me - 170}$$

$$Me = 170 + \frac{0.5 - 0.34}{0.7 - 0.34}(180 - 170) = 170 + \frac{0.16}{0.36}10 = 174.54 \text{ cm}$$

Moda

Definición 5 (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
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
$(170,180]$	**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.

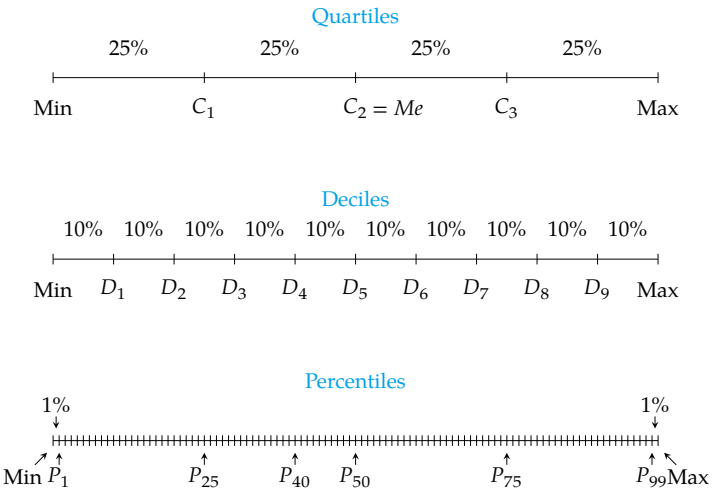
Los más utilizados son:

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).

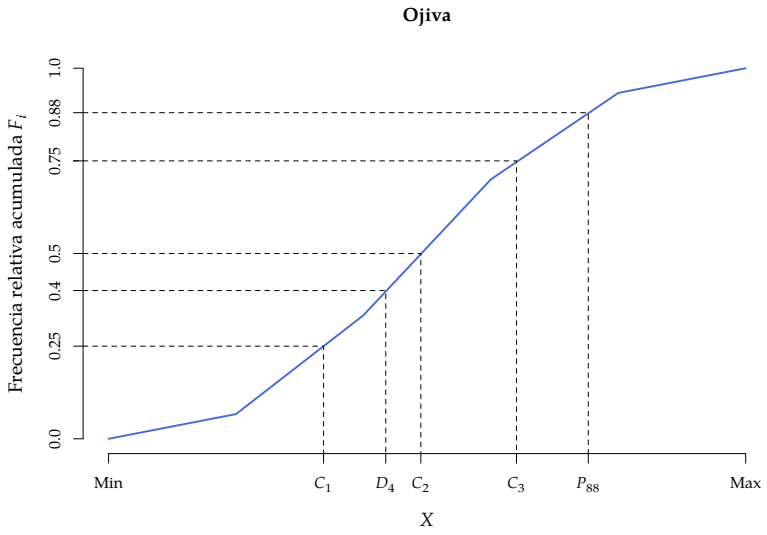
Cuantiles



Observese que hay una correspondencia entre los cuartiles, decilies y percentiles. Por ejemplo, el primer cuartil coincide con el percentil 25, y el cuarto decil coincide con el percentil 4o.

Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. La única diferencia es la frecuencia relativa acumulada que le corresponde a cada cuantil.



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

$$F_{C_1} = 0.25 \Rightarrow C_1 = 1 \text{ hijos},$$
$$F_{C_2} = 0.5 \Rightarrow C_2 = 2 \text{ hijos},$$
$$F_{C_3} = 0.75 \Rightarrow C_3 = 2 \text{ hijos},$$
$$F_{D_3} = 0.3 \Rightarrow D_3 = 1 \text{ hijos},$$
$$F_{P_{92}} = 0.92 \Rightarrow P_{92} = 3 \text{ hijos}.$$

1.5 Estadísticos de dispersión

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.

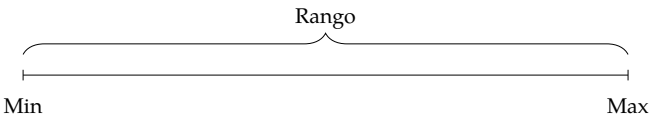
- Desviación Típica.
- Coeficiente de Variación.

Recorrido

Definición 6 (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.

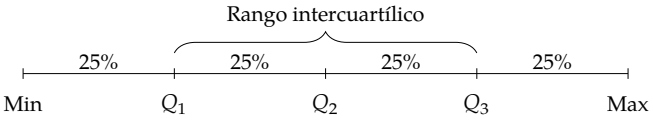


Rango intercuartílico

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 7 (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 mide la variación del 50% de los datos centrales.

Diagrama de caja y bigotes

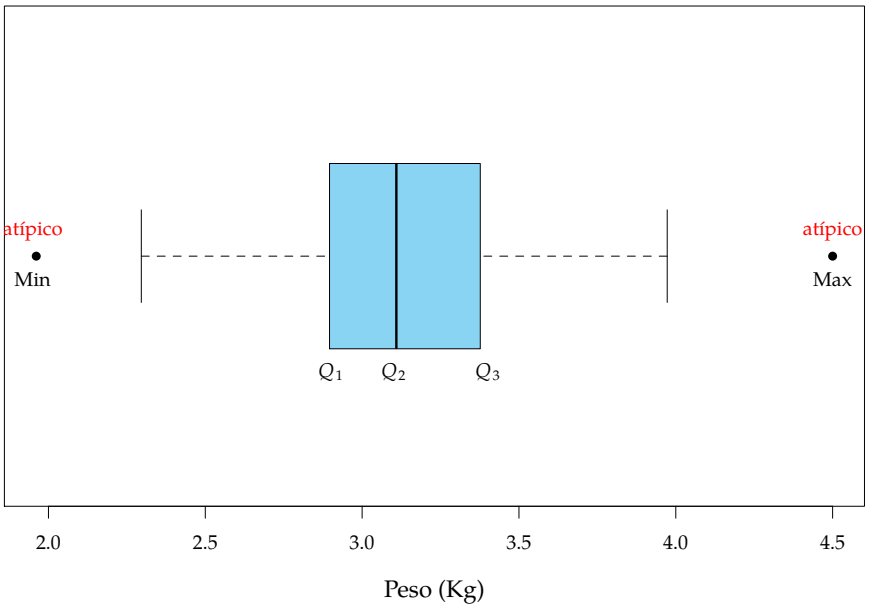
La dispersión de una variable suele representarse gráficamente mediante un **diagrama de caja y bigotes**, que consiste en una caja sobre un eje X donde el borde inferior de la caja es el primer cuartil, y el borde superior el tercer cuartil, y por tanto, la anchura de la caja es el rango intercuartílico.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

Diagrama de caja y bigotes del peso de recien 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:

$$v_1 = C_1 - 1.5RI$$
$$v_2 = C_3 + 1.5RI$$

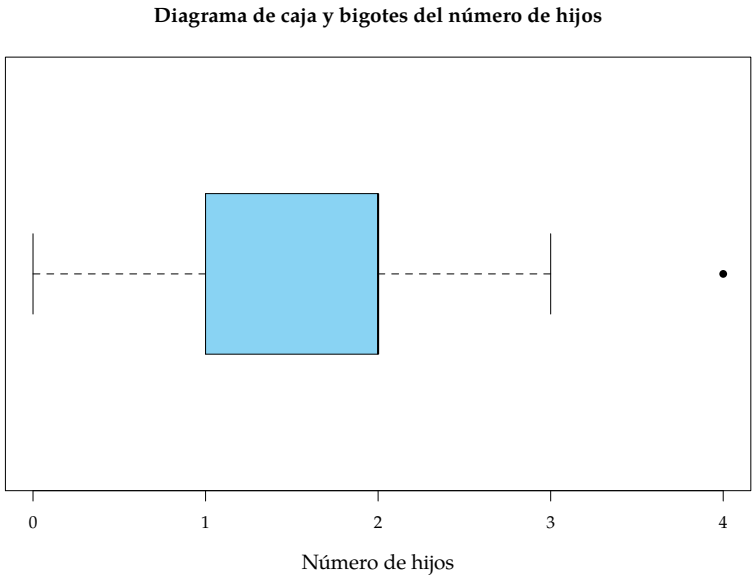
   A partir de las vallas se buscan los valores b_1, que es el mínimo valor de la muestra mayor o igual que v_1, y b_2, que es máximo valor de la muestra menor o igual que v_2. Para el bigote inferior se dibuja un segmento desde el borde inferior de la caja hasta b_1 y para el superior se dibuja un segmento 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

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.

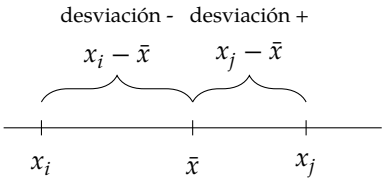


ç

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.

Varianza y desviación típica

Definición 8 (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:

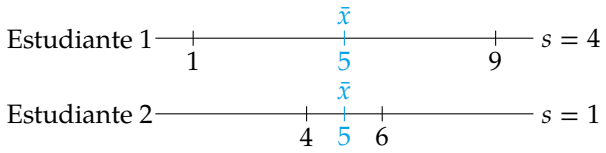
Definición 9 (Desviación típica s)**.** La *desviación típica muestral* de una variable X se define como la raíz cuadrada positiva de su varianza muestral.

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

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. Cuando la varianza o la desviación típica son pequeñas, los datos de la muestra están concentrados en torno a la media, y la media es una buena medida de representatividad. Por contra, cuando la varianza o la desviación típica son grandes, los datos de la muestra están alejados de la media, y la media ya no representa tan bien.

Desviación típica pequeña $\Rightarrow$ *Media representativa*
Desviación típica grande $\Rightarrow$ *Media no representativa*



¿En qué caso es más representativa la media?

Cálculo de la varianza y la desviación típica

Ejemplo con datos no agrupados

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.

Cálculo de la varianza y la desviación típica

Ejemplo con datos agrupados

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, especialmente cuando se compara la dispersión de variables con diferentes unidades.

Por este motivo, es también común utilizar la siguiente medida de dispersión que no tiene unidades.

Definición 10 (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.

El coeficiente de variación es muy útil para comparar la dispersión de distribuciones de variables diferentes, incluso si las variables tienen unidades diferentes.

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

Coeficiente de variación

Ejemplo

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

$$cv = \frac{s}{|\bar{x}|} = \frac{10.1}{|174.67|} = 0.06.$$

Esto significa que 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.

1.6 Estadísticos de forma

Estadísticos de forma

Son medidas que describen la forma de la distribución.

Los aspectos más relevantes son:

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

Apuntamiento: Mide el apuntamiento o el grado de concentración de valores en torno a la media de la distribución de frecuencias. El estadístico más utilizado es el *Coeficiente de Apuntamiento o Curtosis*.

Coeficiente de asimetría

Definición 11 (Coeficiente de asimetría muestral g_1)**.** El *coeficiente de asimetría muestral* de una variable X es 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.

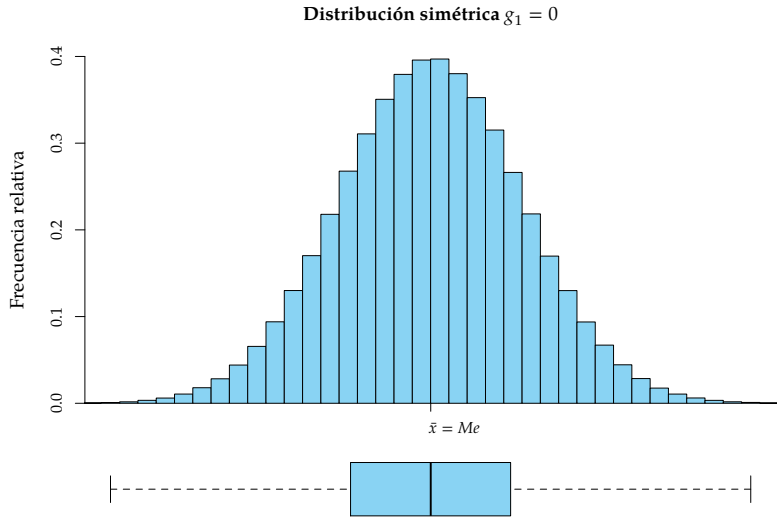
$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i/n}{s^3} = \frac{\sum (x_i - \bar{x})^3 f_i}{s^3}$$

Mide el grado de simetría de los valores de la muestra con respecto a la media muestra, es decir, cuantos valores de la muestra están por encima o por debajo de la media y cómo de alejados de esta.

- $g_1 = 0$ indica que hay el mismo número de valores por encima y por debajo de la media e igualmente alejados de ella (simétrica).
- $g_1 < 0$ indica que la mayoría de los valores son mayores que la media, pero los valores menores están más alejados de ella (asimétrica a la izquierda).
- $g_1 > 0$ indica que la mayoría de los valores son menores que la media, pero los valores mayores están más alejados de ella (asimétrica a la derecha).

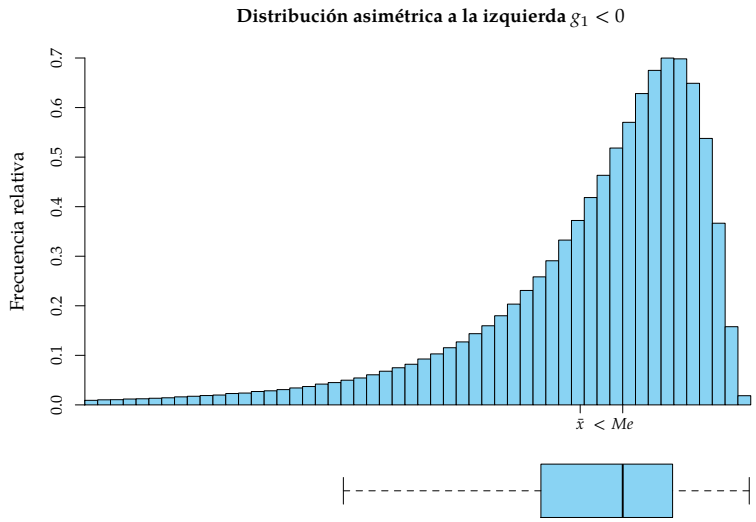
Coeficiente de asimetría

Ejemplo de distribución simétrica



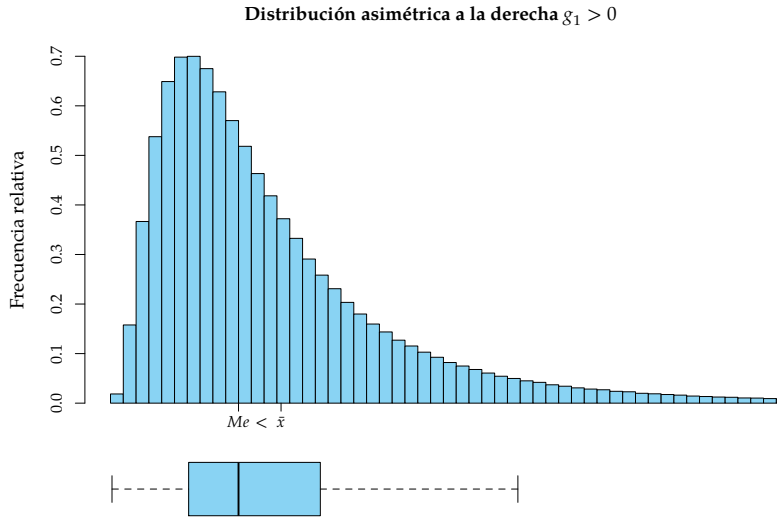
Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la izquierda



Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la derecha



Cálculo del coeficiente de asimetría

Ejemplo con datos agrupados

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 12 (Coeficiente de apuntamiento muestral g_2)**.** El *coeficiente de apuntamiento muestral* de una variable X es 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.

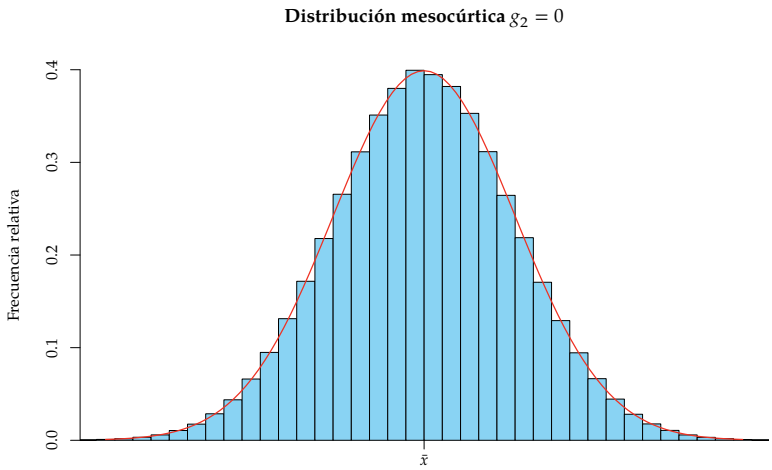
$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{\sum (x_i - \bar{x})^4 f_i}{s^4} - 3$$

El coeficiente de apuntamiento mide la concentración de valores en torno a la media y la longitud de las colas de la distribución. Se toma como referencia la distribución normal

- $g_2 = 0$ 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*).

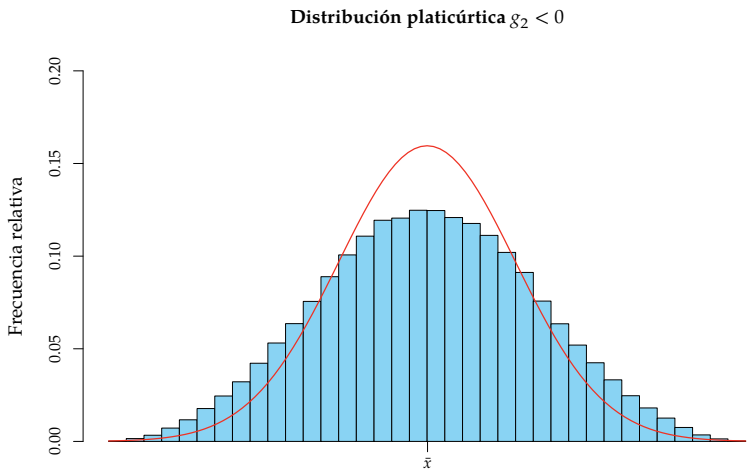
Coeficiente de apuntamiento o curtosis

Ejemplo de distribución mesocúrtica



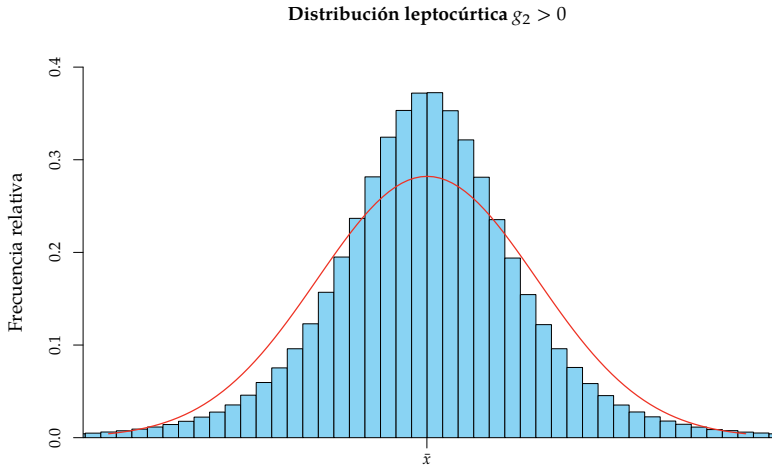
Coeficiente de apuntamiento o curtosis

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.

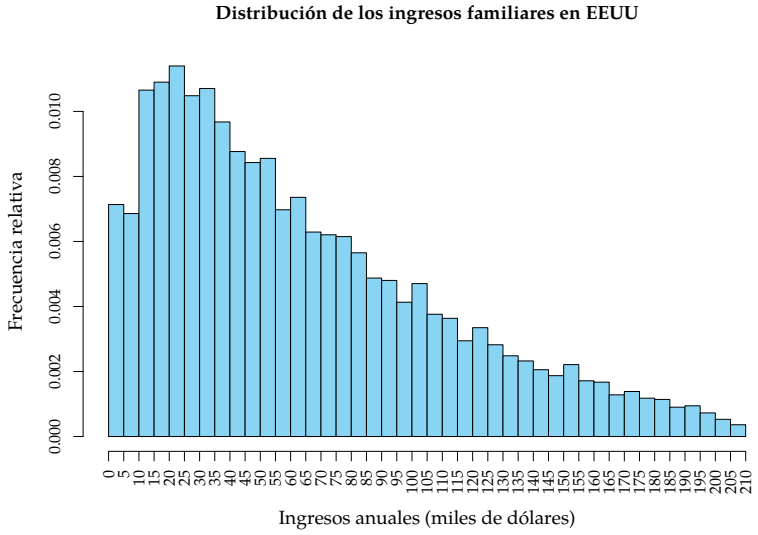
Las poblaciones normales se caracterizan por ser simétricas y mesocúrticas, de manera que, tanto el coeficiente de asimetría como el de apuntamiento pueden utilizarse para contrastar si los datos de la muestra provienen de una población normal.

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.

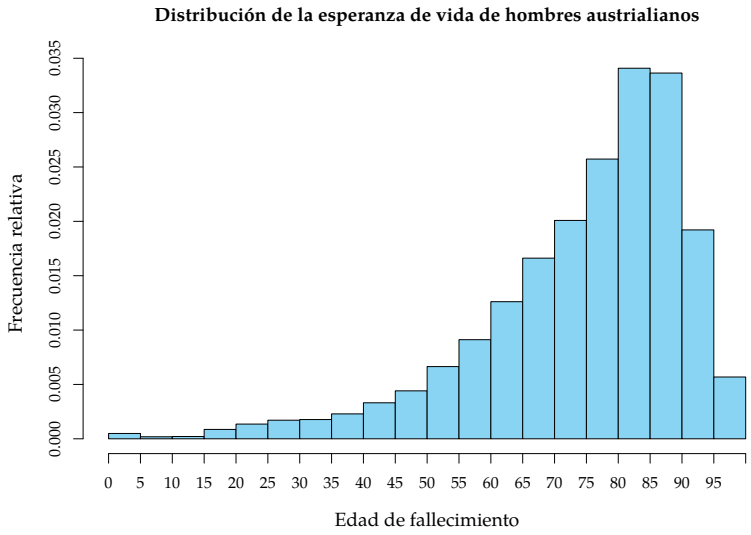
Distribución asimétrica a la derecha no normal

Ingresos por familia



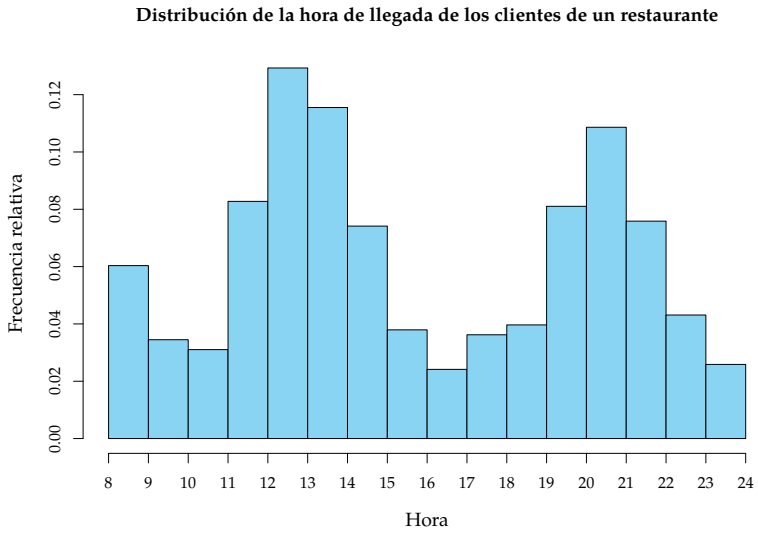
Distribución asimétrica a la izquierda no normal

Edad de fallecimiento



Distribución bimodal no normal

Hora de llegada de los clientes de un restaurante



1.7 Transformaciones de variables

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:

$$\bar{y} = a + b\bar{x},$$
$$s_y = |b|s_x$$

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 13 (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.

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

Para cada valor x_i de la muestra, la *puntuación típica* es el valor que resulta de aplicarle la transformación de tipicación

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

La puntuación típica es el número de desviaciones típicas que un valor está por encima o por debajo de la media, y es útil para evitar 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	$\bar{y} = 5$	$s_y = 3.16$

¿Ha tenido el mismo rendimiento el cuarto alumno en la asignatura X que el tercero en la asignatura Y?

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:

$X:$	-1.5	0	-0.5	1.5	0.5
$Y:$	-1.26	1.26	0.95	0	-0.95

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.

Transformación de tipificación y puntuaciones típicas

Ejemplo

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.

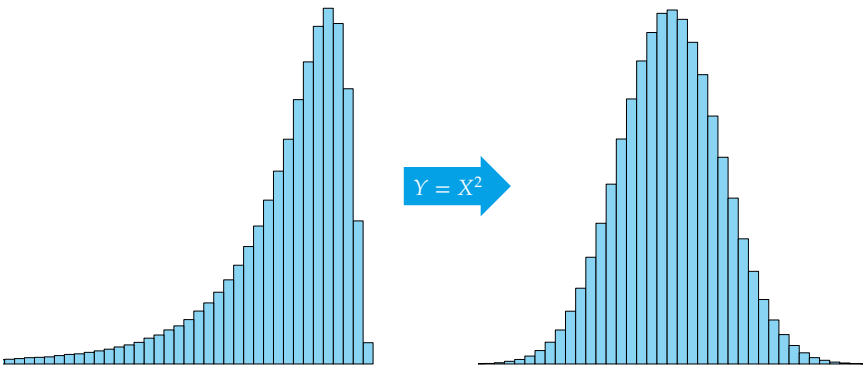
Pero si se considera el rendimiento relativo tomando las puntuaciones típicas se tiene:

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.

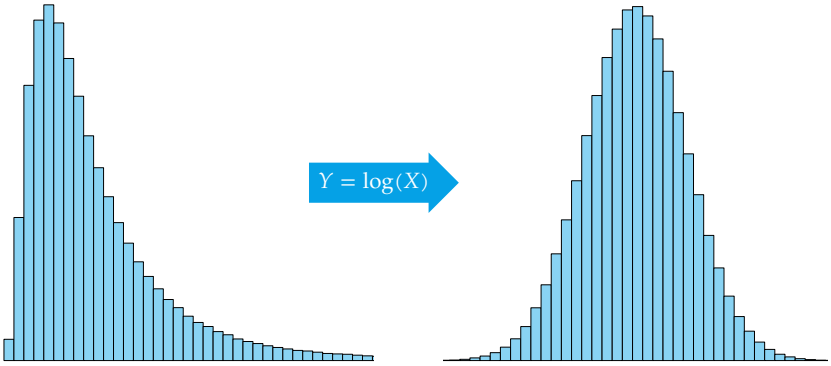
Transformaciones no lineales

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 correspondientes a las categorías de otra variable conocida como **variable clasificadora** o **factor**.

Variables clasificadoras

Dividiendo la muestra de estaturas según el sexo se obtienen dos submuestras:

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.

Comparación de distribuciones según los niveles de un factor