

Curso Básico de Estadística

Santiago Angulo Díaz-Parreño (sangulo@ceu.es)
José Miguel Cárdenas Rebollo (cardenas@ceu.es)
Anselmo Romero Limón (arlimon@ceu.es)
Alfredo Sánchez Alberca (asalber@ceu.es)



CEU
*Universidad
San Pablo*

Curso 2010-2011
©Copyleft

Curso básico de estadística




Alfredo Sánchez Alberca (asalber@gmail.com).

Esta obra está bajo una licencia Reconocimiento-No comercial-Compartir bajo la misma licencia 2.5 España de Creative Commons. Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by-nc-sa/2.5/es/> o envíe una carta a Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Con esta licencia eres libre de:

- Copiar, distribuir y mostrar este trabajo.
- Realizar modificaciones de este trabajo.

Bajo las siguientes condiciones:

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
 -  **No comercial.** No puede utilizar esta obra para fines comerciales.
 -  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.
-
- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
 - Estas condiciones pueden no aplicarse si se obtiene el permiso del titular de los derechos de autor.
 - Nada en esta licencia menoscaba o restringe los derechos morales del autor.

1 Estadística Descriptiva



Estadística Descriptiva

- Variables estadísticas
- Distribución de frecuencias
- Representaciones gráficas
- Estadísticos muestrales
- Estadísticos de posición
- Estadísticos de dispersión
- Estadísticos de forma
- Transformaciones de variables

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 Representar dichos datos gráficamente y en forma de tablas.
- 3 Calcular medidas que resuman la información que contiene la muestra (*estadísticos muestrales*).

Su poder inferencial es mínimo, por lo que nunca deben sacarse conclusiones sobre la población a partir de las medidas resumen que aporta la estadística descriptiva.

Como se vió en el tema introductorio, la estadística trata de conocer las poblaciones a partir del estudio de muestras. Una vez obtenida la muestra de la población, el siguiente paso en cualquier estudio estadístico es el análisis descriptivo de la muestra y este es el objeto la estadística descriptiva, que es la parte de la estadística encargada de representar, analizar y resumir la información contenida en la muestra. El objetivo es exprimir bien la muestra para sacarle toda la información posible y sintetizar dicha información para hacerla comprensible.

Las tareas que realiza la estadística descriptiva suelen ser:

1. Clasificar, agrupar y ordenar los datos de la muestra.
2. Representar dichos datos gráficamente y en forma de tablas.
3. Calcular medidas que resuman la información que contiene la muestra (*estadísticos muestrales*).

Aunque una buena descripción de la muestra facilita el posterior conocimiento de la población, nunca deben sacarse conclusiones sobre la población a partir de las medidas resumen que aporta la estadística descriptiva, ya que de esto se encarga la estadística inferencial que se vera más adelante en este curso.

Variables estadísticas y atributos

La característica objeto de estudio puede ser de dos tipos:

Atributos: De carácter cualitativo.

Variables estadísticas: De carácter cuantitativo.

A su vez, los atributos se dividen en:

Nominales: No existe un orden entre las modalidades.

Ejemplo: El color de ojos o de pelo.

Ordinales: Existe un orden entre las modalidades.

Ejemplo: El grado de gravedad de un paciente o la calificación de un curso.

Y las variables estadísticas en:

Discretas: Reciben valores aislados.

Ejemplo: El número de hijos o el número de coches.

Continuas: Pueden recibir cualquier valor de un intervalo.

Ejemplo: El peso o la estatura.

Todo estudio estadístico comienza por la identificación de las características que interesa estudiar en la población y que se medirán en los individuos de la muestra. Estas características pueden ser de dos tipos según sean de naturaleza cuantitativa o cualitativa, es decir, si miden cantidades o cualidades. Las características de naturaleza cualitativa se conocen como atributos o variables cualitativas, mientras que las de naturaleza cuantitativas se conocen como variables estadísticas o cuantitativas.

Los atributos, a su vez, pueden ser de dos tipos, nominales, cuando no existe un orden natural entre las modalidades o valores que puede tomar el atributo, como por ejemplo el color de ojos, que puede ser negro, azul, verde, etc. pero no tiene sentido ordenar los colores; u ordinales, cuando existe un orden natural entre las modalidades, como por ejemplo el grado de gravedad de un paciente que puede ir de leve, moderado, grave o muy grave siguiendo un orden de menor a mayor gravedad.

Por su parte las variables estadísticas también pueden clasificarse como discretas, cuando toman valores aislados que suelen ser números enteros, como por ejemplo el número de hijos de una familia, que puede ser 0, 1, 2, 3, es decir, un número entero positivo pero no puede tomar valores decimales como 1.27; o continuas cuando si pueden tomar cualquier valor dentro de un intervalo real, como por ejemplo la estatura, donde, entre dos estaturas cualesquiera como 1.60 y 1.70 podemos encontrar infinitas estaturas posibles.

Conocer el tipo de las variables u atributos es fundamental pues las técnicas que se utilizarán para describirlas dependen del tipo de característica.

A la hora de seleccionar las variables que se estudiarán conviene tomar todas las variables que puedan tener relación con el fenómeno que se pretende estudiar, y en esto cabe más pecar por exceso que por defecto, ya que si con el transcurso del estudio se observa que una variable no aporta nada, siempre se puede descartar a posteriori, mientras que si se observa que falta una variable importante, en la mayoría de los casos no se podrá volver atrás para medirla.

En el caso de las variables numéricas, es importante también explicitar las unidades en que van a medirse.

La matriz de datos

Las variables o atributos a estudiar se medirán en cada uno de los individuos de la muestra, obteniendo un conjunto de datos que suele organizarse en forma de matriz que se conoce como **matriz de datos**.

En esta matriz cada columna contiene la información de una variable y cada fila la información de un individuo.

Ejemplo

	Edad (años)	Sexo	Peso (Kg)	Altura (cm)
José Luis Martínez	18	H	85	179
Rosa Díaz	32	M	65	173
Javier García	24	H	71	181
Carmen López	35	M	65	170
Marisa López	46	M	51	158
Antonio Ruiz	68	H	66	174

Una vez seleccionadas las variables, estas se medirán en cada uno de los individuos de la muestra con lo que se obtendrá un conjunto de datos que se organiza en forma de matriz conocida como matriz de datos. Es importante organizar la matriz de manera que en cada columna contenga la información de una variable y cada fila la información de un individuo.

En este ejemplo puede observarse la matriz de datos correspondiente a una muestra de cinco individuos en los que se han medido 4 características, tres variables, que son la Edad, el Peso y la Altura, y un atributo que es el Sexo. Como puede apreciarse la primera columna contiene las edades de todos los individuos de la muestra, la segunda el sexo, la tercera los pesos y la última las alturas, mientras la información de cada fila hace referencia al mismo individuo, de manera que la primera fila, por ejemplo, contiene los datos de José Luis Martínez que tiene 18 años, es hombre, pesa 85 Kg y mide 179 cm.

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. 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 discretas con muchos valores diferentes, y con variables continuas.

La matriz de datos contiene toda la información de la muestra pero resulta difícil de interpretar, por lo que para sintetizar y resumir esa información se realizan varias tareas que empiezan por la clasificación de los valores.

Esta clasificación consiste en ordenar los valores de menor a mayor, cuando existe un orden entre dichos valores, o simplemente en reunir los valores iguales cuando no hay un orden entre ellos.

En las variables cuantitativas, cuando el número de valores distintos es muy grande, estos suelen agruparse en intervalos.

Clasificación de la muestra



$X = \text{Estatura}$

Clasificación



En este ejemplo tenemos una muestra de 10 individuos en los que se ha medido su estatura. Como se trata de una variable cuantitativa, la clasificación consiste, primero en ordenar las estaturas de menor a mayor.

Recuento de frecuencias



X = Estatura

Frecuencias



Y a continuación, si hay valores repetidos, contar la frecuencia de repetición de cada estatura o cuántas estaturas caen en cada uno de los intervalos que hayamos definido en caso de agrupar los datos. Por ejemplo, la primera estatura se repite dos veces y por tanto tiene frecuencia 2, la segunda estatura se repite tres veces y tiene frecuencia 3 y así sucesivamente.

Definición (Frecuencias muestrales)

Dada una muestra de tamaño n de una variable X , para cada valor de la variable x_i observado en la muestra, se define

- **Frecuencia absoluta n_i** : Es el número de individuos de la muestra que presentan el valor x_i .
- **Frecuencia relativa f_i** : Es la proporción de individuos de la muestra que presentan el valor x_i .

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

- **Frecuencia absoluta acumulada N_i** : Es el número de individuos de la muestra que presentan un valor menor o igual que x_i .

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

- **Frecuencia relativa acumulada F_i** : Es la proporción de individuos de la muestra que presentan un valor menor o igual que x_i .

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

Existen distintos tipos de frecuencias que pueden calcularse. Dada una muestra de tamaño n de una variable X , para cada valor de la variable x_i observado en la muestra, se define

- La frecuencia absoluta, que se representa como n_i , es el número de individuos de la muestra que presentan el valor x_i , es decir el número de veces que se repite dicho valor.
- La frecuencia relativa, que se denota f_i , es la proporción de individuos de la muestra que presentan el valor x_i . La frecuencia relativa se calcula dividiendo la frecuencia absoluta entre el tamaño de la muestra.

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

Cuando se multiplica por 100 se convierte en un porcentaje.

- La frecuencia absoluta acumulada, que se escribe N_i , es el número de individuos de la muestra que presentan un valor menor o igual que x_i . Se calcula acumulando las frecuencias absolutas de los valores menores o iguales a x_i y de ahí su nombre.

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

- La frecuencia relativa acumulada, que se escribe F_i , es la proporción de individuos de la muestra que presentan un valor menor o igual que x_i . Puede calcularse acumulando las frecuencias relativas de los valores menores o iguales que x_i o bien dividiendo la frecuencia absoluta acumulada de x_i entre el tamaño de la muestra.

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

Al igual que la frecuencia relativa, si se multiplica por 100 se convierte en el porcentaje acumulado.

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

Normalmente las frecuencias muestrales suelen organizarse en forma de tabla en la que cada fila corresponde a un valor de la variable o un intervalo de valores, que se ordenan, siempre que sea posible, de menor a mayor, y cada columna corresponde a una frecuencia.

En esta tabla siempre se debe cumplir que la suma de las frecuencias absolutas es igual al tamaño de la muestra, la suma de las frecuencias relativas vale 1, la última frecuencia absoluta acumulada es el tamaño de la muestra y la última frecuencia relativa acumulada es 1. De no ser así, habríamos cometido algún error en el cálculo de las frecuencias.

Tabla de frecuencias

Ejemplo de datos sin agrupar

En una encuesta a 25 matrimonios sobre el número de hijos que tenían se obtuvieron los siguientes datos:

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
Σ	25	1		

En este ejemplo se han tomado 25 matrimonios en los que se ha medido el número de hijos que tenían. Se trata de una variable cuantitativa discreta puesto que sólo puede tomar valores enteros positivos y además en la muestra sólo aparece 5 valores distintos que son 0, 1, 2, 3 y 4 hijos, por lo que no es necesario agrupar los datos en intervalos.

Para construir la tabla de frecuencias se comienza por el recuento de las frecuencias absolutas. Como puede observarse en la muestra, hay dos matrimonios que tienen 0 hijos, y por tanto, la frecuencia absoluta del 0 es 2, el 1 aparece 6 veces, el 2, 3 veces, el 3, 2 veces y finalmente el 4 sólo aparece una vez. Obsérvese cómo la suma de las frecuencias absolutas da 25 que es el tamaño muestral.

A continuación se calculan las frecuencias relativas, simplemente dividiendo cada frecuencia absoluta por el tamaño de la muestra que es 25. Por ejemplo, la frecuencia relativa del 0 es 2 entre 25 que es 0.08, y es la proporción de matrimonios en la muestra que tienen 0 hijos. Si se multiplica por 100 da un 8 %, es decir, el 8 % de los matrimonios tienen 0 hijos. Obsérvese cómo la suma de las frecuencias relativas vale 1.

Después se calculan las frecuencias absolutas acumuladas. Así, por ejemplo, la frecuencia absoluta acumulada del 0 es el número de matrimonios que tienen 0 o menos hijos, y al ser el 0 el menor de los valores de la muestra, coincide con la su frecuencia absoluta, que vale 2. La frecuencia absoluta acumulada del 1 es el número de matrimonios que tienen 1 o menos hijos, de manera que habría que acumula la frecuencia absoluta del 1 y del 0, es decir, 2 mas 6, que da un total de 8, y así sucesivamente. En general para calcular cada frecuencia absoluta acumulada se puede tomar la frecuencia absoluta acumulada anterior y sumarle la frecuencia absoluta del valor. $8+14=22$, $22+2=24$ y $24+1=25$. Obsérvese cómo la última frecuencia absoluta acumulada vale 25 que es el tamaño de la muestra.

Finalmente, se calculan las frecuencias relativas acumuladas, que pueden calcularse, bien acumulando las frecuencias relativas del mismo modo que se acumulaban las absolutas, o bien dividiendo las frecuencias absolutas acumuladas por el tamaño de la muestra. Así, por ejemplo, la frecuencia relativa acumulada del 0 es 2 entre 25, que vale 0.08 y coincide con su frecuencia relativa. La frecuencia relativa acumulada del 1 es 8 entre 25 que vale 0.32 y es la proporción de

Tabla de frecuencias

Ejemplo de datos agrupados

Se ha medido la estatura (en cm) de 30 universitarios obteniendo:

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
Σ	30	1		

En este otro ejemplo, se han medido las estaturas de 30 universitarios. Ahora se trata de una variable cuantitativa continua, y como siempre ocurre con este tipo de variables, el número de valores distintos que aparece en la muestra suele ser demasiado grande, por lo que se tiene a agruparlos en intervalos.

En este caso se ha optado por construir 5 intervalos de amplitud 10 cm, empezando en 150 cm y terminando en 200 cm.

El cálculo de frecuencias absolutas es similar al caso anterior, salvo que ahora no se cuenta el número de estaturas que se repiten, sino el número de estaturas que caen en cada intervalo. Por ejemplo, la frecuencia absoluta del intervalo $(150, 160]$ es 2 ya que en la muestra hay dos personas, una que mide 158 y otra que mide 154, que caerían en este intervalo. Una vez calculadas las frecuencias absolutas, el cálculo del resto de frecuencias es idéntico al caso de datos no agrupados.

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 a la raíz cuadrada del tamaño muestral \sqrt{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.

Cada intervalo de agrupación de datos se denomina clase y el centro del intervalo se llama marca de clase. Cuando se decide agrupar los datos en clases, hay que tener en cuenta lo siguiente:

- En primer lugar, el número de intervalos no debe ser muy grande ni muy pequeño. Si hay muchos intervalos, la tabla será enorme y no sintetizará bien la información de la muestra, mientras que si tomamos muy pocos intervalos, su amplitud será muy grande y se perderá gran parte de la información que contiene la muestra, ya que cuando un individuo se cuenta dentro de un intervalo, pierde su valor particular y pasa a ser representado por la marca de clase. Una regla orientativa es tomar un número de intervalos próximo a la raíz cuadrada del tamaño muestral \sqrt{n} .
- En segundo lugar, los intervalos no deben solaparse, ya que de lo contrario se correría el riesgo de que algún individuo cayese en dos intervalos distintos, por eso se suelen construir intervalos abiertos por la izquierda y cerrados por la derecha o al revés, pero siempre siguiendo el mismo criterio. Y también deben cubrir todo el rango de valores, ya que de lo contrario se correría el riesgo de que algún individuo no caería en ningún intervalo y quedaría sin contarse.
- Por último, 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 una muestra de 30 personas son:

A, B, B, A, AB, 0, 0, A, B, B, 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
Σ	30	1

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

En este otro ejemplo, se ha medido el grupo sanguíneo de un grupo de 30 personas. Ahora se trata de un atributo nominal, de manera que como no hay orden entre sus valores, pueden ordenarse de cualquier manera en la tabla de frecuencias, pero el cálculo de frecuencias se realiza como en casos anteriores, con la particularidad de que en este caso no tiene sentido calcular las frecuencias acumuladas. ¿Te imaginas por qué?

Representaciones gráficas

También 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:** Consiste en un diagrama sobre el plano cartesiano en el que en el eje X se representan los valores de la variable y en el eje Y las frecuencias. Sobre cada valor de la variable se levanta una barra de altura la correspondiente frecuencia. Se utiliza con variables discretas no agrupadas.
- **Histograma:** Es similar a un diagrama de barras pero representando en el eje X las clases en que se agrupan los valores de la variable y levantando las barras sobre todo el intervalo de manera que las barras están pegadas unas a otras. Se utiliza con variables discretas agrupadas y con variables continuas.
- **Diagrama de sectores:** Consiste en un círculo dividido en sectores de área proporcional a la frecuencia de cada valor de la variable. Se utiliza sobre todo con atributos.

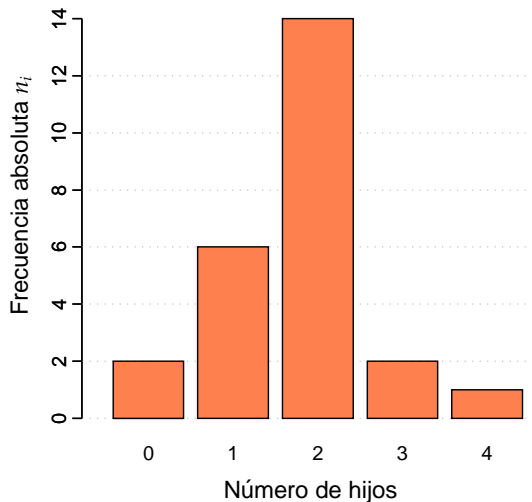
En cada uno de los diagramas pueden representarse los distintos tipos de frecuencias, siempre que estas existan.

Habitualmente las frecuencias también suelen representarse de manera gráfica. Dependiendo del tipo de variable y de si se han agrupado o no los datos, se utilizan distintos tipos de gráficos, entre los que cabe destacar los diagramas de barras, los histogramas y los diagramas de sectores.

Veamos un ejemplo de cada uno de ellos.

Diagrama de barras de frecuencias absolutas

Datos sin agrupar

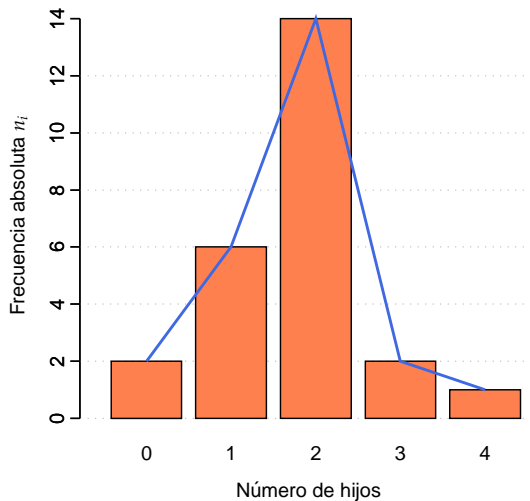


El diagrama de barras se utiliza con variables cuantitativas y datos sin agrupar y consiste en un sistema de ejes cartesianos en el que se representan los valores de la variable en el eje de abscisas y las frecuencias correspondientes en el eje de ordenadas, de manera que sobre cada valor se levanta una barra de altura la correspondiente frecuencia. La frecuencia representada puede ser cualquiera de las cuatro vistas por lo que hay cuatro tipos de diagramas de barras.

En este ejemplo tenemos el diagrama de barras de frecuencias absolutas correspondiente a la muestra de 25 matrimonios en los que se midió el número de hijos. Como puede observarse la barra que hay sobre el 0 tiene altura 2 por que la frecuencia absoluta del 0 es 2, la barra que hay sobre el 1 tiene altura 6 porque el 1 tiene frecuencia 6, etc.

Polígono de frecuencias absolutas

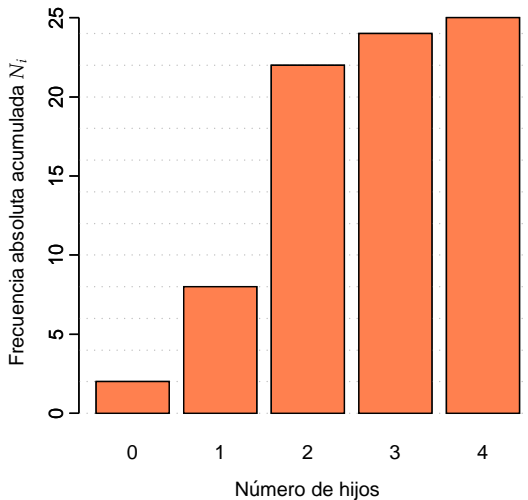
Datos sin agrupar



A veces, sobre el diagrama de barras se suele representar un polígono conocido como polígono de frecuencias, y que en el caso de las frecuencias absolutas se construye uniendo los puntos más altos de las barras mediante segmentos.

Diagrama de barras de frecuencias acumuladas

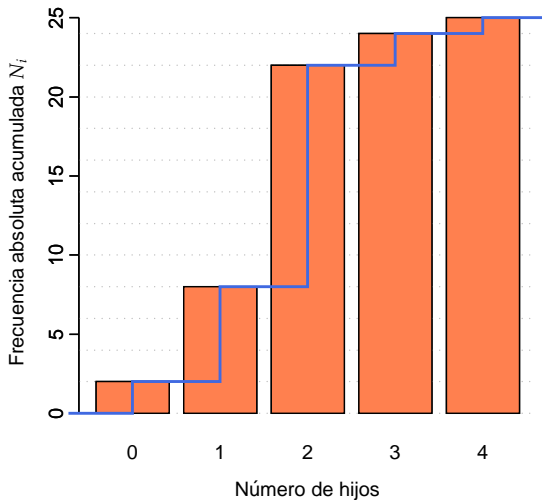
Datos sin agrupar



En este gráfico aparece el diagrama de barras de frecuencias absolutas acumuladas, y como puede apreciarse, al tratarse de frecuencias acumuladas, las barras van creciendo progresivamente hasta que la última barra tiene la altura del tamaño muestral.

Polígono de frecuencias absolutas acumuladas

Datos sin agrupar

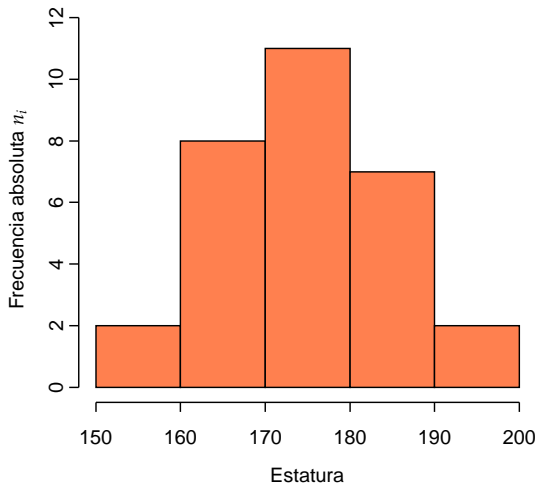


El polígono de frecuencias absolutas acumuladas, a diferencia del de absolutas, tiene forma de escalera, reflejando que la acumulación de individuos se produce a saltos, y no de manera progresiva como podrían entenderse si simplemente se uniesen los puntos más altos de cada barra.

Este polígono refleja que la frecuencia absoluta acumulada de cualquier valor anterior al 0 es 0 ya que no hay matrimonios con menos de 0 hijos. Al llegar al 0 nos encontramos con dos matrimonios que no tienen hijos, y de pronto la frecuencia absoluta acumulada pasa a valer 2. Del 0 al 1, no nos encontramos ningún valor en la muestra, ya que no hay matrimonios que tengan entre 0 y 1 hijos, por lo que el polígono continúa con un segmento horizontal sobre el dos, reflejando que al frecuencia absoluta acumulada se mantiene en 2. Cuando se llega al 1, volvemos a encontrarnos 6 matrimonios con 1 hijo, y el polígono da un salto hasta el 8. Del 1 al dos se mantiene constante en el 8, hasta llegar al 2 donde vuelve a dar un salto hasta el 22 y así sucesivamente hasta que al final se alcanza el nivel del tamaño muestral.

Histograma de frecuencias absolutas

Datos agrupados

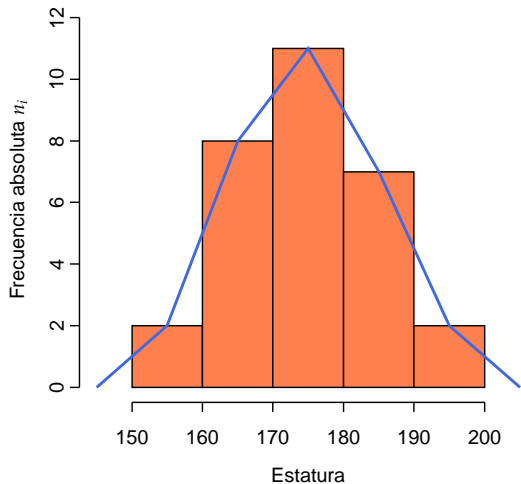


El histograma es parecido al diagrama de barras, solo que se utiliza con variables cuantitativas en las que se han agrupado los datos en intervalos. La idea es la misma salvo que las barras que reflejan las frecuencias se levantan sobre todo el intervalo, en lugar de sobre un valor concreto, de manera que las barras aparecen pegadas unas a otras. Al igual que para el diagrama de barras, existen cuatro tipos de histogramas, uno para cada tipo de frecuencias.

En este ejemplo el gráfico representa el histograma de frecuencias absolutas acumuladas de la muestra de 30 universitarios en los que se había medido la estatura. Como puede observarse la primera barra se levanta sobre el intervalo $(150,160]$ y tiene una altura 2, que es su frecuencia absoluta.

Polígono de frecuencias absolutas

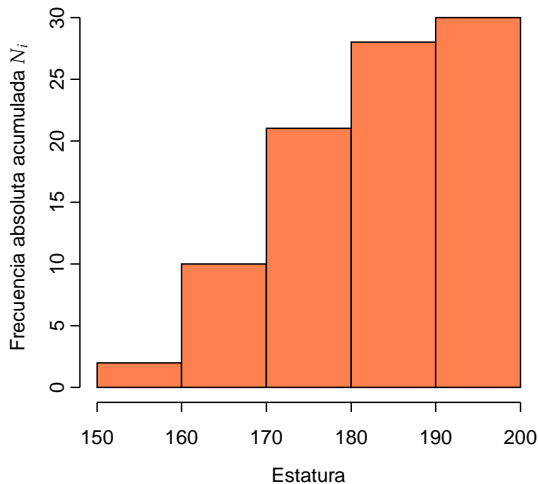
Datos agrupados



Al igual que para el diagrama de barras, para el histograma también se puede construir el polígono de frecuencias absolutas uniendo con segmentos los puntos más altos de cada barra sobre el centro del intervalo.

Histograma de frecuencias absolutas acumuladas

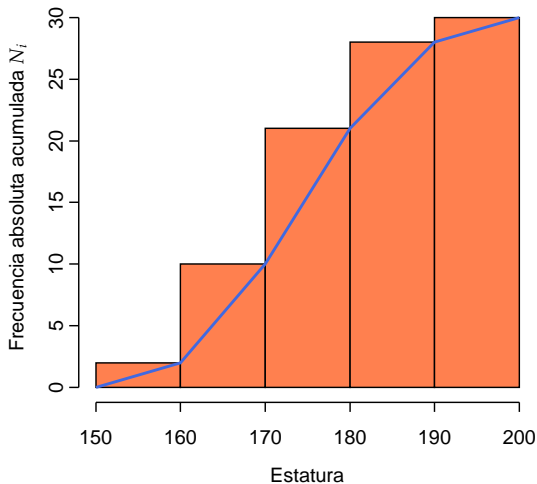
Datos agrupados



Este otro gráfico es el histograma de frecuencias absolutas acumuladas, en el que como puede apreciarse las barras van creciendo progresivamente hasta alcanzar el tamaño de la muestra.

Polígono de frecuencias absolutas acumuladas

Datos agrupados

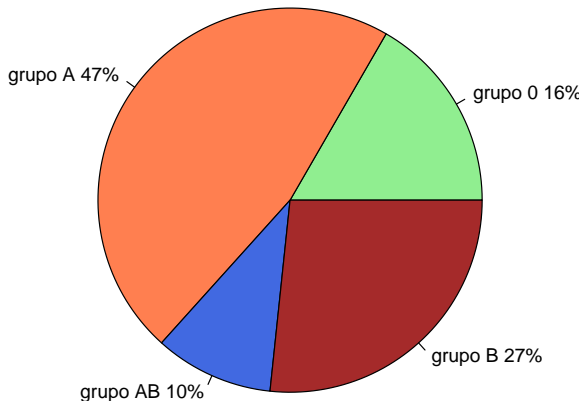


En el caso de histogramas de frecuencias acumuladas, el polígono correspondiente se construye mediante segmentos que unen el vértice inferior izquierdo y el vértice superior derecho de cada barra, reflejando, a diferencia del polígono de frecuencias acumuladas del diagrama de barras que tenía forma de escalera, que la acumulación de individuos es progresiva a medida que se recorre el intervalo.

Diagrama de sectores

Atributos

Distribución del grupo sanguíneo



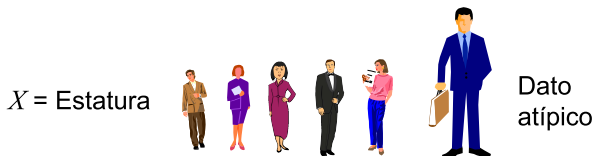
Tanto el diagrama de barras como el histograma se suele utilizar para variables cuantitativas ya que al tratarse de un sistema cartesiano, los valores de las variables deben ser numéricos. En el caso de los atributos el diagrama que se suele utilizar es el diagrama de sectores, que consiste en un círculo dividido en sectores de área o ángulo proporcional a la frecuencia correspondiente.

Para calcular el ángulo correspondiente a cada categoría, basta hacer una simple regla de tres, teniendo en cuenta que los 360 grados de la circunferencia se corresponden con el tamaño muestral y calculando el ángulo correspondiente a cada frecuencia.

En este gráfico se pueden apreciar los sectores correspondientes a cada uno de los grupos sanguíneos en la muestra anterior de 30 personas.

Datos atípicos

Uno de los principales problemas de las muestras son los datos atípicos. Los **datos atípicos** son valores de la variable que se diferencian mucho del resto de los valores.



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, aunque más adelante veremos un diagrama para detectarlos.

Uno de los principales problemas que presentan las muestras es que pueden contener datos atípicos, es decir, valores que se diferencian mucho del resto, bien porque son mucho mayores o menores que los demás.

Es muy importante detectar los datos atípicos antes de realizar cualquier análisis de los datos, pues suelen distorsionar las conclusiones extraídas de la muestra.

Para detectarlos hay que buscar en los extremos de la distribución, entre los valores menores y mayores, aunque más adelante se mostrará un gráfico que permite detectarlos fácilmente.

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:

- Eliminarlo: Siempre que estemos seguros de que se trata de un error de medida.
- Sustituirlo: Si se trata de un individuo real pero que no concuerda con el modelo de distribución de la población. En tal caso se suele reemplazar por el mayor o menor dato no atípico.
- Dejarlo: Si se trata de un individuo real aunque no concuerde con el modelo de distribución. En tal caso se suele modificar el modelo de distribución supuesto.

Si el tamaño de la muestra es grande, los datos atípicos tienen menor influencia que cuando el tamaño muestral es pequeño. En este último caso conviene eliminar el dato atípico si se trata de un error de medida, sustituirlo por otro valor más normal, que suele ser el mínimo o el máximo de los valores que se consideren normales, siempre que sea un valor que corresponda a un individuo real pero que no se ajusta al modelo de distribución de la población, o bien dejarlo y cambiar el modelo de distribución para que pueda explicar la existencia de esos valores atípicos.

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.

La tabla de frecuencias sintetiza la información de la variable estudiada en la muestra, pero en muchas ocasiones resulta insuficiente para describir determinados aspectos de la distribución, como por ejemplo la variabilidad entre los valores o los valores en torno a los cuales se agrupan la mayor parte de los individuos de la muestra.

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.

Dependiendo del aspecto de la distribución que describen se clasifican en:

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

Estadísticos de Dispersión: Miden la heterogeneidad o variabilidad 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.

Veremos primero los estadísticos de posición, que se ocupan de dos tareas: Por un lado, ver en torno a que valores se agrupan los valores de la muestra y qué valores son los más representativos de la muestra. Estos son tres, la media, la mediana y la moda y se conocen como estadísticos de tendencia central. Y por otro lado, ver cómo se reparten los datos a lo largo de la distribución. De esto se encargan los percentiles que dividen la distribución en partes iguales.

Empezaremos viendo los estadísticos de tendencia central

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.

La media aritmética es, sin duda, el estadístico de tendencia central más conocido. Se define como la suma de los valores observados en la muestra dividida por el tamaño muestral.

Si en lugar de la muestra tenemos la tabla de frecuencias, para calcular la media aritmética hay que sumar los productos de cada valor por su frecuencia absoluta y dividir la suma entre el tamaño de la muestra, o bien directamente sumando los productos de cada valor por su frecuencia relativa.

Algo que hay que advertir, es que, aunque la media aritmética es la medida de tendencia central más usada, no puede calcularse para atributos, ya que las categorías no son números que pueden operarse aritméticamente.

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
Σ	25	1	44	1,76

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

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

En el ejemplo del número de hijos en una muestra de 25 matrimonios, para calcular la media a partir de la muestra, basta con sumar los hijos de cada matrimonio, $1 + 2 + 3$, y así hasta el último, lo que da 44 hijos y dividirlo por el número de matrimonios, 25, con lo que se obtiene 1.76 hijos de media.

Si en lugar de la muestra tenemos la tabla de frecuencias, entonces conviene construir una nueva columna donde se vayan calculando los productos de cada valor por su frecuencia absoluta, es decir, $0 * 2 = 0 + 1 * 6 = 6$, etc., luego sumar todos estos productos, lo que de nuevo nos da 44 hijos y finalmente dividir la suma por el tamaño de la muestra que es 25, obteniendo de nuevo 1.76 hijos. La otra opción es construir otra columna donde se vayan calculando los productos de cada valor por su frecuencia relativa, es decir, $0*0.08=0 + 1*0.24 = 0.24$, etc. y finalmente sumar estos productos para obtener, una vez más, 1.76 hijos de media.

Cálculo de la media aritmética

Ejemplo con datos agrupados

En el ejemplo anterior de las estaturas se tiene

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

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
Σ		30	1	5240	174,67

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

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

Cuando se trabaja con datos agrupados, el cálculo de la media aritmética a partir de la tabla de frecuencias, cambia ligeramente, ya que ahora en lugar de valores se tienen intervalos. Lo se hace es tomar como representante de cada clase el valor central, es decir, la marca de clase, y hacer los cálculos con ella. Por ejemplo, el primer intervalo que va de 150 cm a 160 cm, su marca de clase es el centro que vale 155 cm, y entonces se multiplica este valor por su frecuencia absoluta, que vale 2, dando 310 cm. Esto se repite para todos los valores, al igual que antes, luego se suman los productos, obteniendo 5240 cm, y se divide por el tamaño de la muestra que era 30, lo que da 174.67 cm de estatura media.

También podríamos haber hecho el cálculo con las frecuencias relativas.

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:

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

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

La media ponderada para una muestra de tamaño n en la que cada valor x_i tiene asociado un peso p_i se define como la suma de los productos de cada valor por su peso, dividida por la suma de todos los pesos.

En el caso de estar trabajando desde la tabla de frecuencias, los productos de los valores por los pesos deben multiplicarse también por la frecuencia absoluta de cada valor.

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:

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

Veamos un ejemplo. Supongamos que para evaluar el rendimiento académico de un alumno durante un curso queremos calcular la nota media del curso a partir de las notas de todas las asignaturas. Las asignaturas cursadas han sido matemáticas, de 6 créditos, que equivale a 60 horas de clase, donde ha sacado un 5, lengua, de 4 créditos, donde ha sacado un 3 y química, de 8 créditos, y donde ha sacado un 6.

Si se calcula la media aritmética, se obtiene una nota media de 4,67 puntos, lo que supondría un suspenso.

Ahora bien, esta nota no refleja bien el rendimiento del alumno ya que todas las notas no tienen el mismo peso. Por ejemplo la química tiene el doble de créditos que la lengua, y por tanto, debería tener el doble de peso. En consecuencia, es mucho más acertado calcular la media ponderada tomando como pesos los créditos de las asignaturas. Para calcularla multiplicamos la nota de cada asignatura por sus créditos, $5 \cdot 6 + 3 \cdot 4 + 6 \cdot 8$, que vale 90 y lo dividimos por la suma de los créditos, $6 + 4 + 8$, que vale 18, lo que da una nota media de 5.

Esta nota representa mucho mejor el rendimiento del alumno, y además ahora estaría aprobado.

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.

Otro estadístico de tendencia central bastante utilizado es la mediana, que se define como el valor que ocupa el centro de la distribución, una vez ordenados los datos de menor a mayor.

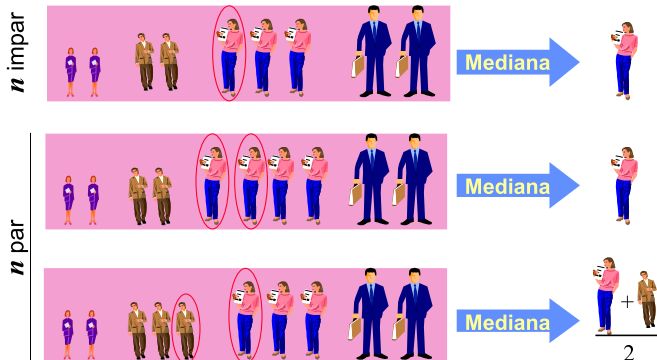
Como está en el centro de la distribución, la mitad de los valores quedarán por debajo de ella y la otra mitad por encima, y por tanto se cumple que la mediana tiene frecuencia absoluta acumulada $n/2$ o, lo que es lo mismo, frecuencia relativa acumulada 0,5.

La mediana resuelve en parte el problema de la media con los atributos, ya que puede calcularse para atributos ordinales donde las categorías pueden ordenarse, pero no se puede calcular 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 $\frac{n}{2} + 1$.



Una vez ordenados los valores de la muestra de menor a mayor, el cálculo de la mediana con datos no agrupados, depende de si el tamaño de la muestra es par o impar. Si es impar, entonces habrá un único valor central que ocupará la posición $\frac{n+1}{2}$ y dicho valor será la mediana. Mientras que si el tamaño muestral es par, entonces habrá dos valores centrales, que ocuparan las posiciones $\frac{n}{2}$ y $\frac{n}{2} + 1$. Si estos valores son iguales, la mediana será ese valor, mientras que si son distintos, la mediana estará entre los valores que ocupan estas posiciones y si puede se calculará su media aritmética.

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, **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:

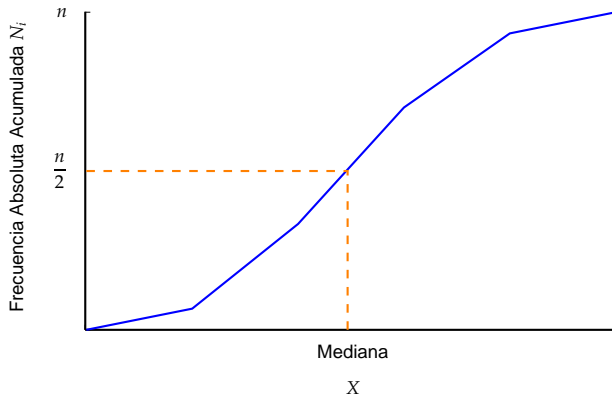
x_i	n_i	f_i	N_i	F_i
0	2	0,08	2	0,08
1	6	0,24	8	0,32
2	14	0,56	22	0,88
3	2	0,08	24	0,96
4	1	0,04	25	1
Σ	25	1		

Volviendo al ejemplo del número de hijos en una muestra de 25 matrimonios, para calcular la mediana, primero se ordenan los datos de menor a mayor, y después se busca el valor central, que en este caso, al ser un tamaño impar, es único y ocupa la posición $\frac{25+1}{2}$ que vale 13. Contamos 1,2, hasta la 13 y observamos que el valor que aparece es un 2. Luego la mediana es 2 hijos.

Para calcular la mediana a partir de la tabla de frecuencias, hay que buscar de nuevo el individuo que ocupa la posición 13, y para ello se busca en la tabla el valor que tiene una frecuencia absoluta acumulada igual o mayor que 13. En la tabla se puede observar que la primera frecuencia absoluta acumulada que iguala o supera al 13 es 22 y corresponde al valor 2 que es la mediana.

Cálculo de la mediana con datos agrupados

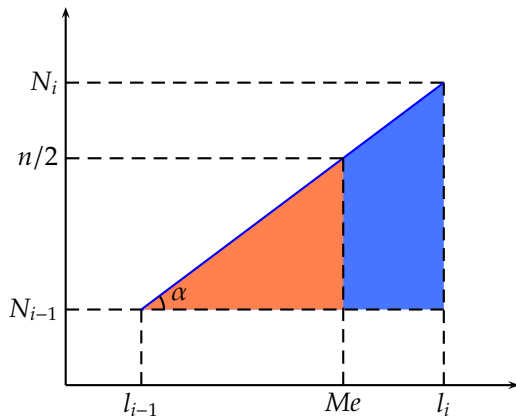
Con datos agrupados la mediana se calcula interpolando en el polígono de frecuencias absolutas acumuladas para el valor $n/2$.



Cuando se trabaja con datos agrupados en clases, la mediana se calcula de forma aproximada, interpolando en el polígono de frecuencias acumuladas para el valor $n/2$.

La interpolación consiste en proyectar sobre el polígono la frecuencia $n/2$ y ver a qué altura del eje de abscisas corta al polígono de frecuencias. Dicho valor es la mediana.

Interpolación en el polígono de frecuencias absolutas acumuladas



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

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

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

La interpolación en realidad consiste en una razón de semejanza de triángulos.

En primer lugar se identifica el intervalo en el que cae la mediana, mirando en la columna de frecuencias acumuladas de la tabla de frecuencias, de igual modo a como se hace para datos no agrupados. Una vez identificado el intervalo, se toma el segmento del polígono de frecuencias acumuladas que corresponde a dicho intervalo. Supongamos que dicho intervalo tiene límite inferior l_{i-1} y límite superior l_i , y que parte de una frecuencia absoluta acumulada N_{i-1} y llega a una frecuencia absoluta acumulada N_i . Este segmento define un triángulo rectángulo de ángulo α cuya tangente es el cateto opuesto, que vale $N_i - N_{i-1}$ entre el cateto contiguo, que es precisamente la amplitud del intervalo $l_i - l_{i-1}$.

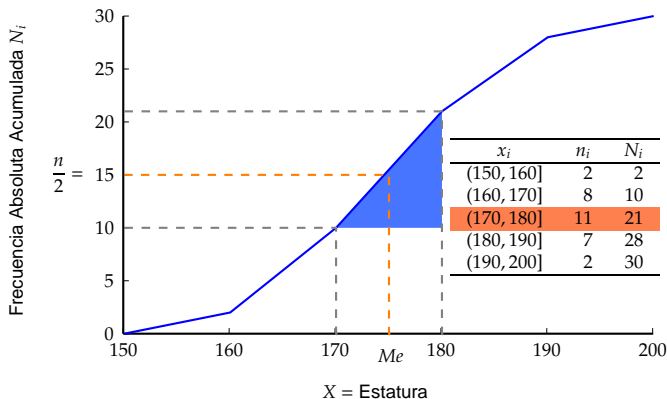
Por otro lado, si proyectamos la frecuencia correspondiente a la media $n/2$ sobre el polígono, en el punto de corte aparecería la mediana, de manera que se tiene otro triángulo rectángulo más pequeño que es semejante al anterior al compartir el mismo ángulo α . Al igual que antes, la tangente de este ángulo será el cateto opuesto, que ahora vale $n/2 - N_{i-1}$ entre el cateto contiguo que ahora vale $Me - l_{i-1}$.

Puesto que se trata del mismo ángulo, su tangente es la misma y se pueden igual ambas expresiones, dando lugar a una ecuación donde la única incógnita es la mediana. Despejándola, se obtiene la fórmula para calcular la mediana.

Cálculo de la mediana

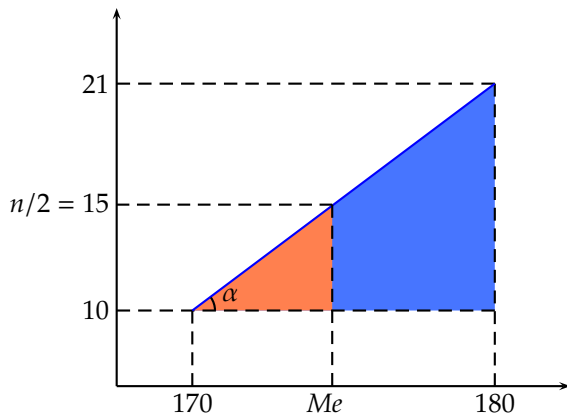
Ejemplo con datos agrupados

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 $(170, 180]$.



Veamos un ejemplo de interpolación para calcular la mediana de la muestra de estaturas. Mirando en la tabla de frecuencias se observa que el primer intervalo con una frecuencia igual o mayor que $n/2 = 30/2 = 15$ es el que va de 170 cm a 180 cm, así que se toma el trozo del polígono de frecuencias absolutas acumuladas correspondiente a este intervalo.

Interpolación en el polígono de frecuencias absolutas acumuladas



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

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

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

Si nos fijamos en el triángulo grande, la tangente de α vale $21 - 10$ que es el cateto opuesto, entre $180 - 170$ que es el cateto contiguo, mientras que si nos fijamos en el triángulo pequeño que aparece al proyectar 15 sobre el polígono, se tiene que la tangente de α vale $15 - 10$ que es cateto opuesto entre $Me - 170$ que es el cateto contiguo. Igualando ambas expresiones y despejando la mediana se obtiene 174,54 cm que es la estatura mediana.

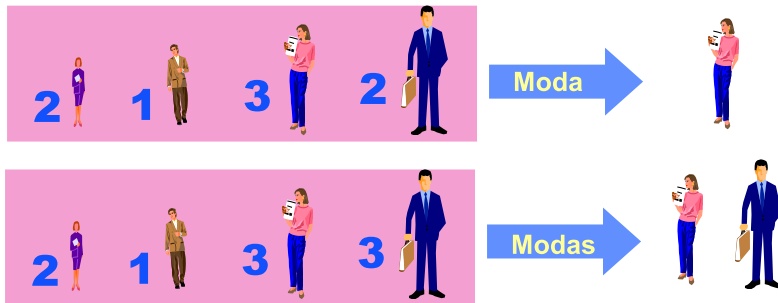
Una comprobación que conviene hacer siempre es ver que el valor obtenido cae efectivamente dentro del intervalo de interpolación.

Definición (Moda muestral M_o)

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.



El último estadístico de tendencia central que veremos se llama moda, y viene a suplir las carencias de la media y mediana, que no podían calcularse para atributos nominales.

La moda se define como el valor más frecuente en la muestra, y puede ocurrir que en algunas distribuciones haya 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

La moda es el estadístico más sencillo de calcular pues sólo hay que fijarse en la columna de frecuencias absolutas de la tabla de frecuencias, buscar la mayor y ver a qué valor le corresponde.

En el ejemplo del número de hijos, la mayor frecuencia es 14, que le corresponde al 2, de manera que la moda es 2 hijos.

En el ejemplo de las estaturas, la mayor frecuencia es 11, que le corresponde al intervalo (170, 180] que es el intervalo modal.

¿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, ¡jojo! 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

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

¿Qué representante de la muestra tomarías?

Hemos visto tres estadísticos de tendencia central, de manera que surge la pregunta de cuál usar en el caso de que puedan calcularse los tres.

En principio, siempre que pueda calcularse se tomará la media, pues es la medida que más información utiliza de la muestra al tener en cuenta la magnitud de los datos. Si no puede calcularse la media, se utilizará la mediana, que aunque no tiene en cuenta la magnitud de los valores, al menos sí tiene en cuenta el orden entre ellos. Y finalmente, si no se pueden calcular ninguna de las dos, se tomará la moda que únicamente tiene en cuenta las frecuencias de los valores.

Una excepción a esta regla es cuando en la muestra haya datos atípicos, ya que la media es muy sensible a datos atípicos, mientras que la mediana no lo es apenas. Si nos fijamos en esta pequeña muestra con el número de hijos de 7 matrimonios, con un matrimonio que es claramente atípico por tener 15 hijos, podemos ver que la media vale 3 hijos, mientras que sin el dato atípico valdría 1 hijo. Es decir, la media se ve muy alterada por el dato atípico. Si embargo la mediana vale 1 hijo tanto con el dato atípico como sin él. En estas circunstancias es mejor utilizar la mediana como medida de representatividad.

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) , \dots , D_9 (90 % acumulado).

Percentiles: Dividen la distribución en 100 partes iguales.

Hay 99 percentiles: P_1 (1 % acumulado), \dots , P_{99} (99 % acumulado).

Una vez vistas las medidas de tendencia central, pasamos al resto de medidas de posición que son los cuantiles y que dividen la muestra en partes iguales. Existen distintos tipos:

Cuartiles: Dividen la distribución en 4 partes iguales.

Hay tres cuartiles: C_1 (25 % acumulado) , C_2 (50 % acumulado), C_3 (75 % acumulado).

Obsérvese que la mediana coincide con el segundo cuartil.

Deciles: Dividen la distribución en 10 partes iguales.

Hay 9 deciles: D_1 (10 % acumulado) , \dots , D_9 (90 % acumulado).

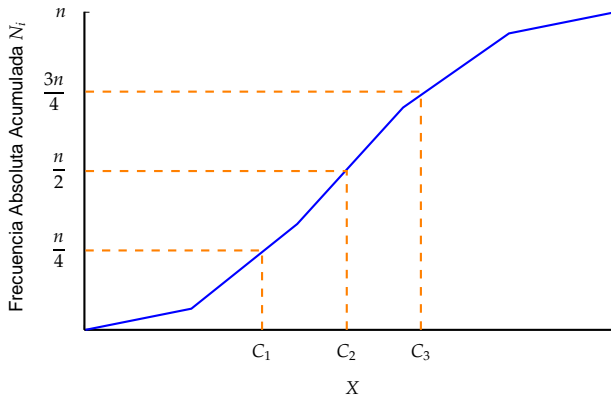
Percentiles: Dividen la distribución en 100 partes iguales.

Hay 99 percentiles: P_1 (1 % acumulado), \dots , P_{99} (99 % acumulado).

Obsérvese la correspondencia que hay entre algunos cuantiles, como por ejemplo el primer cuartil que se corresponde con el percentil 25 o el tercer cuartil que se corresponde con el percentil 75.

Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. Por ejemplo, en el caso de los cuantiles 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.



Los cuantiles se calculan, al igual que la mediana, interpolando sobre el polígono de frecuencias absolutas acumuladas. Por ejemplo, para los cuartiles, proyectando la frecuencia absoluta acumulada del primer cuartil, que es $n/4$, se obtiene el primer cuartil, proyectando $n/2$ se obtiene el segundo cuartil, y proyectando $3n/4$ se obtiene el tercer cuartil, y lo mismo para los otros cuantiles.

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

$$F_{C_1} = 0,25 \Rightarrow C_1 = 1 \text{ hijos,}$$

$$F_{C_2} = 0,5 \Rightarrow C_2 = 2 \text{ hijos,}$$

$$F_{C_3} = 0,75 \Rightarrow C_3 = 2 \text{ hijos,}$$

$$F_{D_3} = 0,3 \Rightarrow D_3 = 1 \text{ hijos,}$$

$$F_{P_{92}} = 0,92 \Rightarrow P_{92} = 3 \text{ hijos.}$$

Los cuantiles se calculan de forma similar a la mediana a partir de las frecuencias acumuladas correspondientes a cada uno de ellos.

En el ejemplo del número de hijos donde se tenían las frecuencias relativas acumuladas que aparecen en esta tabla, para calcular el primer cuartil se parte de su frecuencia relativa acumulada que es 0,25 ya que hasta el primer cuartil se tienen acumulados el 25 % de los individuos y se busca en la tabla el primer valor cuya frecuencia relativa acumulada es igual o superior a 0,25, que en este caso es 1 hijo, ya que su frecuencia relativa acumulada es 0,32 que supera a 0,25 mientras que la frecuencia relativa acumulada del 0 no llega a 0,25. Del mismo modo, el cuartil segundo, cuya frecuencia relativa acumulada es 0,5, es 2 hijos y el tercer cuartil, con frecuencia relativa acumulada 0,75 también es 2 hijos.

Para el decil tercero, su frecuencia relativa acumulada es 0,3, y procediendo como para los cuantiles, se observa que el primer valor con frecuencia relativa acumulada igual o superior a 0,3 es 1 hijo.

Finalmente, para el percentil 92, su frecuencia relativa acumulada es 0,92 y el primer valor con frecuencia relativa acumulada igual o superior a este valor es 3 hijos.

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.

Uno de los aspectos más importantes de una muestra es la variabilidad de los datos, que también se conoce como dispersión de la muestra. Veremos hasta 5 estadísticos para describir la dispersión:

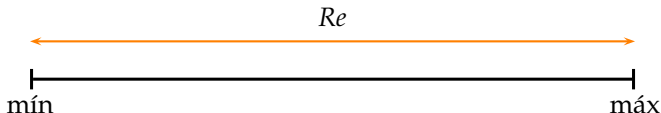
- Recorrido.
- Rango Intercuartílico.
- Varianza.
- Desviación Típica.
- Coeficiente de Variación.

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.



El recorrido es el estadístico de dispersión más natural ya que consiste en medir la distancia que va del mayor al menor de los valores y se calcula restándolos.

Cuanto mayor sea el recorrido, mayor dispersión habrá en la muestra.

El principal problema que presenta el recorrido es que es muy sensible a los datos atípicos, ya que estos precisamente aparecen en los extremos de la distribución.

Rango intercuartílico

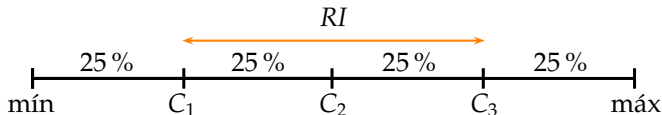
Para evitar el problema de los datos atípicos en el recorrido, se puede utilizar el primer y tercer cuartil en lugar del mínimo y el máximo.

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.



Para evitar el problema que tiene el Recorrido con los datos atípicos, se suele utilizar el Rango Intercuartílico, que utiliza la misma idea del Recorrido pero midiendo la distancia que va del tercer al primer cuartil.

Si recordamos, los cuartiles dividen la distribución en 4 partes iguales, de manera que entre el primer y el tercer cuartil están el 50 % de los datos centrales de la distribución, y quedarían excluidos el 25 % de valores menores y el 25 % de valores mayores, donde posiblemente estarían los datos atípicos.

Por tanto el Rango Intercuartílico mide la dispersión central de la muestra, de manera que cuanto mayor sea, más dispersión central tendrá la muestra, aunque siempre hay que tener en cuenta las unidades de la variable al interpretarlo.

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.

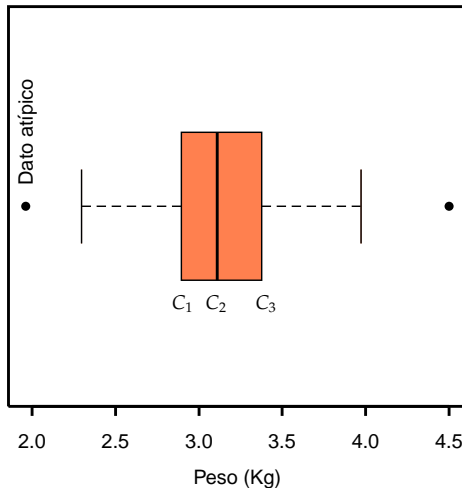
La dispersión de la muestra se representa a menudo mediante un diagrama de caja y bigotes, que como su propio nombre indica, consiste en una caja sobre un eje X donde el borde inferior de la caja es el primer cuartil, y el borde superior el tercer cuartil, y por tanto, la anchura de la caja, por tanto, 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 de la muestra mediante unos segmentos llamados 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 recién nacidos



Aquí podemos ver un ejemplo de un diagrama de caja y bigotes para una muestra de pesos de recién nacidos. El borde inferior de la caja corresponde al primer cuartil, que es aproximadamente 2,8 Kg, el borde superior corresponde al tercer cuartil, que vale aproximadamente 3,4 Kg, de manera que el ancho de la caja es el rango intercuartílico, y vale aproximadamente 0,6 Kg, lo que indica una dispersión central baja, teniendo en cuenta que los pesos de los recién nacidos se suelen mover en unas unidades que van de 1 a 4 Kg, más o menos.

De la caja salen dos sementos, que son los bigotes, el inferior llega aproximadamente hasta 2,3 Kg y el superior hasta 4 Kg, delimitando el intervalo de normalidad de los datos, de manera que cualquier niño que pese menos de 2,3 Kg o más de 4 Kg será un individuo atípico. De hecho en esta muestra aparecen dos pesos atípicos que aparecen marcados por puntos, uno que corresponde a un niño con 2 Kg de peso, y otro que corresponde a otro niño con 4.5 Kg de peso.

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 dato de la muestra más cercano a v_1 por encima de v_1 , y b_2 es el dato de la muestra más cercano a v_2 por debajo de 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.

Para construir el diagrama de cajas, hay que seguir los siguientes pasos:

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 mínimo valor de la muestra mayor o igual que v_1 , y b_2 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 .

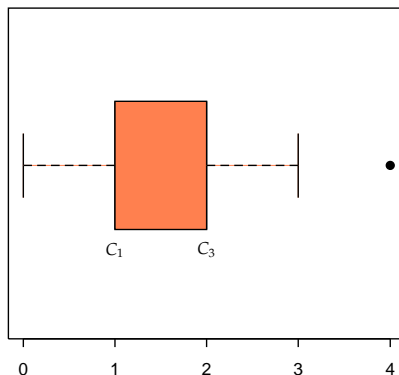
5. Finalmente, si en la muestra hay algún dato por debajo de v_1 o por encima de v_2 se dibuja un punto sobre dicho valor.

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 $b_2 = 3$ hijos.
- 5 Dibujar los datos atípicos: 4 hijos.

Diagrama de caja y bigotes del número de hijos



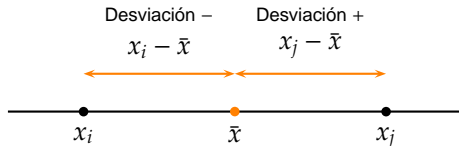
Veamos cómo construir el diagrama de caja y bigotes para el ejemplo del número de hijos:

1. Primero se calcula el primer cuartil que era 1 hijo y se dibuja el borde inferior de la caja. Después se calcula el tercer cuartil que era 3 hijos y se dibuja el borde superior de la caja.
2. Una vez dibujados el borde inferior y superior de la caja se acaba de dibujar esta.
3. También se suele dibujar el cuartil segundo mediante una línea que divide la caja, pero en este ejemplo, al coincidir el cuartil segundo con el tercero, no se dibuja.
4. A continuación se calcula el rango intercuartílico que era $2-1=1$ hijo y con el se calcula la valla inferior v_1 restandole al primer cuartil 1,5 veces el rango intercuartílico, lo que nos da $-0,5$ hijos, y la valla superior v_2 sumandole al tercer cuartil 1,5 veces el rango intercuartílico, lo que nos da $3,5$ hijos.
5. Luego se calculan los extremos de los bigotes. El extremo del bigote inferior es 0 hijos, ya que es el mínimo valor de la muestra por encima de la valla inferior que valía $-0,5$, y el bigote superior es 3 hijos, ya que es el máximo valor de la muestra por debajo de la valla superior que valía $3,5$ hijos.
6. Finalmente se dibujan los datos atípicos. Por debajo de la valla inferior no hay ningún valor en la muestra, pero por encima de la valla superior si hay una familia con 4 hijos, que se trata, por tanto, de una familia atípica y si se dibuja un punto en diagrama sobre dicho valor.

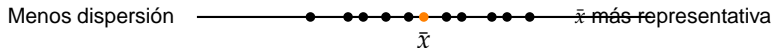
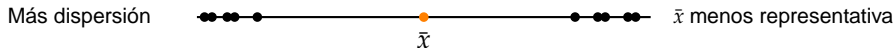
Desviaciones respecto de la media

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.



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 que hay de cada valor a la media, que se conoce como desviación respecto a la media. Cuando el valor sea mayor que la media su desviación será positiva, y cuando sea menor que la media su desviación será negativa.

Resulta evidente que si las desviaciones son grandes, entonces los datos estarán bastante alejados de la media y por tanto la media no será tan representativa de la muestra como cuando los valores estén concentrados en torno a la media y sus desviaciones sean pequeñas.

Aquí tenemos dos casos en los que la dispersión con respecto a la media es mayor en el primer caso es mayor que en el segundo, de manera que la media será más representativa en el último caso que en el primero.

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.

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

A partir de las desviaciones con respecto a la media surge un estadístico conocido como varianza, que se representa como s^2 y que se calcula sumando las desviaciones de los valores a la media elevadas al cuadrado y dividiendo la suma por el tamaño de la muestra. El hecho de considerar los cuadrados es para sumar siempre magnitudes positivas, ya que de lo contrario las desviaciones positivas se compensarían con las negativas. Además, siempre que se trabaje desde la tabla, hay que multiplicar cada desviación al cuadrado por su correspondiente frecuencia absoluta.

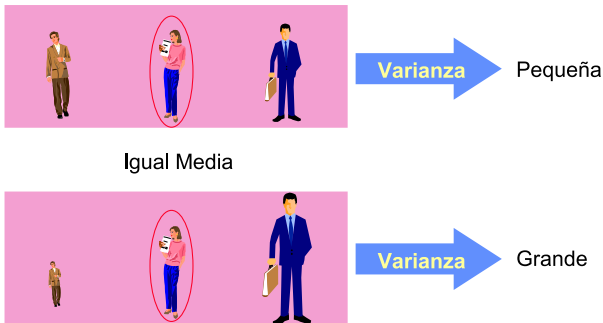
Si se desarrolla el cuadrado de cada desviación y se simplifica, se llega a otra expresión equivalente para calcular la varianza que consiste en sumar los cuadrados de los valores multiplicados por su frecuencia absoluta, dividir la suma por el tamaño de la muestra y al resultado restarle la media al cuadrado. Esta fórmula es un poco más sencilla de calcular y será la que utilizaremos casi siempre.

Como la varianza se calcula a partir de las desviaciones, cuando estas sean grandes, la varianza será grande, lo que indicará gran dispersión y cuando estas sean pequeñas la varianza también será pequeña indicando poca dispersión. El único inconveniente es que la varianza tiene unidades al cuadrado, lo cual dificulta su interpretación.

Para evitar este problema se suele tomar la raíz cuadrada de la varianza que se conoce como desviación típica y que también sirve para medir la dispersión de los datos con respecto a la media, pero ahora con la ventaja de tener las mismas unidades que la variable.

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.



Tanto la varianza como la desviación típica sirven para cuantificar la dispersión de los datos en torno a la media. Si la dispersion con respecto a la media es pequeña, los individuos se parecerán bastante a la media y esta será más representativa que cuando los individuos no se parezcana ella y la dispersión con respecto a la media sea mayor.

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
Σ	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 1 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 del número de hijos, para calcular la varianza conviene añadir una nueva columna a la tabla donde se calculen los productos de los cuadrados de los valores por sus frecuencias absolutas: 0 elevado al cuadrado y por su frecuencia absoluta que es 2, da 0, 1 elevado al cuadrado y por su frecuencia absoluta que es 6, que da 6, y así sucesivamente. Después hay que sumar los valores de esta columna y dividirlos por el tamaño de la muestra que era 25. Finalmente al cociente se le resta el valor de la media que era 1,76 elevada al cuadrado, lo que nos da 0.7424 hijos al cuadrado.

Si sacamos la raíz cuadrada se obtiene una desviación típica de 0.8616 hijos, que no es un valor grande comparado con el recorrido de la variable que va de 1 a 4 hijos, por lo que se puede concluir que la muestra no tiene mucha dispersión y por tanto la media de 1,76 hijos representa muy 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
Σ		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 $s = \sqrt{102,06} = 10,1 \text{ cm}$.

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.

En el ejemplo de las estaturas, como se habían agrupado los datos, como valores se tomarán las marcas de clases, es decir, 155 elevado al cuadrado y por su frecuencia absoluta que es 2, lo que nos da 48050, y así sucesivamente. Después hay que sumar los valores de esta columna y dividirlos por el tamaño de la muestra que era 30. Finalmente al cociente se le resta el valor de la media que era 174,67 elevada al cuadrado, lo que nos da 102.06 cm al cuadrado.

Si sacamos la raíz cuadrada se obtiene una desviación típica de 10.1 cm, que es un valor pequeño comparado con el recorrido de la variable que va de 150 a 200 cm, por lo que se puede concluir que la muestra tiene poca dispersión y por tanto la media representa muy bien al resto de individuos de la muestra.

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.

Tanto la varianza como la desviación típica tienen unidades lo que dificulta a veces su interpretación.

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

El coeficiente de variación muestral, que se representa mediante cv , se define como el cociente entre la desviación típica y el valor absoluto de la media.

Como tanto la desviación típica como la media tienen las unidades de la variable, al hacer su cociente las unidades desaparecen y se obtiene una medida adimensional que resulta más sencilla de interpretar.

Al estar dividido por el valor absoluto de la media, el coeficiente de variación mide la dispersión relativa de los valores de la muestra en torno a la media muestral, y como en el numerador está la desviación típica, cuanto mayor sea esta, mayor será el coeficiente de variación y por tanto mayor será la dispersión relativa de la variable en torno a la media.

Una de las principales utilidades del coeficiente de variación es que, precisamente por no tener unidades, permite la comparación de la dispersión de muestras distintas, incluso si son de variables con distintas unidades.

El único problema de este estadístico es que no vale cuando la media muestral vale 0 o próxima a 0, ya que al estar en el denominador, obtendríamos valores muy grandes.

Coeficiente de variación

Ejemplo

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

Los estadísticos de forma se encargan de describir, como su propio nombre indica, la forma que tiene la distribución de valores en la muestra, en particular se estudian dos aspectos que son la asimetría y el apuntamiento.

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).

La simetría con respecto a la media tiene que ver con la ubicación de los valores a un lado y otro de la media, cuántos valores hay por encima y cuántos por debajo, y cómo están de alejados.

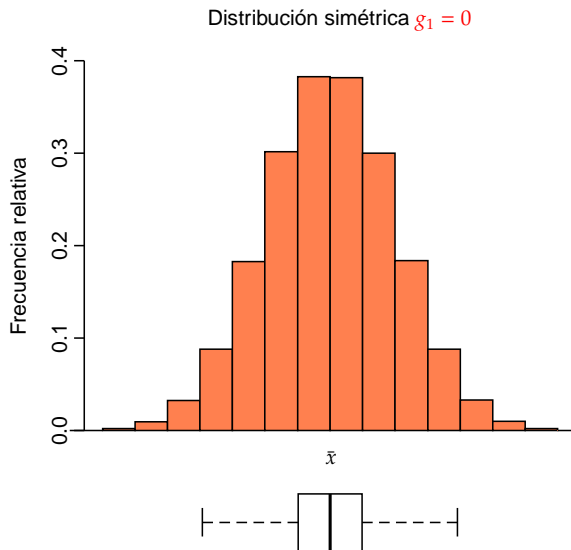
El coeficiente de asimetría muestral, que se representa g_1 se define, como la suma del producto de las desviaciones de los valores de la muestra a la media muestral elevadas al cubo por su frecuencia absoluta, dividida por el tamaño de la muestra, y a su vez todo dividido por la desviación típica al cubo.

Como las desviaciones elevadas al cubo tienen las unidades de la variable al cubo y la desviación típica elevada al cubo también tiene las unidades de la variable al cubo, al realizar el cociente las unidades se cancelan y por tanto el coeficiente de asimetría es una medida adimensional que mide el grado de asimetría de los valores de la muestra con respecto a la media, de manera que:

- $g_1 = 0$ indica que hay el mismo número de valores a la derecha y a la izquierda de la media y por tanto la distribución es simétrica.
- $g_1 < 0$ indica que la mayoría de los valores son mayores que la media y entonces se dice que la distribución es asimétrica hacia la izquierda.
- $g_1 > 0$ indica que la mayoría de los valores son menores que la media y entonces se dice que la distribución es asimétrica hacia la derecha.

Coeficiente de asimetría

Ejemplo de distribución simétrica

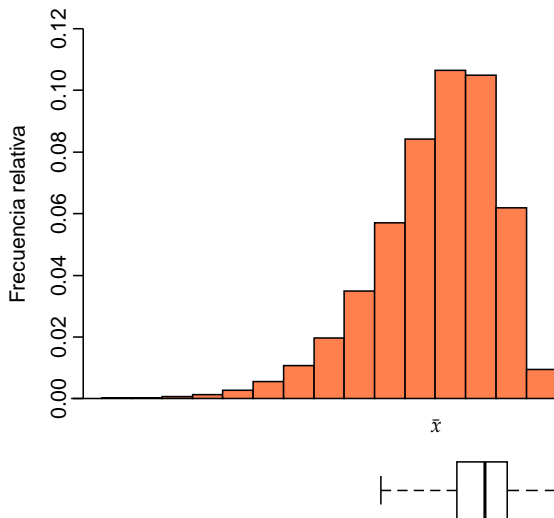


Aquí tenemos el histograma de una distribución simétrica. Como puede observarse, la media queda justo en el centro de la distribución, coincidiendo con la mediana y existe el mismo número de barras y con la misma frecuencia a un lado y a otro de la media.

Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la izquierda

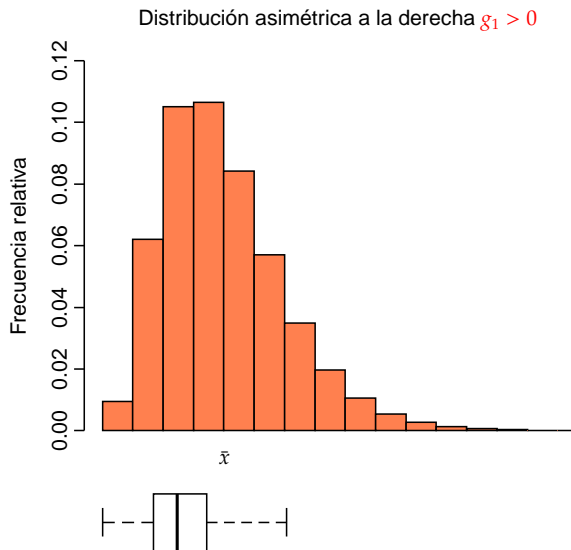
Distribución asimétrica a la izquierda $g_1 < 0$



En este otro caso tenemos una distribución asimétrica hacia la izquierda, donde la media queda por debajo de la mediana y las barras son más altas a la derecha de la media, lo que indica que hay más valores por encima de la media. Por debajo de la media habría menos valores, barras más bajas, pero más alejados.

Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la derecha



Y en este otro caso tenemos una distribución asimétrica hacia la derecha, donde la media queda por encima de la mediana y las barras son más altas a la izquierda de la media, lo que indica que hay más valores por debajo de la media. Por encima de la media habría menos valores, barras más bajas, pero más alejados.

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 $\bar{x} = 174,67$:

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
Σ		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.

Para calcular el coeficiente de asimetría en el ejemplo de las estaturas se puede añadir una nueva columna a la tabla de frecuencias con las desviaciones de los valores a la media que recordemos valía 174,67 cm. Como habíamos agrupado los datos en clases, para calcular las desviaciones a la media se toma la marca de cada clase. Así, la primera desviación es 155 menos la media 174,67 lo que nos da $-19,67$ cm, la segunda es 165 menos 174,67 cm y así sucesivamente. Obsérvese que las desviaciones de los valores menos que la media serán negativas y que las de los valores mayores serán positivas. A continuación se añade otra columna a la tabla con el producto de las desviaciones elevadas al cubo por su frecuencia absoluta, es decir, $-19,67$ al cubo por su frecuencia absoluta que es 2, lo que nos da -15221 , $-9,67$ elevado al cubo y por su frecuencia absoluta que es 8, lo que nos da $-7233,85$, y así sucesivamente. Al final se suman los valores de esta columna y se dividen por el tamaño de la muestra que era 30. Por último el resultado de este cociente se vuelve a dividir por la desviación típica que era 10,1 cm elevada al cubo, y se obtiene 0,07.

Como este valor está muy próximo a 0, se puede concluir que la distribución de las estaturas es prácticamente simétrica.

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*).

El apuntamiento de una distribución muestral tiene que ver con la pendiente su polígono de frecuencias.

El coeficiente de apuntamiento o kurtosis muestral, que se representa g_2 se define, como la suma del producto de las desviaciones de los valores de la muestra a la media muestral elevadas a la cuarta por su frecuencia absoluta, dividida por el tamaño de la muestra, y a su vez todo dividido por la desviación típica a la cuarta, y al final se resta 3 al cociente. Como puede verse, la fórmula es muy parecida a la del coeficiente de asimetría, pero tomando las potencias cuartas en lugar de las potencias al cubo, y restando 3 al cociente.

Al igual que para el coeficiente de asimetría, como las desviaciones elevadas a la cuarta tienen las unidades de la variable a la cuarta y la desviación típica elevada al cubo también tiene las unidades de la variable a la cuarta, al realizar el cociente las unidades se cancelan y por tanto el coeficiente de apuntamiento es una medida adimensional que mide el grado de apuntamiento de la distribución muestral.

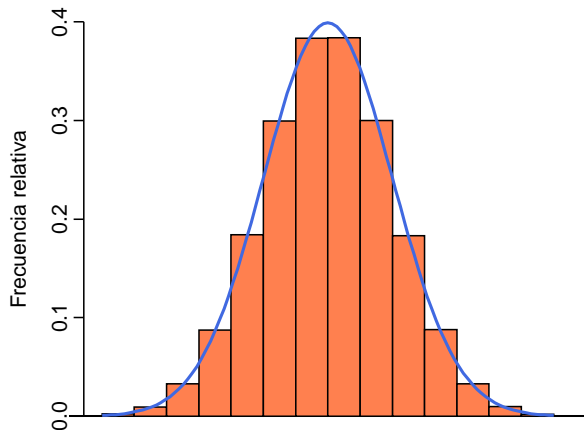
El apuntamiento suele medirse en comparación con un apuntamiento de referencia que es el de una distribución normal. La distribución normal se verá más adelante en el curso, pero baste decir que es la distribución más común que se presenta en la naturaleza, y por lo tanto, está justificado tomarla como referencia y comparar el apuntamiento de cualquier otra distribución con el de la distribución normal que siempre vale 0. Por tanto cuando

- $g_2 = 0$ 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

Distribución mesocúrtica $g_2 = 0$

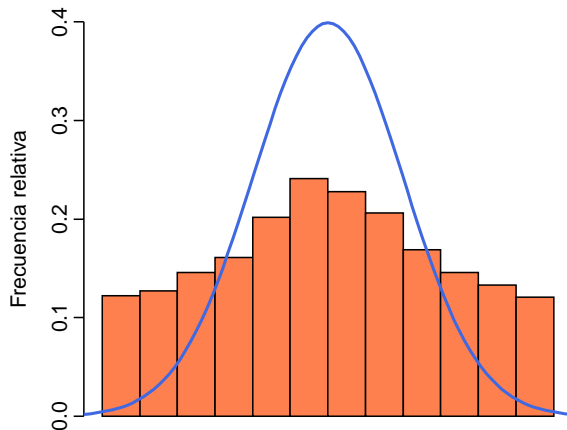


Aquí un histograma con coeficiente de apuntamiento 0 y sobre él una distribución normal, representada por esta curva conocida como campana de Gauss. Obsérvese cómo la altura de las barras coinciden con la campaña de Gauss y se ajustan perfectamente a la distribución normal, lo que indica que la distribución es mesocúrtica.

Coeficiente de apuntamiento o curtosis

Ejemplo de distribución platicúrtica

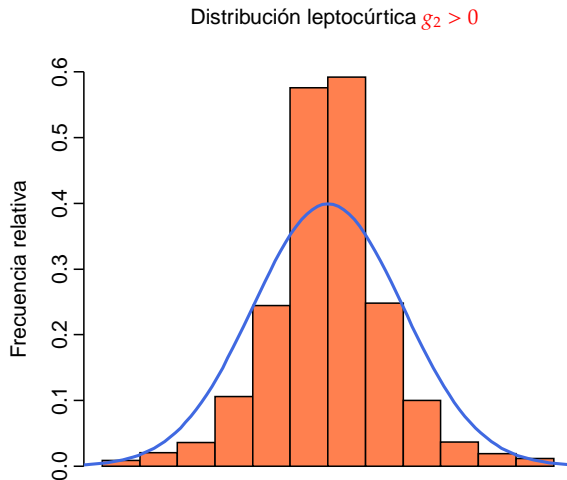
Distribución platicúrtica $g_2 < 0$



Ahora tenemos un histograma con un coeficiente de apuntamiento menor que 0. Como se puede apreciar, en este caso la altura de las barras centrales están por debajo de la curva de Gauss y la distribución tiene menos apuntamiento de lo normal, por lo que se dice que es platicúrtica.

Coeficiente de apuntamiento o curtosis

Ejemplo de distribución leptocúrtica



En este otro caso tenemos un histograma con un coeficiente de apuntamiento mayor que 0. Ahora la altura de las barras centrales están por encima de la campá de Gauss y la distribución tiene más apuntamiento de lo normal, por lo que se dice que es 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 $\bar{x} = 174,67$ 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
Σ		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.

El coeficiente de apuntamiento se calcula de manera similar al coeficiente de asimetría, calculando primero las desviaciones a la media en una columna de la tabla y luego añadiendo otra columna con el producto de las desviaciones elevadas a la cuarta por su frecuencia absoluta, es decir, $-19,67$ a la cuarta por su frecuencia absoluta que es 2, lo que nos da $299396,99$, $-9,67$ elevado a la cuarta y por su frecuencia absoluta que es 8, lo que nos da $69951,31$, y así sucesivamente. Al final se suman los valores de esta columna y se dividen por el tamaño de la muestra que era 30. Por último el resultado de este cociente se vuelve a dividir por la desviación típica que era $10,1$ cm elevada a la cuarta, y al resultado se le resta 3, obteniendo -0.47 .

Como se trata de un valor negativo, aunque próximo a cero, se puede concluir que la distribución de las estaturas 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.

Como se verá más adelante en la parte de inferencia, muchas de las pruebas estadísticas solo pueden aplicarse a poblaciones normales, que se caracterizan por ser simétricas y mesocúrticas, de manera que, tanto el coeficiente de asimetría como el de apuntamiento pueden utilizarse para comprobar 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 anomalía de la distribución.

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

1,75m, 1,65m, 1,80m,

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

175cm, 165cm, 180cm,

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

10cm, 0cm, 15cm,

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:

$$Y = 100X - 165$$

En muchas ocasiones los datos brutos de la muestra suelen transformarse, a veces simplemente para cambiar a una escala más cómoda y otras veces para corregir alguna anormalidad de la distribución.

Si por ejemplo estamos trabajando con estaturas medidas en metros, con dos decimales como los de este ejemplo, podemos evitar el trabajo con decimales multiplicando por 100, es decir, pasando de metros a centímetros. Así tenemos que 1,75 m multiplicado por 100 se transforma en 175 cm, 1,65 m multiplicado por 100 se transforma en 165 cm y 1,80 m se transforma en 180 cm.

Después, si queremos reducir la magnitud de los datos y pasar de centenas a unidades más pequeñas, podemos restarle a todos los datos el mínimo de los valores que es 165 cm. 175 cm menos 165 cm nos da 10 cm, 165 menos 165 nos da 0 cm y 180 menos 165 nos da 15 cm.

Con esto los datos pasan a una escala mucho más fácil de manejar que la original. En el fondo lo que hemos hecho es aplicar a cada dato la transformación lineal $Y = 100x - 165$.

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.

Una de las transformaciones más habituales que suele realizarse es la *transformación lineal* que sigue la ecuación de una recta $Y = a + bX$, donde a es el término independiente y b la pendiente de la recta.

Una propiedad que tiene esta transformación y resulta fácil de comprobar es que la media de la variable transformada se puede obtener aplicando la misma transformación lineal a la media de la variable original, es decir, $\bar{y} = a + b\bar{x}$, y por otro lado, la desviación típica de la variable transformada se puede obtener multiplicando la desviación típica de la variable original por el valor absoluto de la pendiente de la transformación lineal, es decir, $s_y = |b|s_x$.

Además, el coeficiente de curtosis no se altera y el de asimetría sólo cambia de signo si la pendiente es negativa.

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:

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

Entre las transformaciones lineales hay una de especial importancia, y se conoce como transformación de tipificación. La tipificación consiste en dividir las desviaciones de los valores a la media por la desviación típica.

Como las desviaciones a la media tienen las unidades de la variable y la desviación típica también, al hacer el cociente se cancelan las unidades y los valores de la variable tipificada no tienen unidades, por lo que esta transformación es útil para eliminar la dependencia de la variable 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 que se deduce de las propiedades de las transformaciones lineales vistas antes es que la media de una variable tipificada siempre vale 0 y su desviación típica 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:

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.

Para ver la utilidad de la transformación de tipificación, supongamos que tenemos un grupo de 5 alumnos en los que se ha medido la nota en dos asignaturas X e Y . Si calculamos la media y la desviación típica en cada asignatura, se tiene que la nota media en X es 5 con una desviación típica de 2 y que la nota media de Y es también 5 con una desviación típica de 3,16, es decir, hay más dispersión en las notas de Y que en las de X .

Podríamos preguntarnos si sacar un 8 en la asignatura X tiene el mismo mérito que sacar un 8 en la asignatura Y , o dicho de otro modo, el alumno que ha sacado un 8 en la asignatura X , ¿ha tenido el mismo rendimiento que el que ha sacado un 8 en la 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, que son

$X :$	-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 y tiene más mérito sacar un 8 en la asignatura X que en la Y .

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
Σ	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
Σ	-2,76	1,26	0,45	1,5	-0,45

Y el mejor alumno sería el cuarto.

Siguiendo con el ejemplo anterior, también podríamos habernos preguntado ¿cuál es el mejor alumno? Si simplemente sumamos las puntuaciones de cada asignatura tenemos:

Alumno:	1	2	3	4	5
X :	2	5	4	8	6
Y :	1	9	8	5	2
Σ	3	14	12	13	8

El mejor alumno sería el segundo.

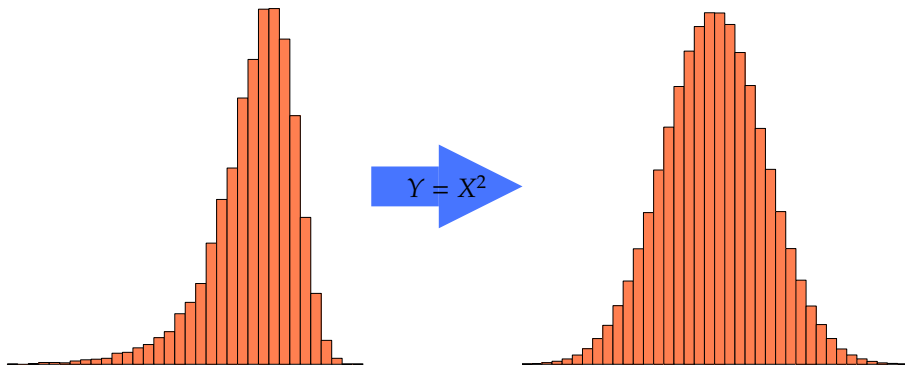
Pero es mucho más razonable considerar el rendimiento relativo tomando las puntuaciones típicas y entonces se tiene que la suma de las puntuaciones típicas es:

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
Σ	-2,76	1,26	0,45	1,5	-0,45

Con lo que realmente el mejor alumno es el cuarto.

Transformaciones no lineales

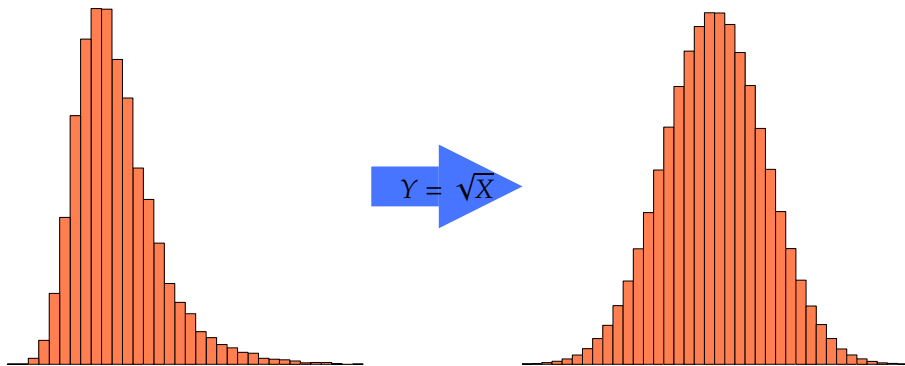
La transformación $Y = X^2$ comprime la escala para valores pequeños y la expande para valores altos, de manera que es muy útil para corregir asimetrías hacia la izquierda.



Otras transformaciones no lineales que son habituales para corregir anormalidades de la muestra son el cuadrado, que comprime la escala para valores pequeños y la expande para valores altos, de manera que es muy útil para corregir asimetrías hacia la izquierda, tal y como puede apreciarse en estos histogramas.

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.



Mientras que para corregir asimetrías hacia la derecha se utilizan o bien la raíz cuadrada, la función logarítmica o la inversa, ya que ambas comprimen la escala para valores altos y la expanden para valores pequeños.

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.

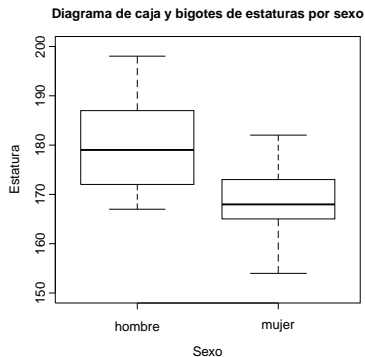
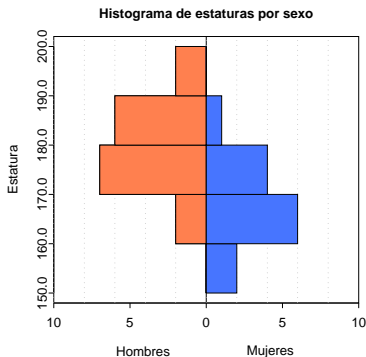
En ocasiones interesa describir el comportamiento de una variable, no para toda la muestra, sino para distintos grupos de individuos, como por ejemplo, estudiar las estaturas en hombres y mujeres por separado.

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.



Si en el ejemplo de las estaturas hubiesemos tomado el sexo como factor, la muestra quedaría dividida en dos grupos, los hombres y las mujeres, y podríamos hacer un estudio descriptivo por separado de cada grupo para luego compararlos.