Curso Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

CEU
Universidad
San Pablo

Licencia creative commons

Curso básico de estadística

Alfredo Sánchez Alberca (asalber@gmail.com).

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.

Estadística descriptiva

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

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

1. Clasificar, agrupar y ordenar los datos de la muestra.
2. 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.

Clasificación de la muestra

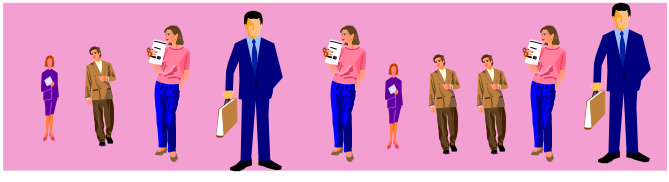
El estudio de una variable estadística comienza por medir la variable en los individuos de la muestra y clasificar los valores obtenidos.

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.

Clasificación de la muestra



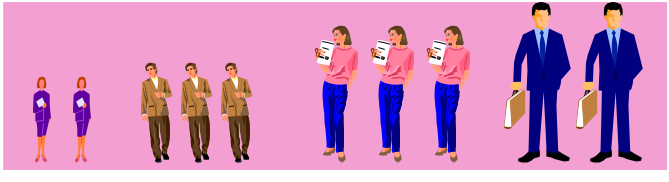
X = **Estatura**

Clasificación

Recuento de frecuencias

X = **Estatura**

Frecuencias



Frecuencias muestrales

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.

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

Tabla de frecuencias

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

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

Tabla de frecuencias

Ejemplo de datos sin agrupar

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
$\sum$	25	1		

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
$\sum$	30	1		

Construcción de clases

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

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

- El número de intervalos no debe ser muy grande ni muy pequeño. Una regla orientativa es tomar un número de intervalos próximo 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.

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

La tabla de frecuencias asociada a esta muestra es

x_i	n_i	f_i
0	5	0,16
A	14	0,47
B	8	0,27
AB	3	0,10
$\sum$	30	1

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

Representaciones gráficas

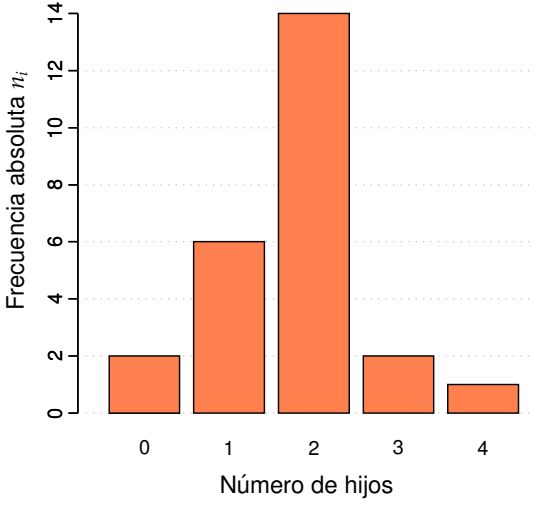
También es habitual representar la distribución muestral de frecuencias de forma gráfica. Dependiendo del tipo de variable y de si se han agrupado o no los datos, se utilizan distintos tipos de gráficos:

- 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.

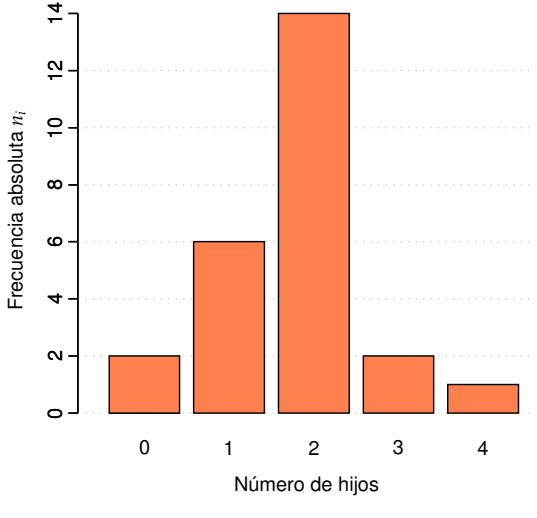
Diagrama de barras de frecuencias absolutas

Datos sin agrupar



Polígono de frecuencias absolutas

Datos sin agrupar



Polígono de frecuencias absolutas

Datos sin agrupar

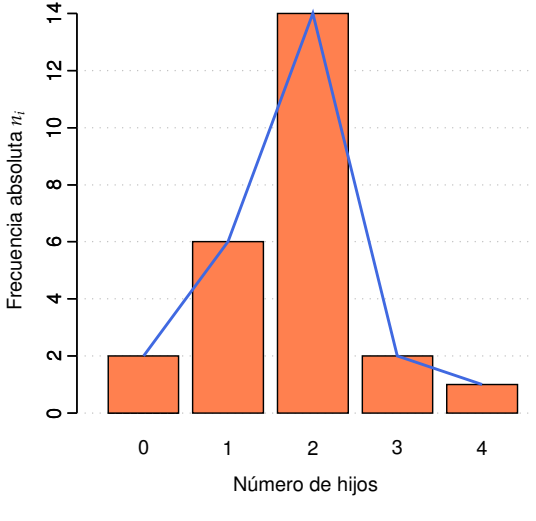
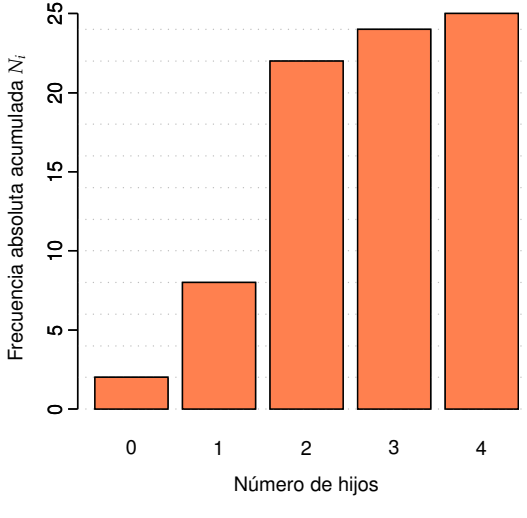


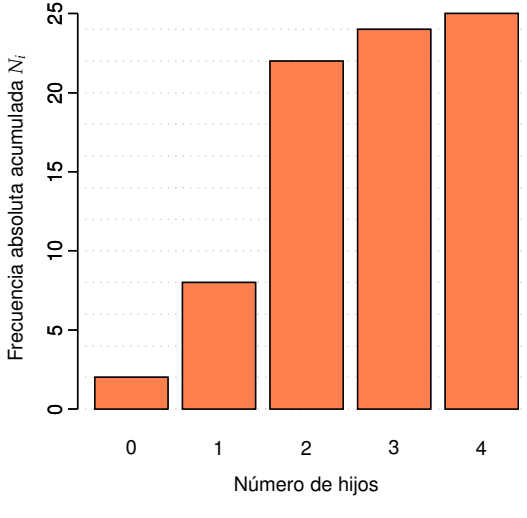
Diagrama de barras de frecuencias acumuladas

Datos sin agrupar



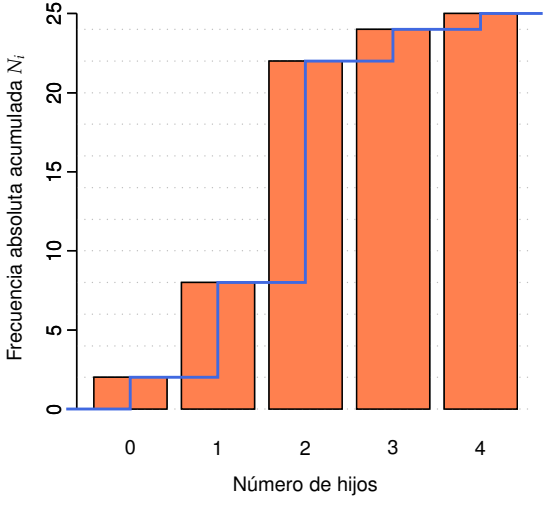
Polígono de frecuencias absolutas acumuladas

Datos sin agrupar



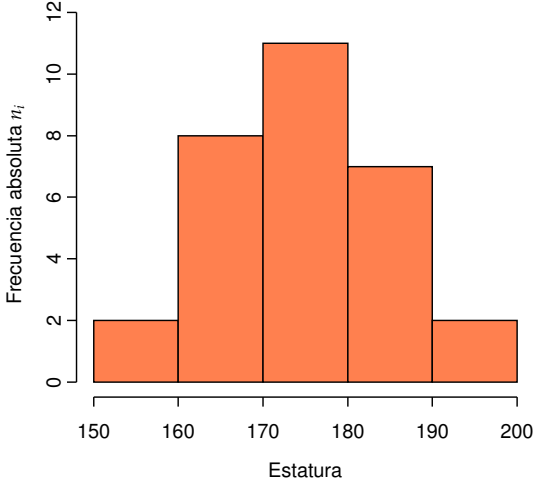
Polígono de frecuencias absolutas acumuladas

Datos sin agrupar



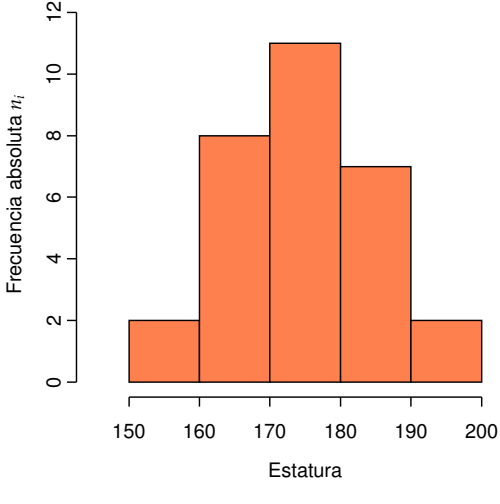
Histograma de frecuencias absolutas

Datos agrupados



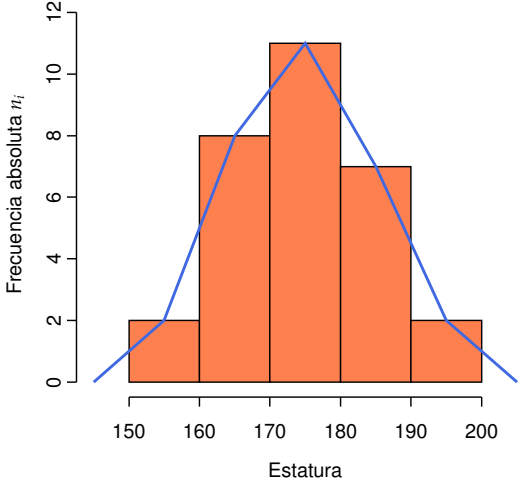
Polígono de frecuencias absolutas

Datos agrupados



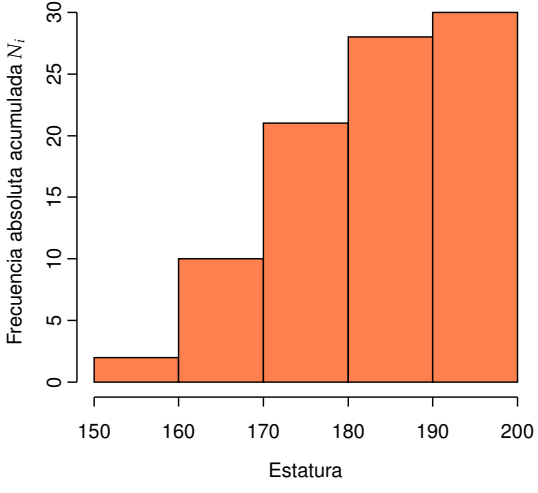
Polígono de frecuencias absolutas

Datos agrupados



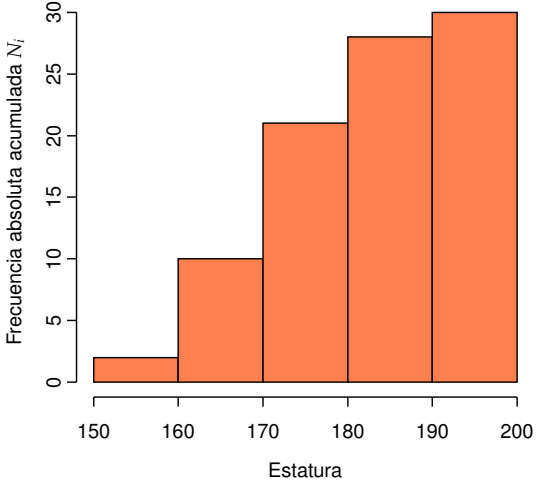
Histograma de frecuencias absolutas acumuladas

Datos agrupados



Polígono de frecuencias absolutas acumuladas

Datos agrupados



Polígono de frecuencias absolutas acumuladas

Datos agrupados

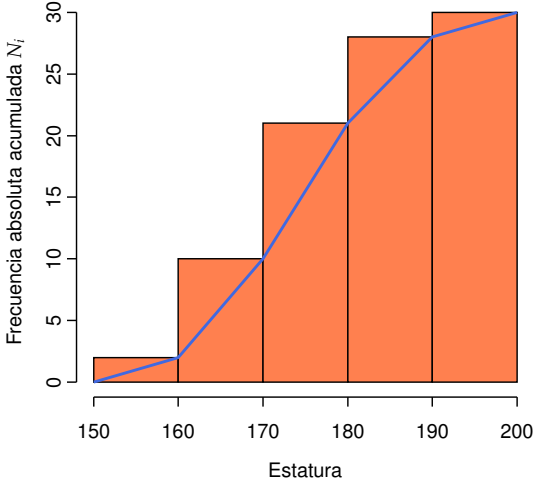
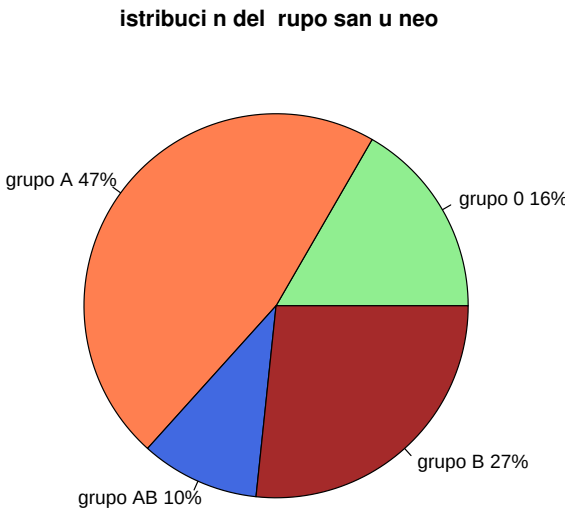


Diagrama de sectores

Atributos



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.

X = Estatura

Dato atípico

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.

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.