Manual Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

Feb 2017

Departamento de Matemática Aplicada y Estadística
CEU San Pablo

Términos de la licencia

Índice general

1 Distribución de frecuencias: Tabulación y gráficos

Estadística descriptiva

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

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

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

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

Clasificación de la muestra

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

Existen dos formas de clasificar estos valores:

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

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

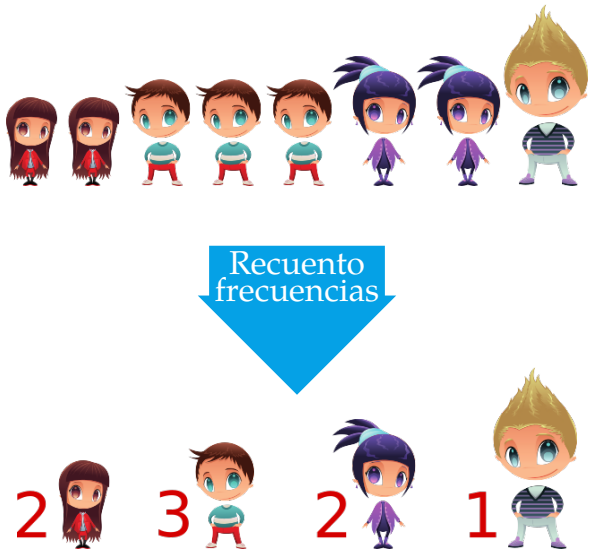
1.1 Distribución de frecuencias

Clasificación de la muestra

X =Estatura

Recuento de frecuencias

X =Estatura



Frecuencias muestrales

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

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

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

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

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

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

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

Tabla de frecuencias

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

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

Tabla de frecuencias

Ejemplo de datos sin agrupar

El número de hijos en 25 familias es

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

La tabla de frecuencias asociada a esta muestra es

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

Tabla de frecuencias

Ejemplo de datos agrupados

Las estaturas (en cm) de 30 estudiantes es

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

La tabla de frecuencias asociada a esta muestra es

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

Construcción de clases

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

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

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

Tabla de frecuencias

Ejemplo con un atributo

Los grupos sanguíneos de 30 personas son

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

La tabla de frecuencias asociada a esta muestra es

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

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

1.2 Representaciones gráficas

Representaciones gráficas

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

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

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

Diagrama de barras

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

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

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

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

Diagrama de barras de frecuencias absolutas

Datos sin agrupar

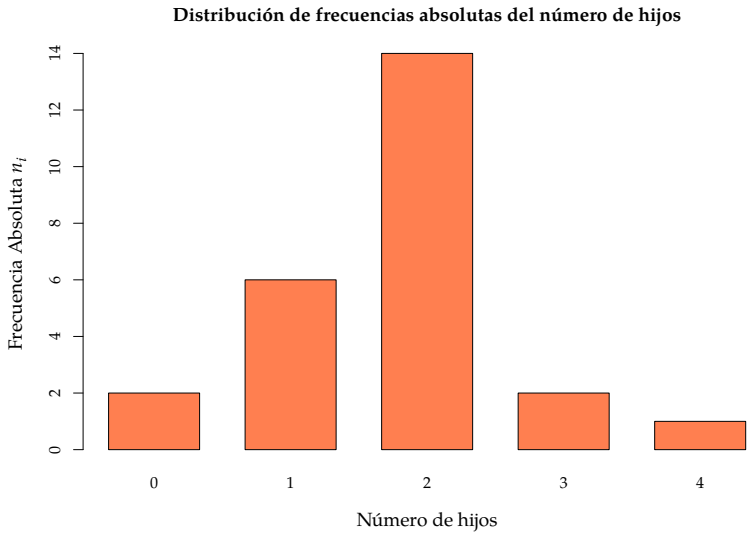


Diagrama de líneas o Polígono de frecuencias absolutas

Datos sin agrupar

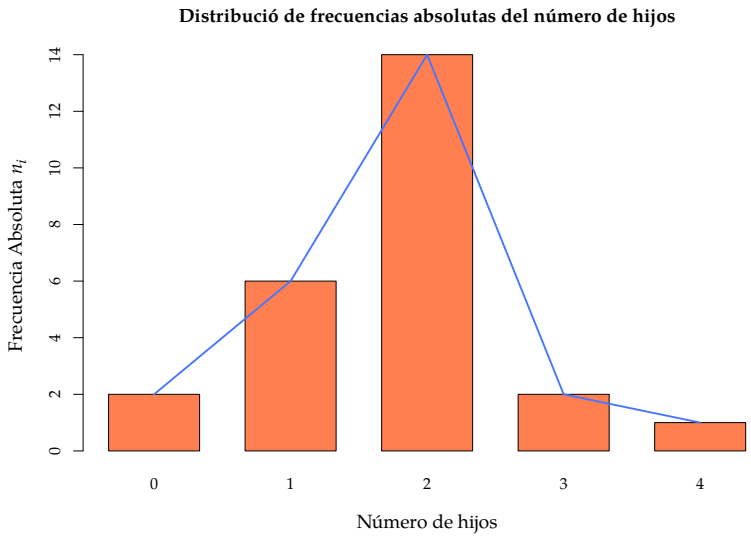


Diagrama de barras de frecuencias acumuladas

Datos sin agrupar

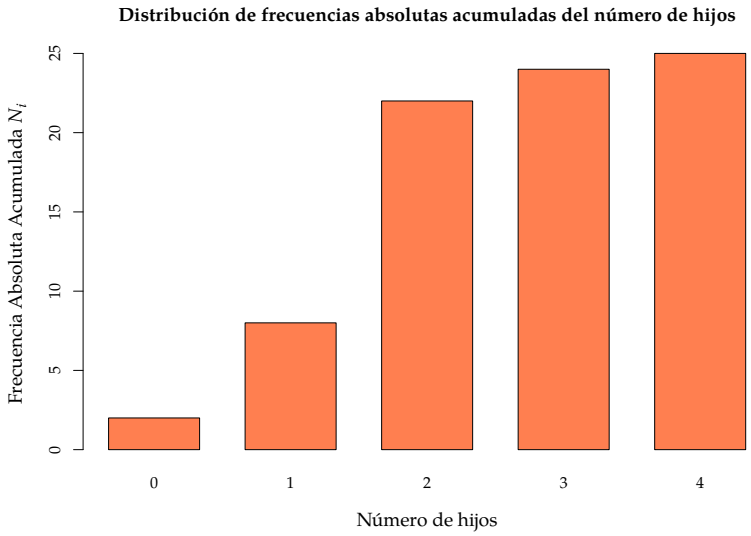
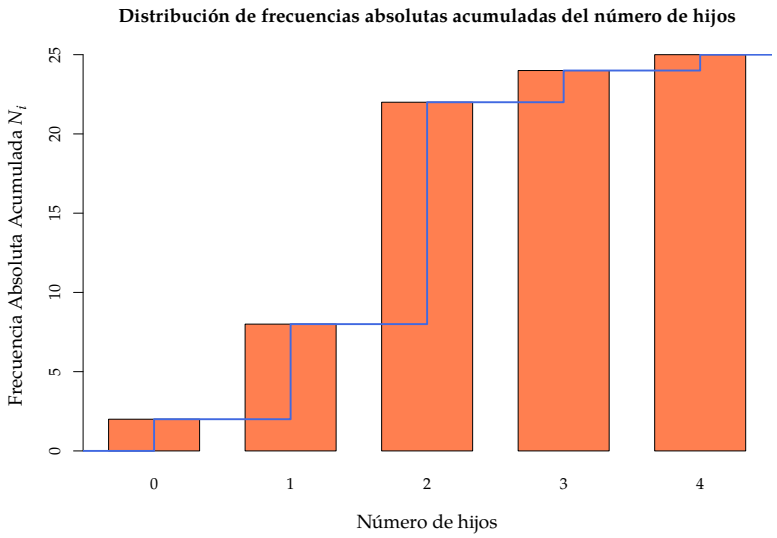


Diagrama de línea o polígono de frecuencias absolutas acumuladas

Datos sin agrupar



Histograma

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

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

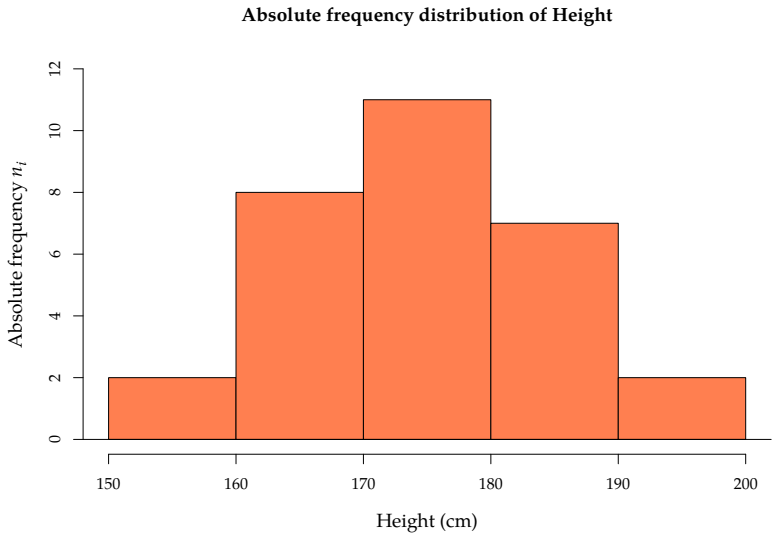
Para cada clase se dibuja una barra de altura la correspondiente frecuencia. A diferencia del diagrama de barras, la anchura del la barra coincide con la anchura de las clases y no hay separación entre dos barras consecutivas.

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

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

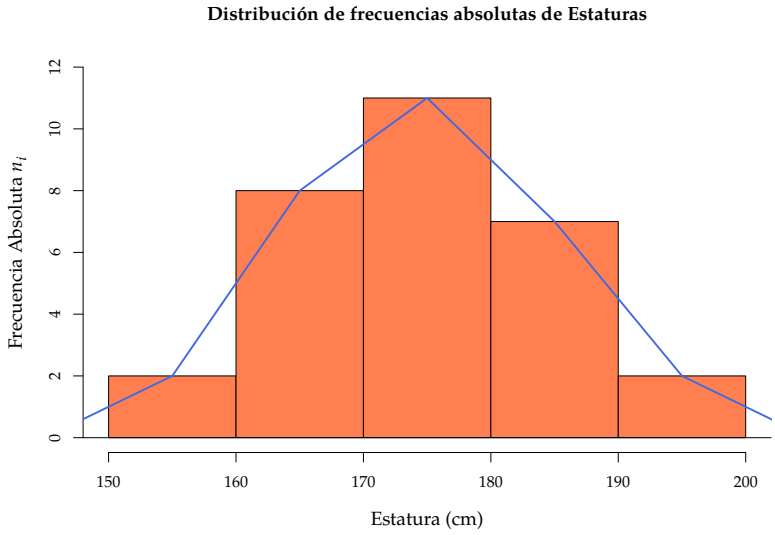
Histograma de frecuencias absolutas

Datos agrupados



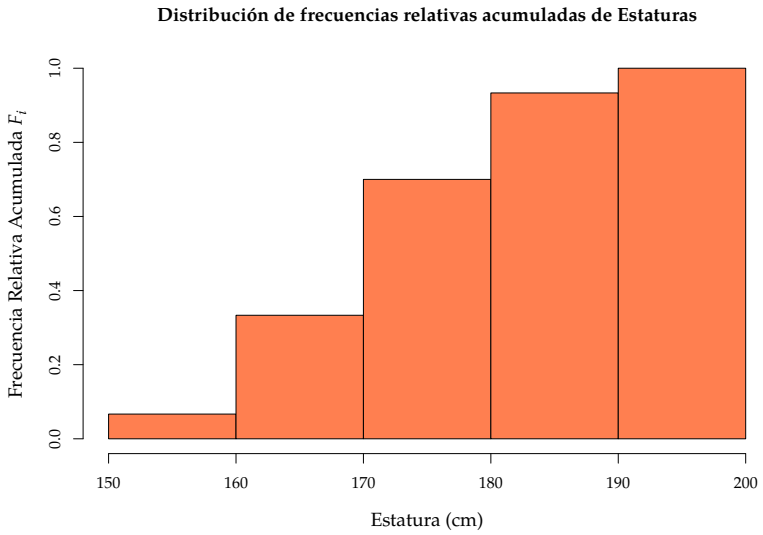
Polígono de frecuencias absolutas

Datos agrupados



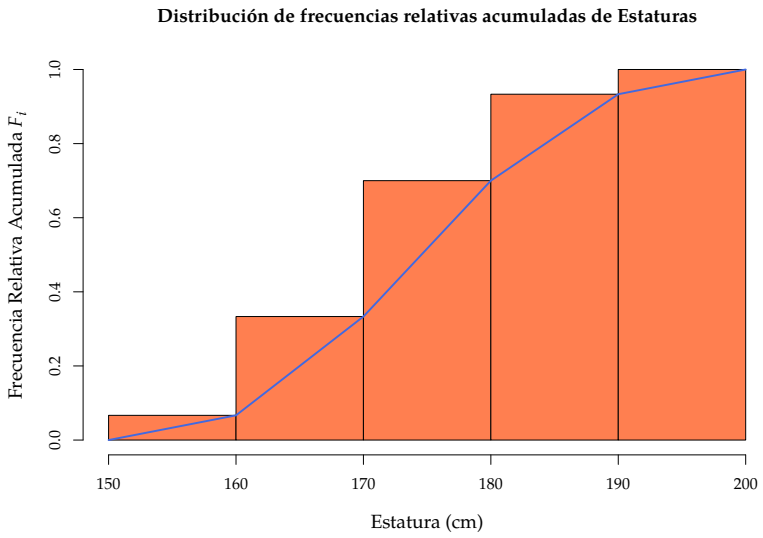
Histograma de frecuencias relativas acumuladas

Datos agrupados



Polígono de frecuencias relativas acumuladas

Datos agrupados



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

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

Diagrama de sectores

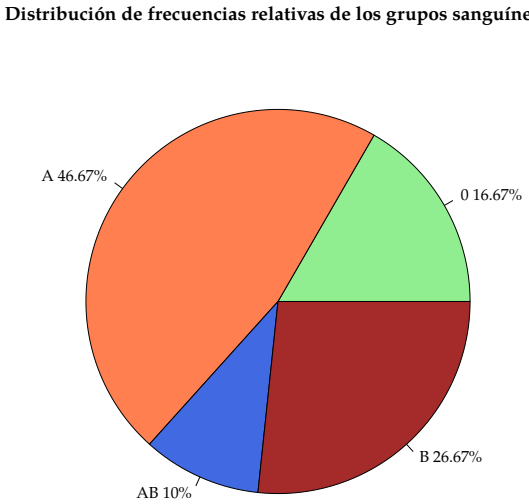
Un **diagrama de sectores** consiste en un círculo divido en porciones, uno por cada valor o categoría de la variable.

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

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

Diagrama de sectores

Atributos



Datos atípicos

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

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

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

Tratamiento de los datos atípicos

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

Cuando trabajemos con muestras pequeñas tenemos varias opciones:

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