Estadística (1)

Jesús García de Jalón de la Fuente

IES Ramiro de Maeztu
Madrid

2020

Estadística

La Estadística trata de describir colectividades formadas por un gran número de objetos.

El conjunto de los objetos que se estudian se denomina **población**.

En ocasiones, el estudio se hace a partir de una **muestra**, esto es, cierto número de objetos tomados aleatoriamente de la población.

El número de objetos de la población o de la muestra es su **tamaño**.

Variable estadística

Sobre la población o sobre una muestra se mide una magnitud. Los valores que toma esta magnitud forman la **variable estadística**.

Si la variable estadística toma valores numéricos se dice que es **cuantitativa**.

Si no es así la variable es **cualitativa**.

Una variable estadística cuantitativa puede tomar un número finito de valores o los infinitos valores comprendidos en un cierto intervalo. En el primer caso hablaremos de variable estadística **discreta** y en el segundo de variable **continua**.

Frecuencias

La **frecuencia** o frecuencia absoluta de un valor x de la variable estadística es el número de objetos de la población que presentan ese valor. Se representa por f.

La frecuencia de un determinado valor dividido por el número de elementos de la población, esto es, la proporción de elementos de la población que presenta este valor, es la **frecuencia relativa**. Se representa por h.

La **frecuencia acumulada** F de un resultado x es el número de elementos de la población en los que la variable toma valores menores o iguales que x.

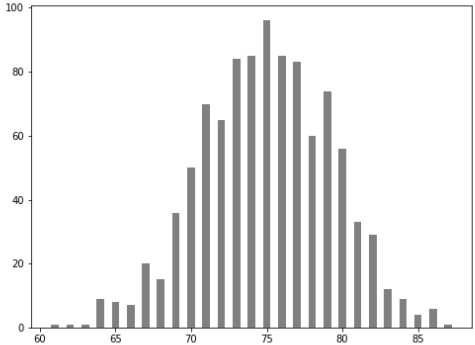
Dividiendo por el número de elementos de la población se obtiene la **frecuencia acumulada relativa** H.

Tablas de frecuencias

x_i	f_i	h_i	F_i	H_i
x_1	f_1	h_1	F_1	H_1
x_2	f_2	h_2	F_2	H_2
x_3	f_3	h_3	F_3	H_3
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_n	f_n	h_n	F_n	H_n

x_i	f_i	h_i	F_i	H_i
$[x_0, x_1)$	f_1	h_1	F_1	H_1
$[x_1, x_2)$	f_2	h_2	F_2	H_2
$[x_2, x_3)$	f_3	h_3	F_3	H_3
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$[x_{n-1}, x_n)$	f_n	h_n	F_n	H_n

Gráficos estadísticos (1)



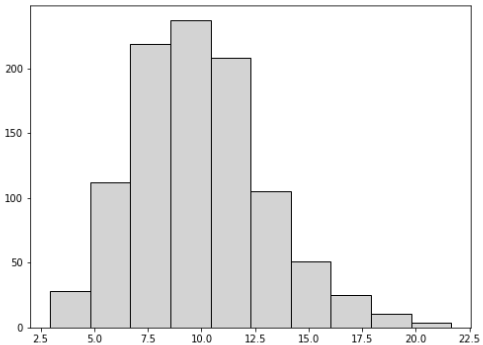
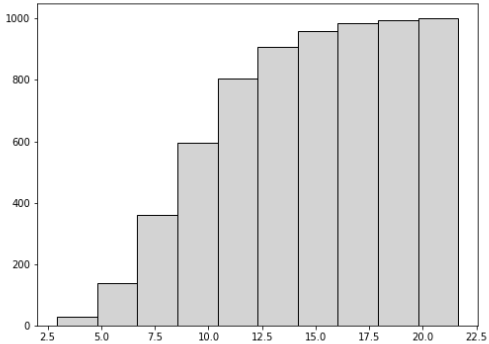
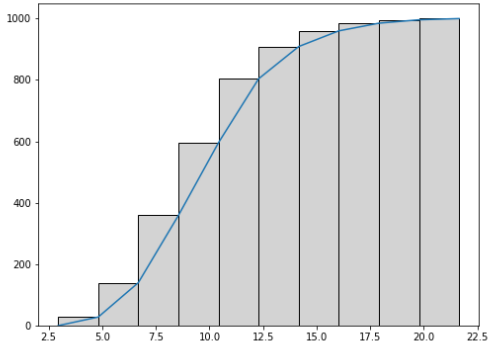


Diagrama de barras e histograma

Gráficos estadísticos (2)





Histograma y polígono de frecuencias acumuladas

Parámetros estadísticos: parámetros de posición

Llamaremos **cuantil** c a un valor de la variable estadística tal que el $c\,\%$ de los valores de la variable estadística son menores o iguales que c.

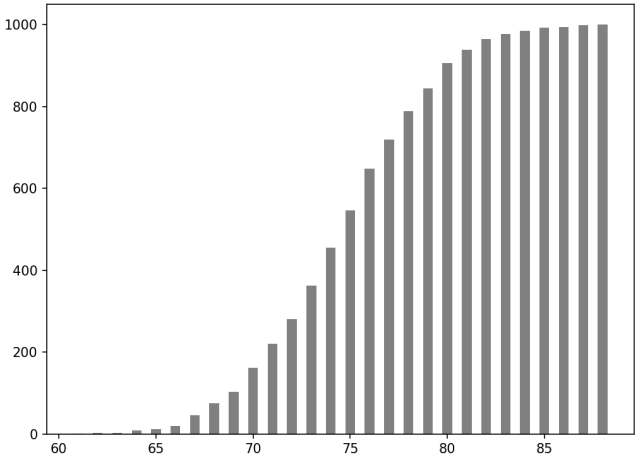
Los cuantiles se definen por grupos de valores que dividen los datos en partes del mismo tamaño. Suelen utilizarse cuantiles que dividen los datos en dos, cuatro, diez o cien partes iguales.

El cuantil correspondiente al $50\,\%$ de los datos se llama **mediana**.

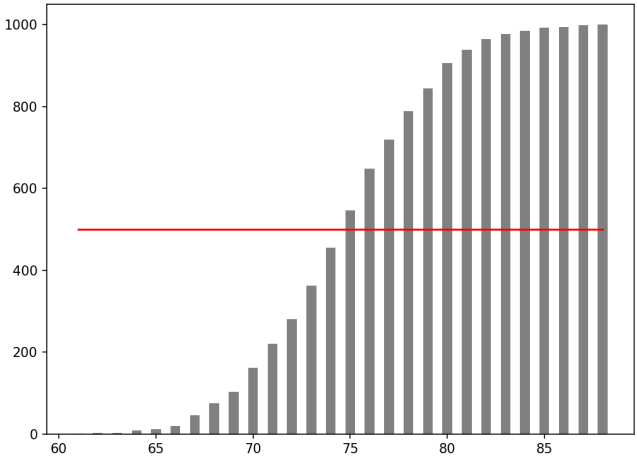
Si los valores obtenidos de la variable estadística se ordenan de menor a mayor, el número que divide los datos en dos partes iguales es la mediana. La mediana es el valor que deja el mismo número de términos a su izquierda y a su derecha. Si el número de términos es par entonces se tomará como mediana la media de los valores centrales.

Mediana y cuartiles

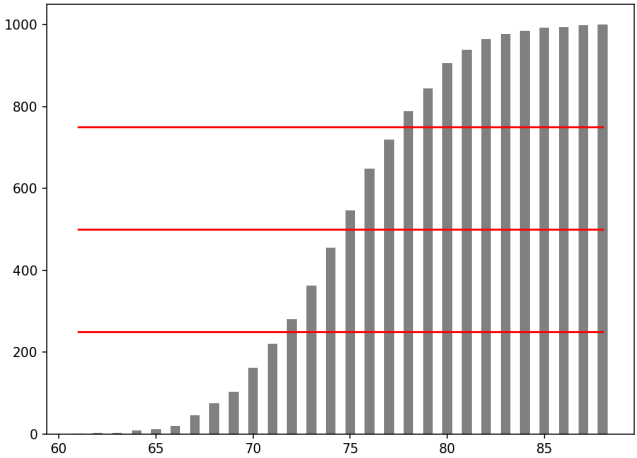
Mediana y cuartiles



Mediana y cuartiles

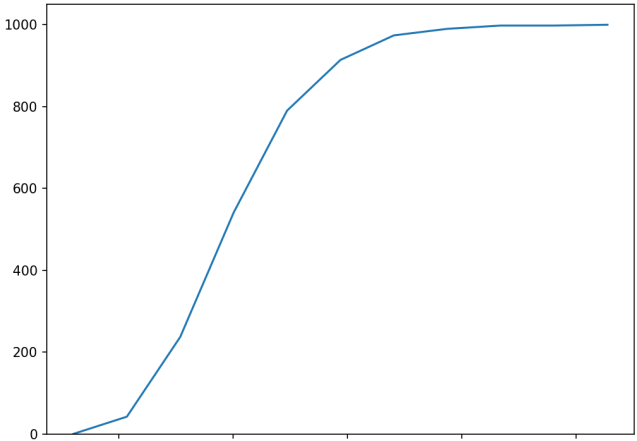


Mediana y cuartiles

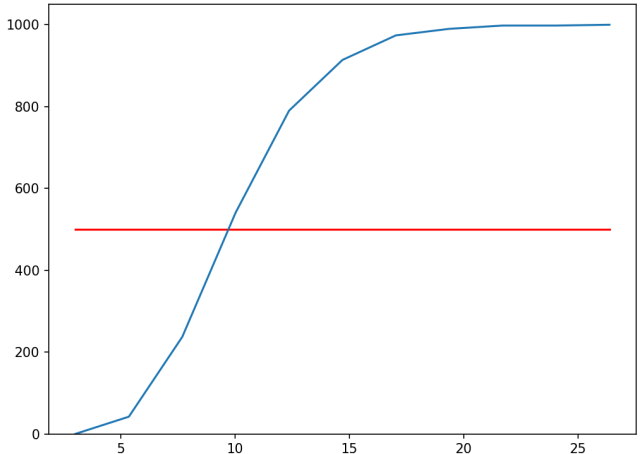


Mediana y cuartiles

Mediana y cuartiles



Mediana y cuartiles



Mediana y cuartiles

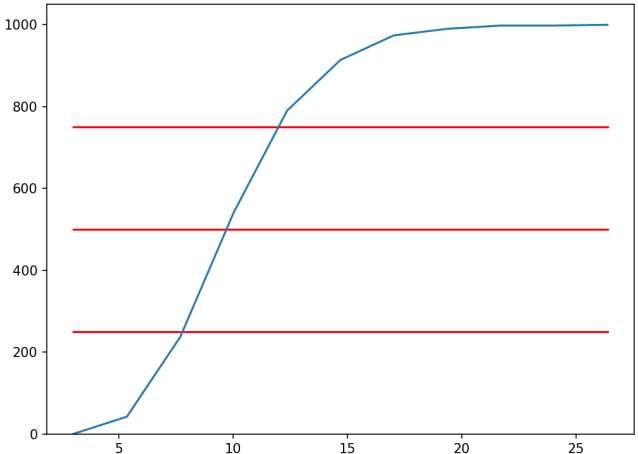


Diagrama de cajas

Los valores mínimo y máximo de los datos y los cuartiles se representan gráficamente mediante el diagrama de cajas:

Diagrama de cajas

Los valores mínimo y máximo de los datos y los cuartiles se representan gráficamente mediante el diagrama de cajas:

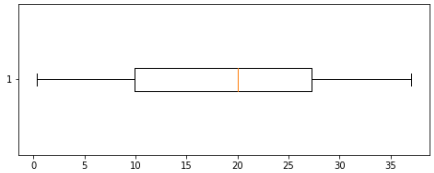
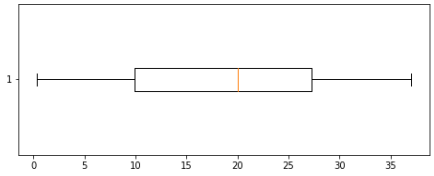
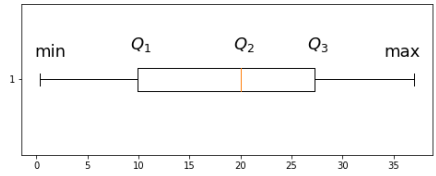


Diagrama de cajas

Los valores mínimo y máximo de los datos y los cuartiles se representan gráficamente mediante el diagrama de cajas:





Cuartiles, deciles y percentiles

Los **cuartiles** dividen los datos en cuatro partes del mismo tamaño. De la misma forma se definen los cuantiles que los dividen en 10 partes (**deciles**) o en 100 partes (**percentiles**).

Se llaman **datos atípicos** aquellos datos que aparecen significativamente distantes del resto de los datos. En muchas ocasiones los datos atípicos no se tienen en cuenta por considerar que se deben a un error o por no considerarlos representativos.

Se suelen tomar como atípicos aquellos datos que son menores que el primer cuartil o mayores que el tercero en 1,5 veces el **rango intercuartílico**, es decir los que no se encuentran en el intervalo:

$$\left[Q_1 - 1{,}5(Q_3 - Q_1), Q_3 + 1{,}5(Q_3 - Q_1)\right]$$

Parámetros de tendencia central

La **moda** es el valor que ocurre más frecuentemente. En una tabla de frecuencias es el valor que se corresponde con la frecuencia más alta.

La **mediana** ya la hemos tratado en el apartado anterior.

Si el intervalo mediano es (x_1, x_2) y a los extremos del intervalo les corresponden unas frecuencias acumuladas relativas H_1 y H_2, el valor de la mediana está dado por:

$$\text{Mediana} = x_1 + \frac{x_2 - x_1}{H_2 - H_1}(0,50 - H_1)$$

La **media o media aritmética** de una variable estadística se define como la suma de todos los valores de la variable dividido por el número de elementos de la población:

$$\bar{x} = \frac{\Sigma x_i}{N}$$

La suma de todos los valores de la variable estadística se puede expresar mediante la suma de cada uno de los valores que toma por sus correspondientes frecuencias. Así:

$$\bar{x} = \frac{\Sigma f_i x_i}{N} = \Sigma h_i x_i$$

En caso de que los datos aparezcan agrupados en intervalos, se toma como valor de la variable la **marca de clase**, es decir, el punto medio del intervalo.

Medidas de dispersión

La media nos permite comparar dos poblaciones sobre las que se ha medido la misma magnitud pero no nos permite saber si los valores de la variable están próximos a la media o no.

Para saber cómo están distribuidos los valores en torno a la media son precisos otros parámetros. Estos son el **rango**, el **rango intercuartílico**, la **varianza** y la **desviación típica**.

El **rango** es la diferencia entre el mayor y el menor valor de la variable estadística.

El **rango intercuartílico** es la diferencia entre el tercer y el primer cuartil.

Varianza y desviación típica

La **varianza** se define por:

$$\sigma^2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{N} = \Sigma h_i (x_i - \bar{x})^2$$

Su raíz cuadrada es la **desviación típica**:

$$\sigma = \sqrt{\frac{\Sigma f_i (x_i - \bar{x})^2}{N}} = \sqrt{\Sigma h_i (x_i - \bar{x})^2}$$

Otra expresión de la varianza

Desarrollando el cuadrado de la diferencia, podemos encontrar otra expresión para la varianza:

$$\begin{aligned}\sigma^2 &= \frac{\Sigma f_i(x_i-\bar{x})^2}{N} = \frac{\Sigma f_i x_i^2 + \Sigma f_i \bar{x}^2 - \Sigma 2 f_i x_i \bar{x}}{N} \\ &= \frac{\Sigma f_i x_i^2}{N} + \frac{\bar{x}^2 \Sigma f_i}{N} - \frac{2\bar{x}\Sigma f_i x_i}{N} \\ &= \frac{\Sigma f_i x_i^2}{N} + \bar{x}^2 - 2\bar{x}\bar{x} \\ &= \frac{\Sigma f_i x_i^2}{N} - \bar{x}^2 = \overline{x^2} - \bar{x}^2\end{aligned}$$

La varianza es igual a la media de los cuadrados menos el cuadrado de la media.

Propiedades de la media y la desviación

La media y la desviación típica tienen las siguientes propiedades:

- Si se suma el mismo número a todos los valores de la variable, la media queda incrementada en esa cantidad pero la desviación típica no varía.
- Si todos los valores de la variable se multiplican por el mismo número, la media y la desviación típica quedan multiplicados por ese número.

El cociente de la desviación típica y la media se llama **coeficiente de variación**:

$$CV = \frac{\sigma}{\bar{x}}$$

Puntuaciones típicas

Para comparar un valor de la variable estadística con el resto de los valores obtenidos en una determinada población se utilizan las **puntuaciones típicas**.

En estas se toma como valor cero el de la media y como unidad la desviación típica.

El paso de la variable x al valor típico z se hace mediante la fórmula:

$$z = \frac{x - \bar{x}}{\sigma} \qquad \text{o, despejando} \qquad x = \bar{x} + z\sigma$$

Ejemplo

En una encuesta sobre tráfico se ha preguntado a 1000 conductores sobre el número de multas recibidas. Se dispone de la siguiente información:

Nº de conductores	180	280	150	200	110	80
Nº de multas	0	1	2	3	4	5

Hacer la tabla de frecuencias con los datos necesarios para calcular:

(a) La mediana.
(b) Los cuartiles y el rango intercuartílico.
(c) La moda.
(d) La media.
(e) La desviación típica.

Solución:

Construimos la tabla con las frecuencias, frecuencias acumuladas, productos de las frecuencias por los datos y productos de las frecuencias por los cuadrados de los datos.

x_i	f_i	F_i	$f_i x_i$	$f_i x_i^2$
0	180	180	0	0
1	280	460	280	280
2	150	610	300	600
3	200	810	600	1800
4	110	920	440	1760
5	80	1000	400	2000
Total	1000		2020	6440

x_i	f_i	F_i	$f_i x_i$	$f_i x_i^2$
0	180	180	0	0
1	280	460	280	280
2	150	610	300	600
3	200	810	600	1800
4	110	920	440	1760
5	80	1000	400	2000
Total	1000		2020	6440

La mediana sería el valor medio de los datos que, ordenados, ocupasen los lugares 500 y 501.

A la vista de la tablas de frecuencias acumuladas, la mediana es $Q_2 = 2$.

x_i	f_i	F_i	$f_i x_i$	$f_i x_i^2$
0	180	180	0	0
1	280	460	280	280
2	150	610	300	600
3	200	810	600	1800
4	110	920	440	1760
5	80	1000	400	2000
Total	1000		2020	6440

De forma similar calculamos el primer cuartil (media entre los datos que ocupan el lugar 250 y 251) $Q_1 = 1$ y el tercer cuartil $Q_3 = 3$.

El rango intercuartílico es $Q_3 - Q_1 = 2$.

x_i	f_i	F_i	$f_i x_i$	$f_i x_i^2$
0	180	180	0	0
1	280	460	280	280
2	150	610	300	600
3	200	810	600	1800
4	110	920	440	1760
5	80	1000	400	2000
Total	1000		2020	6440

La moda es el dato con mayor frecuencia. En este caso 1.

x_i	f_i	F_i	$f_i x_i$	$f_i x_i^2$
0	180	180	0	0
1	280	460	280	280
2	150	610	300	600
3	200	810	600	1800
4	110	920	440	1760
5	80	1000	400	2000
Total	1000		2020	6440

La media es la suma de los $f_i x_i$ dividido por el número de datos que es la suma de las f_i:

$$\bar{x} = \frac{2020}{1000} = 2{,}02$$

x_i	f_i	F_i	$f_i x_i$	$f_i x_i^2$
0	180	180	0	0
1	280	460	280	280
2	150	610	300	600
3	200	810	600	1800
4	110	920	440	1760
5	80	1000	400	2000
Total	1000		2020	6440

La varianza es la media de los cuadrados menos el cuadrado de la media:

$$\sigma^2 = \frac{6440}{1000} - \bar{x}^2 \simeq 2,36$$

y la desviación típica es la raíz de la varianza:

$$\sigma = \sqrt{\sigma^2} \simeq 1,54$$

Agradecimiento

Gracias por vuestra atención