Estadística

La **Estadística** es la ciencia que trata de los métodos y procedimientos para recoger, clasificar, resumir, hallar regularidades y analizar *datos*, así como de realizar *inferencias* a partir de ellos, con la finalidad de ayudar a la toma de *decisiones* y en su caso formular *predicciones*. Podemos por tanto clasificar la Estadística en:

- **Descriptiva o deductiva**, que tiene por objeto la recogida, recopilación, y reducción de datos, su organización en tablas y gráficos y el cálculo de unos valores que representen al conjunto de datos.
- **Inferencial o inductiva** tiene por objeto establecer previsiones o conclusiones sobre una población basándose en los resultados obtenidos de una muestra

Definiciones de términos estadísticos

- **Población:** es el conjunto de elementos, individuos o entes sujetos a estudio y de los cuales queremos obtener un resultado.
- **Variable:** es la característica que estamos midiendo.

Existen dos tipos de variables:

Variable cualitativa: Es aquella que expresa un atributo o característica, ejemplo: Rubio, moreno, etc.

Variable cuantitativa: Es aquella que podemos expresar numéricamente: edad, peso, etc. Esta a su vez la podemos subdividir en:

Variable discreta, aquella que entre dos valores próximos puede tomar a lo sumo un número finito de valores. Ejemplos: el número de obreros de una fabrica, el de alumnos de la universidad, etc.

Variable continua la que puede tomar los infinitos valores de un intervalo. En muchas ocasiones la diferencia es más teórica que práctica, ya que los aparatos de medida dificultan que puedan existir todos los valores del intervalo. Ejemplos, peso, estatura, distancias, etc.

La variable se denota por las mayúsculas de letras finales del alfabeto castellano. A su vez cada una de estas variables puede tomar distintos valores , colocando un subíndice, que indica orden.

$X = (x_1, x_2 , ... X_n)$

- **Muestra:** Conjunto de elementos que forman parte de población . La muestra representa a esta población.
- **Tamaño muestral:** Es le número de elementos u observaciones que tomamos. Se denota por n ó N.
- **Dato:** Cada uno de los individuos, cosas, entes abstractos que integran una población o universo determinado. Dicho de otra forma, cada valor observado de la variable.

Frecuencias absolutas, relativas y acumuladas.

Frecuencia absoluta: Llamaremos así al número de repeticiones que presenta una observación. Se representa por n_i.

Frecuencia relativa: Es la frecuencia absoluta dividida por el número total de datos, se suele expresar en tanto por uno:

La suma de todas las frecuencias relativas, siempre debe ser igual a la unidad.

Frecuencia absoluta acumulada: es la suma de los distintos valores de la frecuencia absoluta tomando como referencia un individuo dado. La última frecuencia absoluta acumulada es igual al nº de casos:

$N_1 = n_1$

$N_2 = n_1 + n_2$

$N_n = n_1 + n_2 + \ldots\ldots + n_{n-1} + n_n = n$

Frecuencia relativa acumulada: es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos, se la suele representar con la notación: F_i

Tabla de frecuencias para una variable discreta.				
x_i	**n_i**	**N_i**	**f_i**	**F_i**
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
x_3	n_3	N	f_3	1
	$Sn_i=N$		1	

EJEMPLO

Queremos hacer un estudio estadístico del número de Técnicos Superiores en Electricidad (TSE) que existen en las empresas eléctricas de una determinada ciudad. Para ello se ha encuestado a 50 empresas y se han obtenido los siguientes datos:

2	4	2	3	1	2	4	2	3	0	2	2	2	3	2	6	2	3	2	2	3	2	3	3	4
3	3	4	5	2	0	3	2	1	2	3	2	2	3	1	4	2	3	2	4	3	3	2	2	1

Se pide:

a) ¿Cuál es la población objeto de estudio?

b) ¿Qué variable estamos estudiando?

c) ¿Qué tipo de variable es?

d) Construir la tabla de frecuencias?

e) ¿Cuál es el número de empresas que tiene como máximo 2 TSE?

f) ¿Cuántas empresas tienen más de 1 TSE, pero como máximo 3?

g) ¿Qué porcentaje de empresas tiene más de 3 TSE ?

SOLUCIÓN:

a) La población objeto de estudio es las **empresas de electricidad** de una ciudad.

b) La variable que estamos estudiando es el **número de TSE** por empresa.

c) El tipo de variable es discreta ya que el número de TSE solo puede tomar determinados valores enteros.

d) Para construir la tabla de frecuencias tenemos que ver cuantas empresas tienen un determinado número de TSE. Podemos ver que el número de TSE, toma los valores existentes entre 0 TSE, los que menos y 6 TSE, los que más y tendremos:

xi	ni	Ni	fi	Fi
0	2	2	0.04	0.04
1	4	6	0.08	0.12
2	21	27	0.42	0.54
3	15	42	0.30	0.84
4	6	48	0.12	0.96
5	1	49	0.02	0.98
6	1	50	0.024	1
	N = 50		1	

e) El número de empresas que tienen dos o menos TSE es: 2+4+21 = 27

f) El número de empresas que tienen más de un TSE pero tres como máximo es: 21 + 15 = 36

Por último el porcentaje de empresas que tiene más de tres TSE, son aquellos que tienen 4; 5 y 6 es decir 6+1+1= 8

El porcentaje será el tanto por uno multiplicado por cien es decir, la frecuencia relativa de dichos valores multiplicado por 100: (0.12+0.02+0.02)* 100 = 0,16 + 100 = 16 %

Marca de Clase

Cuando nos encontramos con una distribución con un gran número de variables, se se suelen agrupar en intervalos para facilitar la comprensión de los datos Se indica por L_{i-1} al extremo inferior del intervalo y por L_i al extremo superior. Cerramos el intervalo por la izquierda y abrimos por la derecha, pero se puede hacer al contrario; **[Li-1 , Li)** Para operar utilizaremos la **marca de clase,** el punto medio de un intervalo

Amplitud del intervalo: la longitud del intervalo, se representa por: **a = Li - Li-1**

Nº de intervalos: A partir de la raíz cuadrada del número de datos, decidimos, redondeando el número de intervalos.

Recorrido: Valor mayor, menos valor menor de los datos. Re= x_n-x_1

Amplitud: División entre el Recorrido y el número de intervalos que hayamos decidido.

Medidas de Centralización

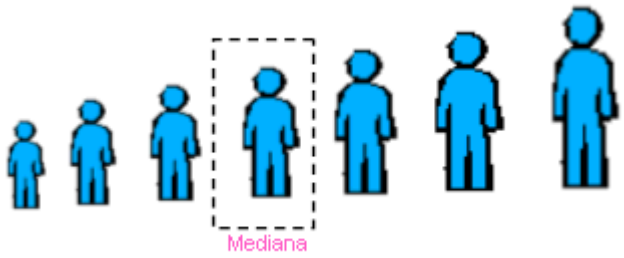
Nos dan un centro de la distribución de frecuencias, es un valor que se puede tomar como representativo de todos los datos. Hay diferentes modos para definir el "centro" de las observaciones en un conjunto de datos. Por orden de importancia, son:

MEDIA : (media aritmética o simplemente media). es el promedio aritmético de las observaciones, es decir, el cociente entre la suma de todos los datos y el numero de ellos. Si x_i es el valor de la variable y n_i su frecuencia, tenemos que:

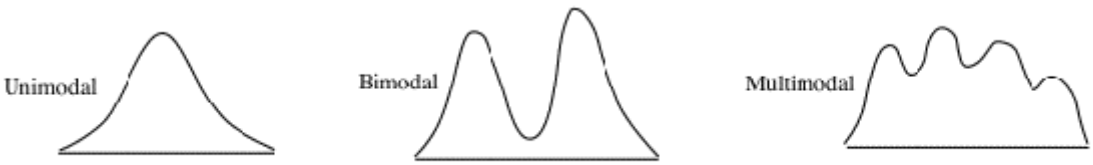
$$\bar{x} = \frac{\sum x_i n_i}{n}$$

Si los datos están agrupados utilizamos las marcas de clase, es decir c_i en vez de x_i.

MEDIANA (Me):es el valor que separa por la mitad las observaciones ordenadas de menor a mayor, de tal forma que el 50% de estas son menores que la mediana y el otro 50% son mayores. Si el número de datos es impar la mediana será el valor central, si es par tomaremos como mediana la media aritmética de los dos valores centrales.



MODA (M_0)*:* es el valor de la variable que más veces se repite, es decir, aquella cuya frecuencia absoluta es mayor. No tiene porque ser única.



Medidas de Dispersión

Las medidas de tendencia central tienen como objetivo el sintetizar los datos en un valor representativo, las medidas de dispersión nos dicen hasta que punto estas medidas de tendencia central son representativas como síntesis de la información. Las medidas de dispersión cuantifican la separación, la dispersión, la variabilidad de los valores de la distribución respecto al valor central. Distinguimos entre medidas de dispersión absolutas, que no son comparables entre diferentes muestras y las relativas que nos permitirán comparar varias muestras.

MEDIDAS DE DISPERSIÓN ABSOLUTAS

VARIANZA (s^2): es el promedio del cuadrado de las distancias entre cada observación y la media aritmética del conjunto de observaciones.

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 n_i}{n}$$

Haciendo operaciones en la fórmula anterior obtenemos otra fórmula para calcular la varianza:

$$s^2 = \frac{\sum_i x_i^2 n_i}{n} - \bar{x}^2$$

Si los datos están agrupados utilizamos las marcas de clase en lugar de X_i.

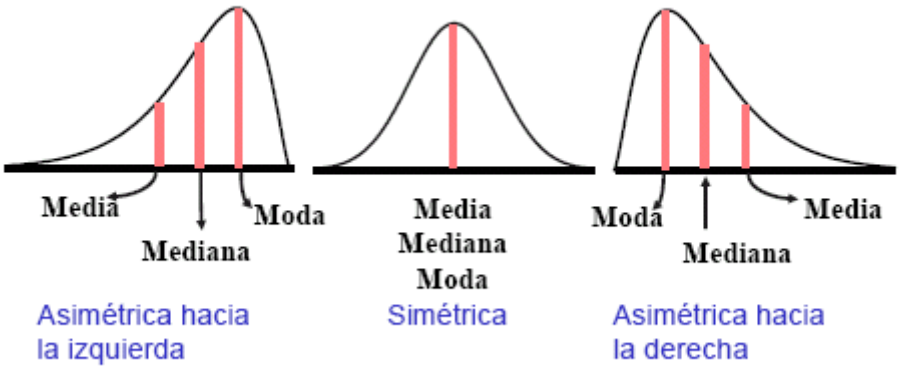
DESVIACIÓN TÍPICA (S)*:* La varianza viene dada por las mismas unidades que la variable pero al cuadrado, para evitar este problema podemos usar como medida de dispersión la desviación típica que se define como la raíz cuadrada positiva de la varianza

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

Para estimar la desviación típica de una población a partir de los datos de una muestra se utiliza la fórmula (**cuasi desviación típica**):

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 n_i}{n-1}}$$

RANGO (R_e). Es la diferencia entre el valor de las observaciones mayor y el menor. $R_e = x_{max} - x_{min}$



EJEMPLO

El número de diás necesarios por 10 equipos de trabajadores para terminar 10 instalaciones de iguales características han sido: 21, 32, 15, 59, 60, 61, 64, 60, 71, y 80 días. Calcular la media, mediana, moda, varianza y desviación típica.

SOLUCIÓN:

La media: suma de todos los valores de una variable dividida entre el número total de datos de los que se dispone:

$$\bar{X} = \frac{21+32+15+59+60+61+64+60+71+80}{10} = 52.3 \textbf{ dias}$$

La mediana: es el valor que deja a la mitad de los datos por encima de dicho valor y a la otra mitad por debajo. Si ordenamos los datos de mayor a menor observamos la secuencia:

15, 21, 32, 59, 60, 60,61, 64, 71, 80.

Como quiera que en este ejemplo el número de observaciones es par (10 individuos), los dos valores que se encuentran en el medio son 60 y 60. Si realizamos el cálculo de la media de estos dos valores nos dará a su vez **60,** que es el valor de **la mediana**.

La moda: el valor de la variable que presenta una mayor frecuencia es **60**

La varianza S^2: Es la media de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética de la distribución.

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 n_i}{n} \qquad S_x^2 = \frac{(15-52.3)^2 + (21-53,2)^2 \ldots (80-53,2)^2}{10} = 427,61$$

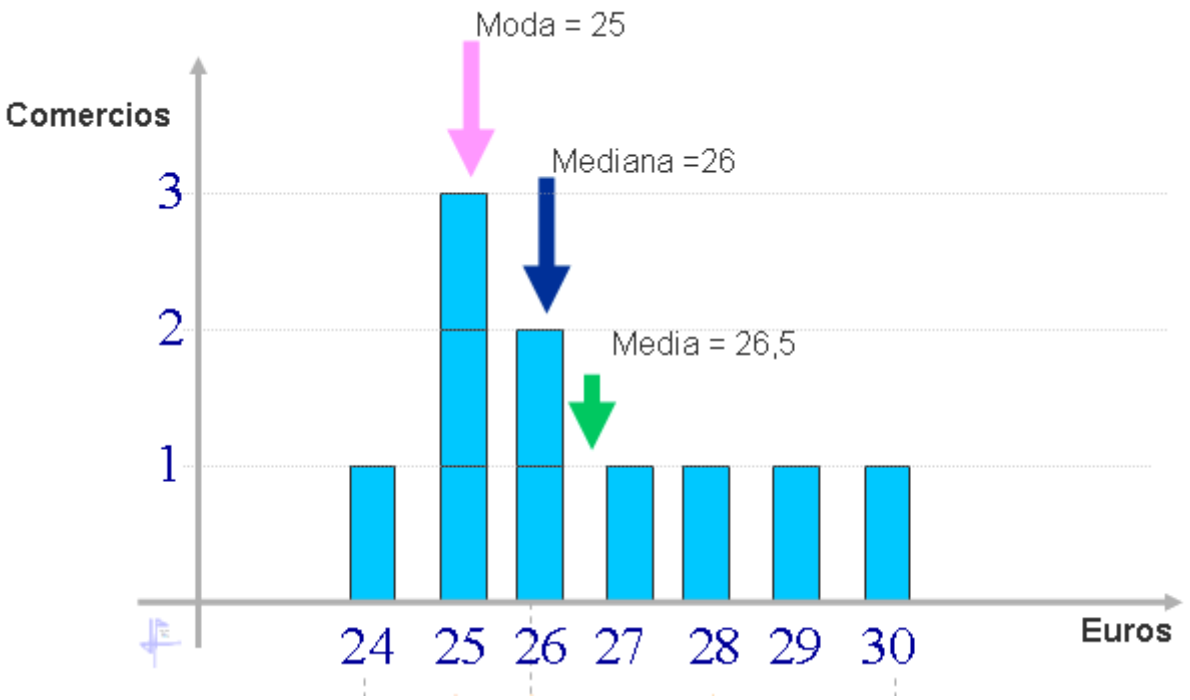
La desviación típica S: es la raíz cuadrada de la varianza. $s = \sqrt{s^2}$ **S** = RAIZ (427,61) = **20.67**

El rango: diferencia entre el valor de las observaciones mayor y el menor : **80 - 15 = 65 días**

EJEMPLO

El precio de un interruptor magentotérmico en 10 comercios de electricidad de una ciudad son : 25, 25, 26, 24, 30, 25, 29, 28, 26, y 27 Euros. Hallar la media, moda, mediana, (abrir la calculadora estadística, más abajo) diagrama de barras y el diagrama de caja.

SOLUCIÓN:



RESUMEN DE FORMULAS

Centralización

Media aritmética: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Mediana: En el conjunto de datos ordenado, valor que ocupa la posición central.

Moda: Es el valor más frecuente.

Dispersión

	Poblaciones	**Inferencia de poblaciones (muestra)**
Varianza:	$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$	$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Desviación típica:	$s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$	$s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
Coef. Variación:	$CV = \dfrac{s}{\bar{x}}$	

Rango: diferencia entre el valor de las observaciones mayor y el menor

Para las medidas de las poblaciones se suele utilizar letras griegas (**μ, σ**)
y para las de muestras letras latinas (**x, s**)