Estadística

CONCEPTOS DE ESTADÍSTICA

POBLACIÓN	*Llamamos población estadística, al conjunto de referencia sobre el cual van a recaer las observaciones.*
INDIVIDUOS	*Se llama unidad estadística o individuo a cada uno de los elementos que componen la población estadística. El individuo es un ente observable que no tiene por qué ser una persona, puede ser un objeto, un ser vivo, o incluso algo abstracto.*
MUESTRA	*Es un subconjunto de elementos de la población. Se suelen tomar muestras cuando es difícil o costosa la observación de todos los elementos de la población estadística.*
CENSO	*Decimos que realizamos un censo cuando se observan todos los elementos de la población estadística.*
CARACTERES	*La observación del individuo la describimos mediante uno o más caracteres. El carácter es, por tanto una cualidad o propiedad inherente en el individuo.* TIPOS DE CARACTERES: **Cualitativos**: aquellos que son categóricos, pero no son numéricos. p. ej. <color de los ojos>, <profesión>, <marca de coche>,... **Ordinales**: aquellos que pueden ordenarse, pero no son numéricos. p. ej. <preguntas de encuesta sobre el grado de satisfacción de algo> Mucho, poco, nada. Bueno, regular, malo,... **Cuantitativos** : son numéricos. p. ej. <peso>, <talla>, <núm. de hijos>, <núm. de libros leídos al mes>,...
MODALIDAD VALOR	*Un carácter puede mostrar distintas modalidades o valores, es decir, son distintas manifestaciones o situaciones posibles que puede presentar un carácter estadístico. Las modalidades o valores son incompatibles y exhaustivos.* Generalmente se utiliza el término <u>modalidad</u> cuando hablamos de caracteres cualitativos y el término <u>valor</u> cuando estudiamos caracteres cuantitativos. p. ej. el carácter cualitativo <Estado Civil> puede adoptar las modalidades: casado, soltero, viudo. El carácter cuantitativo <Edad> puede tomar los valores: diez, once, doce, quince años,...
VARIABLE	*Al conjunto de los distintos valores numéricos que adopta un*

ESTADÍSTICA	*carácter cuantitativo se llama variable estadística.* TIPOS DE VARIABLES ESTADÍSTICAS: **Discretas**: Aquellas que toman valores aislados (números naturales), y que no pueden tomar ningún valor intermedio entre dos consecutivos fijados. p. ej. <núm. de goles marcados>, <núm. de hijos>, <núm., de discos comprados>, <núm. de pulsaciones>,... **Continuas**: Aquellas que toman infinitos valores (números reales) en un intervalo dado, de forma que pueden tomar cualquier valor intermedio, al menos teóricamente, en su rango de variación. p. ej. <talla>, <peso>, <presión sanguínea>, <temperatura>, ...
OBSERVACIONES	*Una observación es el conjunto de modalidades o valores de cada variable estadística medidos en un mismo individuo.* p. ej. en una población de 100 individuos podemos estudiar, de forma individual, tres caracteres: <edad: 18, 19,...>, <sexo: Hombre, Mujer> y <si ha votado en las elecciones: Si, No>. Realizamos 100 observaciones con tres datos cada una, es decir, una de las observaciones podría ser (43, H, S).

ORDENACIÓN DE DATOS

- **CARACTERES CUALITATIVOS**

Consideremos una muestra de tamaño N sacada de una población estadística de la que observamos un carácter cualitativo A que presenta las modalidades siguientes: a_1, a_2, a_3,..., a_k, llamamos

FRECUENCIA ABSOLUTA	n_i	de la modalidad a_i al número de veces que aparece repetida dicha modalidad en el conjunto de las observaciones realizadas. $\sum_{i=1}^{k} n_i = N \quad ; \quad 0 \le n_i \le N \quad ; \quad i = 1,2,3,...$
FRECUENCIA RELATIVA	f_i	de la modalidad a_i al cociente entre la frecuencia absoluta y el número de datos (= tamaño de la muestra N).

		$f_i = \frac{n_i}{N} \quad ; \quad \sum_{i=1}^{k} f_i = 1 \quad ; \quad 0 \le f_i \le 1 \quad ; \quad i = 1,2,3,...$

Los datos de las observaciones se pueden recoger en la siguiente tabla de distribución :

Modalidades del caracter A	n_i	f_i
a_1	n_1	f_1
a_2	n_2	f_2
⋮	⋮	⋮
a_k	n_k	f_k
	$\sum n_i = N$	$\sum f_i = 1$

- **CARACTERES CUANTITATIVOS**

Consideramos una variable estadística X que, en una muestra de tamaño N extraída de una población estadística, toma los valores $x_1 < x_2 < x_3 < ... < x_k$, definimos los siguientes conceptos:

Tamaño de la muestra	N	Llamamos *tamaño muestral* al número de observaciones realizadas, es decir, al número total de datos. $\sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_k = N$
Frecuencia Absoluta	n_i	Llamamos *frecuencia absoluta* de un valor x_i de la variable estadística X al número de veces que aparece repetido dicho valor en el conjunto de las observaciones realizadas. $\sum_{i=1}^{k} n_i = N \quad ; \quad 0 \le n_i \le N \quad ; \quad i = 1,2,3,...$
Frecuencia Absoluta Acumulada	N_i	Llamamos *frecuencia absoluta acumulada* en el valor x_i a la suma de las frecuencias absolutas de los valores inferiores o iguales a él. Evidentemente, los valores x_i han de estar ordenados de forma creciente, como ya se ha indicado, y la frecuencia absoluta acumulada del último valor será igual a N. $N_k = N$

Frecuencia Relativa	f_i	Llamamos *frecuencia relativa* de un valor x_i de la variable estadística X al cociente entre la frecuencia absoluta y el número de observaciones realizadas. $f_i = \frac{n_i}{N}$; $0 \le f_i \le 1$; $\sum_{i=1}^{k} f_i = f_1 + f_2 + \ldots + f_k = 1$
Frecuencia Relativa Acumulada	F_i	Llamamos *frecuencia relativa acumulada* en el punto x_i al cociente entre la frecuencia absoluta acumulada y el número de observaciones realizadas. $F_i = \frac{N_i}{N}$; $F_k = 1$

En las observaciones realizadas en una muestra o población, puede ocurrir:

1. Que la variable estadística tome pocos valores diferentes (ya sea grande o pequeño el tamaño de la muestra).
2. Que, en una muestra de gran tamaño, la variable estadística tome muchos valores diferentes, ya se trate de variable estadística discreta como de variable estadística continua (este último caso es el más habitual).

En el primer caso no es necesario agrupar los datos, y la tabla de distribución presenta el siguiente aspecto (ordenando los datos de menor a mayor) :

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	$N_k = N$	$F_k = 1$
	$\sum n_i = N$	$\sum f_i = 1$		

En el segundo caso por tratarse de variable continua o discreta pero con un número de datos muy grande, es aconsejable **AGRUPAR LOS DATOS EN CLASES**.

- Agrupamos los valores de la variable estadística en **intervalos de clase** contiguos y elegidos convenientemente para no perder mucha información. No existe un criterio claro de cuál debe ser el número de intervalos que debemos escoger, *Norcliffe* establece que el número de clases debe ser, aproximadamente igual a la raíz cuadrada positiva del número de datos. Normalmente, el número de intervalos de clase se suele fijar entre 5 y 15 y de tal manera que en cada clase se tengan, al menos, 5 observaciones. De todas formas el investigador los acomodará a las condiciones especificas del problema estadístico objeto de estudio

(se tomarán tantos intervalos solapados como sean necesarios para recubrir todo el recorrido de la variable).

- Los extremos de los intervalos de clase se denominan **extremos de clase** y sus puntos medios **marcas de clase** (valor que nos representa la información que contiene un intervalo).
- Como cada observación debe quedar perfectamente encasillada en uno y sólo un intervalo de clase, debemos decidir a qué intervalos pertenecen los extremos de las clases, por lo que habrán de tomarse intervalos semiabiertos o tomando el extremo de cada clase con un decimal más que las observaciones. Con el fin de que la clasificación esté bien hecha, los intervalos se deben construir de manera que el límite superior de una clase coincida con el límite inferior de la siguiente, y además, adoptando el criterio de que los intervalos sean cerrados por la izquierda y abiertos por la derecha.

Tabla de frecuencias de una variable estadística agrupada en intervalos.

Intervalos	Marcas x_i	n_i	f_i	N_i	F_i
$[a_0 - a_1)$	x_1	n_1	f_1	N_1	F_1
$[a_1 - a_2)$	x_2	n_2	f_2	N_2	F_2
⋮	⋮	⋮	⋮	⋮	⋮
$[a_k - a_{k+1})$	x_k	n_k	f_k	$N_k = N$	$F_k = 1$
		$\sum n_i = N$	$\sum f_i = 1$		

Media aritmética:

La media aritmética de una variable se define como la suma ponderada de los valores de la variable por sus frecuencias relativas y lo denotaremos por $\bar{x}$ y se calcula mediante la expresión:

$$\bar{x} = \sum_{i=1}^{n} x_i \cdot f_i = \sum_{i=1}^{n} = \frac{x_i \cdot n_i}{N}$$

x_i representa el valor de la variable o en su caso la marca de clase.

Mediana:

La mediana es el valor central de la variable, es decir, supuesta la muestra ordenada en orden creciente o decreciente, el valor que divide en dos partes la muestra.

Para calcular la mediana debemos tener en cuenta si la variable es discreta o continua.

Cálculo de la mediana en el caso discreto:

Tendremos en cuenta el tamaño de la muestra.

Si **N es Impar,** hay un término central, el término $X_{\frac{N+1}{2}}$ que será el valor de la mediana.

Si **N es Par,** hay dos términos centrales, $X_{\frac{N}{2}}, X_{\frac{N}{2}+1}$ la mediana será la media de esos dos valores

Veamos un ejemplo.

N Impar

1,4,6,7,8,9,12,16,20, 24,25,27,30 N=13

Término Central el 7º , 12

Me=12

N par

1,4,6,7,8,9,12,16,20, 24,25,27 N=12

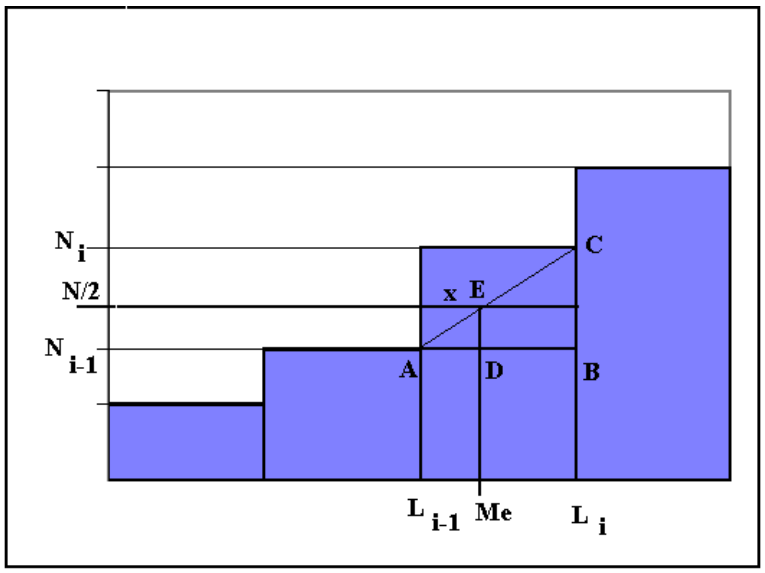
Términos Centrales el 6º y 7º 9 y 12

Me= $\frac{9+12}{2} = 10{,}5$

Cálculo de la mediana en el caso contínuo:

Si la variable es continua, la tabla vendrá en intervalos, por lo que se calcula de la siguiente forma:

Nos vamos a apoyar en un gráfico de un histograma de frecuencias acumuladas.



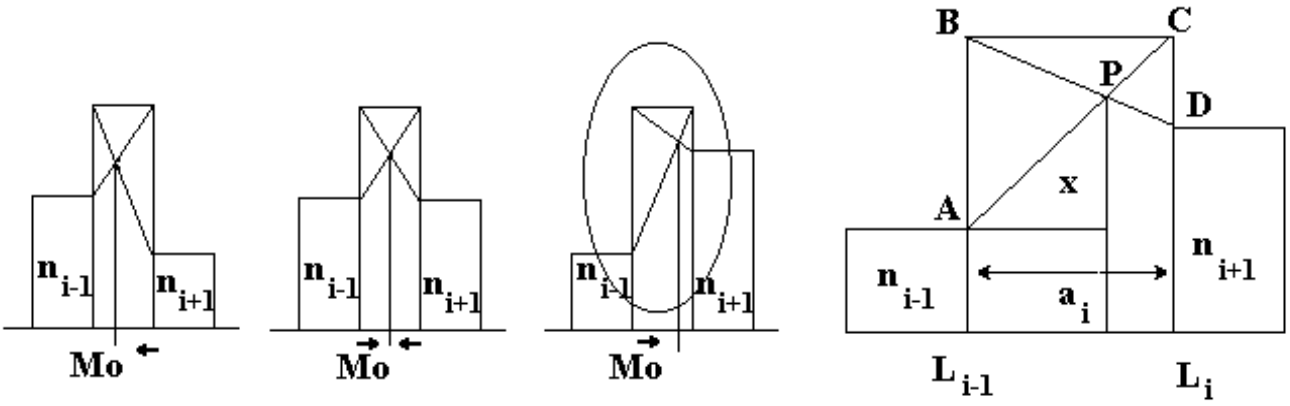
Para el cálculo de la mediana recurriremos a la interpolación de valores.

MODA:

La moda es el valor de la variable que tenga mayor frecuencia absoluta, la que más se repite, es la única medida de centralización que tiene sentido estudiar en una variable cualitativa, pues no precisa la realización de ningún cálculo.

Por su propia definición, la moda no es única, pues puede haber dos o más valores de la variable que tengan la misma frecuencia siendo esta máxima. En cuyo caso tendremos una distribución bimodal o polimodal según el caso.

Por lo tanto el cálculo de la moda en distribuciones discretas o cualitativas no precisa de una explicación mayor; sin embargo, debemos detenernos un poco en el cálculo de la moda para distribuciones cuantitativas continuas.



Para el cálculo de la moda recurriremos a la interpolación de valores.