

1.- ESTADÍSTICA DESCRIPTIVA

El pensamiento estadístico será un día tan necesario para el ciudadano como la capacidad de leer o escribir.

H.G. Wells

1.- Fenómenos aleatorios y determinísticos

1.1.- Introducción. ¿Qué es la Estadística Descriptiva?

¿ Que es la **estadística** ? .

La palabra estadística se emplea con dos significados distintos :

- a) Estadísticas (en plural) selecciones de datos numéricos presentados en forma esquemática y ordenada.
- b) Estadística como ciencia.

Para el alumno la estadística debe tener el significado de la opción b) y desde este punto podemos dar la definición de **estadística** como:

" la ciencia que estudia la técnica o método que se sigue para recoger, organizar, resumir, representar, analizar, generalizar y predecir resultados de las observaciones de fenómenos aleatorios. "

Partes de la estadística, en esquema:

ESTADISTICA	DESCRIPTIVA :	Encuestas. Organización datos. Tabulación. Representaciones. Cálculo de parámetros.
	INFERENCIAL :	Interpretación de resultados. Conclusiones y predicciones.

1.2.- Fenómenos aleatorios y determinísticos. Ejemplos

Decimos que un fenómeno o experimento es **aleatorio** si reúne las siguientes características:

- a) Podemos realizarlo el número de veces que deseemos sin alterar las condiciones del experimento.
- b) No se puede predecir el resultado.

Ejemplos: *lanzar una moneda al aire, un dado, extraer una carta de la baraja, hallar el número de tornillos defectuosos entre 10 elegidos al azar en una caja.*

Si no cumple alguna de las condiciones establecidas, estamos ante un fenómeno o experimento **determinístico**. Son *ejemplos* de este tipo: *tirar una piedra al vacío y medir su aceleración*. Se caracteriza por que podemos prever su resultado, en contra de los fenómenos aleatorios.

Los fenómenos que estudia la estadística son los aleatorios.

Otros conceptos como población estadística, unidad estadística, muestra, tamaño muestral, son estudiados con más profundidad en Métodos Estadísticos.

2.-Variable estadística monodimensional: tipos

2.1.-Variable estadística. Definición y ejemplos.

Consideramos un experimento o muestra de una población cualquiera y realizamos 'n' pruebas o 'n' observaciones, de esta forma obtenemos un conjunto de observaciones que llamaremos muestra aleatoria de tamaño 'n'. Los valores o cualidades que representan los 'n' resultados de las 'n' pruebas realizadas le llamaremos **variable estadística**.

2.2.-Clasificación de las variables estadísticas: cualitativas y cuantitativas (discretas y continuas).

Hemos visto que un carácter estadístico es una propiedad que permite clasificar a los individuos de la población.

Hay dos tipos:

a) Caracteres estadísticos cuantitativos:

Se dice que un carácter estadístico es cuantitativo cuando sus modalidades son medibles (expresables como números y cumpliendo unas propiedades de medida.). *Ejemplos: peso, talla, pulso, edad, etc.*

b) Caracteres estadísticos cualitativos:

Se dice que un carácter estadístico es cualitativo cuando sus modalidades no pueden ser medidas. *Ejemplos: raza, sexo, profesión, estado civil, etc.*

Nota: Es evidente, por ejemplo, que si el carácter es el estado civil, podemos asignarle a sus modalidades los siguientes números: a los casados 1, solteros 0, viudos un 2, etc, pero este carácter no es medible en el sentido de que el $1 > 0$ por ejemplo, expresión que no tiene sentido.

Ejemplos:

La profesión es un carácter cualitativo.

Dentro de él podemos tener modalidades: profesor, peón, abogado, etc.

Lo anterior determina un **atributo** que puede ser observado pero no medido. Podemos contar el número de abogados o profesores, pero no medirlos. En cambio, un carácter cuantitativo determina una variable que llamaremos variable estadística. **Atributos**: se le suele llamar a las variables cualitativas.

La talla es un carácter cuantitativo. Es por lo tanto una variable estadística que podemos medir, puede tomar diversos valores: 1.60 , 1.62 ,, 1.92 ,etc .

Las variables estadísticas cuantitativas pueden ser: continuas o discretas.

Variable estadística

Discreta: es aquella que solo puede tomar un número finito o infinito numerable de valores. Dicho con otras palabras: cuando no puede tomar cualquier valor entre dos valores dados. O bien solo toma valores aislados, generalmente enteros.

Ejemplo: el número de libros en una estantería, las tiradas de un dado, el número de pétalos de una flor, etc.

Continua: cuando puede tomar, al menos teóricamente, todos los valores posibles dentro de un cierto intervalo de la recta real.

Ejemplo: la temperatura de los enfermos entre 35 y 40 grados, aunque en la práctica sea imposible medir temperaturas aproximando hasta la cuarta o quinta cifra decimal. En la práctica son variables estadísticas continuas aquellas que fijamos como suceso elemental las que entren en un intervalo.

VARIABLE ESTADÍSTICA)))1 *	+QCUANTITATIVAS)	+QDiscretas
			.QContinuas
			.QCUALITATIVAS (=atributos)

3.-Tablas De Frecuencias. Representaciones Graficas.

3.1.-Frecuencia absoluta y relativa. Frecuencias acumuladas

Nota sobre la notación con sumatorios: cuando tenemos una serie de sumas podemos utilizar el signo **E** para abreviar la notación:

$$\sum_{j=1}^{j=n} x_j = x_1 + x_2 + x_3 + x_4 + + x_n$$

Si realizamos un experimento o tenemos una muestra de de tamaño **n**, que tiene por variable estadística x_i y el valor de una de las variables es **n'**, o el suceso ha ocurrido **n'** veces, entonces:

Llamamos

frecuencia absoluta del valor x_i al número de veces que se repite dicho valor (**n'**)

$$\text{fr. abs } (x_i) = f_i = n'$$

frecuencia absoluta acumulada del valor x_i a la suma de las frecuencias absolutas de todos los valores anteriores a x_i más la fr. absoluta de x_i .

$$\text{fr.abs.acum}(x_i) = F_i = f_1 + f_2 + \dots + f_n = \sum_{h=1}^{i=n} f_h$$

Llamamos:

frecuencia relativa del valor x_i al cociente entre el número de veces que se repite x_i (frecuencia absoluta) y el número de pruebas realizadas (**n'/n**).

$$\text{Fr. rel } (x_i) = h_i = n'/n$$

frecuencia relativa acumulada del valor x_i a la suma de las frecuencias relativas de todos los valores anteriores a x_i más la fr. relativa de x_i .

$$\text{Fre.rel.acum.}(x_i) = H_i = h_1 + h_2 + \dots + h_n = \sum_{j=1}^{i=n} h_j$$

3.2.-Tabla de la distribución de frecuencias.

Llamamos

Distribución de frecuencias absolutas a la aplicación que asocia a cada valor de la variable estadística su frecuencia absoluta. Análogamente sería para frecuencias relativas.

Tablas estadísticas a una presentación en forma de tabla de la distribución de frecuencias absolutas, que suele ir acompañado de las frecuencias relativas. Este primer ejemplo es una tabla **estadística simple**.

Una tabla estadística simple es la siguiente:

Not x_i	F.abso. f_i	F.Relat. h_i
0	1	0,03
1	2	
2	1	
3	4	
4	6	
5	15	
6	4	
7	3	
8	2	
9	1	
10	1	
	40	1

Tabla estadísticas acumulativas

La tabla la podemos hacer con las frecuencias acumuladas, tanto relativas como absolutas

Var.	Fr. absolutas		Fr. Relativas	
Nota x_i	F.abso f_i	Acumlad F_i	F. rel. h_i	Acumula H_i
0	1	0,03	0,00	0,00
1	2			
2	1			
3	4			
4	6			
5	15			
6	4			
7	3			
8	2			
9	1			
10	1			
	40	40	1	1

Ejemplo 3.2.1: de variable aleatoria continua y con los intervalos de igual tamaño

Experimento:

Muchas personas experimentan reacciones alérgicas a las picaduras de insectos. Estas reacciones difieren de paciente a paciente, no solo en la gravedad sino también en el tiempo de aparición de la reacción. En 40 personas se han obtenido los siguientes resultados:

10'5 11'2 9'9 15'0 11'4 12'7 16'5 10'1

12'7 11'4 11'6 6'2 7'9 8'3 10'9 8'1
 3'8 10'5 11'7 8'4 12'5 11'2 9'1 10'4
 9'1 13'4 12'3 5'9 11'4 8'8 7'4 8'6
 13'6 14'7 11'5 11'5 10'9 9'8 12'9 9'9

Es una variable estadística continua, porque entre el mínimo tiempo de reacción en minutos (3'8) y el máximo (16'5) pueden darse todos los minutos. Como veremos más adelante, tomamos 6 intervalos y por redondeo comenzamos en 3'75 en intervalos de 2'2, hasta 16'95.

Var.	Fr. absolutas		Fr. Relativas	
Nota x_i	F.abso f_i	Acumlad F_i	F. rel. h_i	Acumula H_i
[3'75, 5'95)	2	0,05	0,00	0,00
[5'95, 8'15)	4			
[8'15, 10'35)	10			
[10'35, 12'55)	16			
[12'55, 14'75)	6			
[14'75, 16'95]	2			
	40		1	

¿Aproximadamente qué porcentaje de pacientes han experimentado una reacción cuando han transcurrido diez minutos?

¿En que intervalo se ha presentado la reacción en la mitad de los pacientes?.

¿Qué representa las frecuencias acumuladas?

3.3.-Representación gráfica de las frecuencias.

Aun cuando las tablas estadísticas que hemos visto encierran toda la información, a veces es conveniente traducir esta información mediante la construcción de gráficos con el fin de hacerlos más expresivos.

Los gráficos más habituales son los siguientes, que utilizaremos en un caso u otro, así como pueden hacerse con frecuencias absolutas o con frecuencias relativas

Diagramas de barras o bastones

Diagramas lineales

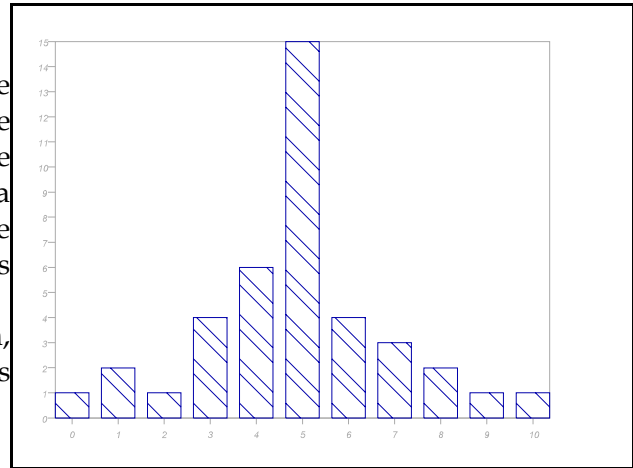
Poligono de frecuencias

Una combinación de los dos anteriores, de manera que se determinan poligonos (trapezios) con la altura de las frecuencias. Hoy con los ordenadores pueden presentarse con aspecto de tres dimensiones, como se ve en la figura.

Histogramas

Utilizado sobre todo para distribuciones de variable estadística continua, donde dividimos en intervalos generalmente de igual amplitud. Si hacemos de distinta amplitud hemos de cuidar en el diagrama que tengan la misma área los rectángulos determinados.

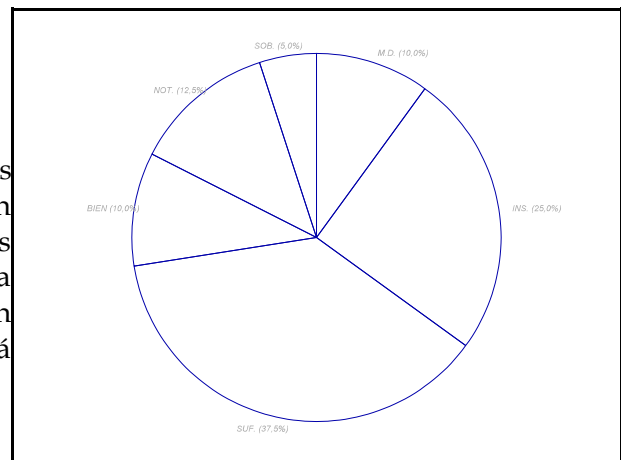
Si representa a una variable discreta, como es este caso, es conveniente que los rectángulos no estén 'pegados'.



Histograma

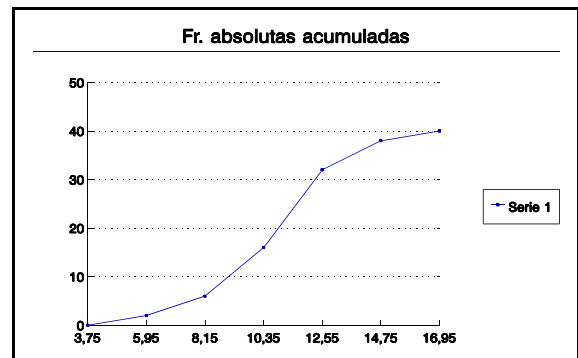
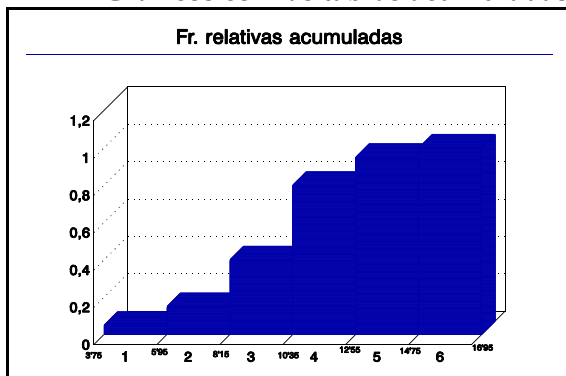
Diagrama de sectores

Consiste en representar, mediante sectores circulares, las distintas modalidades de un carácter, teniendo en cuenta que los sectores han de tener un ángulo central proporcional a la frecuencia absoluta correspondiente. En consecuencia, el área del sector circular será proporcional a la frecuencia absoluta.



D. de Sectores

Gráficos con las tablas acumuladas



4.-Medidas características

Tenemos la representación en forma de tabla de una distribución de frecuencias, hemos visto alguna de sus representaciones gráficas más características, pero todavía no es suficiente.

Por un lado las tablas pueden ser muy costosas para su interpretación y no resumen adecuadamente la información. Por otro lado, es difícil comparar dos distribuciones distintas.

Por otro lado con las gráficas pueden hacerse distorsiones y manipulaciones en:

- Alteración de las escalas.
- Inicio de las escalas
- Mantenimiento de la proporcionalidad de líneas.

Utilizaremos dos tipos de medidas, que llamaremos características.

=Unas de medidas son para medir los valores centrales (**medidas centrales**).

=Otras nos darán valores de cuan dispersos están los datos respecto de los valores centrales (**medidas de dispersión**)

=Y por último, otra para poder comparar distintas distribuciones entre sí.

MEDIDAS CARACTERÍSTICAS

MEDIDAS DE CENTRALIZACION	MEDIDAS DE DISPERSION
De tamaño : Media aritmética	Recorrido
De posición: Mediana.	Desviación Media
De frecuencia : Moda	Varianza
	Desviación Típica.

MEDIDA PARA COMPARAR DISTRIBUCIONES

Coeficiente de variación de Pearson

Objetivo perseguido con las medidas : resumir y sintetizar un conjunto de datos mediante un único número o unos pocos.

4.1.-Medidas de centralizacion

Se llaman **Medidas de centralización** a los valores que tienden a situarse en el centro del conjunto de datos ordenados respecto a su magnitud.

Las medidas centrales más importantes son:

Media aritmética, Mediana, Moda.

MEDIA ARITMÉTICA

Sea X una variable estadística discreta que toma los valores: $x_1, x_2, x_3, \dots, x_n$ con frecuencias absolutas $f_1, f_2, f_3, \dots, f_n$ se llama media aritmética o simplemente media:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \dots + x_n f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{h=1}^n x_h f_h}{\sum_{h=1}^n f_h}$$

se puede operar para transformar a través de las frecuencias relativas.

Expresión mediante sumatorios:

$$\bar{x} = \frac{\sum_{h=1}^n x_h f_h}{\sum_{h=1}^n f_h} = \frac{\sum_{h=1}^n x_h f_h}{N} = \sum_{h=1}^n x_h h_h$$

Ejemplo: en el de notas

$$\bar{x} = \frac{1 \cdot 0 + 2 \cdot 4 + \dots + 9 \cdot 4 + 10 \cdot 4}{40} = 4.875$$

Ventajas e inconvenientes de la Media Aritmética:

Ventajas:

- El cálculo se realiza con todos los valores de la variable.
- Tiene un cálculo sencillo, que aportan las calculadoras actuales.
- Su resultado es único.

Inconvenientes:

- Los efectos que sobre ella producen los valores extremos, que muchas veces son poco significativos por su rareza.

La insuficiencia de la media se ve el ejemplo siguiente:

Ejemplo :

Salarios de las 11 personas de una empresa.

Frecuencia :nº	Sueldos
2	50.000 pts/ mes 70.000
3	75.000 85.000
1	90.000 1.000.000
2	
2	
1	

Salario medio = 157.727 pts/ mes

Nótese lo engañoso del resultado.

Evidentemente este parámetro no es tan representativo como la media, pero es útil en muchas ocasiones. Por ejemplo cuando la moda se destaca preferentemente. En geografía puede ser la expresión de una estructura determinada, caracterizar una región, darnos de un clima dominante, etc. Por otro lado es el único valor central que puede calcularse en las series nominales.

Ejemplo :

En un grupo se procede a la elección del cargo X al que lleve el máximo número de votos:

La elección resulta así:

Personas a elegir:	Número de votos:
Juan Pérez	10
Maria López	20
Carmen Vazquez	5

Aquí en este ejemplo únicamente tendría sentido calcular la moda.

4.2.-Medidas de dispersión

Las medidas de dispersión más importantes son:

Recorrido, desviación media, varianza y desviación típica.

Las medidas centrales de una distribución nos hablan de como es por los valores medios, pudiendo haber dos distribuciones muy distintas que tengan valores medios similares. Queda pues la investigación incompleta, siendo necesario conocer en qué medida los datos numéricos están agrupados o no alrededor de los valores centrales

Ejemplo en donde el alumno puede observar la necesidad de las medidas de dispersión:

Al. <-----Notas ----->

x	9	2	3	0	4	8	2	8
y	6	5	6	4	4	5	4	2

Nota Media de X = 4.5 Y = 4.5

RECORRIDO

Se llama recorrido o rango de una distribución a la diferencia entre el mayor y el menor valor de la variable estadística.

Ejemplo:

En el ejemplo que seguimos desde el principio de notas el recorrido = 10 - 0 = 10

En el ejemplo de la última tabla, el recorrido de x es 9 el recorrido de y es 4 . Creemos que resulta significativo.

DESVIACIÓN MEDIA

Se llama desviación media de una distribución de frecuencias, y se representa por D_x a la media aritmética de las desviaciones respecto de la media tomadas en valor absoluto:

En el caso de que los valores estén repetidos y aparezcan sus frecuencias correspondientes, la expresión de la desviación media es la siguiente:

$$D_x = \frac{f_1|x_1 - \bar{x}| + \dots + f_n|x_n - \bar{x}|}{f_1 + \dots + f_n} = \frac{\sum_{h=1}^{h=n} f_h \cdot |x_h - \bar{x}|}{\sum_{h=1}^{h=n} f_h}$$

fórmula que usaremos en la construcción de las tablas

Ejemplo:

$$D_x = \frac{1 \cdot 0.487 + 2 \cdot 1.487 + \dots + 1 \cdot 9.487 + 1 \cdot 10.487}{40}$$

Inconvenientes de la Desviac. media :

sin lugar a dudas el uso de los valores absolutos (que complica bastante los cálculos)

VARIANZA

Se llama varianza de una distribución de frecuencias y se representa por F^2 , a la media aritmética de los cuadrados de las desviaciones respecto a la media.

En el caso de que los valores estén repetidos y aparezcan sus frecuencias correspondientes, la expresión de la varianza es la siguiente:

$$\sigma^2 = \frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \dots + f_n(x_n - \bar{x})^2}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{h=1}^{h=n} f_h \cdot (x_h - \bar{x})^2}{\sum_{h=1}^{h=n} f_h}$$

fórmula que usaremos en la construcción de las tablas

Consideraciones sobre la varianza:

1: La varianza es siempre un número positivo, por tratarse de la media aritmética de

números positivos.

2: Cuanto mayor es la dispersión le corresponde mayor varianza, y, en consecuencia, menor es la representatividad de los valores centrales.

3: La varianza depende de todos los valores de la variable.

La expresión de la varianza mediante operaciones obtenemos la siguiente, que resulta más fácil para el cálculo.

$$\sigma^2 = \frac{f_1x_1^2 + f_2x_2^2 + \dots + f_nx_n^2}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{h=1}^{h=n} f_h \cdot x_h^2}{\sum_{h=1}^{h=n} f_h} - \bar{x}^2$$

Ejemplo:

$$F^2 = \frac{1(40-4'87)^2 + 2(41-4'87)^2 + \dots + 1(49-4'87)^2 + 1(10-4'87)^2}{40} =$$

Inconvenientes de la varianza:

El inconveniente principal es que utiliza unas medidas distintas a las que tratamos en la variable. Al estar elevado al cuadrado perdemos referencia respecto a las variables.

DESVIACIÓN TÍPICA

Llamamos desviación típica de una distribución de frecuencias y representamos por **F** a la raíz cuadrada positiva de la varianza.

Sus expresiones según los casos es la siguiente:

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{f_1 + f_2 + \dots + f_n}} = \sqrt{\frac{\sum_{h=1}^{h=n} f_h \cdot (x_h - \bar{x})^2}{\sum_{h=1}^{h=n} f_h}}$$

fórmula que usaremos en la construcción de las tablas

$$\sigma = \sqrt{\frac{f_1x_1^2 + f_2x_2^2 + \dots + f_nx_n^2}{f_1 + f_2 + \dots + f_n}} = \sqrt{\frac{\sum_{h=1}^{h=n} f_h \cdot x_h^2}{n} - \bar{x}^2}$$

Ejemplo: $S = \sqrt{S^2} = 2.01$

Otras ventajas de la desviación típica:

Es interesante saber que así como para calcular la desviación típica hemos elegido la

media, si se tomase otro valor, como por ejemplo la moda, la mediana , o un valor m cualquiera puede demostrarse que la media aritmética es el valor que hace mínima la expresión. Dicho con otras palabras , de todas "las posibles desviaciones típicas escogidas" **F** es la mínima.

En el estudio de la estadística inferencial, veremos también la importancia que tiene la desviación típica en las distribuciones normales.

COEFICIENTE DE VARIACION

El coeficiente de variación (de Pearson) $C_v = \text{des.tip} / \text{media}$

$$C_v = \frac{\text{des.tip}}{\text{media}} = \frac{\sigma}{\bar{x}}$$

No tiene unidades y se utiliza para comparar distribuciones con distintas medidas. Por ejemplo tallas y pesos. Suele expresarse en %. También se utiliza cuando al comparar dos distribuciones sobre la misma variable están medidas en distintas unidades, por ejemplo en m y Km. En definitiva, que nos mide la dispersión relativa de una distribución.

Ejemplo: con unas notas

$$\text{notas 1: } \bar{x} = 4'5 \quad F = 3'16 \quad ==> C_v = \frac{3'16}{4'5} = 0'70 \rightarrow 70\%$$

$$\text{notas 2: } \bar{x} = 4'5 \quad F = 1'2 \quad ==> C_v = \frac{1'2}{4'5} = 0'26 \rightarrow 26\%$$

Ventajas

Permite comparar distribuciones distintas, incluso con medidas distintas.

Desventajas

Deja de ser representativa y no debe utilizarse cuando la media de una de las distribuciones sea muy baja.