ESTADÍSTICA

La mayoría de las veces se entiende por estadística los conjuntos de datos distribuidos en tablas y gráficos que aparecen en los periódicos. Ahora bien en la actualidad se entiende como estadística un método de toma de decisiones.

La estadística se puede dividir en dos partes:

- Estadística descriptiva o deductiva.
- Estadística inferencial o inductiva.

La **estadística descriptiva** trata del recuento, ordenación y clasificación de los datos obtenidos por las observaciones. Se construyen **tablas** y se representan **gráficos** que permiten simplificar la distribución y se calculan **parámetros estadísticos** que caracterizan la distribución.

La **estadística inferencial** plantea y resuelve el problema de establecer previsiones y conclusiones generales sobre una población a partir de los resultados obtenidos de una muestra.

POBLACIÓN Y MUESTRA

Supongamos que queremos analizar la estatura de los alumnos de primero de bachillerato de una determinada provincia.

El conjunto formado por todos los alumnos matriculados en dicho curso se llama población , y un subconjunto formado por los alumnos que contestan al formulario sería una muestra.

En general, se llama **POBLACIÓN** al conjunto de todos los elementos que cumplen una determinada característica. Los elementos de la población se llaman individuos.

Se llama **MUESTRA** a cualquier subconjunto de la población. El número de elementos de la muestra se llama tamaño de la muestra.

Tendremos que exigir que la muestra sea representativa de la población.

El proceso mediante el cual se extrae una muestra se llama **MUESTREO ALEATORIO** y en dicho proceso cada individuo de la población tiene que tener la misma probabilidad de ser incluido en la muestra. La muestra así obtenida se llama **MUESTRA ALEATORIA**.

Ejemplos:

1. Si queremos hacer un estudio sobre las preferencias musicales de los jóvenes de entre 15 y 18 años de un cierto país, población será el conjunto de todos los jóvenes de esa edad que haya en el país y muestra será el grupo de jóvenes que escojamos para hacer la encuesta.
2. Si hacemos una encuesta para conocer la intención de voto de los habitantes de un país población será el conjunto de todos los habitantes del país con derecho a voto y muestra será el conjunto de las personas a las que preguntemos.

CARACTERES Y MODALIDADES

Se llama **carácter estadístico** a una propiedad que permite clasificar a los individuos de una población. Hay de dos tipos:

- **Caracteres estadísticos cuantitativos**: son aquellos que se pueden medir, por ejemplo el peso de un individuo, la longitud de una pieza de tela, el sueldo de los obreros de una fábrica, el cociente intelectual de un alumno...
- **Caracteres estadísticos cualitativos**: son aquellos que no se pueden medir por ejemplo la profesión de una persona, el color de pelo, la carrera que piensa estudiar un alumno de segundo de bachillerato, el estado civil...

Se llaman **modalidades** de un carácter estadístico a cada una de las diferencias que se pueden establecer dentro de un mismo carácter estadístico cualitativo. Por ejemplo, modalidades del carácter estadístico "color de pelo" serían rubio, moreno, castaño,...

VARIABLE ESTADÍSTICA

Si tratamos con un carácter estadístico cuantitativo, por ejemplo " el peso de los individuos de una población", dicho carácter tomará distintos valores 65 Kg., 73 Kg., 52'3 Kg.,... El conjunto de estos valores de llama ***VARIABLE ESTADÍSTICA***.

En este curso, dividiremos las variables estadísticas en dos tipos: discretas y continuas.

* **Variable estadística discreta**: cuando puede tomar un número finito de valores o infinito numerable.

- Número de hijos de una familia.
- Número de asignaturas suspendidas por un alumno.
- Número de goles marcados por un equipo de fútbol.
- Número de libros vendidos por una librería en un día.

* **Variable estadística continua**: cuando puede tomar (al menos teóricamente) todos los valores posibles dentro de un intervalo de la recta real.

- Talla de los individuos.
- Temperaturas registradas en un observatorio.
- Litros de agua por metro cuadrado caidos en un observatorio en un día.

Los valores de las variables estadísticas se acostumbran a representar por $x_1, x_2, x_3, ..., x_n, ...$

FRECUENCIAS ABSOLUTAS Y RELATIVAS

Consideremos un ejemplo: un profesor tiene anotadas en su cuaderno las notas de 30 alumnos de una clase. Son las siguientes:

5, 3, 4, 1, 2, 8, 9, 7, 6, 8,
6, 7, 9, 8, 7, 7, 1, 0, 1, 5,
9, 9, 8, 0, 8, 8, 8, 9, 5, 7.

Se trata de una variable estadística cuantitativa discreta que puede tomar los valores $x_1 = 0, x_2 = 1, x_3 = 2, ..., x_{10} = 9$.

Se llama ***frecuencia absoluta*** de un valor EMBED Equation.3 , y se representa por EMBED Equation.3 , al número de veces que se repite dicho valor.

La suma de las frecuencias absolutas es el tamaño de la muestra.

Se llama ***frecuencia absoluta acumulada*** del valor EMBED Equation.3 , y se representa por EMBED Equation.3 , a la suma de las frecuencias absolutas de todos los valores anteriores a EMBED Equation.3 más la frecuencia absoluta de EMBED Equation.3 .

EMBED Equation.3

En el ejemplo anterior:

$f_1 = 2,\ f_2 = 3,\ f_3 = 1, ...$

$F_1 = 2,\ F_2 = 5,\ F_3 = 6,\ ...$

Se llama **frecuencia relativa** de un valor EMBED Equation.3 , y se representa por EMBED Equation.3 , al cociente entre la frecuencia absoluta de EMBED Equation.3 , y el número total de datos.

EMBED Equation.3

Se llama **frecuencia relativa acumulada** del valor EMBED Equation.3 , y se representa por EMBED Equation.3 , al cociente entre la frecuencia absoluta acumulada de EMBED Equation.3 y el número total de datos.

EMBED Equation.3

En el ejemplo anterior:

$h_1 = \frac{2}{30} = \frac{1}{15} = 0'0\hat{6},\ h_2 = \frac{3}{30} = \frac{1}{10} = 0'1,\ h_3 = \frac{1}{30} = 0'0\hat{3}, ...$

$H_1 = \frac{1}{15},\ H_2 = \frac{1}{6},\ H_3 = \frac{1}{5},\ ...$

TRATAMIENTO DE LA INFORMACIÓN. TABLAS ESTADÍSTICAS

A continuación vamos a estudiar cómo debemos proceder ordenadamente para analizar una muestra:

1. *Recogida de datos*.
2. *Ordenación de los datos*: en orden creciente o decreciente.
3. *Recuento de frecuencias*.
4. *Agrupación de los datos*: Si la variable aleatoria es continua, o bien es discreta pero con un gran número de valores es aconsejable agrupar los datos en **CLASES** (intervalos).
   Las clases deben tener la misma amplitud o tamaño.
   A los puntos medios de cada clase se les llama **MARCA DE CLASE**.
5. *Construcción de una tabla estadística*.

En el ejemplo de las notas de los treinta alumnos:

x_i	f_i	F_i	h_i	H_i
0	2	2	0,06666667	0,06666667
1	3	5	0,1	0,16666667
2	1	6	0,03333333	0,2
3	1	7	0,03333333	0,23333333
4	1	8	0,03333333	0,26666667
5	3	11	0,1	0,36666667
6	2	13	0,06666667	0,43333333
7	5	18	0,16666667	0,6
8	7	25	0,23333333	0,83333333
9	5	30	0,16666667	1
	30		1	

En el siguiente ejemplo se muestra como agrupar los datos en clases. No existe un criterio general que nos diga cuál es el número idóneo de clases que debemos escoger a la hora de agrupar. Con carácter muy general podemos enunciar uno de los criterios más sencillos, el de Norcliffe, que establece que el número de clases debe ser aproximadamente igual a la raíz cuadrada positiva del número de datos.

Ejemplo: Se han recogido los siguientes datos sobre el número de personas que acuden a una consulta médica diariamente a lo largo de 36 días:

3, 2, 11, 13, 4, 3, 2, 4, 5, 6, 7, 3,
4, 5, 3, 2, 5, 6, 27, 15, 4, 21, 12, 4,
3, 6, 29, 13, 6, 17, 6, 13, 6, 5, 12, 26.

CLASES	Marca de clase	f_i	F_i	h_i	H_i
[0, 5)	2,5	13	13	0,36111111	0,36111111
[5 , 10)	7,5	11	24	0,30555556	0,66666667
[10 , 15)	12,5	6	30	0,16666667	0,83333333
[15 , 20)	17,5	2	32	0,05555556	0,88888889
[20, 25)	22,5	1	33	0,02777778	0,91666667
[25 , 30)	27,5	3	36	0,08333333	1
		36		1	

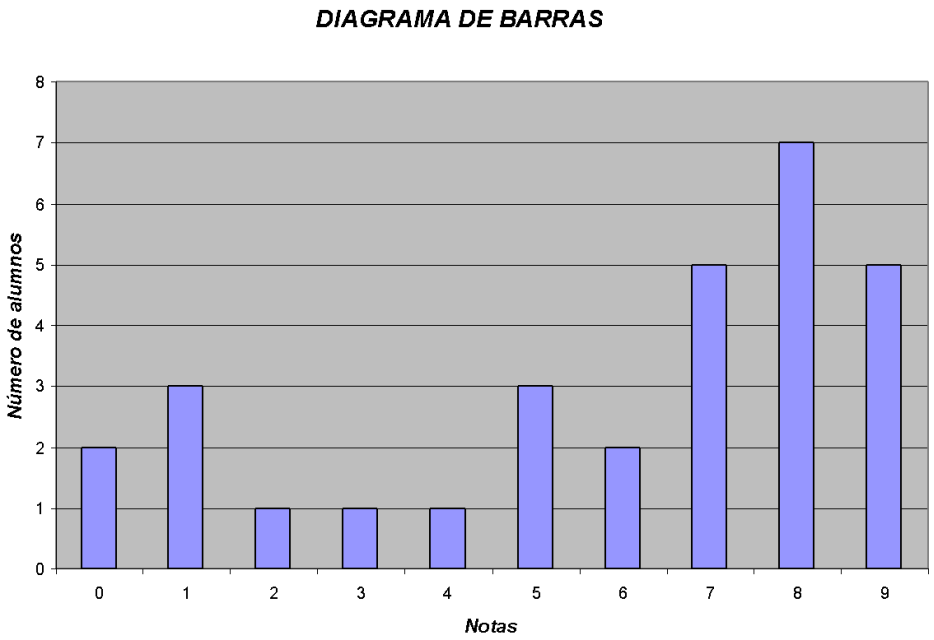
REPRESENTACIONES GRÁFICAS

Aún cuando las tablas estadísticas contienen toda la información, a veces es conveniente expresarla mediante un gráfico, con el fin de hacerla más clara y evidente. Según sea la naturaleza del carácter estudiado, utilizaremos uno u otro tipo de representación gráfica.

Diagrama de barras

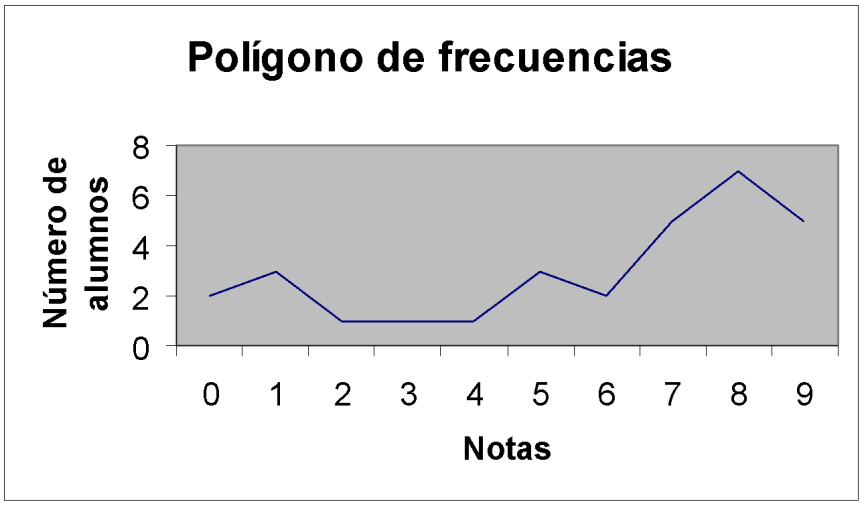
Para trazarlos se representan sobre el eje de abcisas los valores de la variable y sobre el eje de ordenadas las frecuencias absolutas o relativas, según proceda. A

continuación se levantan trazos gruesos o barras, de longitud igual a la frecuencia correspondiente. En el ejemplo de las notas del apartado anterior:



Polígono de frecuencias

Los polígonos de frecuencias se forman uniendo los extremos de las barras mediante una línea quebrada.



Histograma

Se utilizan generalmente para distribuciones de variable estadística continua, o bien para distribuciones de variable estadística discreta , con un gran número de datos que se han agrupado en clases.

Para construir el histograma se representan sobre el eje de abcisas los límites de las clases. Sobre dicho eje se construyen unos rectángulos que tienen por base la amplitud del intervalo y por altura la frecuencia absoluta de cada intervalo siempre que todos los intervalos tengan la misma amplitud. En caso contrario, las alturas de los rectángulos han de ser calculadas teniendo en cuenta que sus áreas deben ser proporcionales a las frecuencias de cada intervalo.

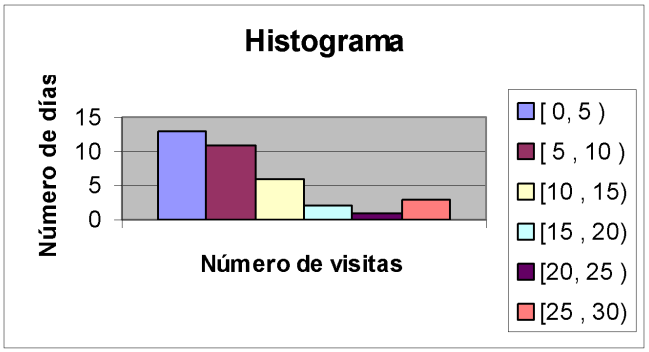
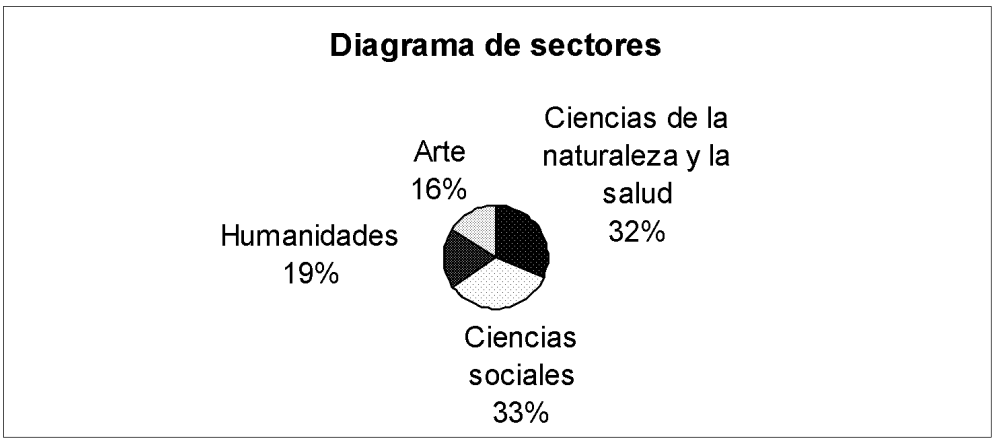


Diagrama de sectores

Los diagramas de sectores representan las distintas modalidades de un carácter mediante sectores circulares. El ángulo central de cada sector ha de ser proporcional a la frecuencia absoluta correspondiente; en consecuencia, el área del sector circular será proporcional a la frecuencia absoluta.

Ejemplo: se ha hecho una encuesta entre los alumnos de 4º de la E.S.O. sobre qué modalidad de bachillerato piensan estudiar.



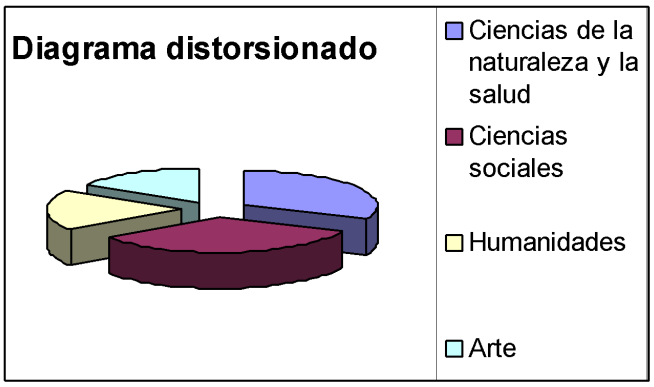
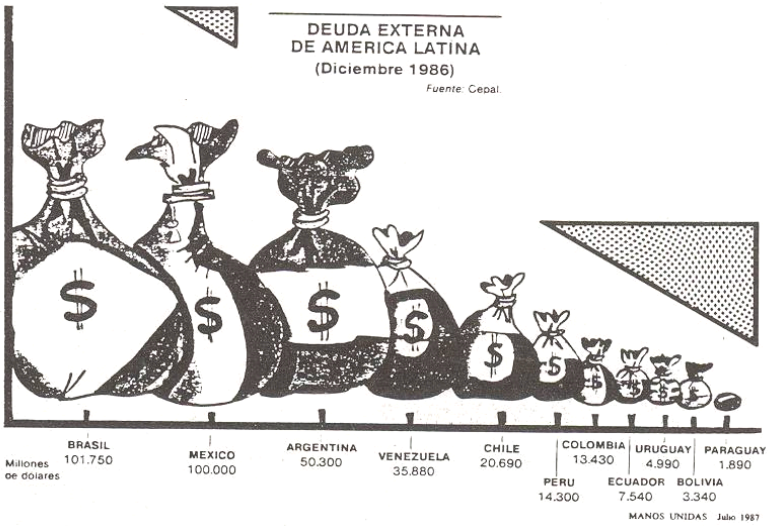


Diagrama distorsionado
Ciencias de la naturaleza y la salud
Ciencias sociales
Humanidades
Arte

Pictogramas

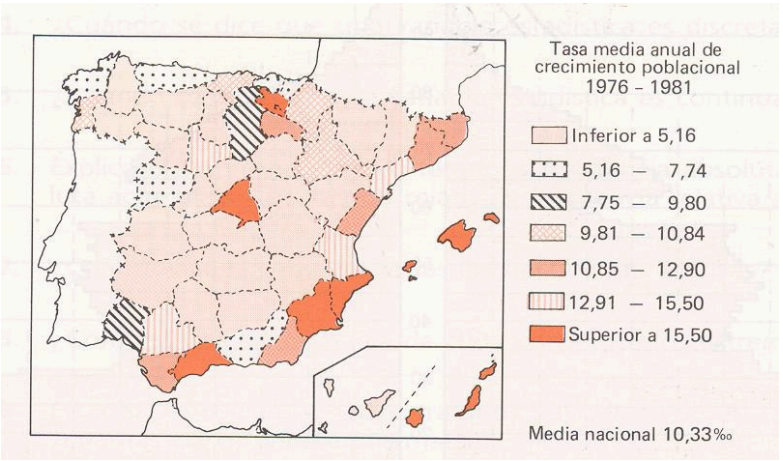
Son dibujos alusivos a la distribución que se pretende estudiar y que mediante su forma, tamaño, etc., ofrecen una descripción lo más expresiva posible de la distribución estadística.



Cartogramas

Se llama cartogramas a los gráficos que se realizan sobre un mapa, señalando sobre determinadas zonas con distintos colores o rayados lo que se trate de poner de manifiesto.

Por ejemplo, se suelen utilizar estos tipos de diagramas para representar la densidad demográfica de una nación, la renta per cápita , las horas de sol anuales, los índices de lluvia,....

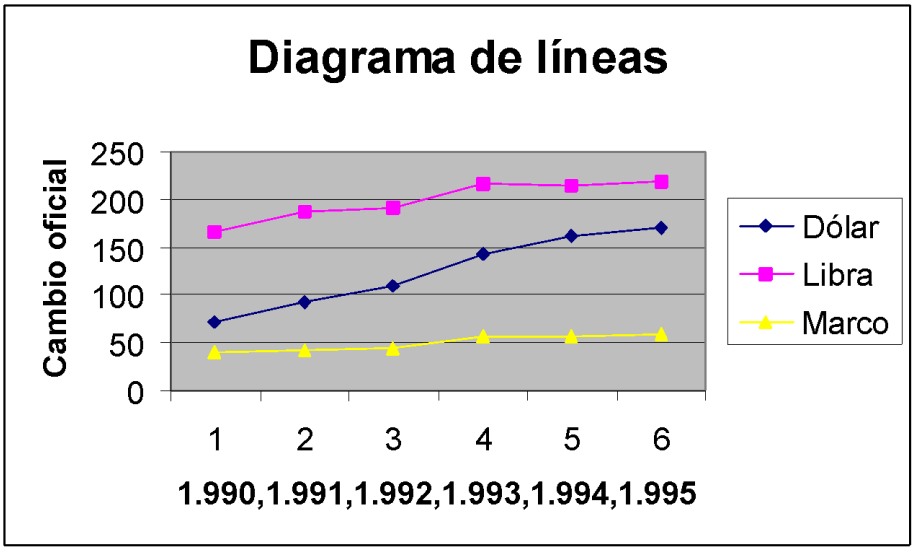


Diagramas lineales

Son muy utilizados para mostrar las fluctuaciones de un determinado carácter estadístico con el paso del tiempo.

Con frecuencia se aprovecha para representar sobre la misma escala varios diagramas lineales. Por ejemplo ingresos y gastos, nacimientos y defunciones...

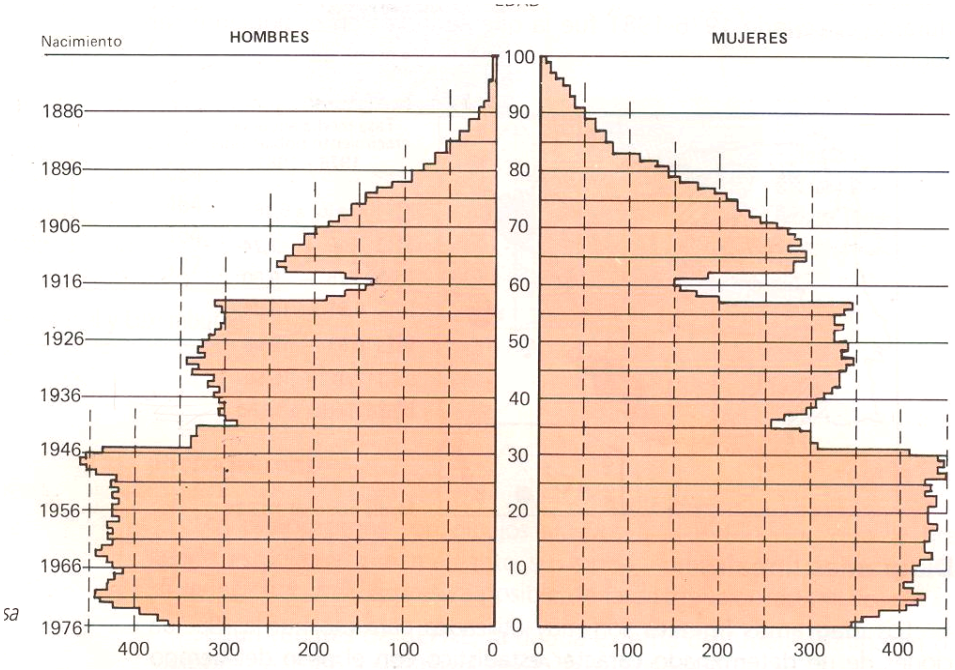
En el siguiente gráfico se muestran los cambios medios anuales para el dólar, la libra esterlina y el marco alemán en el periodo 1.990-1.995:



Pirámides de población

Las pirámides de población se utilizan para estudiar conjuntamente la variable edad y el atributo sexo.

La gráfica se obtiene representando en la ordenada el grupo de edad, y en la abcisa el sexo. Para la modalidad mujer se toma el semieje positivo y para la modalidad hombre el semieje negativo.



Pirámide de población francesa en 1.977

DISTRIBUCIONES UNIDIMENSIONALES.
CÁLCULO DE PARÁMETROS.

Medidas de centralización

Se llama medidas de centralización a las medidas o parámetros que, tienden a situarse hacia el centro del conjunto de datos ordenados.

Las más importantes son: media, moda, mediana, cuartiles, deciles y percentiles.

Media

> Se llama media de una variable estadística a la media aritmética de todos los datos, es decir a la suma de todos los valores de la variable dividida por el número de valores.
>
> La media se representa por **EMBED Equation.3** .

Para calcular la media:

Sea X una variable estadística que toma los valores $x_1, x_2, x_3, ..., x_n$, con frecuencias absolutas $f_1, f_2, f_3, ..., f_n$, respectivamente, la media viene dada por:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + ... + x_n f_n}{f_1 + f_2 + ... + f_n} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i}$$

Si la variable es continua, o aún siendo discreta si están los datos agrupados en clases, se toman como valores $x_1, x_2, x_3, ..., x_n$, las marcas de clase.

Ejemplos:

1.Las calificaciones en la asignatura historia del arte de los 40 alumnos de una clase viene dada por la siguiente tabla

Calificaciones	1	2	3	4	5	6	7	8	9
Núm. de alumnos	2	2	4	5	8	9	3	4	3

Hallar la media.

En la práctica, los cálculos se disponen de la siguiente forma:

x_i	f_i	$x_i \cdot f_i$
1	2	2
2	2	4
3	4	12
4	5	20
5	8	40
6	9	54
7	3	21
8	4	32
9	3	27
	40	212

$$\bar{x} = \frac{212}{40} = 5{,}3$$

2.Se ha aplicado un test sobre satisfacción en el trabajo a 88 empleados de una fábrica, obteniéndose los siguientes resultados:

Puntuaciones	[38,44)	[44,50)	[50,56)	[56,62)	[62,68)	[68,74)	[74,80)
Nº de trabajadores	7	8	15	25	18	9	6

Se completa la tabla estadística calculando la marca de clase:

Clases	Marca	f_i	$x_i \cdot f_i$
[38,44)	41	7	287
[44,50)	47	8	376
[50,56)	53	15	795
[56,62)	59	25	1475
[62,68)	65	18	1170
[68,74)	71	9	639
[74,80)	77	6	462
		88	5204

$$\bar{x} = \frac{5204}{88} = 59{,}1$$

La media es el parámetro de centralización más utilizado.

Tiene en cuenta todos los datos y es fácil de calcular.

Su inconveniente es que los datos extremos y poco significativos distorsionan su valor.

No siempre se puede calcular; si los datos son cualitativos o están agrupados en clases siendo una de ellas abierta como por ejemplo mayores de 18 años.

Moda

Se llama moda de una variable estadística al valor de dicha variable que presenta mayor frecuencia absoluta.

La moda se representa por **Mo**

Como consecuencia de su definición, el calculo de la moda es muy sencillo en el caso de variables discretas con los datos sin agrupar. Ahora bien, en el caso de datos agrupados en intervalos, es fácil determinar la clase modal (clase con mayor frecuencia), pero el valor dentro del intervalo que se presume tenga mayor frecuencia se obtiene a partir de la siguiente expresión:

$$M_o = L_i + c \cdot \frac{D_1}{D_1 + D_2}$$

L_i = límite inferior de la clase modal.

c = amplitud de los intervalos.

D_1 = diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase anterior.

D_2 = diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase siguiente.

En el ejemplo del "test de satisfacción en el trabajo" sería:

$M_o = 56 + 6 \cdot \frac{10}{10+7} = 59'5$ Este es el valor que, teóricamente se supone tiene mayor frecuencia.

Mediana

> Se llama mediana de una variable estadística al valor de dicha variable tal que el número de observaciones menores que él es igual al número de observaciones mayores.
>
> La mediana se representa por **M.**

Cálculo de la mediana:

Si la variable es discreta, la mediana es el primer valor de la variable cuya frecuencia absoluta acumulada es mayor que la mitad del número de datos. En el caso de que la mitad del número de datos coincida con la frecuencia acumulada de un valor, la mediana será la semisuma de ese valor y el siguiente.

x_i	f_i	F_i	
0	2	2	
1	3	5	
2	1	6	
3	1	7	
4	1	8	M = 7
5	3	11	
6	2	13	
7	5	18	
8	7	25	
9	5	30	
	30		

Otro ejemplo:

x_i	f_i	F_i
3	15	15
6	20	35
7	15	50
8	40	90
9	10	100
	100	

$M = \frac{7+8}{2} = 7'5$

Si la variable es continua o es discreta pero tiene los datos agrupados, se busca primero la *clase mediana* (donde se alcanzan la mitad de los datos), pero para obtener el valor concreto de la variable que deja a su izquierda igual número de datos que a su derecha, aplicaremos la siguiente expresión:

$$M = L_i + c \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i}$$

L_i = límite inferior de la clase mediana

c = amplitud del intervalo

N = número total de datos

F_{i-1} = frecuencia absoluta acumulada de la clase anterior a la mediana

f_i = frecuencia absoluta de la clase mediana

Ejemplo: En el "test de satisfacción en el trabajo"

Clases	Marca	f_i	F_i
[38,44)	41	7	7
[44,50)	47	8	15
[50,56)	53	15	30
[56,62)	59	25	55
[62,68)	65	18	73
[68,74)	71	9	82
[74,80)	77	6	88
		88	

$$M = 56 + 6 \cdot \frac{44 - 30}{25} = 59'36$$

Como consecuencia de la definición de mediana, el 50% de los datos son menores o iguales que ella y el 50% de los datos son mayores o iguales.

En las variables que se pueden representar con un histograma, la mediana es el valor de la variable tal que la vertical levantada sobre el mismo divide el histograma en dos partes de igual área.

Cuantiles

La mediana divide a la distribución en dos partes iguales, los cuantiles son parámetros que dividen los datos de la distribución en partes iguales.

Los más usados son:

Cuartiles:

Se llaman cuartiles a tres valores que dividen a la serie de datos en cuatro partes iguales.

Q_1 , Q_2 y Q_3 (cuartil primero, cuartil segundo y cuartil tercero)

Quintiles:

Se llaman quintiles a cuatro valores que dividen a la serie en cinco partes iguales.

K_1 , K_2 , K_3 y K_4 (quintil primero,...)

Deciles:

Nueve valores iguales que dividen la distribución en 10 partes iguales.

D_1 , D_2 , ... y D_9 (decil primero,...)

Percentiles:

Noventa y nueve valores que dividen la serie en 100 partes iguales.

P_1 , P_2 , ... y P_{99} (percentil primero,...)

El cálculo es análogo al de la mediana.

Medidas de dispersión

Consideremos el siguiente ejemplo:

Se ha aplicado a dos grupos de ocho alumnos de 2º de la E.S.O. un test de 100 preguntas sobre capacidad numérica, obteniéndose los siguientes resultados:

Grupo A	Grupo B
46	10
48	18
49	30
50	50
50	50
51	70
52	82
54	90

Si calculamos la media, la mediana y la moda de ambas distribuciones, observaremos que todas son iguales a 50. Sin embargo, los dos grupos de alumnos son bien distintos. Las puntuaciones del grupo A están muy concentradas, poco dispersas; en cambio, las del grupo B se encuentran poco concentradas respecto a la media y diremos que se encuentran muy dispersas.

Así pues la investigación acerca de una distribución queda incompleta si sólo se estudian las medidas de centralización, siendo imprescindible conocer si los datos numéricos están agrupados o no respecto a los valores centrales. A esto se le llama dispersión y los parámetros que miden esta desviaciones respecto a la media se les llama medidas de dispersión o parámetros de dispersión.

Las medidas de dispersión más importantes son: el **recorrido**, la **varianza** y la **desviación típica**.

Rango o recorrido

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

En el ejemplo anterior:

Recorrido grupo A = 54-46=8

Recorrido grupo B = 90-10 = 80

Cuanto menor es el recorrido, mayor es la representatividad de los valores centrales. Son parámetros más estables el rango intercuartílico y el rango entre percentiles ($P_{90} - P_{10}$) .

Varianza y desviación típica

Se llama desviaciones respecto a la media a las diferencias entre cada valor de la variable y la media.

$x_1 - \bar{x}$, $x_2 - \bar{x}$, $x_3 - \bar{x}$, ... , $x_n - \bar{x}$

Se llama **varianza** de una variable a la media aritmética de los cuadrados de las desviaciones respecto a la media.

Se llama **desviación típica** de una variable a la raíz cuadrada positiva de la varianza.

La varianza se representa por s^2 , y la desviación típica se representa por s.

La varianza viene dada por la fórmula:

$$s^2 = \frac{(x_1 - \bar{x})^2 \cdot f_1 + (x_2 - \bar{x})^2 \cdot f_2 + ... + (x_n - \bar{x})^2 \cdot f_n}{f_1 + f_2 + ... + f_n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1}^{n} f_i}$$

Con frecuencia, se simplifican los cálculos utilizando la siguiente expresión:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{\sum_{i=1}^{n} f_i} - \bar{x}^2$$

La desviación típica viene dada por la raíz cuadrada positiva de dicha expresión.

Ejemplo: Calculamos la varianza y la desviación típica en el ejemplo de las calificaciones de 40 alumnos:

x_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
1	2	2	2
2	2	4	8
3	4	12	36
4	5	20	80
5	8	40	200
6	9	54	324
7	3	21	147
8	4	32	256
9	3	27	243
	40	212	1296

$$\bar{x} = \frac{212}{40} = 5'3$$

$$s^2 = \frac{1296}{40} - (5'3)^2 = 4'31$$

$$s = \sqrt{4'31} = 2'08$$

Utilización conjunta de la media y la desviación típica

La media, se encuentra aproximadamente hacia el centro de la distribución. La desviación típica informa sobre la dispersión de los datos respecto a la media.

En distribuciones unimodales, simétricas o ligeramente asimétricas suele cumplirse que:

1. En el intervalo $(\bar{x}-s, \bar{x}+s)$ se encuentran el 68% de los datos.
2. En el intervalo $(\bar{x}-2s, \bar{x}+2s)$ se encuentran el 95% de los datos.
3. En el intervalo $(\bar{x}-3s, \bar{x}+3s)$ se encuentran el 98% de los datos.

Comparación de puntuaciones. Puntuaciones típicas

Sea X una variable estadística que toma los valores $x_1, x_2, x_3, ..., x_n$ y sean $\bar{x}$ y s respectivamente la media y la desviación típica de dicha variable. Se llaman puntuaciones típicas de la variable X a los valores:

$$z_1 = \frac{x_1 - \bar{x}}{s}\ ,\ z_2 = \frac{x_2 - \bar{x}}{s}\ ,\ \ldots\ ,\ z_n = \frac{x_n - \bar{x}}{s}$$

Las puntuaciones típicas son muy utilizadas en las ciencias sociales y se usan para comparar las puntuaciones obtenidas en distintas distribuciones.

Ejemplo: El señor López y el señor Pérez van a pasar un examen físico. El grupo de hombres de la edad, altura y complexión del Sr. López tiene un peso medio de 77 Kg. Y una desviación típica de 6 Kg., y el grupo del Sr. Pérez tiene un peso medio de 91'5 Kg. Y una desviación típica de 8 Kg.. Si el Sr. López pesa 88 Kg. Y el Sr. Pérez pesa 106 Kg., ¿ cuál de ellos es más grueso en relación con su grupo?.

$$z_l = \frac{88-77}{6} = 1'83$$

$$z_p = \frac{106-91'5}{8} = 1'81$$

Es pues más grueso en relación a su grupo el señor López.

Por último, El **coeficiente de Variación de Pearson** es:

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