ESTADÍSTICA DESCRIPTIVA

Introducción

La estadística es una rama de las matemáticas que trata de la recogida, ordenación, análisis y presentación adecuada de datos recogidos sobre cierta población (no necesariamente humana) con el fin de extraer conclusiones a partir de ellos.

Así por ejemplo, podría interesarnos hacer un estudio sobre la estatura de los alumnos del instituto; sobre el peso de los recién nacidos en España; sobre las marcas de leche más vendidas; sobre los estudios que piensan abordar en el futuro los estudiantes de bachillerato; sobre la producción de maíz en los diferentes países europeos; etc.

La estadística trata de obtener resultados globales, busca las características generales de un colectivo y prescinde de las particulares de cada individuo.

La estadística se divide tradicionalmente en dos ramas:

Estadística descriptiva o deductiva de "describir" y analizar algunas características de los elementos de un grupo dado con el fin de describir dicho grupo, sin extraer conclusiones para un grupo mayor. No hace uso del cálculo de probabilidades.

Estadística inferencial o inductiva: Trabaja con *muestras* y mediante sus técnicas se obtienen conclusiones y/o previsiones para toda la *población* a partir de los resultados de la *muestra*. Es decir se "infieren" características de toda la *población* a partir de los resultados obtenidos en sólo una parte de ella. Se utilizan resultados obtenidos mediante estadística descriptiva y se hace uso del cálculo de probabilidades. Es la rama más interesante y tiene infinidad de aplicaciones, aunque se debe proceder con mucha cautela, y aspectos tales como el tamaño y la forma de elección de la *muestra* son fundamentales para la fiabilidad de las conclusiones que se obtengan.

Supongamos que queremos estudiar la talla de los alumnos de 1º de bachillerato de nuestro centro. En tal caso, mediríamos a todos los alumnos de dicho nivel y, con los datos

obtenidos, elaboraríamos gráficos y tablas y hallaríamos parámetros que describiesen y resumiesen la información obtenida. Esto es estadística descriptiva.

Imaginemos ahora que deseásemos estudiar la talla de todos los españoles de 16 años. Podríamos medirlos a todos, pero esto sería muy costoso y nos llevaría mucho tiempo. Lo que se hace es escoger a una parte (muestra) de, por ejemplo, 1.000 individuos, medirlos a todos, analizar los datos obtenidos (mediante estadística descriptiva) y a partir de aquí obtener conclusiones para el total. Esto es estadística inferencial. Es un proceso delicado, piénsese por ejemplo cómo se debe elegir a los 1.000 individuos. Deberían representar lo más fielmente posible al conjunto total. La composición de la muestra debe estar en proporción con la composición del total de la población. Tendríamos en cuenta el sexo, lugar de residencia (población urbana y rural, población de las distintas comunidades autónomas, etc.). Y, ¿por qué 1.000?, ¿por qué no 2.000? ¿ó 750?. (Algunas de estas cuestiones se tratan en 2º de bachillerato)

Como cualquier disciplina, la estadística tiene una terminología que debemos conocer, en este caso viene heredada de sus orígenes, que fueron trabajos de tipo demográfico (tablas de mortalidad y censos):

Población (o universo): es el conjunto de todos los elementos cuyo conocimiento nos interesa, es decir, el conjunto de personas, animales u objetos que se desea estudiar.

Muestra: es un subconjunto, extraído de la población, cuyo estudio sirve para deducir las características de toda la población.

La conveniencia o necesidad de trabajar con muestras se ilustra en el ejemplo de un párrafo anterior. Otros supuestos en los que esto es así son por ejemplo: estudio sobre la duración de las bombillas de cierta marca, estudio sobre los efectos de un nuevo medicamento para una determinada enfermedad, estudio sobre la inversión en ropa interior de los habitantes de los distintos países europeos, etc. (piénsese en cada caso por qué es conveniente elegir muestras).

Individuo: Es cada uno de los elementos de una población o de una muestra.

Tamaño: Es el número de individuos que componen una población o muestra. Se denota con N.

Caracteres: son las propiedades, cualidades o características de los individuos que se desea estudiar.(talla, peso, color de ojos, estado civil, número de hijos, etc).

Los caracteres estadísticos se clasifican en:

- Cuantitativos :Se denominan variables estadísticas (v.e) son los que toman valores numéricos (presión sanguínea, peso, número de hijos, etc.).Medibles y/o contables.
- Cualitativos: se denominan atributos, son los que no toman valores numéricos (estado civil, profesión, color de ojos, etc.)

A su vez las variables estadísticas pueden ser:

Discretas: las que toman valores puntuales, concretos (número de hijos, número de empleados de una fábrica, etc.)

Continuas: las que, al menos teóricamente, pueden tomar cualquier valor dentro de ciertos intervalos (temperatura, talla, longitud de tornillos, etc.)

Por otra parte los caracteres cualitativos se dividen en:

Ordinales: si sus posibles valores admiten un orden implícito (opinión que merece un futbolista: pésimo, malo, regular, bueno, excelente)

Nominales: si no admiten ningún orden (carreras universitarias elegidas por los estudiantes).

En general, se llaman modalidades de un caracter a cada una de sus diferentes opciones. Así son modalidades del atributo profesión: economista, chapista, psicólogo, panadero, etc.

Es decir: $Carácter \begin{cases} \text{Cuantitativo} = \text{Variable Estadística} \begin{cases} \text{Discreta} \\ \text{Contínua} \end{cases} \\ \text{Cualitativo} = \text{Atributo} \begin{cases} \text{Ordinal} \\ \text{Nominal} \end{cases} \end{cases}$

NOTA:

En matemáticas y en general en otras muchas disciplinas se usa el símbolo $\sum$ (letra griega sigma mayúscula) para designar de forma abreviada una suma. Así, por ejemplo:

$1+2+3+4+5+6+7+8=\sum_{i=1}^{8} i \quad \sum_{i=1}^{8} i$ se lee *"sumatorio de i desde i=1 hasta 8"*

$x_1+x_2+x_3+x_4=\sum_{k=1}^{4} x_k \quad \sum_{k=1}^{4} x_k$ se lee *"sumatorio de x sub k desde k=1 hasta 4"*

$\sum_{i=1}^{n} x_i = x_1+x_2+x_3+x_4+\ldots+x_n \qquad 2^3+2^4+2^5+2^6=\sum_{n=3}^{6} 2^n$

ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL

Como ya se ha dicho, trata de "describir" y analizar algunas características de los elementos de un grupo dado con el fin de describir dicho grupo. Se elaboran tablas y se representan gráficos que permiten simplificar en gran medida la complejidad de los datos. Se calculan parámetros que resumen la información obtenida.

Ilustraremos los distintos conceptos con un ejemplo (de aquí en adelante **Ejemplo A**):

El número de suspensos de cada uno de los alumnos de bachillerato en la primera evaluación fue:

4	0	2	0	6	0	5	1	4	7	5	3	2	3	1	5	3	0	4	0	3	2	4
2	5	1	0	6	3	0	3	0	4	0	3	1	7	0	1	6	2	4	2	6	2	0
1	0	4	0	1	1	4	3	1	4	0	5	1	3	4	2	1	0	2	4	1	2	6
6	7	1	3	5	1	3	4	2	0	4	2	2	0	5	2	6	3	1	2	4	7	0

La lista de los valores así dispuestos no permite observar lo que ocurre con la variable cuantitativa discreta

X=*"nº de suspensos de los alumnos de Bachillerato en la primera evaluación"*

Si hacemos un recuento y los ordenamos de menor a mayor obtenemos la siguiente tabla:

x_i	f_i
0	18
1	15
2	15
3	12
4	14
5	7
6	7
7	4
	92

El tamaño de la población es 92. Se denota por N=92

$\mathbf{x_i}$ son los distintos valores que toma la variable X.

En nuestro caso $x_1=0$, $x_2=1$, $x_3=2$, etc.

Se llama **frecuencia absoluta** del valor x_i al número de veces que se repite dicho valor en las N observaciones. Se representa por $\mathbf{f_i}$.

Evidentemente las sumas de todas las f_i tiene que ser N.

$$f_1 + f_2 + f_3 + ... + f_n = \sum_{i=1}^{n} f_i = N$$

Así por ejemplo f_5 será las veces que aparece x_5 en el total de las N observaciones. En nuestro caso se dan 7 casos con 5 suspensos.

Se llama **frecuencia absoluta acumulada** del valor x_i a la suma de las frecuencias absolutas de todos los valores menores o iguales que x_i. Se representa por $\mathbf{F_i}$

Es decir $F_i = f_1 + f_2 + ... + f_i$ o lo que lo mismo $F_i = \sum_{k=1}^{i} f_k$.

La última frecuencia absoluta acumulada coincide con el tamaño: $F_n = N$

La frecuencia absoluta no es suficiente para reflejar la intensidad con que se repite un valor. Por ejemplo, decir que un valor se repite 3 de cada 100 veces no es lo mismo que decir que se repite 3 de cada 1.000 veces. Por esto para determinar si un valor es muy frecuente o no, es mejor utilizar la frecuencia relativa.

Se llama **frecuencia relativa** de un valor x_i y la representamos por $\mathbf{h_i}$, al cociente de la frecuencia absoluta de x_i y el número total de observaciones. Es decir $h_i = \frac{f_i}{N}$.

Obviamente la suma de todas las h_i es 1. $\sum_{i=1}^{n} h_i = 1$

Se llama **frecuencia relativa acumulada** del valor x_i y la representamos por $\mathbf{H_i}$, al cociente entre la frecuencia absoluta acumulada de x_i y el total de datos que intervienen (N). Es decir $H_i = \frac{F_i}{N}$.

La última frecuencia relativa acumulada vale 1.

Las frecuencias relativas y relativas acumuladas se pueden expresar en forma de fracción, en forma decimal y en %.. (para expresarla en forma de porcentaje se hace $h_i \cdot 100$).

Con todo lo anterior la tabla quedaría así:

x_i	f_i	F_i	h_i	H_i
0	18	18	0,196	0,196
1	15	33	0,163	0,359
2	15	48	0,163	0,522
3	12	60	0,130	0,652
4	14	74	0,152	0,804
5	7	81	0,076	0,880
6	7	88	0,076	0,957
7	4	92	0,043	1,000
	92		1,000	

Aún cuando las tablas estadísticas contienen toda la información, es conveniente expresarla mediante un gráfico con el fin de hacerla más clara:

Diagramas de barras: Se representan en el eje de abscisas los valores de la variable, y sobre el eje de ordenadas las frecuencias absolutas o relativas, según proceda. A continuación, por los puntos marcados en el eje de abscisas se levantan trazos gruesos (o barras finas), de longitud igual a la frecuencia correspondiente.

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

Siguiendo con nuestro ejemplo tendríamos:

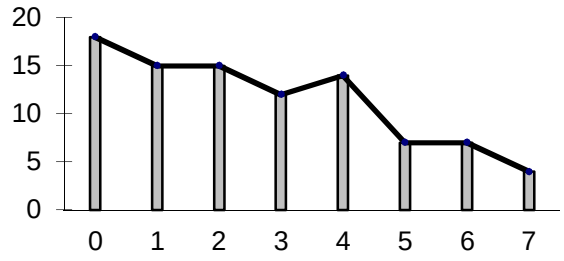


Diagrama de barras y polígono de frecuencias absolutas.

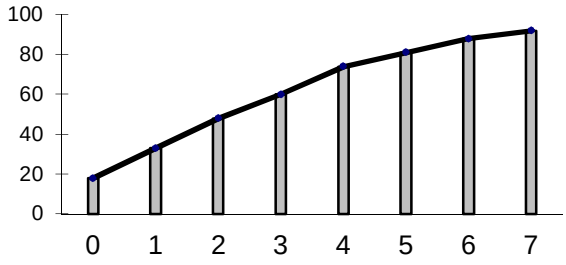


Diagrama de barras y polígono de frecuencias absolutas acumuladas.

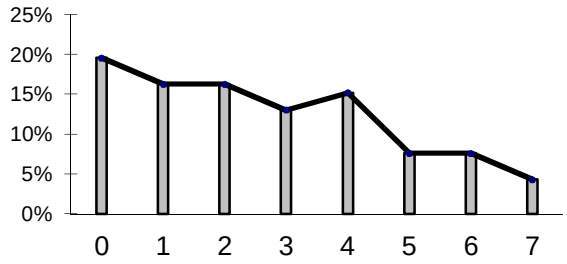


Diagrama de barras y polígono de frecuencias relativas (formato de %).

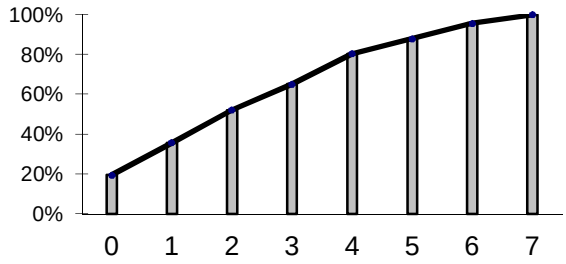


Diagrama de barras y polígono de frecuencias relativas acumuladas (en formato de %).

Se observa que los gráficos de las frecuencias absolutas y absolutas acumuladas son idénticos respectivamente a los de frecuencias relativas y relativas acumuladas (salvo por la escala vertical). Por tanto bastará con representar dos de los cuatro gráficos.

En nuestro **Ejemplo A** tratamos con una variable estadística discreta que toma pocos valores (de 0 a 9 suspensos). Sin embargo la situación no es siempre tan simple. Imaginemos que deseamos estudiar el peso de los alumnos del centro. Los pesamos a todos y obtenemos con toda seguridad varias decenas de valores distintos para la variable peso. Si en estas condiciones repitiéramos el proceso seguido en el **Ejemplo A** deberíamos manejar una tabla con varias decenas de filas, lo cual es muy engorroso.

Es fácil pensar en situaciones aún peores con tablas de miles de filas. Es por esto, que cuando se trata con variables continuas (o discretas con gran número de valores distintos) es muy útil agrupar los datos en intervalos y determinar el número de individuos pertenecientes a cada intervalo.

Ilustraremos esta situación con otro ejemplo (de aquí en adelante **Ejemplo B**):

Se ha hecho un estudio sobre el retraso (en minutos) de 120 trenes de cierta línea de largo recorrido obteniéndose los siguientes resultados

26	5	11	16	3	20	11	19	0	15	14	2	25	9	28	2	27	13	5	16
12	7	12	2	17	8	10	21	6	31	11	8	12	22	0	19	3	18	6	22
11	8	20	5	16	13	32	0	11	10	7	18	14	6	11	6	11	18	9	19
17	0	10	11	6	22	1	19	26	8	33	14	4	17	12	15	1	24	9	17
23	5	15	3	16	7	8	15	13	16	9	10	23	8	12	4	21	10	0	28
13	9	5	14	11	9	14	15	1	13	2	16	16	8	29	1	17	34	10	25

Aparecen todos los valores entre 0 y 34 salvo el 30. En una tabla normal (como la del Ejemplo A) esto daría lugar a 34 filas. En lugar de eso agrupamos los datos en intervalos de tal forma que la tabla quedaría así:

I.C.	c_i	f_i	F_i	h_i	H_i
[0,5)	2,5	18	18	0,150	0,150
[5,10)	7,5	26	44	0,217	0,367
[10,15)	12,5	30	74	0,250	0,617
[15,20)	17,5	24	98	0,200	0,817
[20,25)	22,5	10	108	0,083	0,900
[25,30)	27,5	8	116	0,067	0,967
[30,35)	32,5	4	120	0,033	1,000
		120		1,000	

Los intervalos en los que se agrupan los valores de la variable reciben el nombre de **Intervalos de clase**, los puntos medios de dichos intervalos (los c_i) se denominan **Marcas de clase**; el significado y nombre de f_i, F_i, h_i y H_i no varían.

En general, cuando se trata de una variable estadística es continua, la información obtenida no se puede valorar, de hecho, en un punto o en un instante dado, sino que ha de ser valorada en un intervalo de espacio o de tiempo.

En este caso es útil agrupar los datos en intervalos y determinar el número de individuos pertenecientes a cada intervalo.

Procederemos así:

1. Ordenamos los datos de menor a mayor.
2. Dividimos el recorrido (diferencia entre el valor mayor y menor de la variable) en una serie de intervalos, generalmente de la misma amplitud, aunque no siempre es posible. Se consigue con esto una clasificación de los datos; por eso se denominan intervalos de clase. El punto medio de cada intervalo se denomina marca de clase y se denota por c_i.

   Los intervalos son cerrados por la izquierda y abiertos por la derecha. No se deben de ser menos de 6 y sobrepasar los 15 intervalos.

   Y una aproximación será $\text{Número de intervalos} = \sqrt{\text{número de datos}}$

3. Recuento. Se cuenta el número de valores de la variable comprendidos en cada intervalo, logrando así las frecuencias absolutas.
4. Creamos la tabla siguiente:

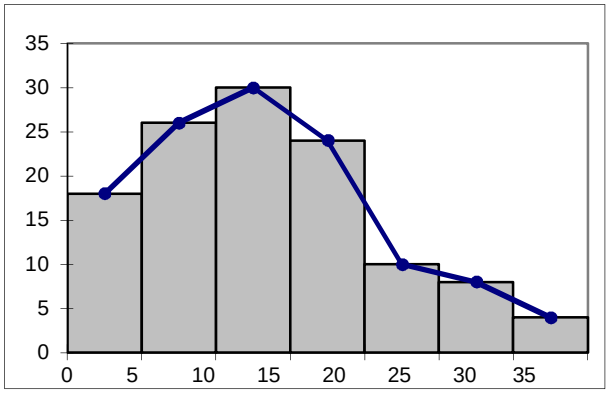
Intervalos de clase	**Marcas de clase c_i**	**Amplitudes**	**f_i**	**F_i**	**h_i**	**H_i**
$[L_0 , L_1)$	c_1	a_1	f_1	F_1	h_1	H_1
$[L_1 , L_2)$	c_2	a_2	f_2	F_2	h_2	H_2
...	...	...	...	...	...	...
$[L_{k-1} , L_k)$	c_k	a_k	f_k	$F_k=N$	h_k	$H_k=1$
Total			N		1	

Si tenemos un número muy grande de datos para variables discretas, se pueden agrupar como las variables continuas, trabajando luego con ella como si lo fuese.

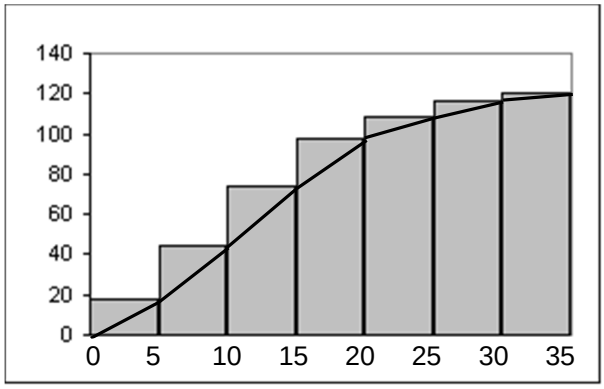
Histogramas y polígonos de frecuencias: A la hora de representar gráficamente la información que proporciona una tabla de datos agrupados en intervalos recurriremos a los **histogramas** y a los **polígonos de frecuencias.**

En un histograma se utilizan rectángulos de tal forma que las bases sean las amplitudes de los intervalos de clase y las alturas las frecuencias de cada intervalo. (Más tarde veremos que en realidad lo que debe ocurrir es que las áreas de los rectángulos sean iguales o proporcionales a las frecuencias).

Así, por ejemplo el histograma y el polígono de frecuencias absolutas correspondiente al Ejemplo 2 sería:



El histograma y el polígono de frecuencias absolutas acumuladas quedaría así:



IMPORTANTE: Obsérvese que el polígono de frecuencias absolutas (o relativas) se obtiene uniendo los puntos correspondientes a las marcas de clase, mientras que el polígono de frecuencias absolutas acumuladas (o relativas acumuladas) se obtiene uniendo el extremo inferior izquierdo con el extremo superior derecho del rectángulo que se añade con respecto al anterior.

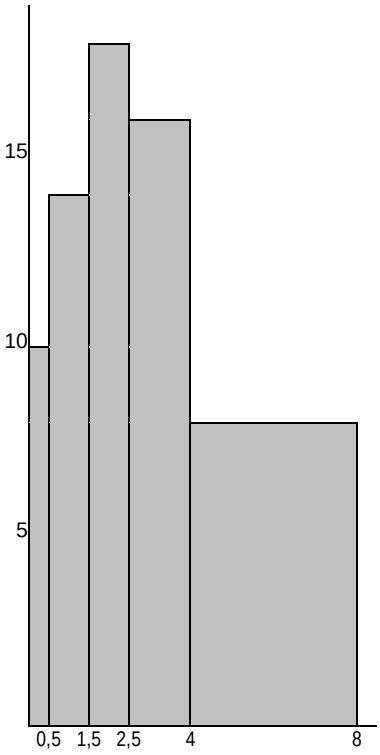
Un caso particular de tablas de datos agrupados en intervalos se presenta cuando dichos intervalos no son todos de la misma longitud. Consideremos el siguiente ejemplo:

Al preguntar a un grupo de personas cuánto tiempo dedicaron a ver la televisión durante un fin de semana se obtuvieron los siguientes resultados:

	c_i	a_i	f_i	F_i	h_i	H_i
[0; 0,5)	0,25	0,5	10	10	0,152	0,152
[0,5; 1,5)	1	1	14	24	0,212	0,364
[1,5; 2;5)	2	1	18	42	0,273	0,636
[2,5; 4)	3,25	1,5	16	58	0,242	0,879
[4,8)	6	4	8	66	0,121	1,000
			66		1,000	

Hemos insertado una nueva columna (a_i) que contiene la amplitud de cada intervalo.

Si construimos el histograma de (por ejemplo) frecuencias absolutas resulta:



Uno de los objetivos de un gráfico estadístico (quizá el fundamental) es proporcionar una visión clara y fiel de la situación representada. Si alguien, desconociendo los datos, viese el gráfico anterior sacaría conclusiones erróneas. Por ejemplo si se fijase en la primera y en la última columna pensaría que hay muchas más personas en el intervalo [4,8) que en el intervalo [0; 0,5); O que hay más personas en [4,8) que en [1,5; 2,5), etc

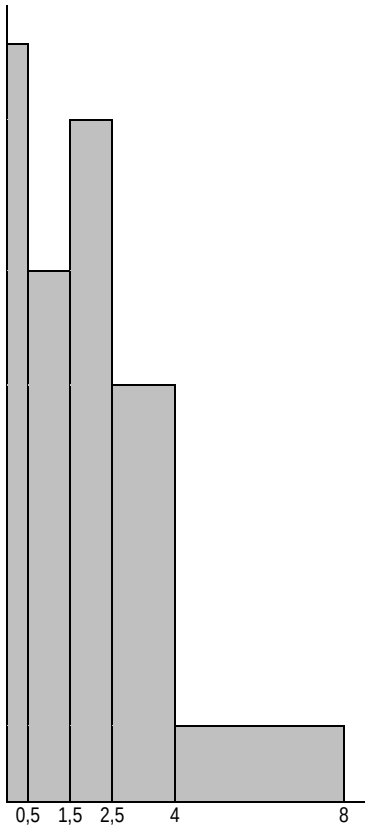
Ya se ha comentado que en un histograma las áreas (no las alturas) de los rectángulos deben ser iguales o proporcionales a las frecuencias. Cuando los intervalos son todos de la misma longitud no se da este problema pero en nuestro caso debemos hacer algunos cambios:

Se construyen las frecuencias medias que son las $\frac{f_i}{a_i}$ que nos darán las alturas de los rectángulos de tal manera que, ahora sí, las áreas de los rectángulos son iguales a sus frecuencias.

Y se tendría :

	c_i	a_i	f_i	f_i/a_i **Altura**
[0; 0,5)	0,25	0,5	10	*20*
[0,5; 1,5)	1	1	14	*14*
[1,5; 2;5)	2	1	18	*18*
[2,5; 4)	3,25	1,5	16	*10,7*
[4,8)	6	4	8	*2*

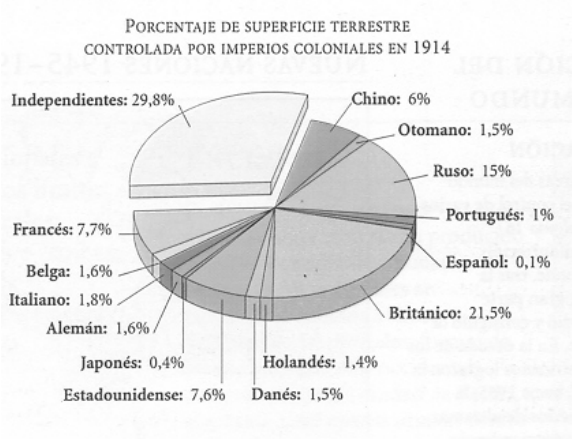
El histograma correcto sería:



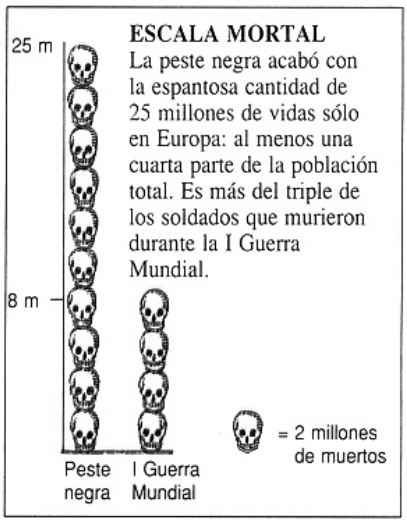
El proceso sería análogo en el caso de frecuencias absolutas acumuladas, frecuencias relativas y relativas acumuladas.

Para terminar este apartado mencionaremos que, además de los vistos, existen otros muchos tipos de gráficos estadísticos. Algunos de los más comunes son:

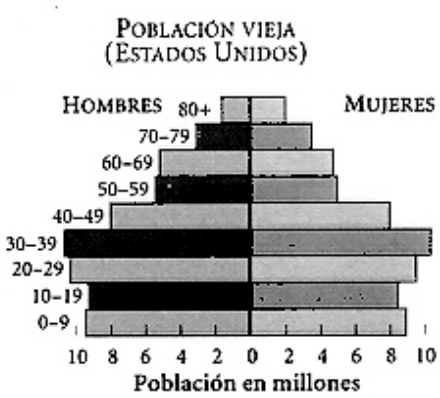
Diagrama de sectores (indicado para caracteres cualitativos):



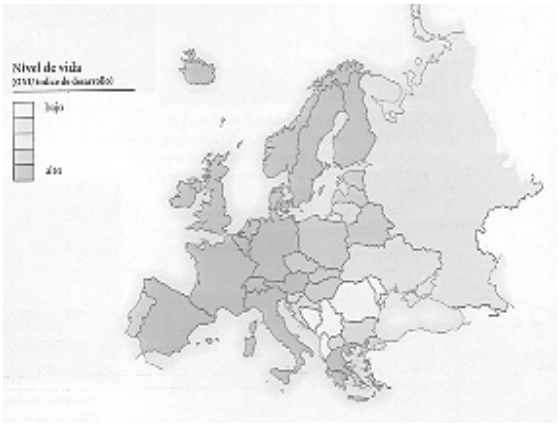
Pictograma (gráfico con dibujos alusivos):



Pirámides de población:



Cartogramas (gráficos sobre mapas donde se pinta cada zona de un color o con un rayado para indicar densidades demográficas, renta per cápita, etc.):



PARÁMETROS ESTADÍSTICOS

Aunque la observación visual de cualquier representación gráfica de una distribución de frecuencias proporciona una primera aproximación al análisis de los datos, se hace necesario estudiar procedimientos numéricos para obtener, a partir de los datos de la distribución, unos valores que permitan obtener una información cuantitativa.

La idea de resumir en unos pocos datos la información del comportamiento global del fenómeno estudiado, se realiza calculando algunos parámetros. Estos se clasifican en (solamente se nombran los que vamos a estudiar):

Medidas de centralización (o de posición): Son las que se sitúan hacia el centro o una parte cualquiera previamente determinada de la distribución. A su vez se clasifican en:

Medidas de posición centrales: **media aritmética, moda y mediana**

Medidas de posición no centrales: **cuartiles, quintiles, deciles y percentiles**

Medidas de dispersión: Son las que miden el grado de dispersión (y de alejamiento del centro) de los elementos de la distribución: **recorrido, recorrido intercuartílico, desviación media absoluta, desviación típica, varianza.**

Medidas de centralización

Media aritmética

Se llama media aritmética de una variable estadística X a la suma de todos los valores de dicha variable dividido por el número total de valores.

La media aritmética de X se representa por $\overline{x}$.

Si X es 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 fórmula para obtener la media aritmética es:

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

Si la variable X es continua, o aun siendo discreta, y por tratarse de muchos datos se encuentran agrupados en clases, se toman como valores x_1, x_2, x_3, ..., x_n las marcas de clase.

A la vista de la fórmula de $\bar{x}$, añadiremos a nuestras tablas una nueva columna: $x_i \cdot f_i$

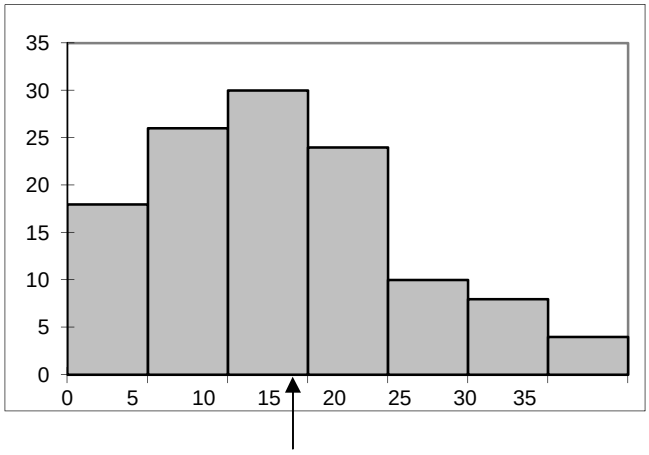
La media aritmética es el parámetro de centralización más usado.

Conviene tener en cuenta las siguientes cuestiones:

1) Tiene en cuenta todos los datos de la distribución. Aunque no tiene por qué coincidir con alguno. Esto tiene el inconveniente, de que si la distribución presenta valores extremos, excepcionalmente raros y poco significativos, éstos producen una distorsión sobre el valor de la media, alterando el significado matemático de ésta.

2) Puede expresarse en las mismas unidades que la variable estudiada. Lo cuál es muy significativo.
3) Es el centro de gravedad de la distribución y por tanto es única para cada distribución.

Si recordamos el histograma de frecuencias absolutas (también servirían las relativas, aunque no las acumuladas):



Si el histograma fuese un sólido (por ejemplo una plancha de madera) la media aritmética sería el lugar en el que la plancha de madera apoyada sobre él permanecería en equilibrio.
Este no en un método de cálculo de la media, pero puede servir para hacernos una idea de su valor o para descubrir errores “gruesos”.

“ La suma de las desviaciones de los valores de la variable respecto a la media es cero

$$\sum (x_i - \bar{x}) f_i = 0"$$

Es una consecuencia inmediata de la definición:

$$\sum (x_i - \bar{x}) f_i = \sum (x_i f_i - \bar{x} f_i) = \sum x_i f_i - \bar{x} \sum f_i = N\bar{x} - \bar{x}N = 0$$

4) No siempre se puede realizar el cálculo de la media:

- Si los datos de la distribución no son cuantitativos sino cualitativos. Solo tiene sentido para datos cuantitativos.

- Si los datos están agrupados en clases, estando alguna de ellas abierta. Por ejemplo, en una encuesta sobre lectores de la prensa diaria, se obtuvo la siguiente distribución:

Grupos de edad	Núm. de personas
Menores de 18 años	264
Entre 18 y 40 años	1376
Entre 40 y 60 años	825
Mayores de 60 años	341

En estos casos en los que no es posible calcular la media, se utilizan otros parámetros, como la moda y la mediana.

5) Si se suma una constante a todos los valores x_i, la media aumenta en la misma constante.

6) Si se multiplican todos los valores x_i por un mismo número, la media resulta también multiplicada por ese número

Moda

Se llama moda de una variable estadística al valor con mayor frecuencia absoluta. Se denota por Mo.

La moda no tiene por que ser única, puede haber varios valores de la variable con la mayor frecuencia. En este caso se llamarán distribuciones unimodales, bimodales, trimodales, ... dependiendo de que tengan una, dos, tres, modas.

El cálculo de la moda resulta sencillo en el caso de una variable estadística discreta. Con datos no agrupados, sólo se necesita observar las frecuencias absolutas en la tabla estadística, y ver a que valor corresponde la mayor frecuencia.

Si la variable estadística es continua, o discreta con datos agrupados en intervalos, es fácil encontrar el intervalo o clase en el que se encuentra la moda; dicho intervalo se denomina **clase modal**. Una vez determinada la clase modal para calcular la moda se aplica las fórmula siguiente:

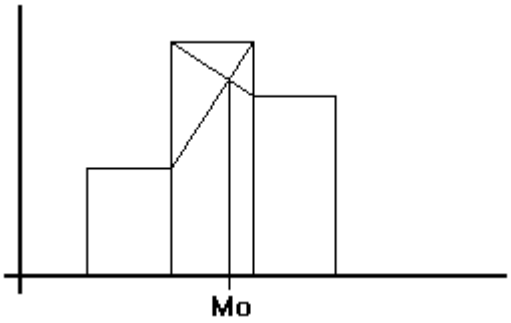
$$Mo = L_i + a_i \cdot \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})}$$

donde: L_i es el límite inferior de la clase modal

a_i es la longitud de la clase modal

f_i ; f_{i-1} ; f_{i+1} son las frecuencias de la clase modal y de las clases anterior y posterior, respectivamente.

Esta fórmula se obtiene del cálculo de la moda mediante el método gráfico. Para ello se representa el histograma de frecuencias absolutas. Seguidamente se unen los extremos de la clase modal con los contiguos como en el diagrama adjunto. La moda viene dada por la abscisa del punto de corte.



Hay que recordar que si los intervalos en los que está dividida la variable no son todos de la misma longitud debe tenerse especial cuidado a la hora de realizar los histogramas.

Observaciones:

1) Puede ocurrir que existan distribuciones que no tengan moda; esto ocurre cuando las frecuencias de todos los datos son iguales.
2) La moda es menos representativa que otros parámetros, pero en ocasiones es más útil, por ejemplo cuando se trata de datos cualitativos.
3) En la moda no intervienen todos los datos de la distribución.
4) Aún siendo un parámetro de centralización, es frecuente encontrar la moda próxima a los extremos de la distribución.

Mediana

Es el valor de la variable estadística, tal que el número de observaciones menores que él es igual al número de observaciones mayores que él. Es decir, está situada de modo que antes que ella está el 50% de la población y, detrás, el otro 50%. Se denota por M o Me.

En el caso de una variable estadística discreta, si se considera la serie íntegra de los datos ordenados de menor a mayor, poniendo .tantas copias de cada dato como indica su frecuencia, la mediana es el valor de la variable que está en la posición central en esa ordenación.

Si tenemos un número impar de datos, el cálculo de la mediana es sencillo: será el valor que ocupe el lugar central. Por ejemplo la mediana de la serie 3,4,5,5,6,7,8 es 5.

Si el número de datos es par, se toma como mediana la media aritmética de los dos valores que ocupan los lugares centrales. Por ejemplo la mediana de 3,4,5,5,6,7,8,9 es 5,5.

En la práctica, para el cálculo de la mediana, tenemos que tener en cuenta las frecuencias absolutas acumuladas, pues F_i nos da los lugares que ocupan el valor x_i en la distribución. Si queremos calcular la mediana nos tendremos que fijar en la columna de las F_i, y buscar el valor de la variable estadística que tenga asociada una frecuencia acumulada que supere, exactamente, el 50% (es decir $N/2$)

Ejemplo: La siguiente tabla se corresponde con el número de caramelos que comen en un sábado un grupo de 20 niños:

Número de caramelos	2	5	7	8	10
Niños	2	8	4	4	2

x_i	f_i	F_i
2	2	2
5	8	10
7	4	14
8	4	18
10	2	20
	20	

Como tenemos un número par de datos (20) la mediana será la media aritmética de los que ocupen los dos lugares centrales (lugares10 y 11).
Si nos fijamos en la columna F_i el lugar 10 lo ocupa x_2=5 y el lugar 11 x_3=7, por lo tanto la mediana será la media de 5 y 7, es decir, 6.
Esto quiere decir que hay un 50% de los 20 niños que comen menos de 6 caramelos y otro 50% que comen más de 6.

En el caso de una variable estadística continua, o una discreta agrupada en intervalos, hablaremos del intervalo o clase mediana (clase donde se alcanza la mitad de los datos).

Para calcular el valor concreto de la mediana aplicaremos la fórmula:

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

donde: L_i es el límite inferior de la clase mediana

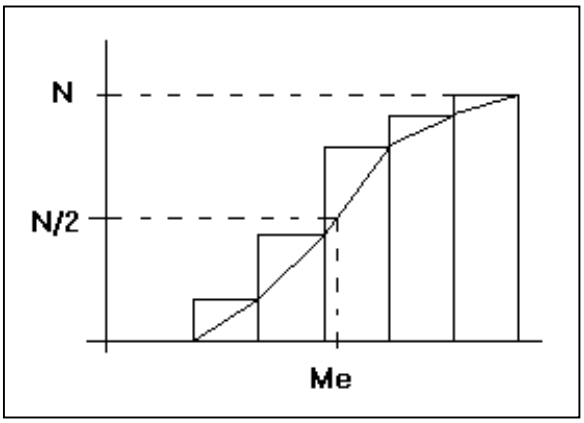
a_i es la longitud de la clase modal

N número total de datos

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

f_i frecuencia absoluta de la clase mediana

Al igual que ocurría con la moda, esta fórmula se obtiene del cálculo gráfico de la mediana.



El cálculo gráfico de la mediana se realiza considerando el polígono de frecuencias absolutas acumuladas, y sobre él se representa una recta paralela al eje de abscisas que pase por $N/2$, para posteriormente calcular la abscisa correspondiente al punto de corte de las dos curvas, que nos da el valor de la mediana.

Observaciones:

1) No influyen en ella los valores extremos de la distribución. Solamente los datos centrales y sus frecuencias. Por lo tanto es especialmente útil cuando tengamos una distribución con algunos valores extremos excesivamente exagerados.
2) Depende del orden y número de los datos, pero no de su valor.
3) Geométricamente, y para distribuciones que se puedan representar mediante un histograma de frecuencias absolutas (o relativas), la mediana es el valor de la variable, tal que la vertical levantada sobre el mismo divide al histograma en dos partes de igual área.

Cuartiles, quintiles, deciles y percentiles

Los cuartiles, quintiles, deciles y percentiles son parámetros que se corresponden al mismo tipo de problema que la mediana; por esto se les incluye dentro de los parámetros centrales, aunque como veremos no se sitúan en el centro de la serie estadística.

Si la mediana es el valor de la variable estadística que divide en dos partes iguales el número de observaciones, análogamente:

Los cuartiles son los tres valores que dividen la serie estadística en cuatro partes iguales. (dichas partes comprenden cada una el 25% de los datos). A estos valores se les denomina Q_1(primer cuartil), Q_2 (segundo cuartil) y Q_3 (tercer cuartil).

El cálculo es semejante al de la mediana. Así:

$$Q_j = L_i + a_i \frac{j\frac{N}{4} - F_{i-1}}{f_i} \quad \text{con} \quad j = 1,2,3$$

Los quintiles son los cuatro valores que dividen la serie estadística en cinco partes iguales, y siguen, lógicamente, la expresión:

$$K_j = L_i + a_i \frac{j\frac{N}{5} - F_{i-1}}{f_i} \quad \text{con} \quad j = 1,2,3,4$$

Los deciles son los nueve valores que dividen el número de datos en diez partes iguales (cada una comprenderá el 10% de los datos). Su expresión será:

$$D_j = L_i + a_i \frac{j\frac{N}{10} - F_{i-1}}{f_i} \quad \text{con} \quad j = 1,2,...,9$$

Los percentiles son los 99 valores que dividen la distribución en 100 partes iguales. Se calculan mediante la fórmula:

$$P_j = L_i + a_i \frac{j\frac{N}{100} - F_{i-1}}{f_i} \quad \text{con} \quad j = 1,2,...,99$$

Hay que tener en cuenta que existe una relación entre estos parámetros de posición, en ocasiones los valores coinciden. Así, por ejemplo: Q_2=Me, K_1=D_2, P_{75}=Q_3, etc.

Para realizar el cálculo gráficamente de los cuartiles (ídem. para los demás) se trabaja igual que para la mediana.

Se representa el polígono de frecuencias absolutas acumuladas, después se traza una recta paralela al eje de abscisas pasando por el punto de frecuencia acumulada que corresponda al cuartil que estamos calculando y el valor buscado es la abscisa del punto de intersección de la recta y la poligonal.

Relación entre media, moda y mediana

Si una distribución fuese completamente simétrica los valores de la media, moda y mediana coincidirían.

Para las distribuciones unimodales que no son demasiado asimétricas hay una relación aproximada entre estos tres parámetros $\bar{x} - Mo = 3\cdot(\bar{x} - Me)$

Gracias a esta relación se puede obtener, con un cierto error, alguno de estos parámetros en función de los otros (siempre que se den las condiciones nombradas)

Asimismo, para distribuciones unimodales lo más frecuente es que la mediana esté comprendida entre la moda y la media $M_0 \le M_e \le \bar{x}$

Para terminar este apartado vamos a calcular los parámetros descritos para nuestros ejemplos

Ejemplo A: *El número de suspensos de cada uno de los alumnos de bachillerato en la primera evaluación fue:*

x_i	f_i	F_i	h_i	H_i	$x_i \cdot f_i$
0	18	18	0,196	0,196	0
1	15	33	0,163	0,359	15
2	15	48	0,163	0,522	30
3	12	60	0,130	0,652	36
4	14	74	0,152	0,804	56
5	7	81	0,076	0,880	35
6	7	88	0,076	0,957	42
7	4	92	0,043	1,000	28
	92		1,000		242

Media: $\bar{x} = \frac{242}{92} = 2,63$

Moda: El valor con mayor frecuencia absoluta es 0. Mo=0

Mediana Hay 92 valores (par), luego la mediana será la media aritmética de los dos valores centrales (lugares 46 y 47). Si se observa la columna N_i, se comprueba que ambos lugares están ocupados por x_i=2. Por lo tanto Me=2

Cuartiles Hay 92 valores. Si queremos dividirlos en 4 grupos, cada grupo tendrá exactamente, 23 valores. Por lo tanto, Q_1 será la media aritmética de los valores que ocupan los lugares 23 y 24. Estos valores son 1 en ambos casos. Luego Q_1=1

Q_3 será la media de los valores que ocupan los lugares 69(23+23+23) y 70. Estos valores son 4 en ambos casos. Así pues Q_3=4

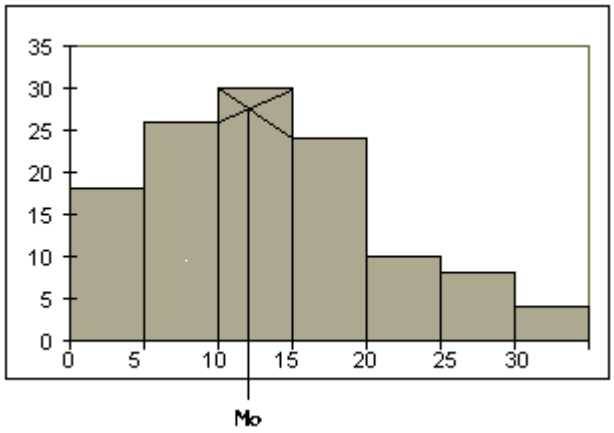
Ejemplo B: *Se ha hecho un estudio sobre el retraso (en minutos) de 120 trenes de cierta línea de largo recorrido obteniéndose los siguientes resultados*

	$c_i=x_i$	f_i	F_i	h_i	H_i	$x_i \cdot f_i$
[0,5)	2,5	18	18	0,150	0,150	45
[5,10)	7,5	26	44	0,217	0,367	195
[10,15)	12,5	30	74	0,250	0,617	375
[15,20)	17,5	24	98	0,200	0,817	420
[20,25)	22,5	10	108	0,083	0,900	225
[25,30)	27,5	8	116	0,067	0,967	220
[30,35)	32,5	4	120	0,033	1,000	130
		120		1,000		1610

Media: $\overline{x} = \frac{1610}{120} = 13,42$

Moda: El intervalo con mayor frecuencia absoluta (intervalo modal) es [10,15)

$$Mo = L_i + a_i \cdot \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} = 10 + 5 \cdot \frac{30-26}{(30-26)+(30-24)} = 12$$



Mediana Hay 120 valores , luego hay que buscar en la columna de las frecuencias absolutas acumuladas que intervalo contiene al valor situado en el lugar 60. Dicho intervalo es otra vez [10,15). Entonces

$$Me = L_i + a_i \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i} = 10 + 5\frac{60-44}{30} = 12,67$$

Cuartiles

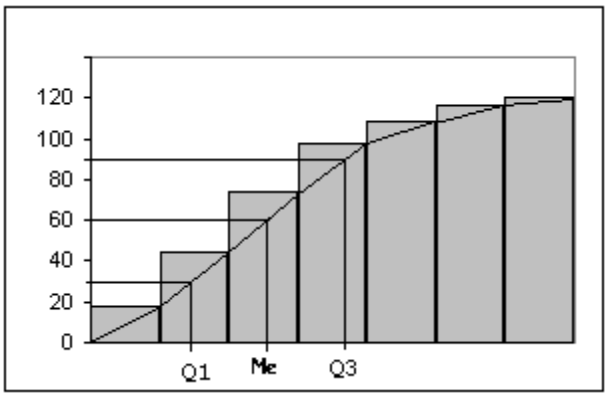
Hay 120 valores. Si queremos dividirlos en 4 grupos, cada grupo tendrá exactamente, 30 valores. Por lo tanto, para Q_1 debemos buscar el intervalo que contiene al valor situado en el lugar 30 -resulta ser [5,10) y para Q3 el intervalo que contiene el valor situado en el lugar 90, que es [15,20).

Por lo tanto:

$$Q_1 = L_i + a_i \frac{\frac{N}{4} - F_{i-1}}{f_i} = 5 + 5\frac{30-18}{26} = 7,31$$

$$Q_2 = M_e = L_i + a_i \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i} = 10 + 5\frac{60-44}{30} = 12,67$$

$$Q_3 = L_i + a_i \frac{3\frac{N}{4} - F_{i-1}}{f_i} = 15 + 5\frac{90-74}{24} = 18,33$$



Los quintiles, deciles y percentiles se calcularían de la misma manera.

Medidas de dispersión

Puede ocurrir que varias distribuciones tengan los mismos parámetros de centralización, pero que sean de aspecto muy diferente. Así por ejemplo:

(A)	8	8	9	9	9	9	9	10	10
(B)	6	6	7	9	9	9	11	12	12
(C)	1	3	6	9	9	11	13	14	15

La moda, mediana y media de las tres distribuciones anteriores es 9, y sin embargo dichas distribuciones son muy diferentes.

Conviene entonces buscar otras medidas que nos permitan obtener información sobre la forma de la distribución o sobre lo separados que están los datos respecto a las medidas de centralización. Estas son las medidas (o parámetros) de dispersión, que completan nuestro análisis numérico de una distribución estadística. Dan una idea del alejamiento de los datos respecto a las medidas de centralización.

Recorrido

Se llama recorrido (o rango, o amplitud) a la diferencia entre el mayor y el menor valor de la variable.

Cuando los datos estén agrupados en intervalos se tomará como recorrido la diferencia entre la mayor y la menor marca de clase.

Aunque el recorrido es fácil de calcular y sus unidades son las mismas que las de la variable, posee varios inconvenientes:

1) No utiliza todas las observaciones (sólo dos de ellas);
2) Se puede ver muy afectada por alguna observación extrema;

Recorrido intercuartílico

Recibe este nombre la diferencia entre el tercer y el primer cuartil (Q_3-Q_1). Es interesante por que nos da la amplitud de la banda en la que se encuentra el 50% central de la población, "*despreciando*" los valores extremos.

De forma similar se podrían definir recorridos interdecílicos, interquintílicos, ...

Veremos a continuación medidas de dispersión mejores que las anteriores.

Estas se determinan en función de la *distancia* entre las observaciones y algún estadístico de tendencia central; en nuestro caso, la media.

Una primera idea bastante natural sería calcular la diferencia entre cada valor de la variable y la media de la distribución y a continuación hallar la media aritmética de estas cantidades. Tendríamos entonces la media de las diferencias respecto de la media, es decir:

$$\frac{\sum (x_i - \overline{x}) \cdot f_i}{N}$$

Pero ya vimos que $\sum (x_i - \overline{x}) \cdot f_i = 0$

A poco que se piense esto es lógico, recuérdese que la media aritmética es el centro de gravedad de la distribución. Se puede comprobar con las tres distribuciones del principio de esta sección.

Para solucionar este problema se puede recurrir a varios procedimientos:

Desviación media

Se llama desviación media (o también desviación media absoluta) a la media aritmética de los valores absolutos de las desviaciones de todos los valores respecto de la media. Se denota por DM. Entonces:

$$DM = \frac{\sum |x_i - \overline{x}| \cdot f_i}{N}$$

Este parámetro tiene un difícil tratamiento algebraico (por el valor absoluto) y en el desarrollo superior de la estadística no se utiliza.

Si en lugar de recurrir al valor absoluto, se recurre a los cuadrados de las desviaciones tenemos lo siguiente:

Varianza

Se llama varianza de una variable estadística a la media aritmética de los cuadrados de las desviaciones respecto de la media. Se denota por σ^2 o s^2. Su expresión es:

$$s^2 = \frac{\sum (x_i - \overline{x})^2 \cdot f_i}{N} = \frac{\sum x_i^{\,2} \cdot f_i}{N} - \overline{x}^2$$

La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en metros, la varianza lo hace en metros^2). Si queremos que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada.

Desviación Típica

Se llama desviación típica de una variable estadística a la raíz cuadrada positiva de la varianza. Se representa por s o σ.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^{\,2} \cdot f_i}{N} - \overline{x}^2}$$

Conviene tener en cuenta las siguientes observaciones:

1) Tanto la varianza como la desviación típica dependen de todos los valores de la distribución, en función de que dependen de la media.
2) En los casos en los que no era posible calcular la media tampoco, se podrá obtener ni la varianza ni la desviación típica.
3) La desviación típica está expresada en las mismas unidades que los datos (la varianza está expresada en unidades cuadradas).
4) Bajo buenas condiciones (gran cantidad de datos en distribuciones unimodales, simétricas o ligeramente asimétricas), se llega a verificar que:

a) En $(\overline{x}-\sigma, \overline{x}+\sigma)$ se hallan aproximadamente el 68% de los datos.

b) En $(\overline{x}-2\sigma, \overline{x}+2\sigma)$ se hallan aproximadamente el 95% de los datos.

c) En $(\overline{x}-3\sigma, \overline{x}+3\sigma)$ se hallan aproximadamente el 99% de los datos.

A la hora de calcular los parámetros de dispersión añadiremos a nuestras tablas dos nuevas columnas: $x_i^2 \cdot f_i$ para calcular la varianza y la desviación típica y $|x_i - \overline{x}| \cdot f_i$ para la desviación media.

x_i	f_i	F_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$\lvert x_i - \overline{x}\rvert \cdot f_i$
…….	…….	…….	…….	……	…….
Totales					

Comparación de distribuciones

Tipificación

Se conoce con este nombre al proceso de restar la media y dividir por su desviación típica a una variable *X*. De este modo se obtiene una nueva variable

$$Z = \frac{X - \overline{x}}{s}$$

Esta nueva variable Z, que llamamos **variable tipificada,** tiene media 0 y desviación típica 1.

Si la variable X toma los valores $x_1, x_2, ..., x_n$ se llaman **puntuaciones típicas** a los valores que se obtienen para Z: $z_1 = \frac{x_1 - \overline{x}}{s}$, $z_2 = \frac{x_2 - \overline{x}}{s}, ..., z_n = \frac{x_n - \overline{x}}{s}$

Z carece de unidades y permite hacer comparables dos medidas que en un principio no lo son, por aludir a conceptos diferentes.

Así por ejemplo nos podemos preguntar si un elefante es más grueso que una hormiga determinada, cada uno en relación a su población. También es aplicable al caso en que se quieran comparar individuos semejantes de poblaciones diferentes.

Por ejemplo si deseamos comparar el nivel académico de dos estudiantes de diferentes Universidades para la concesión de una beca de estudios, en principio sería injusto concederla directamente al que posea una nota media más elevada, ya que la dificultad para conseguir una buena calificación puede ser mucho mayor en un centro que en el otro, lo que limita las posibilidades de uno de los estudiante y favorece al otro. En este caso, lo más correcto es comparar las calificaciones de ambos estudiantes, pero tipificadas cada una de ellas por las medias y desviaciones típicas respectivas de las notas de los alumnos de cada Universidad.

Ejemplo: *"Un profesor ha realizado dos exámenes a un grupo de alumnos obteniendo que la media y la desviación típica son para el primer examen 6 y 1,5 respectivamente y para el segundo examen 4 y 0,5 respectivamente. Un alumno obtuvo un 7 en el primer examen y un 5 en el segundo. ¿Cuál de las dos notas tiene más mérito?"*

La dificultad de esta cuestión radica en que estamos comparando 2 poblaciones con distinta media y distinta desviación típica. Para solventar este problema tipificamos las notas:

En el primer examen: $z_1 = \frac{7-6}{1,5} = 0,67$

En el segundo examen: $z_2 = \frac{5-4}{0,5} = 2$

Estos resultados indican que la nota del alumno en el primer examen se halla "0,67 desviaciones sobre la media" mientras que la nota del segundo examen se halla "2 desviaciones sobre la media". Por lo tanto la nota del segundo examen es más meritoria que la del segundo.

Hemos visto que las medidas de centralización y dispersión nos dan información sobre una muestra. Nos podemos preguntar si tiene sentido usar estas magnitudes para comparar dos poblaciones.

¿Qué ocurre si, por ejemplo, comparamos la altura de un grupo de personas con su peso? Tanto la media como la desviación típica se expresan en las mismas unidades que la variable. Por ejemplo, en la variable altura podemos usar como unidad de longitud el metro y en la variable peso, el kilogramo. Comparar una desviación (con respecto a la media) medida en metros con otra en kilogramos no tiene ningún sentido.

El problema no se resuelve teniendo las mismas unidades para ambas poblaciones. Por ejemplo, se nos puede ocurrir pesar a las hormigas con las mismas unidades que a los elefantes (ambos en kilogramos). Es evidente que la dispersión de la variable *peso de las hormigas* será

prácticamente nula, mientras que la de la variable *peso de los elefantes* será de varios cientos de kilos. Pero, dado que la diferencia entre las medias es enorme, de esto no se deduce que la dispersión del peso de las hormigas sea menor que la dispersión del peso de los elefantes.

Para solventar estos problemas se recurre al llamado

Coeficiente de variación (de Pearson)

El coeficiente de variación elimina la dimensionalidad de las variables y tiene en cuenta la proporción existente entre medias y desviación típica. Se define del siguiente modo:

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

Se deben tener en cuenta las siguientes consideraciones:

1) Sólo se debe calcular para variables que no tengan medias próximas a 0.
2) El coeficiente de variación no tiene unidades. Se suele expresar en %.
3) El coeficiente de variación sirve para comparar las dispersiones de dos conjuntos de valores, mientras que si deseamos comparar a dos individuosde cada uno de esos conjuntos, es necesario usar los valores tipificados.

Ejemplo: *"Los pesos de los toros de lidia de una ganadería se distribuyen con una media* $\bar{x}_1 = 510\,\text{Kg}$ *y una desviación típica* $s_1 = 25\,\text{Kg}$. *mientras que los pesos de los perros de una exposición canina se distribuyen con una media* $\bar{x}_2 = 19\ \text{Kg}$ *y una desviación típica* $s_2 = 10\,\text{Kg}$. *"*

La desviación típica de los pesos de la manada de toros bravos es superior que la de los perros ($s_1 > s_2$). Sin embargo, esos 25 Kg son poca cosa para el enorme peso de los toros (es decir, los toros de esa manada son muy parecidos en peso), mientras que 10 Kg en relación con el peso de un perro es mucho (imaginamos que en la exposición canina habrá perros muy dispares: caniches, "salchichas", dogos, mastines,...).

Calculamos los coeficientes de variación:

$$CV_1 = \frac{25}{510} = 0,049 \equiv 4,9\% \qquad CV_2 = \frac{10}{19} = 0,526 \equiv 52,6\%$$

Con este parámetro se ve claramente que el peso de los perros de la exposición canina es mucho más disperso que el de los toros de la manada.

Para terminar este apartado vamos a completar el estudio de nuestros ejemplos A y B

Ejemplo A: *El número de suspensos de cada uno de los alumnos de bachillerato en la primera evaluación fue:*

x_i	f_i	F_i	h_i	H_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$\lvert x_i - \bar{x}\rvert \cdot f_i$
0	18	18	0,196	0,196	0	0	47,35
1	15	33	0,163	0,359	15	15	24,46
2	15	48	0,163	0,522	30	60	9,46
3	12	60	0,130	0,652	36	108	4,43
4	14	74	0,152	0,804	56	224	19,17
5	7	81	0,076	0,880	35	175	16,59
6	7	88	0,076	0,957	42	252	23,59
7	4	92	0,043	1,000	28	196	17,48
	92		1,000		242	1.030	162,52

Media: $\bar{x} = \frac{242}{92} = 2,63$

Moda: El valor con mayor frecuencia absoluta es 0. Mo=0

Mediana Hay 92 valores (par), luego la mediana será la media aritmética de los dos valores centrales (lugares 46 y 47). Si se observa la columna N_i, se comprueba que ambos lugares están ocupados por x_i=2. Por lo tanto Me=2

Cuartiles Hay 92 valores. Si queremos dividirlos en 4 grupos, cada grupo tendrá exactamente, 23 valores. Por lo tanto, Q_1 será la media aritmética de los valores que ocupan los lugares 23 y 24. Estos valores son 1 en ambos casos. Luego Q_1=1

Q_3 será la media de los valores que ocupan los lugares 69(23+23+23) y 70. Estos valores son 4 en ambos casos. Así pues Q_3=4

Recorrido $7 - 0 = 7$

Recorrido intercuartílico $Q_3 - Q_1 = 4 - 1 = 3$

Desviación media $DM = \frac{\sum \lvert x_i - \bar{x}\rvert \cdot f_i}{N} = \frac{162,52}{92} = 1,77$

Varianza $s^2 = \frac{\sum x_i^2 \cdot f_i}{N} - \bar{x}^2 = \frac{1030}{92} - (2,63)^2 = 4,28$

Desviación típica $s = \sqrt{s^2} = \sqrt{4,28} = 2,07$

Coeficiente de variación $CV = \frac{s}{\bar{x}} = \frac{2,07}{2,63} = 0,79$

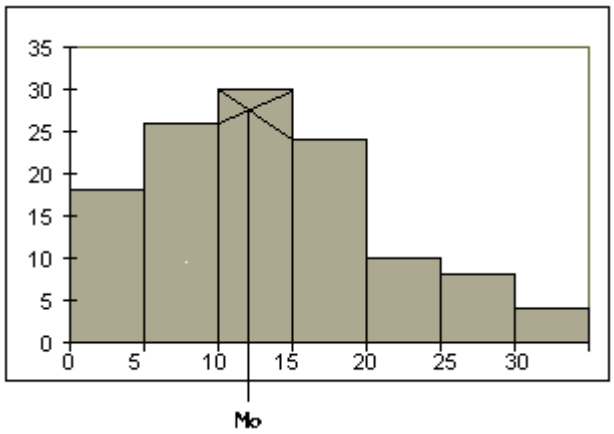
Ejemplo B: *Se ha hecho un estudio sobre el retraso (en minutos) de 120 trenes de cierta línea de largo recorrido obteniéndose los siguientes resultados*

	$c_i=x_i$	f_i	F_i	h_i	H_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$\lvert x_i - \bar{x}\rvert \cdot f_i$
[0,5)	2,5	18	18	0,150	0,150	45	112,50	196,50
[5,10)	7,5	26	44	0,217	0,367	195	1.462,50	153,83
[10,15)	12,5	30	74	0,250	0,617	375	4.687,50	27,50
[15,20)	17,5	24	98	0,200	0,817	420	7.350,00	98,00
[20,25)	22,5	10	108	0,083	0,900	225	5.062,50	90,83
[25,30)	27,5	8	116	0,067	0,967	220	6.050,00	112,67
[30,35)	32,5	4	120	0,033	1,000	130	4.225,00	76,33
		120		1,000		1610	28.950,00	755,67

Media: $\bar{x} = \frac{1610}{120} = 13,42$

Moda: El intervalo con mayor frecuencia absoluta (intervalo modal) es [10,15)

$$Mo = L_i + a_i \cdot \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} = 10 + 5 \cdot \frac{30-26}{(30-26)+(30-24)} = 12$$



Mediana Hay 120 valores , luego hay que buscar en la columna de las frecuencias absolutas acumuladas que intervalo contiene al valor situado en el lugar 60. Dicho intervalo es otra vez [10,15). Entonces

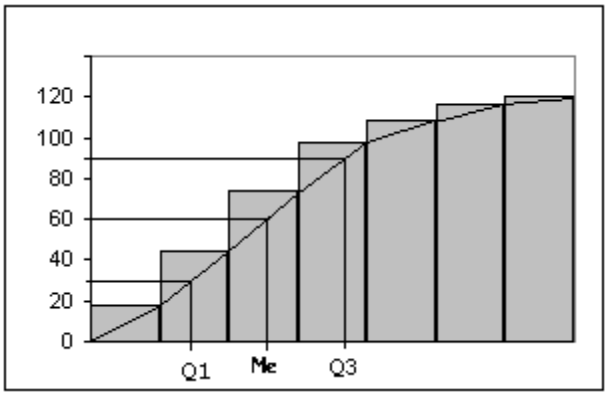
$$Me = L_i + a_i \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i} = 10 + 5\frac{60-44}{30} = 12,67$$

Cuartiles Hay 120 valores. Si queremos dividirlos en 4 grupos, cada grupo tendrá exactamente, 30 valores. Por lo tanto, para Q_1 debemos buscar el intervalo que contiene al valor situado en el lugar 30 -resulta ser [5,10) y para Q_3 el intervalo que contiene el valor situado en el lugar 90, que es [15,20). Por lo tanto:

$$Q_1 = L_i + a_i \frac{\frac{N}{4} - F_{i-1}}{f_i} = 5 + 5\frac{30-18}{26} = 7,31$$

$$Q_2 = M_e = L_i + a_i \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i} = 10 + 5\frac{60-44}{30} = 12,67$$

$$Q_3 = L_i + a_i \frac{3\frac{N}{4} - F_{i-1}}{f_i} = 15 + 5\frac{90-74}{24} = 18,33$$



Los quintiles, deciles y percentiles se calcularían de la misma manera.

Recorrido $32,5 - 2,5 = 30$

Recorrido intercuartílico $Q_3 - Q_1 = 18,33 - 7,31 = 11,02$

Desviación media $DM = \frac{\sum |x_i - \overline{x}| \cdot f_i}{N} = \frac{755,67}{120} = 6,30$

Varianza $s^2 = \frac{\sum x_i^2 \cdot f_i}{N} - \overline{x}^2 = \frac{28950}{120} - (13,42)^2 = 61,16$

Desviación típica $s = \sqrt{s^2} = \sqrt{61,16} = 7,82$

Coeficiente de variación $CV = \frac{s}{\overline{x}} = \frac{7,82}{13,42} = 0,58$

Ejemplo C: *Una empresa debe de cubrir cierto número de puestos de trabajo de dos tipos, A y B. Se somete a los aspirantes a dos pruebas, ambas puntuables de 0 a 50, diseñadas para valorar sus aptitudes en uno y otro tipo de trabajo. En la prueba A, la media de las calificaciones ha sido* $\overline{x}_A = 28$ *y la desviación típica* $s_A = 3,4$. *En la B han sido respectivamente,* $\overline{x}_B = 24$ *y* $s_B = 2,1$ *¿qué tipo de puesto asignarías a un estudiante que hubiese obtenido 33 puntos en la prueba A y 28 en la B?.*

Solución:

En ambos casos se halla por encima de la media.

Su puntuación es más alta en la prueba A (33 frente a 28), así como su desviación respecto de las medias (+5 frente a +4). No obstante, valorar igual los puntos obtenidos en ambas pruebas puede ser un "error de apreciación".

Las desviaciones típicas indican que los resultados de la prueba B se hallan más agrupados que los de la A. En estas condiciones "+4 puntos sobre la media" en la prueba B puede indicar mayor aptitud para el trabajo B, frente lo que indican "+5 puntos sobre la media" en la prueba para el trabajo A.

Para salir de dudas calculamos las puntuaciones típicas del aspirante en ambas pruebas:

$$z_A = \frac{33-28}{3,4} = 1,476$$

$$z_B = \frac{28-24}{2,1} = 1,905$$

Esto significa que su calificación en la prueba A se halla 1,476 desviaciones sobre la media y, en la B, 1,905 desviaciones sobre la media.

Por tanto, está más cualificado para ocupar un puesto de trabajo de tipo B que un puesto de tipo A.