La Estadística trata del recuento, ordenación y clasificación de los datos obtenidos por las observaciones, para poder hacer comparaciones y sacar conclusiones.
Un estudio estadístico consta de las siguientes fases:

- Recogida de datos.
- Organización y representación de datos.
- Análisis de datos.
- Obtención de conclusiones.

- **Población:** Conjunto de todos los individuos (personas, animales, objetos, etc) que porten información sobre el fenómeno que se estudia. Por ejemplo, si estudiamos el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.
- **Muestra:** subconjunto que seleccionamos de la población. Así, si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de la ciudad (sería una labor muy compleja), sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

- **Tamaño muestral:** número de individuos que tiene la muestra.

- **Tipos de Muestreos**

  - **Muestreo Aleatorio Simple**: Todos los elementos de la población tienen la misma probabilidad de ser elegidos para formar parte de la muestra.
  - **Muestreo aleatorio estratificado**: En ocasiones, la característica que se estudia en la población varía según diferentes grupos, como pueden ser los delimitados por el sexo o la edad.
    En este caso la población se divide en grupos homogéneos que se llaman estratos, y posteriormente se extrae una muestra aleatoria de cada estrato de forma que en la muestra, cada estrato mantenga la misma proporción que en la población.

- **Caracteres y variables:** Caracteres son los aspectos que deseamos estudiar en los individuos de una población.
  Cada carácter puede tomar distintos valores o modalidades. Una variable estadística recorre todos los valores de un cierto carácter.
  Clasificación de las variables estadísticas:

  - Cualitativas: No toman valores numéricos. Ejemplo (color de ojos, sexo de las personas..)
  - Cuantitativas discretas: Toman valores numéricos aislados. Ejemplo (número de hermanos, hijos,..)
  - Cuantitativas continuas: Pueden tomar todos los valores de un intervalo. Ejemplo (altura, peso..)

Tablas de Frecuencias: variables cuantitativas y cualitativas discretas

Las tablas de frecuencias sirven para ordenar y organizar los datos estadísticos.

Con los datos que nos aportan se construye la tabla de frecuencias:

- En la primera columna, la variable xi, con todos sus posibles valores.
- En la segunda columna, la correspondiente frecuencia, fi: número de veces que aparece cada valor.

Ejemplo: en una clase de 20 alumnos, las notas en la asignatura de Matemáticas (X) son las siguientes:

4 5 5 6 7 8 7 6 5 6 8 9 10 4 4 5 6 10 5 7

x_i	f_i
4	3
5	5
6	4
7	3
8	2
9	1
10	2
	20

Frecuencias Relativas

Cuando se desea comparar varias distribuciones similares con distinto número de elementos, se debe recurrir a las frecuencias relativas **(h_i).** Estas vienen dadas en "tanto por uno" o en"tantos por ciento" (%). Si N es el número de individuos: $h_i = \frac{f}{N}$ % = 100· h_i

Frecuencias Acumuladas

En una distribución de frecuencias, se llama frecuencia acumulada, Fi, correspondiente al valor i-ésimo, xi, a la suma de la frecuencia de ese valor con todas las anteriores: Fi = f1 + f2 +.... + fi

Análogamente se puede definir frecuencia relativa acumulada o porcentaje acumulado.

Ejemplo: Usamos el ejemplo anterior.

x_i	f_i	h_i	F_i	H_i
4	3	3/20= 0,15	3	0,15
5	5	0,25	8	0,4
6	4	0,2	12	0,6
7	3	0,15	15	0,75
8	2	0,1	17	0,85
9	1	0,05	18	0,9
10	2	0,1	20	1
	20	1		

$f_i \rightarrow$ frecuencias absolutas.

$h_i \rightarrow$ frecuencias Relativas.

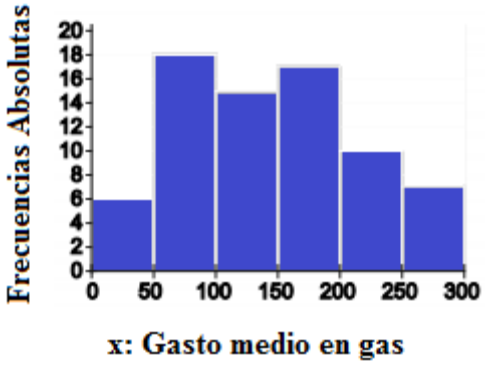
$F_i \rightarrow$ frecuencias Absolutas Acumuladas.

$H_i \rightarrow$ frecuencias Relativas Acumuladas

Tablas de Frecuencias: variable cuantitativa continua

Al poder una variable continua tomar cualquier valor en un cierto intervalo de la recta real, los valores se agrupan en intervalos o clases. Vamos a tomar como representante de cada clase su valor central o marca de clase.

Ejemplo: el gasto medio mensual en gas (x), en euros, de 73 viviendas lo mostramos en el siguiente gráfico:



Completamos la tabla de frecuencias:

Clases	Marca de clase x_i	f_i	h_i	F_i	H_i
[0,50)	25	6	0,08219	6	0,08219
[50,100)	75	18	0,24658	24	0,32877
[100,150)	125	15	0,20548	39	0,53425
[150,200)	175	17	0,23288	56	0,76713
[200,250)	225	10	0,13699	66	0,90412
[250,300]	275	7	0,09589	73	1
		73	1		

Parámetros Estadísticos

Los parámetros estadísticos sirven para sintetizar la información dada por una tabla. Los hay de dos tipos: de **centralización y de dispersión.**

- **Los parámetros de centralización** nos indican en torno a qué valor (centro) se distribuyen los datos. Son de este tipo la media, la mediana y la moda.
- **Los parámetros de dispersión** informan sobre cuánto se alejan del centro, los valores de la distribución. La mayoría de los parámetros de dispersión están asociados con un parámetro central al cual complementan:
  - Mediana: Me ⇒ Cuarteles (Q_1 y Q_3) y recorrido intercuartílico ($Q_3 - Q_1$)
  - Media: x ⇒ Desviación media (D.M.), varianza (s^2), desviación típica (s), coeficiente de variación (C.V.)

El recorrido (diferencia entre los valores extremos) es una medida de dispersión asociable tanto a la mediana como a la media.

La Media

Se llama media aritmética muestral o media muestral de una variable estadística cuantitativa a la suma de todos sus valores dividida por el número total de datos, es decir el tamaño de la muestra.

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

La Moda

La moda (M_o) de una variable discreta representa el valor que más veces se repite en el conjunto de datos.
Si la variable es continua, la clase modal será la clase que presenta una mayor frecuencia absoluta.

La Mediana

La mediana, M, es el valor de la variable que supera al 50% de los individuos.
Ordenamos los datos de menor a mayor, la mediana es el valor que ocupa el lugar central, de modo que la mitad de las observaciones es inferior a ella y la otra mitad es superior.
Para calcular la mediana utilizando la tabla de distribución de frecuencias tenemos que coger el primer valor de las frecuencias acumulas que supere a $\frac{N}{2}$.

Los cuantiles

Sabemos que la mediana divide a los datos en dos partes iguales, también tiene interés estudiar otros parámetros , llamados cuantiles, que dividen los datos de la distribución en función de otras cantidades. Los más importantes son los cuartiles, quintiles, deciles y percentiles.

- Cuartiles: son tres valores que divides la serie de datos en cuatro partes iguales. Se representan por Q_1(cuartil primero), Q_2 (cuartil segundo) y Q_3(cuartil tercero)
- Deciles: Son nueve valores que dividen la serie de datos en 10 partes iguales: D_1, D_2, ..., D_9.
- Percentiles: son 99 valores que dividen la serie de datos en 100 partes iguales: P_1, P_2, ..., P_{99}.

Ejemplo: Se ha realizado un test, compuesto de 10 preguntas, a 40 alumnos de un grupo con los siguientes resultados:

Respuestas	[0,2)	[2,4)	[4,6)	[6,8)	[8,10)
Alumnos	4	9	15	7	5

a) Calcular el valor de la media aritmética, la moda y la mediana.

b) ¿A partir de qué dato se encuentra el 70% (percentil 70) de los alumnos que han obtenido la mejor nota?

Calcula el decil 9 y cuartil 3

(L_i,L_s)	x_i	f_i	F_i	h_i	H_i	$x_i.f_i$
[0,2)	1	4	4	0,1	0,1	4
[2,4)	3	9	13	0,225	0,325	27
[4,6)	5	15	28	0,375	0,7	75
[6,8)	7	7	35	0,175	0,875	49
[8,10)	9	5	40	0,125	1	45
		40		1		200

a) La media aritmética :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N} = \frac{200}{40} = 5$$

- La moda es el valor de la variable que presenta mayor frecuencia absoluta.
  La mayor f_i= 15 que corresponde a la marca de la clase xi=5. Por lo tanto Mo = 5

- Para calcular la mediana utilizando la tabla de distribución de frecuencias tenemos que coger el primer valor de las frecuencias acumulas que supere a $\frac{N}{2} = \frac{40}{2}$ = 20 este valor correspondería al intervalo [4,6) pero cogemos la marca de la clase, xi= 5. Por lo tanto la mediana, M=5.

b)

- Calculamos el percentil 30.
  $\frac{30N}{100} = \frac{1200}{100}$ =12, El percentil 30 se encuentra en la clase [2,4). Tomamos como P_{30} la marca de clase P_{30} = 3

- Calculamos el decil 9
  $\frac{9N}{10} = \frac{360}{10} = 36$ el decil 9 se encuentra en la clase [8,10). Tomamos como D_9 la marca de clase D_9= 9

- Calculamos el cuartil 3

  $\frac{3N}{4} = \frac{120}{4} = 30$ el cuartil 3 se encuentra en la clase [6,8). Tomamos como Q_3 la marca de clase Q_3 = 7

Medidas de dispersión

Las medidas de dispersión o medidas de variabilidad muestran la variabilidad de un conjunto de datos, indicando la mayor o menor concentración de datos respecto a las medias de centralización.

Rango

El rango (R) o recorrido estadístico es la diferencia entre el valor máximo y el mínimo de un conjunto de elementos. Ejemplo: $D_1 = \{1,3,5,7\}$

Rango = Dato máximo – Dato mínimo = 7 -1 = 6

Rango Intercuartílico

El rango intercuartílico (RIC) es una estimación estadística de la dispersión de una distribución de datos. Consiste en la diferencia entre el tercer y el primer cuartil. Mediante esta medida se eliminan los valores extremadamente alejados.

$$\mathbf{RIC = Q_3 - Q_1}$$

Varianza

La varianza (S^2) mide la dispersión de los datos de una muestra respecto a la media, calculando la media de los cuadrados de las distancias de todos los datos.

$$\mathbf{s^2 = \frac{1}{n}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2 \rightarrow \ s^2 = \frac{1}{n}\sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2}$$

Siempre se cumple que la varianza es mayor o igual que cero ($S^2 \geq 0$). La varianza es cero cuando todos los datos son el mismo (ejemplo: {5,5,5,5,5}).

Observaciones a la varianza

1) Tanto la varianza como la desviación típica dependen de todos los valores de la distribución, así como de la media.
2) En los casos en los que no sea posible calcular la media aritmética, no será posible tampoco obtener la varianza y la desviación típica, por ser funciones de la media.
3) La varianza tiene el inconveniente de que no viene expresada en las mismas unidades que los datos, debido a que las desviaciones están elevadas al cuadrado. Si los datos fueran en metros, la varianza vendría dada en

metros cuadrados. En cambio, la desviación típica sí viene expresada en las mismas unidades que los datos, de ahí que resulte más interesante que la varianza.

Desviación típica

La desviación típica , es la raíz cuadrada positiva de la varianza

$$\mathbf{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} f_i x_i^2 - \overline{x}^2}$$

Coeficiente de Variación (solo para variables positivas)

Si queremos comparar la variabilidad de sos muestras con diferente media o de dos conjuntos de datos que no están formados en las misma unidades o en la misma unidad pero distinta medida, se utiliza el coeficiente de variación, que mide la variabilidad de la muestra respecto de su media.

$$\mathbf{CV} = \frac{s}{\overline{x}}$$

El coeficiente de variación es adimensional, no depende de las unidades de medida. Cuanto más pequeño sea , más homogéneos serán los datos.

Ejemplo: Se ha contado el número de oficinas municipales de información al consumidor abiertas al público en 25 ciudades. Estos son los datos:

3 6 4 2 3 4 5 4 7 3 5 4 5 4 3 3 4 3 2 4 6 1 8 2 5

a) Construye la tabla de frecuencias.

b) Halla la media.

c) Calcula la varianza, la desviación típica y el coeficiente de variación.

d) Calcula la moda, la mediana y el rango intercuartílico.

a)

xi	fi	Fi	hi	Hi	xi · fi	$x_i^2 \cdot f_i$
1	1	1	0,04	0,04	1	1
2	3	4	12	0,16	6	12
3	6	10	0,24	0,40	18	54
4	7	17	0,28	0,68	28	112
5	4	21	0,16	0,84	20	100
6	2	23	0,08	0,92	12	72
7	1	24	0,04	0,96	7	49
8	1	25	0,04	1	8	64
	25		1		100	464

b) $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i f_i}{N} = \dfrac{100}{25} = 4$

c) $s^2 = \dfrac{\sum_{i=1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - \bar{x}^2 = \dfrac{464}{25} - 4^2 = 2{,}56$ $\qquad s = \sqrt{2{,}56} = 1{,}6$ $\qquad CV = \dfrac{s}{\bar{x}} = \dfrac{1{,}6}{4} = 0{,}4$

d) La moda es el valor de la variable que presenta mayor frecuencia absoluta. La mayor fi= 7 que corresponde a xi=4. Por lo tanto Mo = 4.

- Para calcular la mediana utilizando la tabla de distribución de frecuencias tenemos que coger el primer valor de las frecuencias acumulas que supere a $\frac{N}{2} = \frac{25}{2}$ este valor correspondería a 17, con lo que iríamos a la columna xi y cogeríamos ese valor, por lo tanto la mediana M=4.

Si existe algún valor igual a $\frac{N}{2}$, la mediana sería la media entre este valor y el siguiente.

$\frac{N}{4} = \frac{25}{4}$ = 6,25, el primer cuartel corresponde al séptimo dato de la serie ordenada, por tanto Q1=3.

$\frac{3N}{4} = \frac{25 \cdot 3}{4}$ = 18,75, el tercer cuartel corresponde al dato decimonoveno de la serie, por tanto Q3=5

RIC = Q_3-Q_1 = 5 - 3 = 2