ESTADÍSTICA

LA ESTADÍSTICA es la rama de las matemáticas que estudia los fenómenos aleatorios, también llamados de azar, por no saber con anterioridad qué es lo que va a ocurrir. Es decir, no estudia fenómenos determinísticos, donde se sabe de antemano el resultado.

La palabra *estadística* tiene que ver con Estado, con el país, y es que tradicionalmente se relaciona con la información que tiene un estado para su organización. Aunque los primeros testimonios escritos de estadísticas datan del 3000 a.C. en Babilonia, pasando hasta el siglo XVI (Grecia, Roma, Edad Media…) la estadística sólo consistía en la recopilación de datos. El primer trabajo estadístico serio no llega hasta el s. XVII en Inglaterra, pero será un siglo más tarde, en Alemania, cuando empezó a sistematizarse y estudiarse seraimente.

La estadística es un conjunto de métodos científicos de recogida, organización, resumen, presentación y análisis de datos que permiten extraer conclusiones válidas y tomar decisiones acertadas basadas en esos datos.

Muchas veces, aunque incorrectamente, también solemos llamar *Estadística* a los propios datos, o a números derivados de esos datos, como por ejemplo, la media aritmética.

Un estudio estadístico consiste en recoger mucha información y ordenarla para sacar conclusiones. La forma más reducida y clara de ordenar información es mediante una **tabla**.

Una **serie estadística** es el conjunto de todos los resultados de un fenómeno aleatorio.

Población o **universo** es el conjunto de todos los elementos o individuos sometidos a un estudio. La población puede ser finita o infinita.

Una **muestra** es el subconjunto de población sobre el que se realiza el estudio cuando no es posible hacerlo sobre la población entera. Las muestras han de ser representativas.

Individuo es cada uno de los elementos que forman la población o la muestra.

Si la muestra es representativa de una población, se pueden sacar conclusiones importantes sobre esa población, derivadas del análisis de la muestra (por ejemplo, un sondeo electoral ante unas elecciones). La parte estadística que analiza las condiciones bajo las cuales tales conclusiones son válidas se llama *Estadística Inferencial* o *Inferencia Estadística*.

La parte de la Estadística que sólo describe y analiza un grupo determinado, se sacar conclusiones o inferencias sobre un grupo más amplio, se llama *Estadística Descriptiva* o *Deductiva*.

Una **variable** x_i es la característica que deseamos estudiar y representar. Una variable puede ser *cuantitativa*, cuando puede ser representada por números (número de hermanos), o *cualitativa*, cuando no se puede (color preferido). Sin embargo, para un estudio estadístico también podríamos trabajar con variables cualitativas asignando un número a cada cualidad; por ejemplo, si preguntamos "cuál es tu color preferido", podemos asignar para las respuestas los valores 1 para "rojo", 2 para "verde", 3 para "amarillo", etc.

Una *variable discreta* sólo toma valores aislados, mientras que una *variable continua* toma todos los valores posibles del intervalo.

Por ejemplo, si preguntamos cuántos hermanos tienen los individuos de una muestra, nos dirán que ninguno, que 1, que 2, que 3, etc., pero no hay valores intermedios: nadie tiene 2,6 hermanos; el número de hermanos es una variable discreta. En cambio, si les preguntamos cuál es su sueldo mensual, puede ser 700 euros, 800, 900, 1000, etc. Pero también puede situarse mejor que entre 900 y 1000, entre 900 y 950; pero todavía podemos ser más exactos, entre 920 y 930; y acercarnos más, y más… llegar hasta céntimos… Estamos ante variables continuas, que podemos agrupar en intervalos. Otros ejemplos de variables continuas serían la altura, el peso de los individuos, las notas de alumnos, la distancia entre ciudades…

En general, los conteos dan origen a variables discretas, y las mediciones, a variables continuas.

Si la variable, continua o discreta, conllevan un número grande de datos, para trabajar más cómodamente esos datos se agrupan en **intervalos** o **clases**. Un intervalo viene delimitado por las *cotas* inferior y superior, y la diferencia entre esas cotas es la *amplitud* del intervalo. De cada intervalo se toma un valor representativo llamado **marca de clase** que en muchas ocasiones se hace coincidir con el valor medio del intervalo, es decir, sumando las cotas superior e inferior y dividiendo entre 2.

No necesariamente todos los intervalos han de tener la misma amplitud.

Cuando un caso esté en el límite de 2 intervalos, *se incluirá siempre en el mayor de ellos*. Es decir, los intervalos son cerrados por la izquierda y abiertos por la derecha → [a,b). El último intervalo, el que recoge los valores más grandes, será también, lógicamente, cerrado por la derecha, a no ser que su límite sea el infinito → ∞. También el primer intervalo puede tener -∞ como límite inferior.

El número de individuos correspondiente a cada valor de la variable se llama **frecuencia** o **frecuencia absoluta** f_i de ese valor; es el número de veces que se repite esa modalidad o valor. La suma de las frecuencias absolutas de todas las variables da como resultado el total de individuos que forman la muestra.

La **frecuencia relativa** h_i de un valor es la proporción de veces que se presenta, y se obtiene dividiendo su frecuencia absoluta por el número total de datos o individuos, $N = \sum f_i$. El resultado de sumar todas las frecuencias relativas da como resultado la unidad, 1. Para obtener las frecuencias relativas da igual si trabajamos con variable discreta o continua; pero es necesario conocer la frecuencia absoluta.

Se suele expresar en porcentaje*. El porcentaje* resulta de multiplicar la frecuencia relativa por 100. La suma de todos los porcentajes debe ser 100%.

La **frecuencia absoluta acumulada** F_i de un valor x_i de una variable estadística es la suma de las frecuencias absolutas de todos los valores anteriores, los menores o iguales a x_i. Los valores de la variable han de estar ordenados de menor a mayor. La frecuencia absoluta acumulada correspondiente al último valor de la variable debe coincidir con el número de individuos de la muestra.

La **frecuencia relativa acumulada**, H_i, de un valor x_i de una variable estadística es el cociente entre su variable absoluta acumulada, F_i, y el número total de datos, N.

PARÁMETROS ESTADÍSTICOS

Hay 2 tipos de parámetros estadísticos: de **centralización** y de **dispersión**; y **medidas de posición**.

Por los parámetros de centralización podemos calcular en torno a qué valores centrales podemos resumir los datos; y los de dispersión, cuánto se alejan del centro los datos.

Une *media* es un valor típico, representativo, de un conjunto de datos. Como los valores representativos tienen tendencia a estar en el centro del conjunto de datos, los solemos llamar *parámetros de centralización*. En cambio, el grado con que los datos numéricos tienden a dispersarse en torno a un valor central se mide con los *parámetros de dispersión*.

Por ejemplo, tenemos dos grupos en 3º de la ESO; en 3ºA, la nota de matemáticas de la mayoría de los alumnos está entre 4 y 6; y en 3ºB, más o menos la mitad de los alumnos está entre 8 y 9, y la otra mitad, entre 1 y 2. Si sólo usáramos las medidas de centralización, los dos grupos parecerían similares, cuando en realidad son muy distintos; también hemos de usar las medidas de dispersión para darnos cuenta de lo diferentes que son ambos grupos.

1. PARÁMETROS DE CENTRALIZACIÓN

Indican en torno a qué valores se agrupan la mayoría de los datos. Son 3:

- La **Moda**, *Mo*: es el valor de la variable de mayor frecuencia absoluta: puede haber más de una Moda, o no existir. Una distribución con sólo una moda es *unimodal*.
- La **Media Aritmética**, $\bar{x}$, es el resultado de dividir la suma de todos los valores de la variable por el número total de observaciones, teniendo en cuenta las veces que se repite cada valor, es decir, su frecuencia o *peso*; matemáticamente:

$$\bar{x} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 \cdot f_3 \cdot x_3 + \cdots + f_n \cdot x_n}{f_1 + f_2 + f_3 + \cdots + f_n} = \frac{\sum f_i x_i}{\sum f_i}$$

- La **Mediana**, *Me*, es el valor central de un conjunto de datos numéricos ordenados. Cuando se trata de un número par de datos, la Mediana es la media aritmética de los dos datos centrales; en una serie de datos, SÓLO hay una mediana.

2. MEDIDAS DE POSICIÓN: LOS CUARTILES

Los **cuartiles** de una variable estadística son tres valores de la variable que dividen los datos en cuatro partes iguales:

El primer cuartil, Q_1, deja por debajo la cuarta parte de los datos.

El segundo cuartil, Q_2, coincide con la mediana: $Q_2=Me$

El tercer cuartil, Q_3, deja por debajo tres cuartas partes de los datos.

Para datos agrupados, los cuartiles se aproximan por las marcas de clase.

3. PARÁMETROS DE DISPERSIÓN

Los parámetros de dispersión permiten conocer el grado de mayor o menor agrupamiento de los datos entre sí o con respecto a un valor central; son:

– El **Rango** o **Recorrido** de una serie estadística es la diferencia entre el mayor y el menor de los datos de la serie.

– La **desviación respecto a la media** de un dato es el valor absoluto de la diferencia entre dicho dato y la media aritmética del conjunto de datos:

$$D_M(x_i) = |x_i - \bar{x}|$$

– La **varianza**, σ^2, es el promedio de los cuadrados de las desviaciones:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2 \cdot f_i}{\sum f_i} = \frac{\sum x_i^2 \cdot f_i}{\sum f_i} - \bar{x}^2$$

Y se puede calcular con cualquiera de las dos ecuaciones anteriores.

– La **desviación típica**, σ, es la raíz cuadrada positiva de la varianza.[1]

En una distribución estadística, con una muestra grande, y que no sea muy extraña, aproximadamente las 2/3 partes de las variables x_i están en el intervalo $(\bar{x} - \boldsymbol{\sigma}, \bar{x} + \boldsymbol{\sigma})$ (***¡RECUERDA!*** siendo $\bar{x}$ la media aritmética y $\boldsymbol{\sigma}$ la desviación típica).

– El **coeficiente de variación**, CV, es la razón (cociente) entre la desviación típica y la media aritmética. El CV permite comparar la dispersión entre 2 series estadísticas distintas. $CV = \frac{\sigma}{\bar{x}}$

El *coeficiente de variación* es una ***medida de dispersión relativa***. Pensemos que, por ejemplo, una dispersión de 10 centímetros no tiene la misma importancia en una medida de 1 metro que en otra de 100 metros.

El *coeficiente de variación* es independiente de las unidades de medida, es decir, no tiene unidades, lo que lo convierte en muy útil para comparar distribuciones con unidades de medida muy diferentes. En cambio, su inconveniente reside en no ser conveniente para valores de las variables próximos a cero.

Como el CV no tiene unidad (numerador y denominador tiene la misma, y al dividir se van), se suele expresar como un porcentaje:

a.	Si: CV < 30% → la dispersión es baja
b.	Si: CV > 60% → la dispersión es alta
c.	Si: 30% < CV < 60% → la dispersión es media

Por ejemplo, si CV=31%, nos indica que el valor de la desviación típica, σ, es el 31% de la media.

Los cálculos de estas medidas de dispersión son complejos para hacerse a mano, y se suele usar una hoja de cálculo, o en su defecto, la calculadora científica.

---

[1] El ejemplo más importante de distribuciones continuas de probabilidad es la *distribución normal*, *curva normal* o *campana de Gauss*, que verás en el Bachillerato, y para las que las propiedades de la desviación típica en *Distribuciones Normales* son:

1)	El 68,27% de las observaciones están entre $\bar{x} - \sigma$ y $\bar{x} + \sigma$
2)	El 95,45% de las observaciones están entre $\bar{x} - 2\sigma$ y $\bar{x} + 2\sigma$
3)	El 99,73% de las observaciones están entre $\bar{x} - 3\sigma$ y $\bar{x} + 3\sigma$