3 Estadística Descriptiva

Nociones Generales

La **Estadística** es una ciencia que se dedica al estudio de ciertas técnicas numéricas que permiten conocer y analizar características de toda clase de objetos, partiendo de datos obtenidos de forma empírica: a través de experimentos o de encuestas.

La **población** es el conjunto de seres u objetos acerca de los que se desea obtener información.

Una **muestra** es una parte o subconjunto de la población que es examinada para obtener los datos, y cuyo estudio nos permitirá inferir las características que estemos estudiando de la población.

Se llama **individuo** a cada uno de los miembros de la población – o la muestra –. El **tamaño** de la población o de la muestra es el número de individuos que la forman.

Variables Estadísticas

Se denomina **variable estadística** a cada una de las características que se están estudiando en la población.

Las variables se clasifican en

Cualitativas: si toman valores que no son numéricos.

Cuantitativas discretas: si toman valores numéricos que pueden enumerarse.

Cuantitativas continuas: si toman valores numéricos que pueden tomar todos los valores de un cierto intervalo.

Los valores de las variables continuas se suelen agrupar en intervalos, denominados **clases**; al punto medio de cada intervalos se le llama **marca de clase**.

Frecuencia

La **frecuencia absoluta** – n – de un dato es el número de veces que aparece entre los valores que toma la variable estadística.

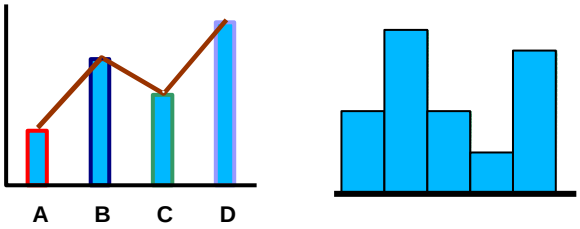
La **frecuencia relativa** – f – es el cociente de la absoluta entre el número total de datos – puede expresarse mediante porcentajes – .

Las **frecuencias acumuladas** – N y F – son las sumas de las frecuencias de todos los datos hasta el que consideramos (incluido).

Tablas y Gráficos

Las tablas y gráficas persiguen clasificar y organizar los datos obtenidos, de modo que se visualice con facilidad la información.

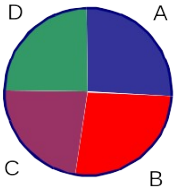
Los **diagramas de barras** son adecuados para las variables cualitativas y discretas.



Los **histogramas de frecuencias** son adecuados para representar variables continuas.

El **polígono de frecuencias** es la poligonal que une el centro de cada barra rectangular con el de la siguiente.

Los **diagramas de sectores** sirven para comparar distintas distribuciones dadas en términos de porcentajes.



3 Estadística Descriptiva

Nos referiremos a una distribución cuyos valores

$$x_1\,, x_2, \ldots, x_l$$

tienen unas frecuencias

$$n_1\,, n_2, \ldots, n_l$$

Si los datos están agrupados en clases, x_i y n_i son las marcas y frecuencias de las respectivas clases.

Medidas de Centralización

Una medida de centralización es un valor que pretende resumir la diversidad de los datos, intentando ser lo más representativo posible del conjunto de los datos.

La **media aritmética** es el número dado por:

$$\bar{x} = \frac{\sum x_i n_i}{\sum n_i}$$

Si colocáramos los datos sobre una barra horizontal sería su punto de equilibrio o centro de gravedad.

La **moda** es el dato con mayor frecuencia. Con datos agrupados en clases, diremos que la moda es la marca de la clase con mayor frecuencia –la clase modal – .

Medidas de Posición

Los valores de centralización de variables cuantitativas dependen de la ordenación de los datos y no sólo del valor de ellos.

Procedimiento de cálculo:

1. Calculamos las frecuencias relativas acumuladas F_i y las expresamos en porcentajes.
2. Llamamos "**centil** k" o "**percentil** k" al valor p_k que, tras ordenar los datos de menor a mayor, deja antes de él al k % de ellos.
3. En las variables discretas, buscamos el primer dato x_i que cumpla $F_i > k$ y entonces pueden darse dos casos:
   a) $F_i > k$ y $F_{i-1} < k \longrightarrow p_k = x_i$
   b) $F_i > k$ y $F_{i-1} = k \longrightarrow p_k = \dfrac{x_{i-1} + x_i}{2}$
4. Si están los datos agrupados en clases, dicho percentil p_k se encuentra en la primera clase $[a_i\,, b_i)$ con $F_i > k$ y usaremos la fórmula:

$$p_k = a_i + (b_i - a_i)\,\frac{k - F_{i-1}}{F_i - F_{i-1}}$$

Se llaman **cuartiles** a los siguientes percentiles:

$$Q_1 = p_{25}\ ,\ Q_2 = p_{50}\ ,\ Q_3 = p_{75}$$

Y a p_{50} se le llama **mediana**.

Medidas de Dispersión

Las **medidas de dispersión** tratan de medir el grado de cercanía de los datos a su media.

El **recorrido** o **rango** es la diferencia entre el mayor y el menor de los datos.

La **desviación típica** viene dada por:

$$\sigma = \sqrt{\frac{\sum x_i^2 n_i}{\sum n_i} - \bar{x}^2}$$

A su cuadrado se le llama **varianza**. Así:

$$\mathrm{VAR} = \sigma^2 \quad \text{ó} \quad \sigma = \sqrt{\mathrm{VAR}}$$

Interpretación Conjunta

En una distribución:

- La media $\bar{x}$ nos dice dónde está su centro.
- La desviación típica σ nos señala el alejamiento, la dispersión de los datos.

Para poder comparar la dispersión de dos distribuciones distintas se usa el llamado **coeficiente de variación**:

$$\mathrm{C.V.} = \frac{\sigma}{\bar{x}}$$