IES JOVELLANOS CURSO 2017/2018
APUNTES ESTADÍSTICA UNIDIMENSIONAL 3º ESO ACADÉMICAS

Aunque las tablas estadísticas y las representaciones gráficas contienen toda la información relativa a un problema, muchas veces interesa simplificar ese conjunto de datos por uno o varios valores que caractericen de la mejor forma posible esa distribución de frecuencias y que, además nos permita comparar unas distribuciones con otras.
En este sentido hay:
Medidas de centralización, que tienden a situarse en el centro de la distribución.
Medidas dispersión cuyo valor indica si los datos están concentrados o dispersos alrededor de un valor prefijado.
Medidas de posición que tienden a situarse en un determinado lugar de la distribución.

MEDIDAS DE CENTRALIZACIÓN

A) MEDIA ARITMÉTICA O MEDIA

La **media aritmética** de un conjunto de N valores x_1, x_2, x_3, ..., x_N es el cociente entre la suma de todos los valores observados (valores de la variable) y el número total de observaciones (tamaño poblacional); se representa por $\bar{x}$ y su expresión aritmética es:

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

Si los datos están **agrupados en intervalos**, consideraremos como valor de variable x_i al punto medio de cada intervalo, es decir, la **marca de clase**.
Ventajas de la media:

- La media es el valor medio o promedio de las observaciones.
- La media es la medida de centralización más utilizada.
- Es sensible a cualquier cambio en los datos

Desventajas:

- Es sensible a los valores extremos

B) MEDIANA

La **mediana** de una distribución es un valor $\mathbf{M_e}$ que divide a la distribución en dos partes iguales; es decir, deja tantas observaciones a la izquierda como a la derecha.
Para su cálculo distinguimos las siguientes situaciones:

1) Pocos datos y sin agrupar se colocan estos en orden creciente de magnitud.
Si el número de datos es **impar l**a mediana coincide con el valor central.
Si el número de datos es **par**, se suele tomar el valor medio de los dos valores centrales.

2)Muchos datos y sin agrupar, se construye la tabla de frecuencias acumuladas F_i, y se toma la mediana como aquel valor de la variable x_i para el cual F_i sea igual o supere $\frac{N}{2}$

3) Datos agrupados en intervalos primero buscaremos el **intervalo mediano**, que es el primer intervalo de clase cuya frecuencia acumulada es igual o superior a la mitad del número de observaciones, $\frac{N}{2}$. Después se usará semejanza de triángulos.

C)MODA

La **moda M_o** es el dato que más se repite, es decir el valor de la variable con mayor frecuencia absoluta. Es la única medida de centralización que tiene sentido estudiar en una variable cualitativa, pues no precisa la realización de ningún cálculo. La moda no tiene por qué ser única, sino que puede haber distribuciones multimodales.

Si los datos están agrupados en intervalos elegimos el intervalo modal, que es aquel con mayor frecuencia absoluta.

MEDIDAS DE DISPERSIÓN

¿Por qué las medidas de dispersión?

Las medidas de centralización representan bien a un conjunto de datos cuando están agrupados en torno a ellas, pero no cuando hay bastantes observaciones alejadas de ellas. Las medidas de dispersión miden, por tanto, el grado de alejamiento de los datos respecto a las medidas de centralización, fundamentalmente respecto de la media. Esas medidas son:

A) RANGO O RECORRIDO

El recorrido de una distribución es la diferencia entre el dato mayor y el dato menor, obtenidos al observar los valores de la variable.

B) VARIANZA

Se llama varianza de una serie de datos x_1, x_2, x_3, ..., x_n, que tienen frecuencias f_1, f_2, f_3, ..., f_n respectivamente, y se representa por σ^2 (o s^2), a la media aritmética de los cuadrados de las desviaciones respecto de la media, esto es:

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \overline{x})^2}{N}$$

La varianza también puede calcularse como la media de los cuadrados menos el cuadrado de la media.

$$\sigma^2 = \overline{x^2} - \left(\overline{x}\right)^2$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - \left(\frac{\sum_{i=n}^{n} x_i \cdot f_i}{N} \right)^2$$

C) DESVIACIÓN TÍPICA

Es la raíz cuadrada positiva de la varianza y se denota por σ (o **s**).

D) COEFICIENTE DE VARIACIÓN

Se llama coeficiente de variación y se representa por **C.V.** al cociente entre la desviación típica y el valor absoluto de la media.

$$C.V. = \frac{\sigma}{|\overline{x}|}$$

Consideraciones:

- Tanto la varianza como la desviación típica miden la dispersión de los datos respecto de la media. La varianza tiene el inconveniente que la unidad de medida en la que viene expresada es el cuadrado de la unidad en que se expresan los datos; sin embargo, la desviación típica viene expresada en las mismas unidades que los datos, por eso es más utilizada.
- El CV es un número real positivo que no tiene dimensiones, es decir no depende de las escalas usadas al medir, y se utiliza para comparar dispersiones de dos variables estadísticas diferentes. En ocasiones se suele expresar en tanto por ciento.

- El CV mide la dispersión relativa de los datos en relación con la media. Cuanto **más pequeño** sea más concentrados estarán los datos alrededor de la media, siendo por tanto **la media más representativa**.

- Si X e Y son dos variables de medias $\bar{x}$ e $\bar{y}$ y desviaciones típicas σ_x y σ_y :

a) Si $\bar{x} = \bar{y}$, $\sigma_x < \sigma_y \Rightarrow \bar{x}$ es más representativa.

b) Si $\bar{x} \neq \bar{y}$, será más representativa la que tenga menor CV

MEDIDAS DE POSICIÓN

Las medidas de posición pretenden localizar el lugar que ocupa un cierto elemento en la distribución. También las utilizaremos para responder a preguntas tales como ¿entre qué límites se encuentra el 50% central de los datos?, ¿cuál es el valor por debajo del cual están el 90% de los datos?, etc.

Para calcularlos hacemos algo similar a lo que hacíamos en el cálculo de la mediana.

A) CUARTILES

Son cada uno de los valores que divide la distribución en 4 partes iguales. Reciben los nombres de primer, segundo y tercer cuartil respectivamente y se representan por Q_1, Q_2 y Q_3.

De modo que debajo del primer cuartil queda el 25% de la distribución, debajo del segundo el 50% y debajo del tercero el 75% de la misma. Es claro que Q_2 es precisamente la mediana.

B) DECILES

Son los valores de la distribución que dividen a esta en diez partes iguales, y los denotamos por D_1, D_2, D_3, ..., D_9. Así D_1 deja por debajo el 10% de los valores de la distribución, D_2 deja por debajo el 20% de los valores de la distribución, y así sucesivamente. El valor de D_5 coincide con la media.

C) PERCENTILES

Son los valores que dividen la serie de datos en cien partes iguales. Los denotamos por P_1, P_2, P_3, ..., P_{99}. Así P_1 deja por debajo el 1% de los valores de la distribución, P_2 deja por debajo el 2% de los valores de la distribución, y así sucesivamente. Es claro que P_{50} coincide con la mediana.