Estadística Descriptiva y Análisis de Datos

Diplomatura en Estadística. Curso 2007/2008

Formulario Tema 2.

Consideramos una muestra de tamaño n de una variable X. Sean $x_1, x_2, \ldots, x_k$, $k \leq n$, los diferentes valores que ha tomado esta variable sobre los n individuos de la muestra. A continuación se definen una serie de conceptos:

- **Frecuencia absoluta**, n_i, es el número de veces que se ha observado x_i en la muestra. Se cumple que $\sum_{i=1}^{k} n_i = n$.
- **Frecuencia relativa**, f_i, es la proporción de individuos para los cuales se ha observado el valor x_i, es decir $f_i = n_i/n$. Se cumple que $\sum_{i=1}^{k} f_i = 1$.
- **Frecuencia absoluta acumulada**, N_i, es el número de individuos para los cuales se ha observado un valor menor o igual que x_i, es decir: $N_i = \sum_{j=1}^{i} n_j$.
- **Frecuencia relativa acumulada**, f_i, es la proporción de individuos para los cuales se ha observado un valor menor o igual que x_i, es decir: $F_i = \sum_{j=1}^{i} f_j = N_i/n$.
- Cuando los datos están agrupados en intervalos de clase, llamamos l_{i-1} y l_i a los **extremos inferior** y **superior**, respectivamente, del intervalo i-ésimo.
- La **longitud** o **amplitud** del intervalo i-ésimo es $L_i = l_i - l_{i-1}$.
- La **marca de clase** del intervalo i-ésimo es $x_i = \frac{l_i + l_{i-1}}{2}$.

Medidas de centralización o de tendencia central:

- **Media aritmética**
$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i \, n_i.$$
- **Media artimética ponderada**. Si $x_1, x_2, \ldots, x_n$ son los valores distintos observados en una muestra de tamaño n y $p_1, p_2, \ldots, p_n$ son los pesos correspondientes,
$$\overline{x}_P = \frac{\sum_{i=1}^{n} x_i \, p_i}{\sum_{i=1}^{n} p_i}.$$

- **Media geométrica**

$$\overline{x}_G = \left(\prod_{i=1}^{k} x_i^{n_i}\right)^{1/n} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}.$$

- **Mediana**.
a) Ordenación simple de los datos:

$$Me = \begin{cases} \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \text{si } n \text{ es par,} \\ x_{(\frac{n+1}{2})}, & \text{si } n \text{ es impar.} \end{cases}$$

b) Ordenación agrupada de los datos: El **intervalo mediano** es aquel intervalo de clase que contiene a $x_{(n/2)}$. Si $[l_{i-1}, l_i)$ es el intervalo mediano, entonces:

$$Me = l_{i-1} + (l_i - l_{i-1}) \cdot \frac{(n/2 - N_{i-1})}{N_i - N_{i-1}}.$$

- **Moda**.
a) Ordenación simple de los datos: La moda es el valor x_i que presenta una frecuencia absoluta (o relativa) más alta.

b) Ordenación agrupada de los datos: El **intervalo modal** es aquel al que le corresponde una altura mayor en el histograma de frecuencias. Si $[l_{i-1}, l_i)$ es el intervalo modal, entonces:

$$Mo = l_{i-1} + L_i \frac{n_{i+1}/L_{i+1}}{n_{i-1}/L_{i-1} + n_{i+1}/L_{i+1}}.$$

Medidas de posición:

- **Cuartil** k-ésimo, para $k = 1, 2, 3$, es el número Q_k que deja k cuartas partes de la muestra por debajo de él y $4-k$ cuartes partes por encima.

- **Decil** k-ésimo, para $k = 1, 2, \ldots, 9$, es el número D_k que deja k décimas partes de la muestra por debajo de él y $10-k$ décimas partes por encima.

- **Percentil** k-ésimo, para $k = 1, 2, \ldots, 99$, es el número P_k que deja k centésimas partes de la muestra por debajo de él y $100-k$ centésimas partes por encima.

  Para calcular los cuartiles, deciles y percentiles para datos agrupados en intervalos de clase, se utiliza la fórmula de cálculo de la mediana, pero substituyendo $n/2$ por $k\,n/4$ para el cuartil Q_k, $k\,n/10$ para el decil D_k y $k\,n/100$ para el percentil P_k.

Medidas de dispersión o de variabilidad:

- Rango o amplitud: $R = x_{max} - x_{min}$.
- Rango intercuartílico: $RI = Q_3 - Q_1$.
- Desviación media: $D_m = \frac{1}{n}\sum_{i=1}^{k} |x_i - \overline{x}|\, n_i$.
- Varianza muestral: $s_n^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2\, n_i = \overline{x^2} - \overline{x}^2$.
- Desviación típica (o estándar) muestral: $s_n = \sqrt{s_n^2}$.
- Cuasivarianza muestral o varianza muestral corregida:

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \overline{x})^2\, n_i.$$

- Cuasidesviación típica muestral: $s_{n-1} = \sqrt{s_{n-1}^2}$.
- Mediana de las desviaciones absolutas:

$$MEDA = Me\left(|x_1 - Me(X)|, |x_2 - Me(X)|, \ldots, |x_k - Me(X)|\right).$$

- Coeficiente de variación de Pearson: $CV = s_n/\overline{x}$.

Medidas de asimetría y de apuntamiento (o curtosis):

- Coeficiente de asimetría de Pearson

$$As_P = \frac{\overline{x} - Mo}{s_n}.$$

- Coeficiente de asimetría de Fisher

$$As_F = \frac{\frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^3\, n_i}{s_n^3}.$$

- Coeficiente de apuntamiento o de curtosis

$$Ap = \frac{\frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^4\, n_i}{s_n^4}.$$