Estadística

Tema 1: Estadística descriptiva. Descripción de una variable

La estadística es la ciencia de los datos; implica la colección, clasificación, síntesis, organización, análisis e interpretación de datos

Suele aplicarse a dos tipos de problemas:

- Resumir, describir y explorar datos referidos a un colectivo.
- Utilizar datos de muestras para deducir conclusiones sobre un colectivo más amplio del cual se escogieron las muestras

La estadística descriptiva se dedica a la organización, síntesis y descripción de conjunto de datos.
La estadística inferencial se ocupa de utilizar datos de muestras, para inferir algo acerca de la población de la que provienen

Errores básicos mostrando datos estadísticos

- La muestra de un porcentaje suelto nunca puede servir para inferir una relación entre 2 variables.
- Mostrar Rankings absolutos, no relativos, para intentar clasificar

Conceptos previos

Población: Conjunto de elementos que son objeto de estudio.

Individuo: Cada uno de los elementos de la población descrito mediante una serie de características a las que se refiere el estudio estadístico.

Muestra: Una muestra es un subconjunto no vacío de individuos de la población. El número de elementos que componen la muestra se denomina tamaño muestral (N)

Caracteres o variables: las cualidades de los individuos de la población que son objetos de estudio. Pueden ser cualitativas o cuantitativas

Modalidades: las diferentes situaciones posibles de una variable cualitativa. Un individuo debe pertenecer a una y sólo una modalidad.

Tipos de variables:

Cualitativa nominal: País, color;

Cualitativa ordinal: Todo, mucho, regular, poco, nada.

Cuantitativa discreta: Nº hijos, Nº de mensajes.

Cuantitativa continua: Altura en cm, peso en Kg, ruido en dB.

Frecuencias

Frecuencia absoluta: (n_i) de la modalidad x_i es el número de individuos observados que presentan esa modalidad.

Frecuencia relativa: (f_i) de la modalidad x_i es el cociente entre la frecuencia absoluta y el número total de individuos

$$f_i = \frac{n_i}{N}$$

Frecuencia absoluta acumulada: (N_i) de una modalidad x_i de la variable X es la suma de las frecuencias de los valores que son inferiores o iguales a él.

$$N_i = \sum_{j=1}^{j=i} n_j$$

Frecuencia relativa acumulada: (F_i) de una modalidad x_i de X es el cociente entre la frecuencia absoluta acumulada y el número total de individuos

$$F_i = \frac{N_i}{N}$$

Medidas de tendencia central: Medias, mediana y moda.

Media: la media aritmética simple es la suma de todos los valores divididos por el número total de datos.

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \sum_{i=1}^{k} x_i f_i$$

Media ponderada: la media ponderada de los datos x_i por los pesos w_i se define como:

$$\bar{x}_w = \frac{\sum_i x_i w_i}{\sum_i w_i}$$

G: {2'6; 3'7; 5'1; 4'9 y 6'4} Peso: 1, 1, 1, 2, 3

$$\bar{x}_w = \frac{2'6 \cdot 1 + 3'7 \cdot 1 + 5'1 \cdot 1 + 4'9 \cdot 2 + 6'4 \cdot 3}{1+1+1+2+3} = 5'05$$

Media cuadrática o Valor Cuadrático Medio (RMS)

La media cuadrática de los datos x_i se obtiene mediante la expresión:

$$\bar{x}_c = \sqrt{\frac{\sum_i x_i^2}{N} n_i}, \text{ o bien para datos agrupados: } \bar{x}_c = \sqrt{\frac{\sum_i n_i x_i^2}{N}}$$

Media armónica

$$H = \frac{N}{\sum \frac{n_i}{x_i}}$$

$$\binom{K}{N} \quad \binom{K}{n} = \frac{K!}{n!(K-n)!}$$

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} \cdot \frac{5!}{3!\,2!} = \frac{5 \cdot 4 \cdot 3!}{3!\,2!} = \frac{5 \cdot 4}{2} = 10$$

Moda: La moda (Mo) de un conjunto de datos es el valor de la variable que presenta mayor frecuencia. Puede no ser única o puede que no exista si todos los valores tienen la misma frecuencia.

Mediana: La mediana (Me) es aquel valor que divide a la población en dos partes de igual tamaño. Si N es impar la mediana coincidirá con un término de la población, si N es par, se toman los dos valores centrales y se calcula su media.

Cuantiles: Constituyen una generalización del concepto de mediana.
Dado un valor $c \in (0,1)$ se define el cuantil c como el valor de $X(c)$ que divide a la variable dejando una proporción c menor y una proporción $1-c$ mayor que él
Evidentemente la mediana coincide con el cuantil $c=0'5$.

Cuartiles: Son tres valores con las siguientes características:
$Q_1 = X(0'25)$: Valor que deja por debajo 1/4 de la población.
$Q_2 = X(0'5) = Me$: Deja por debajo la mitad de la población
$Q_3 = X(0'75)$: Deja por debajo 3/4 de la población.

$$Q(K) = L_{i-1} + \frac{N \cdot K - N_{i-1}}{n_i} a_i$$

Deciles: Hay 9 deciles que dividen la población en 10 partes iguales
$D_K = X\left(\frac{K}{10}\right)$

Percentiles: Hay 99 percentiles que dividen en 100 partes iguales a la población. Se denotan por $P_K = X\left(\frac{K}{100}\right)$ que será el valor que divide a la población dejando por debajo el K% de los valores y por encima el (100-K)%.

Cálculo del cuantil.
Realizamos la descomposición de cN en su parte entera (E) y decimal (D): $cN = (E) + (D)$
- Si $D \neq 0$, $X_{(E+1)}$ es el valor que ocupa el lugar $(E+1)$
- Si $D = 0$, $X(c) = \frac{\text{Valor de lugar } (E) + \text{valor lugar } (E+1)}{2}$

Medidas de desviación y dispersión

Rango: Recorrido o intervalo (R) es la diferencia entre el mayor y el menor valor observado de la variable.

Otros rangos son:

Rango intercuartílico: $R_q = Q_3 - Q_1$

Rango intercentílico: $R_p = P_{99} - P_1$

El rango es muy sensible a un error en los datos, no así los rangos intercuartílicos y intercentílicos.

Desviación media

La desviación d_i de un valor x_i de la variable respecto a un parámetro p es la diferencia $d_i = |x_i - p|$ entre esos valores.

La desviación media respecto a un promedio p es la media del valor absoluto de las desviaciones a una determinada medida de tendencia central p.

$$DM(p) = \frac{\sum_{i=1}^{k} |x_i - p| \cdot n_i}{N} = \sum_{i=1}^{i=k} |x_i - p| \cdot f_i$$

Si el parámetro p es la media aritmética simple lo llamamos desviación media:

$$DM(\bar{x}) = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| \cdot n_i}{N} = \sum_{i=1}^{i=k} |x_i - \bar{x}| \cdot f_i$$

Con el valor absoluto no se puede derivar

Error cuadrático medio

Llamamos error cuadrático medio a la media de las desviaciones al cuadrado:

$$ECM(p) = \frac{\sum_i n_i (x_i - p)^2}{p}$$

Ej: Dados los valores $\{5, 2, 3, 3, 3, 5, 7\}$ hallar la DM y el ECM respecto a la media.

$\bar{x} = 4$, las desviaciones absolutas $|\vec{d_i}|$ son: $\{1, 2, 1, 1, 1, 1, 3\}$ luego

$$DM = \frac{5 \cdot 1 + 1 \cdot 2 + 1 \cdot 3}{7} = \boxed{\frac{10}{7}} \quad \text{y} \quad ECM = \frac{5 \cdot (1)^2 + 1 \cdot 2^2 + 1 \cdot 3^2}{7} = \boxed{\frac{18}{7}}$$

La varianza y la desviación típica

La varianza de un conjunto de datos viene dada por:

$$V = \sigma^2 = \frac{\sum_{i=1}^{K} (x_i - \bar{x})^2 n_i}{N} = \sum_{i=1}^{K} (x_i - \bar{x})^2 f_i$$

Es la media de los cuadrados de las desviaciones respecto a la media.

Otra forma es:

$$V = \sum_{i=1}^{K} x_i^2 \cdot f_i - \bar{x}^2 = \frac{\sum_i^K n_i x_i^2}{N} - \bar{x}$$

La desviación típica o estándar es la raíz cuadrada de la varianza.

$$\sigma = +\sqrt{V} = \sqrt{\sum_{i=1}^{K} (x_i - \bar{x})^2 f_i}$$

Desviación media respecto a C

$$DM(C)_i = \sum_{i=1}^{K} |x_i - c| f_i \rightarrow \text{Si } c = \bar{x} \Rightarrow \text{Desviación media}$$

Error cuadrático medio respecto a C

$$ECM(C)_i = \sum_{i=1}^{K} (x_i - c)^2 f_i \rightarrow \text{Si } c = \bar{x} \rightarrow \text{Varianza}$$

Momento de orden r respecto a C

$$Mr(C) = \sum_{i=1}^{K} (x_i - c)^r f_i = \sum_i^K \frac{n_i (x_i - c)^r}{N}$$

• $C = 0$ **Momento ordinario de orden r**

$$m_r := m_r(0) = \sum_{i=1}^{K} x_i^r \cdot f_i$$

$\rightarrow m_0 = \sum_{i=1}^{K} x_i^0 \cdot f_i = \sum_{i=1}^{K} f_i = 1$

$\rightarrow m_1 = \sum_{i=1}^{K} x_i^1 \cdot f_i = \bar{x}$

$\rightarrow m_2 = \sum_{i=1}^{K} x_i^2 \cdot f_i \rightarrow \sqrt{m_2} \rightarrow$ media cuadrática

$\rightarrow \sigma^2 = m_2 - \bar{x}^2 \rightarrow$ sabiendo que $V = \sum_{i=1}^{K} (x_i - \bar{x})^2 f_i = \underbrace{\left(\sum_{i=1}^{K} x_i^2 f_i\right)}_{m_2} - \bar{x}^2$

$$\boxed{\sigma^2 = m_2 - \bar{x}^2}$$

• Momento central de orden r

$m_{r,i} = m_r(\bar{x})$

$\rightarrow m_0 = \sum_{i=1}^{k} (x_i - \bar{x})^0 \cdot f_i = 1$ $\quad \mu_0 = 1$

$\rightarrow m_1 = \sum_{i=1}^{n} |x_i - \bar{x}|^1 f_i = 1$ $\quad \mu_1 = 0$

$\rightarrow \mu_2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 f_i \rightarrow$ Varianza $\rightarrow \mu_2 = m_2 - \bar{x}^2$ // $\mu_2 = \left(\sum_{i=1}^{k} (x_i^2 f_i)\right) - \bar{x}^2$

$\rightarrow \mu_3 = m_3 - 3m_2\bar{x} + 2\bar{x}^3$

$\rightarrow \mu_4 = m_4 - 4m_3\bar{x} + 6m_2\bar{x}^2 - 3\bar{x}^4$

$m_r = \sum_{i=1}^{k} x_i^r \cdot f_i = \sum_{i=1}^{k} x_i^r \frac{n_i}{N}$

$N \cdot m_r = \sum_{i=1}^{k} x_i^r n_i$

Medidas de comparación.

Haciendo uso de la media y de la desviación típica de la variable X, podemos considerar una nueva variable dada por:

$Z = \frac{X - \bar{x}}{\sigma}$ con valores $z_i = \frac{x_i - \bar{x}}{\sigma}$ $\quad i = 1, 2, \ldots, k$

La variable tipificada es adimensional y, por tanto, independiente de las unidades usadas. Mide la desviación de la variable respecto a su media en términos de la desviación típica.

Coeficiente de variación de Pearson

Es el cociente entre la desviación típica y el valor absoluto de la media

$CV = \frac{\sigma}{|\bar{x}|} \times 100$ (en tanto por ciento) $\quad \uparrow$ mayor $\uparrow$ disperso

Coeficientes de asimetría

Coeficiente de asimetría de Pearson:

$A_p = \frac{\bar{x} - Mo}{\sigma}$ $\quad A_p = 0 \rightarrow$ Simetría

- $A_p > 0 \rightarrow$ Asimetría a la derecha o positiva
- $A_p < 0 \rightarrow$ Asimetría a la izquierda o negativa

Coeficiente de asimetría de Fisher (Sesgo)

$g_1 = \frac{\mu_3}{\sigma^3}$ $\quad g_1 = 0 \rightarrow$ Simetría (o insesgada)

- $g_1 > 0 \rightarrow$ Asimetría (o sesgada) a la derecha o positiva
- $g_1 < 0 \rightarrow$ Asimetría (o sesgada) a la izqda o negativa.

Coeficiente de Apuntamiento

El coeficiente de aplastamiento de Fisher es:

$g_2 = \frac{\mu_4}{\sigma^4} - 3$

$g_2 < 0 \rightarrow$ Menos aplastamiento que la normal (PLATICURTICA)

$g_2 = 0 \Rightarrow$ igual (MESOCURTICA)

$g_2 > 0 \rightarrow$ mayor apuntamiento que la normal (LEPTOCURTICA)