Tema 1: Análisis de una variable

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

Aplicaciones:
- Resumen y Descripción de los datos acerca de un colectivo (Estadística Descriptiva)
- Utilización de datos de muestra para deducir conclusiones (Estadística Inferencial) sobre un colectivo más amplio del cual se escogieron las muestras

Conceptos Previos

1.- Población: Conjunto de elementos en los que se estudia

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

3.- Muestra: Es un subconjunto no vacío de individuos de la población. El cardinal o número de elementos que la conforman se denomina tamaño muestral (N)

4.- Caracteres o Variables: Las cualidades de los individuos de la población que son objeto de estudio.

Pueden ser:
- Cualitativas (Nominales u Ordinales) { Nom: Verde, Amarillo,... Ord: Muy alto, Alto,...
- Cuantitativas (Discretas o Continuas) { Dis: Número de hijos (0, 1, ...) Cont: Altura en cm.

5.- Modalidades: Las diferentes situaciones posibles de una variable cualitativa. Un individuo debe pertenecer a una, y solo una, modalidad.

Definiciones

a) Frecuencia Absoluta

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

b) Frecuencia Relativa

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

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

c) Sumatoria de Frecuencia Absoluta (Frecuencia Absoluta Acumulada)

La 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.

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

d) Sumatoria de Frecuencias Relativas (Frecuencia Relativa Acumulada)

La 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.

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

Distribuciones de Frecuencia

La distribución de frecuencias de un carácter, sea cualitativo (atributo) o sea cuantitativo (Variable Estadística), está constituida por las distintas modalidades del carácter junto a las correspondientes frecuencias. Estas distribuciones se presentan en forma de tabla estadística o de frecuencia.

⇒ Cuando el número de observaciones y el número de modalidades es muy grande, es común mostrar los datos agrupados en intervalos (clases) y se determina el número de individuos que pertenecen a cada intervalo. Usualmente, (por convenio) los intervalos se expresan de la forma:

$$I_i =]L_{i-1}, L_i]$$

Donde sus extremos son: $]-\infty, L_1]$ y $]L_{i-1}, \infty[$

- Amplitud de Intervalo:

$$a_i = L_i - L_{i-1}$$

- Marca de clase:

$$x_i = \frac{L_{i-1} + L_i}{2}$$

- Los puntos medios de las clases son llamados marcas de clases.

Ojo

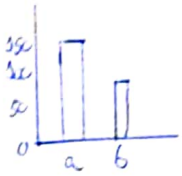
Hay que tener en cuenta el número de intervalos, cuanta más particiones tenga el intervalo, mayor información habrá. Por ejemplo, es lo mismo decir que hay cuatro platos que cuestan entre 0 y 4 €, y hay dos platos que cuestan entre 0 y 2 € y dos platos que cuestan entre 2 € y 4 €. Es lo mismo, sin embargo, en la segunda opción hay más información.

Representaciones Gráficas

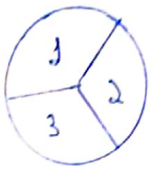
Las representaciones tratan de facilitar una síntesis visual y conviene cuidar la presentación (colores, formas, ...). El tipo de carácter establece una clasificación de las representaciones gráficas.

Caracteres Cualitativos

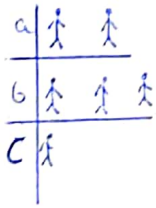
1.- Diagrama de Rectángulos o barras.



2.- Diagrama de Sectores.



3.- Pictograma y Cartograma. (Uso de Símbolos)

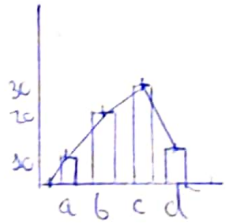


= 10 personas

Caracteres Cuantitativos

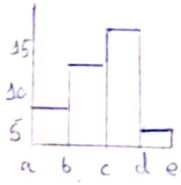
1.- Diagrama de Barras o puntos.

Caso Discreto. La longitud de la barra queda determinada por la frecuencia y el valor de la variable determina el lugar donde se apoya en el eje horizontal

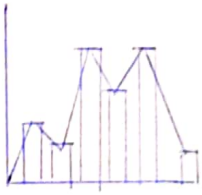


2.- Histograma

(Es conveniente usar la proporcionalidad de las f.r. o f.a.)



3.- Polígono de Frecuencias



Definiciones

a) Media aritmética Simple

Es la suma de todos los valores dividido por el número total de datos.

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

$\bar{x}$ = Media Muestral

μ = Media Aritmética de la población

b) Outliers

Son datos extraños que se separan bastante del resto, causados por una recolección errónea de datos o casos excepcionales.

Inciso

Cuando un estadístico es poco sensible a modificar su valor ante la presencia de un outlier decimos que se trata de un estadístico robusto

c) Transformación Afín

Si, a partir de los valores de una variable X, construimos otra $Y = aX + b$, entonces.

$$\bar{y} = a\bar{x} + b$$

Para realizar la transformación, necesitamos realizar la inversa.

$$f^{-1} \Rightarrow \bar{y} = \frac{\bar{x} - b}{a}$$

d) Media Ponderada

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

$$\bar{x}_w = \frac{\sum_{j=1}^{i} x_i \cdot w_i}{\sum_{j=1}^{i} w_i}$$

e) Media Cuadrática o Raíz Cuadrática Media (RMS)

$$\bar{x}_c = \sqrt{\frac{\sum_{j=1}^{i} x_i^2}{N}}$$

Para datos agrupados

$$\bar{x}_c = \sqrt{\frac{\sum_{j=1}^{i} n_i x_i^2}{N}}$$

f) Moda

La moda es el valor de la variable (en un conjunto) que presenta mayor frecuencia.

Puede no ser única o puede que no exista si todos los valores tienen la misma frecuencia.

g) Mediana

La mediana 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 la media.

h) Cuantiles

Dado un valor $c \in]0,1[$ se define cuantil c como el valor $X(c)$ que divide a la variable dejando una proporción c menor y una proporción $1-c$ mayor que él.

i) Cuartiles

Son tres valores que dividen a la población en 4 partes. $Q_i = X\left(\frac{i}{4}\right)$

j) Deciles

Son nueve valores que dividen a la población en 10 partes. $D_i = X\left(\frac{i}{10}\right)$

k) Percentiles

Son 99 valores que dividen a la población en 100 partes. $P_i = X\left(\frac{i}{100}\right)$

Cálculo del cuantil c

a) Caso Discreto

Realizamos la descomposición de cN en su parte entera (E) y decimal (D):

$$cN = E + D$$

+ Si $D \neq 0$, $X(c)$ es el valor que ocupa el lugar $(E+1)$
+ Si $D = 0$, $X(c) = \frac{\text{Valor del lugar}(E) + \text{Valor del lugar}(E+1)}{2}$

b) Caso Continuo

En la columna de las frecuencias acumuladas N_i busco la primera que rebasa ese valor $N_{i-1} < cN < N_i$, a continuación aplico:

$$X(c) = L_{i-1} + \frac{cN + N_{i-1}}{n_i} a_i$$

L_{i-1} = Límite inferior del intervalo
N_{i-1} = Frecuencia Absoluta acumulada correspondiente al intervalo anterior.
a_i = Amplitud del intervalo
n_i = Frecuencia Absoluta del intervalo

Medidas de Desviación y Dispersión

Ayudan a determinar la variación de los datos. Sirven para determinar lo agrupada o dispersa que está una población y si la medida de tendencia central calculada es representativa.

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

- Rango Intercuartílico: $R_Q = Q_3 - Q_1$
- Rango Intercentílico: $R_P = P_{99} - P_1$

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.

⇒ Normalmente, p es una medida de tendencia central

1| La desviación media respecto a un promedio p es la media de los valores absolutos de las desviaciones a una medida de tendencia central p.

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

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

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

Error Cuadrático Medio

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

$$ECM(p) = \frac{\sum_{i=1}^{k} n_i \cdot (x_i - p)^2}{N}$$

La Varianza y La Desviación Típica

La varianza poblacional o varianza de un conjunto de datos viene dada por:

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

Esto es la media de los cuadrados de las desviaciones respecto a la media

Otra forma equivalente para calcular la varianza es:

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

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

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

Media y Varianza Muestral

La media muestral es el mejor estimador para realizar la media poblacional (μ), ya que normalmente no podemos medir toda la población y nos conformamos con una muestra.

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

Sin embargo, el mejor estimador de la varianza de una población no es la varianza de la muestra, es la cuasivarianza de la muestra:

La varianza muestral o cuasi-varianza (s^2), para una muestra de tamaño N de

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

Inciso

No confundir "varianza muestral" $\left(s^2 = \frac{N}{N-1} \cdot \sigma_M^2\right)$, con "varianza de la muestra"

$$\left(\sigma_M^2 = \frac{\sum_{i=1}^{K} (x_i - \bar{x})^2 \cdot n_i}{N}\right)$$

Medias de Comparación

Se usan para comparar información obtenidas de distintas muestras o distintas poblaciones.

⇒ Variable Tipificada

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

$$Z = \frac{X - \bar{x}}{\sigma} \qquad \text{con valores} \qquad z_i = \frac{x_i - \bar{x}}{\sigma} \qquad i = 1, 2, \dots, K$$

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

Coeficiente de Variación de Pearson (Dispersión Relativa)

Un problema de la desviación típica como medida de dispersión es que depende de las unidades de la variable y de la muestra. Por tanto no resulta útil para comparar dispersiones entre dos muestras distintas o expresadas con unidades distintas.

Por ello se define el coeficiente de variación de Pearson, como el cociente entre la desviación típica y el valor absoluto de la media:

$$CV = \frac{\sigma}{|\bar{x}|}$$

Normalmente se expresa en tanto por ciento, para ello basta multiplicar el cociente por 100.

+ Tiene el problema de no estar definido cuando $\bar{x} = 0$

Momentos Ordinarios respecto a un punto

→ Se define el momento de orden r respecto al punto c como:

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

Momentos Ordinarios

→ Se define el momento ordinario de orden r como la media aritmética de las potencias de orden r de los datos de la variable:

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

Inciso

Se verifica que:

+ El momento ordinario de orden 0 vale 1, $m_0 = 1$
+ El momento ordinario de orden 1 es la media aritmética. $m_1 = \bar{x}$
+ El momento ordinario de orden 2, $m_2 = \sigma^2 + \bar{x}^2$

Momentos Centrales

Se define el momento central de orden r como la media aritmética de las potencias de orden r de las desviaciones de los datos respecto de la media:

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

Inciso

Propiedades:

1.- Los momentos centrales $\mu_0 = 1$ y $\mu_1 = 0$

2.- El momento central de orden 2 es la varianza

$$\mu_2 = V = \sigma^2 = m_2 - \bar{x}^2$$

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

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

Medidas de Forma: Simetría y Apuntamiento

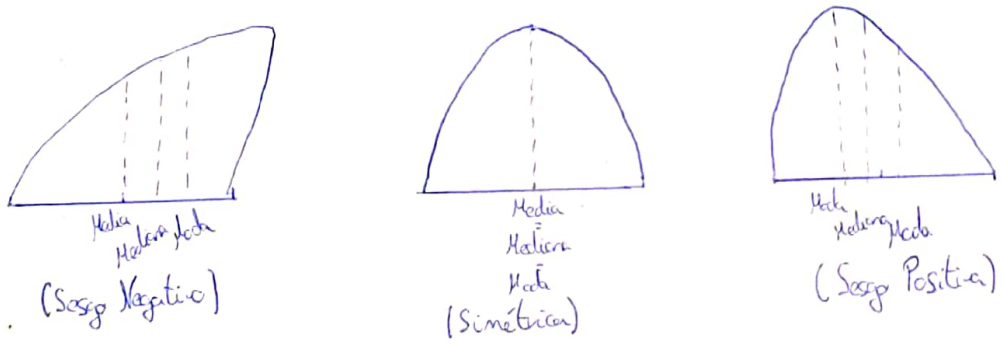
Otras medidas que nos permiten clasificar la forma de una ~~puta~~ distribución son las medias de asimetría (o sesgo) y las medidas de apuntamiento (o curtosis)

⇒ Medidas de Asimetría

Una distribución de frecuencias es simétrica cuando los valores de la variable que equidistan de un valor central tienen las mismas frecuencias.

+ Las distribuciones simétricas verifican:

a) $\bar{x} = Me$ b) $\bar{x} = Me = Mo$



⇒ Coeficientes de Asimetría

Gracias a los coeficientes de asimetría de Fisher o Pearson, sabemos si existe asimetría o no:

Pearson ⇒ $Ap = \dfrac{\bar{x} - Moda}{\sigma}$

- $Ap > 0$ Asimetría a la derecha o sesgo positivo
- $Ap = 0$ Simetría
- $Ap < 0$ Asimetría a la izquierda o negativa (sesgo)

Fisher ⇒ $g_1 = \dfrac{\mu_3}{\sigma^3}$

- $g_1 > 0$ Asimetría a la derecha o sesgo positivo
- $g_1 = 0$ Asimetría insesgada o simetría
- $g_1 < 0$ Asimetría a la izquierda o sesgo negativo

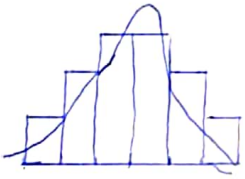
⇒ Coeficiente de Apuntamiento

El aplastamiento, apuntamiento o curtosis de una distribución es el grado de achatamiento o afilamiento en comparación con la distribución normal con igual media y varianza.

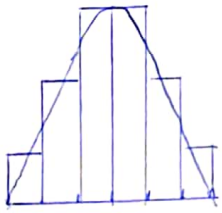
El coeficiente de aplastamiento de Fisher es:

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

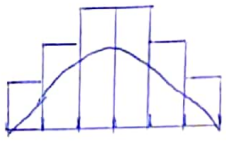
- $g_2 < 0$ Menos apuntamiento que la normal (Platicúrtica)
- $g_2 = 0$ Igual apuntamiento que la normal (Mesocúrtica)
- $g_2 > 0$ Más apuntamiento que la normal (Leptocúrtica)



Platicúrtica



Mesocúrtica



Leptocúrtica

⇒ Media Armónica

La media armónica es la media recíproca de la media aritmética.

$$H = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$