Tema2

Alejandro Zubiri

Wed Oct 16 2024

Contents

Este tema desarrollará dos variables simultáneas.

$$\begin{aligned} x &= \{x_1, \dots, x_n\} \\ y &= \{y_1, \dots, y_n\} \end{aligned} \tag{1}$$

1 Definiciones

- Frecuencia absoluta del par: f_{ij}
- Frecuencia relativa del par: $fr_{ij} = \frac{f_{ij}}{n}$
- Distribución conjunta: valores y frecuencias de los pares.
- Distribución marginal: distribución de las variables por separado:

$$\begin{aligned} f(x_i) &= \sum^{s} f_{ij} \\ f(y_j) &= \sum^{r} f_{ij} \end{aligned} \tag{2}$$

1.1 Distribución condicionada

Condicionada a que x o y tome un determinado valor:

$$fr(y_j | x = x_i) = \frac{f(x_i, y_j)}{f(x_i)} \tag{3}$$

1.2 Variables independientes

El conocimiento de una no afecta a la otra. Se cumple si

$$fr(y_j | x = x_i) = fr(y_j) \tag{4}$$

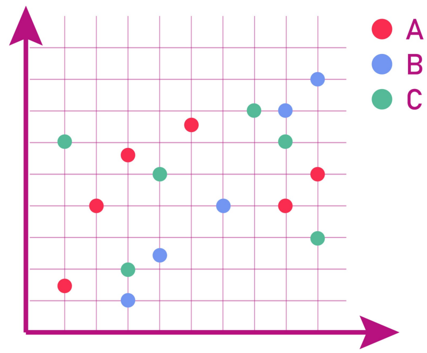
Además, se cumplirá que

$$fr(x_i, y_j) = fr(x_i) \cdot fr(y_j) \tag{5}$$

Esto es la **condición de independencia**.

2 Representación gráfica

Utilizaremos el diagrama de dispersión:



3 Tipos de covariación

3.1 Relaciones de dependencia

- Causal o unilateral: x afecta a y, pero no viceversa.
- Interdependencia: x afecta a y y viceversa.
- Dependencia indirecta: covariación en función de otras variables.
- Concordancia: tener la misma calificación bajo un mismo hecho.
- Covariación casual.

3.2 Covarianza

Relación lineal entre variables.

$$cov(x, y) = s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})f(x_i, y_i)}{n} = \frac{\sum x_i y_i f(x_i y_i)}{n} - \bar{x}\bar{y} \tag{6}$$

Si es

- Positiva: pendiente positiva.
- Negativa: pendiente negativa.
- Cero o casi: no hay relación o no es lineal.

3.3 Coeficiente de correlación

$$r = \frac{S_{xy}}{S_x S_y} \implies -1 \leq r \leq 1 \tag{7}$$

4 Medidas características

4.1 Vector de medias

Para variables k-dimensionales. Con k observaciones de un individuo $\mathbf{x}$. Sus componentes son las medias:

$$\mathbf{x} = \begin{bmatrix} \bar{x} \\ \bar{y} \\ . \end{bmatrix} = \frac{1}{n} \sum \vec{x}_i \tag{8}$$

4.2 Matriz de varianzas

$$M = \frac{1}{n} \sum \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \\ z_i - \bar{z} \end{pmatrix} \begin{pmatrix} x_i - \bar{x} & y_i - \bar{y} & z_i - \bar{z} \end{pmatrix} \tag{9}$$

4.3 Varianza efectiva

La raíz de orden k del determinante de la matriz de varianzas

$$VE = \sqrt[k]{|M|} \tag{10}$$

Definimos la desviación típica afectiva como

$$DM = \sqrt{VE} = |M|^{\frac{1}{2k}} \tag{11}$$

5 Relación entre variables cualitativas

5.1 Q de Yule

Oscila entre -1 y 1. Se puede utilizar cuando tenemos dos categorías:

$$Q = \frac{f_{11} f_{22} - f_{12} f_{21}}{f_{11} f_{22} + f_{12} f_{21}} \tag{12}$$

1 es asociación perfecta, 0 no hay, y -1 perfecta negativa.

5.2 Coeficiente de contingencia de la χ^2

Comparamos las frecuencias observadas f_{ij} con las esperadas como si fueran independientes:

$$e_{ij} = \frac{f_{i\cdot} f_{\cdot j}}{n} \tag{13}$$

Entonces

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \tag{14}$$

Si son independientes, $\chi^2 = 0$.
Si $r = s$:

$$max\chi^2 = n(s-1) \tag{15}$$

5.3 Coeficiente de contingencia de Pearson

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{16}$$

Que varía entre $[0, 1)$ según la dimensión de la tabla. El valor máchimo es:

$$C_{\max} = \sqrt{\frac{k-1}{k}} \tag{17}$$

Siendo $k = min(r, s)$

5.4 Coeficiente de contingencia corregido de Pawlik

$$C_{\text{corr}} = K^* = \frac{C}{C_{\max}} \tag{18}$$

5.5 V de Cramer

$$V = \sqrt{\frac{\chi^2}{n[min(r,s) - 1]}} \tag{19}$$

Teniendo una escala indicando:

- $V \leq 0.2$: débil
- $0.2 \leq V \leq 0.6$: moderada
- $0.6 \leq V$: fuerte

5.6 T de Tschuprov

$$T^2 = \frac{\chi^2}{n\sqrt{(r-1)(s-1)}} \in [0, 1] \tag{20}$$

5.7 Coeficiente de correlación por rangos de Spearman

En escala ordinal, oscila entre $[-1, 1]$:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{21}$$

Donde d_i es la diferencia de cada par ordenado.

5.8 Razón de correlación

Mide el grado entre cualitatica y cuantitativa:

$$\eta^2_{y/x} = \frac{\sum f_{\cdot j}(\bar{y_x} - \bar{y})^2}{n\sigma^2_y} \tag{22}$$

donde

- $f_{\cdot j}$: número de datos por categoría.
- $\bar{y_x}$: media de la categoría.