Capítulo 2

Análisis de datos cualitativos

DEFINICIÓN DE VARIABLES CUALITATIVAS

Son aquellas variables cuyos valores son un conjunto de cualidades no numéricas a las que se llama categorías o modalidades.

CLASIFICACIÓN DE VARIABLES CUALITATIVAS

- **Escala nominal:** No se puede definir un orden natural entre sus categorías. (Ejemplo: la raza, el color del pelo, o la religión)

- **Escala ordinal:** Se pueden establecer relaciones de orden entre las categorías. (Ejemplo: el rango militar, la clase social o el nivel de estudios)

- **Por intervalos:** Pueden tratarse como ordinales y se pueden calcular distancias numéricas entre dos niveles. (Ejemplo: El número de años de educación recibidos (0, 1, 2, ...) es una variable cuantitativa que puede ser agrupada por intervalos)

TABLAS DE CONTINGENCIA

Una tabla de contingencia es una tabla bidimensional en la que las variables objeto de estudio no son cuantitativas.

Ejemplo. Tabla de contingencia para estudiar la asociación entre color y fragancia de las flores azaleas:

	Color de la flor		
Fragancia	Blanca	Rosa	Naranja
Sí	12	60	58
No	50	10	10

INDEPENDENCIA DE VARIABLES CUALITATIVAS

Contrastamos

$$\begin{cases} H_0: & \text{A y B son independientes} \\ H_1: & \text{A y B no son independientes} \end{cases}$$

Estadístico de contraste

$$\chi^2_{\text{exp}} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

- $t_{ij} = \dfrac{n_{i\cdot} n_{\cdot j}}{N}$, y que bajo la hipótesis nula sigue una distribución $\chi^2_{(p-1)(q-1)}$
- p y q : Número de filas y columnas, respectivamente, de la tabla
- $n_{i\cdot}$: Total de las frecuencias observadas de la $i-$ésima fila (modalidad i del carácter A)
- $n_{\cdot j}$: Total de las frecuencias observadas de la $j-$ésima columna (modalidad j el carácter B)
- N : Número de individuos observados

ASOCIACIÓN DE VARIABLES CUALITATIVAS

Estudiamos algunas de las medidas de asociación más empleadas en la práctica.

MEDIDAS DE ASOCIACIÓN

■ Chi-cuadrado

Medida que compara los valores (n_{ij}) observados en la tabla con los que teóricamente se obtendrían (t_{ij}) bajo la hipótesis nula

$$\chi^2_{\text{exp}} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

Toma valores:

★ Entre 0 y N para tablas de contingencia 2×2 y

★ Entre 0 y $N \times \text{mín}\{p-1, q-1\}$ en tablas $p \times q$, con $p, q \geq 2$.

★ Un valor igual a 0 indica independencia de A y B.

MEDIDAS EN ESCALA NOMINAL

En escala nominal podemos considerar las siguientes **medidas de asociación**:

1) **Coeficiente** ϕ

2) **Coeficiente de contingencia o C de Pearson** (C)

3) **Coeficiented**

4) **Coeficiente V de Cramer** (**V**)

5) **Coeficiente Lambda** ($\boldsymbol{\lambda}$)

■ En tablas de contingencia 2×2 el coeficiente ϕ y el coeficiente **V** de Cramer toman valores entre 0 y 1:

★ Un valor 0 implica independencia de los atributos.

★ Un valor 1 denota asociación perfecta.

★ Valores cercanos a 1 indican un grado de asociación fuerte mientras que valores próximos a 0 implican un grado de asociación débil.

■ El Coeficiente de contingencia o **C** de Pearson toma en tablas 2×2 valores comprendidos entre 0 y $\dfrac{\sqrt{2}}{2}$, siendo:

★ El valor $\dfrac{\sqrt{2}}{2}$ denota asociación perfecta.

★ Un valor 0 indica independencia.

- Los valores del coeficiente lambda están comprendidos entre 0 y 1 para tablas $p \times q$, con $p, q \geq 2$:
  - ★ Valores próximos a 0 implican baja asociación
  - ★ Valores próximos a 1 denotan fuerte asociación.
  - ★ Sin embargo un valor $\boldsymbol{\lambda = 0}$ no implica independencia de los atributos.

Tabla $\mathbf{2 \times 2}$ para medidas en escala nominal			
Medida	Valores	Independencia	Asociación perfecta
Coeficiente $\boldsymbol{\phi}$	$0 \leq \boldsymbol{\phi} \leq 1$	0	1
Coeficiente **V** de Cramer	$0 \leq \mathbf{V} \leq 1$	0	1
Coeficiente de contingencia **C** de Pearson	$0 \leq \mathbf{C} \leq \frac{\sqrt{2}}{2}$	0	$\frac{\sqrt{2}}{2}$
Coeficiente Lambda	$0 \leq \boldsymbol{\lambda} \leq 1$	–	1

- Los valores de estas medidas no dependen del número de filas ni de columnas de la tabla, por lo que permiten la comparación entre tablas.

Tabla $\mathbf{p \times q}$ con $\mathbf{p, q > 2}$ para medidas en escala nominal			
Medida	Valores	Indepenpendencia	Asociación perfecta
Coeficiente $\boldsymbol{\phi}$	$0 \leq \boldsymbol{\phi} \leq$ A	0	A
Coef. **V** de Cramer	$0 \leq \mathbf{V} \leq 1$	0	1
Coef. de contingencia **C** de Pearson	$0 \leq \mathbf{C} \leq$ B	0	B
Coeficiente Lambda	$0 \leq \boldsymbol{\lambda} \leq 1$	–	1

donde:

- ♣ $\mathrm{A} = \sqrt{\text{mín}\{p-1, q-1\}}$
- ♣ $\mathrm{B} = \sqrt{\dfrac{\text{mín}\{p-1, q-1\}}{\text{mín}\{p-1, q-1\}+1}}$

- Los valores de $\boldsymbol{\phi}$ y de **C** dependen de **p** y **q**, por lo que no permiten realizar comparaciones entre tablas.

MEDIDAS EN ESCALA ORDINAL

Para variables en escala ordinal, puede considerarse además del grado de asociación la dirección de ésta.

- Se dice que dos variables están relacionadas positivamente si a valores altos (bajos) de una de ellas le corresponden valores altos (bajos) en la otra.
- Se dice que están relacionadas negativamente si a valores altos (bajos) de una de ellas le corresponden valores bajos (altos) en la otra.

★ Si A y B son medidas a escala ordinal pueden aplicarse las medidas de asociación válidas para escala nominal.

★ Además en escala ordinal pueden considerarse:

1) **Coeficiente Gamma de Goodman y Kruskal** (γ)

2) **Coeficiente d de Somers** (**d**)

3) **Coeficiente Tau−B de Kendall (Tau−B**)

4) **Coeficiente Tau−C de Kendall (Tau−C**)

$$\boxed{-1 \leq \gamma, \mathbf{d},\ \mathbf{Tau{-}B}\ ,\ \mathbf{Tau{-}C} \leq 1}$$

EN GENERAL, PARA ESTAS MEDIDAS SE TIENE:

- Cuanto más próximos estén los valores de estas medidas a 0 más débil será la asociación entre las variables.
- Cuanto más cercanos a 1 (o a -1) sean los valores de todas estas medidas mayor será la asociación positiva (negativa) entre las variables.

Tabla $\mathbf{p \times q}$ con $\mathbf{p, q > 2}$ para medidas en escala ordinal				
Medida	Valores	Independencia	Asociación perfecta positiva	Asociación perfecta negativa
d de Somers	$-1 \leq \mathbf{d} \leq 1$	0	1	-1
Tau−B (Kendall)	$-1 \leq \mathbf{Tau{-}B} \leq 1$	0	1	-1
Tau−C (Kendall)	$0 \leq \mathbf{Tau{-}C} \leq 1$	0	1	-1

- En tablas no cuadradas la medida **Tau−B** de Kendall no alcanza los límites.
- Si las variables son independientes entonces $\boldsymbol{\gamma} = 0$, sin embargo el recíproco no es cierto.
- Además $|\boldsymbol{\gamma}| = 1$ no implica asociación perfecta.

Bibliografía utilizada:

★ **Abad Montes, F. y Vargas Jiménez, M. (2002).** *"Análisis de datos para las Ciencias Sociales". Ed.: Proyecto Sur.*

★ **Aguilera del Pino, A. M. (2001).** *"Tablas de contingencia bidimensionales". Ed.: La Muralla, S.A.*

★ **Milton, Susan (2001).** *"Estadística para Biología y Ciencias de la Salud". Ed.: Mc Graw-Hill.*

♦ **Temporalización:** Una hora