TEMA9: ASOCIACIÓN ENTRE VARIABLES ESTADÍSTICAS

En el tema ocho se estudió el Coeficiente de correlación lineal, que es un coeficiente que nos indica el grado de asociación entre dos variables cuantitativas.

En este tema vamos a estudiar coeficientes para determinar el grado de asociación de variables cualitativas, es decir, aquellas variables cuyas modalidades no se pueden cuantificar.

Recordemos que a las variables cualitativas también se las denomina atributos, por ello a veces se denomina a este estudio, estudio de asociación entre atributos.

Se denomina tabla de contingencia a la tabla de doble entrada que se representa para una distribución bidimensional de atributos. En realidad, es exactamente una tabla de correlación donde en la primera fila y columna, en lugar de aparecer las cuantificaciones numéricas de las variables, se escriben las definiciones de cada una de las modalidades asociadas a los atributos.

Debe notarse que el concepto de independencia que se estudió para variables cuantitativas, es el mismo para atributos, y que por tanto el concepto de asociación también coincide sólo que los coeficientes que se utilizan para medir la asociación son distintos.

Ejemplo de atributos:

- Sexo: Hombre y mujer.
- Color del pelo: Moreno, castaño, rubio.
- Color de ojos: azul, marrón, negro.

- **Coeficiente de contingencia χ^2**

Estudia la diferencia entre las frecuencias observadas, n_{ij}, y las frecuencias esperadas, e_{ij},

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{(e_{ij} - n_{ij})^2}{e_{ij}}$$

donde $e_{ij} = \frac{n_{i.}n_{.j}}{n}$.

Cuando hay independencia estadística, o no hay asociación entre los atributos,se tiene que $n_{ij} = e_{ij}$, resultando $\chi^2 = 0$.

El coeficiente puede reescribirse de forma que facilite su cálculo,

$$\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{p} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right],$$

donde, si el coeficiente vale cero es que hay independencia entre las variables, y cuanto más grande salga, mayor asociación hay entre las variables.

El mayor inconveniente que tiene este coeficiente es que es proporcional al número de observaciones, y por tanto no tiene una cota, por lo que no es muy adecuado su uso.

Ejemplo: Calcular el coeficiente de contingencia χ^2 para las siguientes tablas de contingencia.

X/Y	Casado	Soltero	Viudo	$n_{i.}$
Hombre	2	0	1	3
Mujer	0	3	4	7
$n_{.j}$	2	3	5	10

X/Y	Casado	Soltero	Viudo	$n_{i.}$
Hombre	20	0	10	30
Mujer	0	30	40	70
$n_{.j}$	20	30	50	100

Son las mismas variables, sólo que en el primer caso $n = 10$ y en el segundo $n = 100$, por lo que debe haber el mismo grado de asociación ya que además las frecuencias relativas coinciden. sin embargo, el coeficiente de contingencia χ^2 toma valores distintos, lo que nos da una idea de su problema.

En la primera tabla se tiene:

e_{ij}			
	0,6	0,9	1,5
	1,4	2,1	3,5

$\left(\frac{e_{ij}-n_{ij}}{e_{ij}}\right)^2$			
	3,27	0,9	0,17
	1,4	0,39	0,07

Luego $\chi^2 = 6{,}2$.

En la segunda tabla se obtiene:

e_{ij}			
	6	9	15
	14	21	35

$\left(\frac{e_{ij}-n_{ij}}{e_{ij}}\right)^2$			
	32,7	9	1,7
	14	3,9	0,7

Luego $\chi^2 = 62$.

- **Coeficiente de contingencia de Pearson**

Se define a partir del coeficiente de contingencia χ^2:

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

Si $C = 0$, hay independencia entre las variables, y si $C = C_{max}$ existe asociación máxima entre las variables.

Este valor máximo, C_{max}, depende de las dimensiones de la tabla de contingencia.

- Si la tabla es cuadrada de dimensión kxk

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

- Si la tabla no es cuadrada, sino de dimensión kxp, entonces sea $h = \text{mín}\{k, p\}$

$$C_{max} = \sqrt{\frac{h-1}{h}}$$

El inconveniente que tiene este coeficiente es que, ya que su cota máxima depende de las dimensiones de la tabla de contingencia, no puede usarse para comparar tablas de dimensiones distintas.

Este problema se resuelve usando el coeficiente de contingencia corregido de Pawlik, el cual varía entre cero y uno:

$$C_c = \frac{C}{C_{max}}$$

Ejemplo: Sobre las tablas anteriores, calcular el coeficiente de contingencia de Pearson.

$$\text{Tabla 1: } C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{6{,}2}{10+6{,}2}} = 0{,}6186$$

$$\text{Tabla 2: } C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{62}{100+62}} = 0{,}6186$$

Mediante ese coeficiente hemos obtenido igual valor para ambas tablas, lo cual es lo deseado ya que ambas tablas contienen datos proporcionales de las mismas variables. Por tanto se ve claramente como este coeficiente mejora al coeficiente de contingencia χ^2.

- **Coeficiente de Tschuprow**

Se define como:

$$T^2 = \frac{\chi^2}{n\sqrt{(k-1)(p-1)}}$$

siendo k y p las dimensiones de la tabla de contingencia. Este coeficiente varía entre cero y uno:

- Si $T^2 = 0$, hay independencia entre las variables bajo estudio.
- si $T^2 = 1$, existe asociación total entre las variables.

Ejemplo: Calcular el coeficiente de Tshuprow para los datos del ejemplo anterior.

$$T^2 = \frac{\chi^2}{n\sqrt{(k-1)(p-1)}} = \frac{6{,}2}{10\sqrt{1*2}} = 0{,}4384$$

$$T^2 = \frac{\chi^2}{n\sqrt{(k-1)(p-1)}} = \frac{62}{100\sqrt{1*2}} = 0{,}4384$$

Existe una asociación intermedia entre ambas variables.

- **Coeficiente de correlación por rangos de Spearman**

Existen experimentos que generan respuestas que no son cuantificables pero que pueden ordenarse. Supongamos que se desea observar el atractivo de cuatro nuevos modelos de automóvil o el sabor de tres nuevas salsas. No es posible, en dichos casos, dar una medida exacta del fenómeno que se desea estudiar, pero si se pueden ordenar las preferencias entre las modalidades.

Las observaciones de este tipo pueden definirse sobre una escala ordinal, dado que la distancia entre dos puntos no es de consecuencia y sólo es importante el orden o rango de los números.

Veamos un método para poner de manifiesto la existencia de una relación monótona (creciente o decreciente) entre dos caracteres a cuyas modalidades se les han asignado rangos u órdenes.

Sean X e Y dos características de interés y supongamos que existe una muestra aleatoria de n pares de valores que consisten sólo en los rangos de X e Y. El coeficiente de correlación del rango de Spearman es el coeficiente ordinario de correlación de la muestra, excepto que ahora se emplean los rangos asignados en lugar de las observaciones de X e Y.

Al igual que el coeficiente de correlación lineal r, el de Spearman, r_s, varía entre -1 y 1:

- Si $r_s = 1$, entonces hay asociación monótona creciente.

- Si $r_s = -1$, entonces hay asociación monótona decreciente.

- Si $r_s = 0$, entonces no hay asociación.

Una expresión para su cálculo rápido es:

$$r_s = 1 - 6\,\frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

donde $d_i = x_i - y_i,\ i = 1, \ldots, n$ y n =número de modalidades.

Ejemplo: En relación a cuatro marcas de automóvil, se ha realizado una encuesta sobre las preferencias de compra, mecánica y diseño, obteniéndose:

Marca	Compra	Mecánica	Diseño
Seat	2	4	3
Renault	1	2	1
Citroen	3	3	2
Opel	4	1	4

Nota: los valores de la tabla no son valores de las variables, sólo es una ordenación de las marcar según preferencia.

Analizar el grado de asociación que hay entre la preferencia de comprar y cada una de las dos facetas consideradas, mecánica y diseño.

a) X = Compra Y = Mecánica n = número de modalidades= número de marcas= 4

X	Y	d_i	d_i^2
2	4	-2	4
1	2	-1	1
3	3	0	0
4	1	3	9

$$r_s = 1 - 6\,\frac{14}{4 * 15} = -0{,}4$$

Existe relación inversa entre la preferencia de compra y mecánica.

b) X = Compra Y = diseño

X	Y	d_i	d_i^2
2	3	-1	1
1	1	0	0
3	2	1	1
4	4	0	0

$$r_s = 1 - 6\,\frac{2}{4 * 15} = 0{,}8$$

Existe relación directa notable entre la preferencia de compra y diseño.