CAPITULO 4: ANALISIS CONJUNTO DE DOS VARIABLES

1. INTRODUCCIÓN

Este tema se centra en el estudio conjunto de dos variables.

Dos variables cualitativas

- Tabla de datos
- Tabla de contingencia
- Diagrama de barras
- Tabla de diferencias entre frecuencias empíricas y teóricas
- Calculo de coeficiente X^2
- Cálculo del coeficiente de contingencia

Dos variables cuantitativas

- Tabla de datos conjuntos
- Diagrama de dispersión
- Cálculo de covarianza
- Cálculo del coeficiente de correlación de Pearson

Además…

Si dos variables cuantitativas están relacionadas linealmente utilizaremos la recta de regresión.

2. CONCEPTOS PREVIOS

Asociación y/o relación entre dos variables: Dos variables están relacionadas entre sí cuando ciertos valores de una de las variables se asocian con ciertos valores de la otra variable.

3. ASOCIACIÓN ENTRE DOS VARIABLES CUALITATIVAS

Recordamos que la variable cualitativa era aquella que estaba medida en una escala nominal o de clasificación (tema 1). Además pueden ser:

Dicotómicas: Cuando solo representan dos categorías
Politómicas: Cuando representan un mayor número

Cuando se dispone de los datos de dos variables cualitativas para todos los sujetos de una muestra, se puede elaborar la **Tabla de contingencia** y su correspondiente **diagrama de barras (página 125).** Los datos de esta tabla son las frecuencias empíricas u observadas y se representan por (n_e)

Ahora tenemos que construir una nueva tabla con las frecuencias teóricas (n_t). Para ello utilizaremos la fórmula:

$$\text{Frecuencia teórica} = n_t = \frac{Totalfila\_x\_totalcolumna}{n}$$

Una vez creada esta segunda tabla (página 126) tenemos que crear una tercera tabla que muestra las diferencias entre la tabla 1 y la tabla 2. Es decir, la **tabla de diferencias** entre las frecuencias empíricas menos las frecuencias teóricas. (página 127)

- Es importante quedarnos con el dato de que la suma de las filas y las columnas de esta tercera tabla siempre es igual a 0, si sale otra cosa es que algo hemos hecho mal.

Una vez que tenemos la tabla debemos interpretarla: La interpretación que hace el libro se basa en analizar los **valores positivos** (8) como fuente de información. (parece ser que los valores negativos no nos aportan información) .Así tenemos un 8 en Sí-V y en No-M. Por lo tanto concluiremos que los varones tienen mayor tendencia a padecer estrés (Sí-V) y las mujeres tiene menos tendencia a padecer estrés (No-M).

Y ahora…

Calculamos un **estadístico X^2**

$$\text{Estadístico } X^2 = \sum \frac{(n_e - n_t)^2}{n_t}$$

n_e = frecuencia empírica
n_t = frecuencia teórica

Para calcular el estadístico no hace falta información nueva, ya que extraemos todos los números de las tablas anteriores.

Sin embargo este estadístico nos da poca información porque desconocemos su límite superior. Sólo sabemos que si nos da valor 0 no hay relación entre las dos variables. Sin embargo si nos da un valor cualquiera como por ejemplo 10,78 (página 128) no sabemos que interpretar ya que el límite podría ser 20, 50 , 100 etc y lo desconocemos. Para resolver este problema se calcula algo que sí que sabemos sus límites y es el **índice o Coeficiente de Contingencia, C.** (da valores entre 0 y 1)

$$\textbf{Coeficiente de contingencia = C} = \sqrt{\frac{X^2}{X^2 + n}}$$

Además del Coeficiente de Contingencia tenemos también que calcular su **máximo** (para posteriormente poder comparar uno con otro)

$$C_{máx} = \sqrt{\frac{k-1}{k}}$$

k = Número de filas y número de columnas (en el ejemplo que vamos a ver a continuación K=2 porque tenemos mismo número de filas (2) que de columnas (2)

Siguiendo el ejemplo del libro, el Coeficiente de contingencia nos da 0,312 y su máximo 0,707. Por lo tanto el coeficiente de contingencia está prácticamente a la mitad de su máximo y por ello diremos que la relación entre las dos variables es de **tipo medio.**

También tenemos el ejemplo de tablas con distinto número de filas y columnas, por lo tanto no podremos calcular el $C_{máx.}$ Y la información la extraeremos directamente de C (ejemplo página 129-130) En este ejemplo el procedimiento para calcular las tablas es el mismo que el explicado en la primera parte, la única diferencia es cuando llegamos a C ya que no podemos calcular su $C_{máx.}$

Para concluir:

Características del Coeficiente C

- Tiene valores entre 0 y 1
- Cuando C = 0 diremos que no existe relación entre ellas
- C = 1 nunca se puede dar

- Cuanto mayor es C, mayor es la relación entre las dos variables y viceversa

- Cuando utilicemos C para comparar la relación entre dos variables cuyos datos tenemos en dos tablas de contingencia diferentes, tenemos que vigilar que tienen el mismo número de filas y de columnas. De lo contrario los valores de C no permiten una comparación válida.

- Cuando existe un valor elevado de C, no podemos afirmar con rotundidad que una de las variables es causa de la otra, ya que puede haber una tercera variable que está relacionando a ambas.

- Cuando la tabla de contingencia tiene igual número de filas que de columnas, podemos estimar un valor máximo que alcanzará C.

4. CORRELACIÓN ENTRE DOS VARIABLES CUANTITATIVAS

Nos presentan una **tabla de datos conjuntos** (página 132)

Lo primero que hacemos es elaborar el diagrama de **dispersión o nube de puntos** (página 133)

Una vez realizado el diagrama y tan sólo observándolo, podemos decir que existe una **relación lineal** en las variables X e Y. Es decir, a valores mayores de X corresponderán valores mayores de Y y viceversa.

Una vez llegados a este punto calculamos 2 índices que nos permiten ponerle números a todo esto que llevamos analizado:

El primero de estos índices es la **covarianza** y hace referencia a la variación conjunta de dos variables.

$$\textbf{Covarianza} = S_{XY} = \frac{\sum X_i Y_i}{n} - \overline{X}\overline{Y}$$

X_i = Valor de la variable X en el caso i
Y_i = Valor de la variable Y en el caso i
$\overline{X}$ = Media de la variable X
$\overline{Y}$ = Media de la variable Y
n = número de casos de la muestra

Si el signo de la covarianza es **positivo**, diremos que existe **relación lineal directa.**
Si el signo de la covarianza es **negativo**, diremos que existe **relación lineal inversa.**

En el ejemplo de la página 134 observamos que la covarianza da 6,4 (signo positivo) por lo tanto se cumple la relación lineal directa que ya habíamos observado en el diagrama de dispersión.

Sin embargo la covarianza tiene un problema y es que no conocemos su rango (de la misma manera que con el estadístico X^2 no sabíamos su límite superior y teníamos que calcular el coeficiente de contingencia) , por lo tanto para la covarianza calcularemos algo llamado **Coeficiente de Correlación de Pearson (r_{xy})**

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

S_X = Desviación típica de la variable X (tema 3)
S_Y = Desviación típica de la variable Y (tema 3)
S_{XY} = Covarianza entre X e Y

Propiedades del coeficiente de Correlación de Pearson

- Toma valores comprendidos entre -1 y +1
- Cuando vale 0 no existe relación lineal entre X e Y
- Cuando vale exactamente +1 o -1 diremos que una variable es una transformación lineal de la otra
- Cuanto mayor es el valor absoluto del coeficiente nos está indicando que la relación lineal entre las dos variables es más fuerte.
- Cuando el signo es positivo, indica que a valores mayores de la variable X, tienden a corresponder valores mayores de la variable Y y a valores menores de la variable X tienden a corresponder valores menores de la variable Y. Es una **relación directa**.
- Cuando el signo es negativo, indica que a valores mayores de la variable X, tienden a corresponder valores menores de la variable Y, y a valores menores de la variable X tienden a corresponder valores mayores de la variable Y. Es una **relación inversa.**

Página 137 y 138: Ejemplos de diagramas de dispersión y nubes de puntos con sus correspondientes explicaciones.

Caso A

- Coeficiente de correlación positivo
- Relación lineal directa bastante clara

Caso B

- Coeficiente de correlación negativo
- Relación lineal inversa

Caso C

- Coeficiente de correlación lineal cercano a 0
- No existe correlación lineal

Caso D

- Coeficiente de correlación lineal cercano a 0
- No existe una relación lineal pero sí existe una relación curvilínea entre las dos variables. (sin embargo el coeficiente de correlación no puede detectar esto por lo tanto diremos que es una de sus limitaciones)

Como hemos dicho antes…

Cuando r_{XY}= +1 o -1 , existe correlación lineal perfecta
Cuando r_{XY}= 0 , existe ausencia total de correlación lineal

¿pero qué pasa cuando tenemos valores intermedios como por ejemplo 0,55?

En ese caso no podemos afirmar que ese valor indica correlación alta o baja ya que dependerá del tipo de datos que estemos analizando

- Será baja si se trata de dos tests similares que estemos aplicando a los mismos sujetos o si tenemos pocos sujetos
- Será alta si se trata de tests bastante diferenciados o si tenemos muchos sujetos.

4. REGRESIÓN LINEAL

Cuando existe relación lineal podemos utilizar la **recta de regresión** para efectuar pronósticos de los valores de una variable a partir de otra variable.

$$Y = a + bX$$

Para hallar la recta tenemos que calcular **a** y **b** con las siguientes fórmulas:

$$b = \frac{n\sum(XY) - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$a = \overline{Y} - b\overline{X}$$

La recta pasa por el punto $\overline{X},\overline{Y}$. Las puntuaciones obtenidas mediante la recta de regresión las denominaremos **puntuaciones pronosticadas.**

A la diferencia entre la:

puntuación real o verdadera Y_i
y su pronóstico Y_i'

lo llameremos **error** y lo representaremos por $\mathbf{E_i}$

$$E = (Y - Y')$$

Propiedades de las puntuaciones pronosticadas y de los errores

- La media de los errores es cero $\overline{E} = 0$
- La media de las puntuaciones pronosticadas coincide con la media de las verdaderas puntuaciones en Y. $\overline{Y}' = \overline{Y}$
- La varianza de las puntuaciones en Y es igual a la suma de la varianza de los pronósticos, más la varianza de los errores:

$$S_Y^2 = S_{Y'}^2 + S_{YX}^2$$

Además también se pueden comprobar las siguientes igualdades:

$$b = r_{XY}\frac{S_Y}{S_X} \qquad r_{XY}^2 = \frac{S_{Y'}^2}{S_Y^2} \qquad 1 - r_{XY}^2 = \frac{S_{YX}^2}{S_Y^2}$$