Índice general

Capítulo 1

Análisis de atributos

1.1. Análisis de la asociación entre atributos

Hasta ahora hemos estudiado la denominada *Estadística de variables*, que incluye las diferentes técnicas para analizar la información disponible acerca de un determinado fenómeno colectivo, cuyos sucesos vienen expresados en términos cuantitativos o numéricos (renta, salarios, precios, notas, etc.)

Sin embargo, cuando esos sucesos vienen referidos a cualidades o características no medibles del fenómeno estudiado (color, nacionalidad, enfermedades, sexo, afiliación política, etc.) dará lugar a lo que definiremos como *Estadística de atributos*. En ella se distinguen los tipos de escala *nominal* (sexo, estado civil, distintas ramas de actividad económica, profesión, ideología política, ..) y *ordinal* (nivel de estudios, estratificación de familias por su capacidad de consumo, nivel de autoestima, ..)

En la Estadística de atributos, bien establecemos un determinado orden o *rango* entre las observaciones, cuando estas son susceptibles de aparecer en una determinada escala ordinal, o bien procedemos al simple recuento de las distintas modalidades en que se divide el atributo o cualidad, cuando la información aparezca en escala nominal.

En este último caso, el carácter numérico surge al efectuar el recuento, obteniéndose de este modo la distribución de frecuencias del atributo correspondiente.

Como ya comentábamos, en estos casos no tiene sentido el empleo de promedios, tales como la media aritmética, y que cuando las observaciones se nos ofrecen en una escala nominal, sólo la *moda* podía utilizarse como medida resumen; y si éstas respondían a una escala ordinal, podría determinarse, además del valor modal, también la mediana.

El problema que ahora se nos plantea es el de estudiar la posibilidad de establecer medidas similares a las de correlación, que se han estudiado anteriormente, para estos casos en los que las variables no son estrictamente métricas.

Cuando los caracteres estudiados pueden ordenarse de acuerdo con una cierta escala, se puede llegar a unos coeficientes de correlación que midan el grado de asociación entre ellos de manera parecida a como lo hicimos para medir la asociación entre variables (caracteres cuantitativos). Estos coeficientes están basados en los rangos u órdenes de las observaciones.

Si las observaciones son nominales, entonces estableceremos los llamados *coeficientes de asociación y contingencia*.

1.2. Asociación entre caracteres nominales

De forma análoga al caso de dos variables, la observación simultánea de dos atributos da lugar a una tabla de doble entrada, en donde n_{ij} indica el número de individuos de la población que poseen conjuntamente las modalidades indicadas en la fila i-ésima y en la columna j-ésima de la tabla de doble entrada. Esta tabla recibe el nombre de **tabla de contingencia**. su representación es la siguiente:

A B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_p	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.p}$	n

Las distribuciones que se refieren a uno sólo de los dos atributos también se denominan *distribuciones marginales*. Éstas están reflejadas en la última fila reseñada para el atributo B, y en la última columna para el atributo A, siendo

$$\sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{p} n_{.j} = n$$

Se dirá que los dos atributos A y B son independientes cuando entre ellos no exista ningún tipo de influencia mútua.

Si tenemos dos atributos A y B, no habrá relación entre ambos atributos si la proporción de individuos que presentan conjuntamente A_iB_j entre los que presentan A_i es la misma para cualquier valor de i y de j, como ya estudiamos.

Con esto queremos señalar que si los dos atributos A y B son estadísticamente independientes, las frecuencias relativas conjuntas serán igual al producto de las marginales

respectivas; esto es

$$n_{ij} = \frac{n_{i.}n_{.j}}{n};\ \forall i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, p.$$

En la práctica, basta que se verifique para $(k-1)(p-1)$ valores de n_{ij}, ya que entonces se verificará para los restantes.

1.2.1. Coeficientes de asociación

Como concepto contrario al de independencia tenemos el de **asociación**.

Se dice que dos atributos A y B están asociados cuando aparecen juntos en mayor número de casos que el que cabría esperar si fuesen independientes.

Según que esa tendencia a coincidir o no coincidir esté más o menos marcada, tendremos distintos grados de asociación. Para medirlos se han ideado diversos coeficientes de asociación, entre los que destacamos los siguientes.

Coeficiente de contingencia χ^2

Si designamos por n_{ij} la frecuencia conjunta de las modalidades A_i y B_j de A y B, respectivamente, y por n'_{ij} la frecuencia teórica que correspondería en el caso de que ambos atributos fueran independientes; esto es,

$$n'_{ij} = \frac{n_{i.}n_{.j}}{n};\ \forall i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, p,$$

y n al total de elementos que se estudian, definimos el **coeficiente de contingencia** χ^2 como

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

Algunos autores lo denominan **cuadrado de contingencia**.

Si desarrollamos este cuadrado, tenemos:

$$\begin{aligned}
\chi^2 = & \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{(n'_{ij}-n_{ij})^2}{n'_{ij}} \\
= & \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{({n'_{ij}}^2+n_{ij}^2-2n'_{ij}n_{ij})}{n'_{ij}} \\
= & \sum_{i=1}^{k}\sum_{j=1}^{p}n'_{ij}+\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}}-2\sum_{i=1}^{k}\sum_{j=1}^{p}n_{ij} \\
= & \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{i.}n_{.j}}{n}+\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}}-2n \\
= & \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}}-n.
\end{aligned}$$

También se suele utilizar la expresión

$$\varphi^2=\frac{\chi^2}{n}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}}-1,$$

denominado **cuadrado medio de contingencia.**

Tanto χ^2 como φ^2 son sumas de cuadrados, con lo que no podrán ser nunca negativos. Si los atributos fuesen independientes, ambos serían nulos, puesto que las frecuencias teóricas coincidirían con las observadas.

Coeficiente de contingencia de Pearson

El coeficiente de contingencia χ^2 y el cuadrado medio de contingencia φ^2 no son muy apropiados para constituir por sí mismos un coeficiente adecuado, dado que sus límites varían en cada caso.

Por este motivo, Pearson propuso el coeficiente que lleva su nombre

$$C=\sqrt{\frac{\chi^2}{n+\chi^2}}=\sqrt{\frac{\varphi^2}{1+\varphi^2}},$$

que se puede transformar en

$$C^2=\frac{\chi^2}{n+\chi^2}=\frac{\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}}-n}{\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}}}=1-\frac{n}{\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}}}.$$

Este coeficiente varía entre 0 y 1, de manera que cuando existe una carencia absoluta de asociación entre los atributos (cuando éstos son independientes), entonces todos los n_{ij} serán iguales a sus respectivos n'_{ij}, con lo que $C = 0$. Cuando los atributos muestren una total asociación entre sí, el coeficiente debería ser igual a 1, pero esto último no se logra nada más que en el caso ideal de infinitas modalidades.

Se puede demostrar que, en el caso de una tabla cuadrada ($k = p$), el límite superior de C es $\sqrt{\frac{k-1}{k}}$, que difiere de la unidad, dependiendo de dicho límite superior, en el caso más general de $k \neq p$, precisamente de estos parámetros.

En cualquier caso, un coeficiente C nos revelará un menor grado de asociación entre los atributos cuanto más próximo esté a 0.

Coeficiente de Tschuprow

Para evitar los inconvenientes del coeficiente C, Tschuprow propuso un coeficiente que depende de χ^2 (más concretamente, de φ^2), del número de filas y columnas, y del total de elementos, n. Este coeficiente se suele denotar por T^2, y se define como

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

que varía entre 0 y 1.

La relación entre los coeficientes C y T^2 se obtiene fácilmente a partir de las expresiones

$$C^2 = \frac{\varphi^2}{\varphi^2 + 1} = \frac{T^2\sqrt{(k-1)(p-1)}}{T^2\sqrt{(k-1)(p-1)} + 1}.$$

$$T^2 = \frac{\varphi^2}{\sqrt{(k-1)(p-1)}} = \frac{C^2}{(1-C^2)\sqrt{(k-1)(p-1)}}.$$

Aunque estos coeficientes son muy útiles y de fácil aplicación, su utilización no debe inducirnos a olvidar métodos más detallados de análisis. Toda tabla de contingencia debe ser examinada con cuidado para observar si presenta particularidades significativas en la distribución de sus frecuencias, antes de comenzar los cálculos de estos coeficientes.

1.3. Asociación entre caracteres ordinales. Correlación por rangos

Sean A_i y B_i los caracteres que representan las observaciones, y consideremos que x_i es el rango o número de orden que le correspondería a A_i si ordenáramos esta característica, con la escala que se determine, de menor a mayor. Análogamente, y_i representaría el rango

de cada B_i (no hay pérdida de generalidad en suponer que las observaciones son de la forma (A_i, B_i), todas con frecuencia 1).

Queremos estudiar el grado de asociación entre los caracteres A y B, basándonos en la concordancia o discordancia de las *clasificaciones por rangos* x_i e y_i.

1.3.1. Coeficiente de correlación por rangos de Spearman

Si designamos por A y B los criterios de ordenación y por x_i e y_i sus rangos correspondientes, el coeficiente de correlación por rangos de Spearman se obtendrá fácilmente a partir del coeficiente de correlación lineal

$$r = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\frac{1}{n}\sum_i (x_i - \bar{x})^2)(\frac{1}{n}\sum_i (y_i - \bar{y})^2)}}.$$

Teniendo en cuenta que tanto x_i como y_i son rangos, se tiene:

$$\sum_i x_i = \sum_i y_i = 1 + 2 + \ldots + n = \frac{1+n}{2}n$$

$$\sum_i x_i^2 = \sum_i y_i^2 = 1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6},$$

y, por tanto

$$\bar{x} = \bar{y} = \frac{\frac{1+n}{2}n}{n} = \frac{1+n}{2}$$

$$\sum_i (x_i - \bar{x})^2 = \sum_i (y_i - \bar{y})^2 = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \frac{n(n+1)(2n+1)}{6} - \frac{(\frac{1+n}{2}n)^2}{n} = \frac{n^3 - n}{12}.$$

Por otra parte, si denotamos por $d_i = x_i - y_i$, y teniendo en cuenta que en este caso $\bar{x} = \bar{y}$, se tiene

$$\begin{aligned} \sum_i d_i^2 = & \sum_i (x_i - y_i)^2 = \sum_i [(x_i - \bar{x}) - (y_i - \bar{y})]^2 = \\ = & \sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2 - 2\sum_i (x_i - \bar{x}) - (y_i - \bar{y}), \end{aligned}$$

de donde

$$\sum_i (x_i - \bar{x}) - (y_i - \bar{y}) = \frac{\frac{n^3-n}{12} + \frac{n^3-n}{12} - \sum_i d_i^2}{2} = \frac{n^3-n}{12} - \frac{\sum_i d_i^2}{2},$$

con lo que el **coeficiente de correlación por rangos**, que notaremos por ρ, será

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)(\sum_i (y_i - \bar{y})^2)}} = \frac{\frac{n^3-n}{12} - \frac{\sum_i d_i^2}{2}}{\sqrt{(\frac{n^3-n}{12})(\frac{n^3-n}{12})}} = 1 - \frac{6\sum_i d_i^2}{n^3-n}.$$

Este coeficiente, también denominado **coeficiente de correlación ordinal** varía entre -1 y 1. Cuando la concordancia entre los rangos es perfecta, entonces $d_i = x_i - y_i = 0$, y, por tanto, $\rho = 1$. Por el contrario, cuando la discordancia es perfecta, entonces los pares de rangos (x_i, y_i) vienen dados por

x_i	1	2	3	$\cdots$	i	$\cdots$	$(n-1)$	n
y_i	n	$(n-1)$	$(n-2)$	$\cdots$	$\{n-(i-1)\}$	$\cdots$	2	1

con lo que

$$\sum_i d_i^2 = \sum_i (x_i - y_i)^2 = \sum_i [(2i-1) - n]^2 = \frac{n^3-n}{3},$$

y, entonces, $\rho = -1$.

Ejemplo:
Los rangos de 5 *estudiantes, según sus calificaciones de Estadística y Álgebra fueron:*

$$(1,3), (2,2), (3,1), (4,5), (5,4).$$

¿Qué relación existe entre las calificaciones de ambas asignaturas?
Solución:
La tabla que recoge la información que necesitamos para hacer los cálculos necesarios a sustituir en la fórmula del coeficiente de correlación por rangos es la siguiente

Alumno	Estad.(x_i)	Álgebra(y_i)	d_i	d_i^2
A	1	3	-2	4
B	2	2	0	0
C	3	1	2	4
D	4	5	-1	1
E	5	4	1	1

$$\rho = 1 - \frac{6\sum_i d_i^2}{n^3 - n} = 1 - \frac{60}{120} = 0{,}5,$$

con la misma interpretación que el coeficiente de correlación lineal r (existe cierta asociación positiva, o concordancia).

1.3.2. Coeficiente τ de Kendall

Un segundo coeficiente de correlación por rangos, que tiene ciertas ventajas sobre el de Spearman es el coeficiente τ **de Kendall**.

Consideremos unos atributos A y B, cuyas modalidades presentan, como en el caso anterior, los rangos x_i e y_i, respectivamente.

Se clasifican las observaciones de acuerdo con el orden natural de los rangos x_i, con lo que se obtiene una nueva secuencia de rangos para las observaciones y_i, que representamos por y_i^*. Así, por ejemplo, si los pares de rangos para las modalidades eran

x_i	3	4	2	1
y_i	3	1	4	2

clasificando los rangos x_i por el orden natural se tiene

x_i	1	2	3	4
y_i^*	2	4	3	1

Nos fijamos ahora en cada y_i^* y sus rangos consecutivos, comparando cada y_i^* con cada uno de los rangos siguientes de manera que si no existe inversión en el orden natural de esos rangos, se asigna un $+1$, y si existe inversión, un -1.

Podemos utilizar para ello la función indicadora

$$i < j \quad \psi_{ij}(y_i^*, y_j^*) = \begin{cases} 1 & \text{si no hay inversión} \\ -1 & \text{si hay inversión} \end{cases}$$

teniendo en cuenta que las comparaciones se hacen siempre con los rangos siguientes y no con los anteriores; es decir, y_1^* se compararía con $y_2^*, y_3^*, \ldots, y_n^*$; el rango y_2^* con $y_3^*, y_4^*, \ldots, y_n^*$; y así hasta llegar a y_{n-1}^*, que sólo se compara con y_n^*.

Se representa por S la suma total de todos los indicadores; es decir:

$$S = \sum_{\substack{i=1 \\ i<j}} \psi_{ij} \quad \text{para} j = 2, 3, \ldots, n.$$

El máximo valor que puede tener S será

$$\text{máx}(S) = (n-1) + (n-2) + (n-3) + \cdots + 2 + 1 = \frac{n(n-1)}{2}.$$

Teniendo esto en cuenta, Kendall definió el coeficiente

$$\tau = \frac{S}{\frac{n(n-1)}{2}},$$

que relaciona S con su valor máximo y que puede interpretarse como la relación por cociente entre el desorden existente entre los rangos y el desorden máximo que pudieran tener entre sí.

El coeficiente τ varía entre -1 y 1 de manera que si $\tau = 1$, la concordancia es total entre los atributos, siendo la discordancia total en el caso de que $\tau = -1$. En el caso en que $\tau = 0$, no existe asociación entre los atributos en estudio.

Por tanto, $|\tau| \leq 1$, siendo el grado de asociación tanto mayor cuanto más próximo esté τ de 1 o -1.

En el ejemplo, si tenemos en cuenta la reclasificación, se tiene:

$$\begin{array}{rcr} \psi_{12}(y_1^*, y_2^*) & = & 1 \\ \psi_{13}(y_1^*, y_3^*) & = & 1 \\ \psi_{14}(y_1^*, y_4^*) & = & -1 \\ \psi_{23}(y_2^*, y_3^*) & = & -1 \\ \psi_{24}(y_2^*, y_4^*) & = & -1 \\ \psi_{34}(y_3^*, y_4^*) & = & -1 \\ \hline S & = & -2 \end{array}$$

Por tanto,

$$\tau = \frac{S}{\frac{n(n-1)}{2}} = \frac{-2}{\frac{4,3}{2}} = -\frac{1}{3} = -0{,}33.$$