Asociación de variables cualitativas nominales y ordinales

Preparado por Luis M. Molinero (Alce Ingeniería) *Abril 2004*
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF

www.seh−lelha.org/stat1.htm

Introducción

Las tablas de contingencia son una de las herramientas más antiguas y conocidas de la estadística, por lo que su utilización rutinaria puede llevar aparejada una cierta despreocupación, que es contraria al cuidado y meticulosidad con el que siempre deben analizarse los datos, sin abandonarnos a la tarea simple de introducir datos en un programa informático y limitarnos a transcribir mecánicamente los resultados obtenidos, sin mayor análisis, restringiendo además nuestra mirada a los resultados con los que estamos familiarizados, y olvidándonos del resto de información que quizás no entendemos.

Este artículo viene motivado por la pregunta que recientemente me planteó un amigo, respecto de la siguiente tabla, producto de una salida de un programa estadístico, y me hacía esta pregunta: ¿Cuál de estos test [Pearson, Likelihood, etc, me indica la probabilidad de que a peor estado Hoehn y Yahr haya más pacientes con fatiga?

Hoehn y Yahr	Fatiga 1=SI o 2= NO por el Neurólogo 1	2	Total
1	1	5	6
1.5	6	4	10
2	38	19	57
2.5	15	8	23
3	24	8	32
4	9	1	10
Total	93	45	138

Pearson chi2(5)	10.5058	Pr = 0.062
likelihood−ratio chi2(5)	10.6180	Pr = 0.060
Cramér's V	0.2759	
gamma	−0.3183	ASE = 0.127
Kendall's tau−b	−0.1831	ASE = 0.074

Tabla 1

De los resultados que aparecen en la tabla, el parámetro que todo el mundo conoce es el denominado **chi2 de Pearson**, que se calcula como:

$$\chi^2 = \sum_{i}^{I}\sum_{j}^{J}(E_{ij} - O_{ij})^2 / E_{ij}$$

donde E representa la frecuencia esperada para una celda, en la hipótesis de independencia entre las variables, y O la frecuencia efectivamente observada. Si existe independencia entre las variables, los valores esperados y observados serán parecidos y el valor del χ^2 será pequeño, mientras que si dicha hipótesis no se cumple, y existe asociación entre las variables, éste valor será tanto mayor. Las frecuencias esperadas bajo la hipótesis de no asociación se calculan considerando fijas las frecuencias denominadas marginales (por representarse en

los márgenes de la tabla), y que corresponden a las frecuencias de las categorías de cada una de las variables ignorando la existencia de la otra variable (totales de filas y columnas). De esta manera dado que en la tabla anterior la proporción de FATIGA=SI ignorando la existencia de la otra variable es 93/138, la frecuencia esperada para la primera celda, puesto que hay 6 sujetos con HY=1 (olvidando la otra variable), bajo la hipótesis de no asociación, se espera que sea 6 x 93/138=4, mientras que la frecuencia que hemos observado es 1.

Pero ¿Qué significado tiene el siguiente resultado que nos presenta la salida del programa con el nombre "***likelihood−ratio chi2***"?

Se trata del resultado de contrastar un modelo estimado para nuestros datos, considerando que no existe asociación entre las variables y que por tanto se puede utilizar una distribución multinomial, y utilizando para la estimación el **método de máxima verosimilitud** (es lo que significa *likelihood*, descrito en otro artículo), comparando el modelo estimado con el modelo que se ajusta perfectamente a nuestros datos.

El chi^2 de Pearson y el chi^2 del cociente de máxima verosimilitud son asintóticamente equivalentes y por ello siempre suelen ser parecidos. Cuando se comuniquen los resultados basta con indicar cualquiera de ellos.

Diferencia entre significación y asociación

Las pruebas de significación del chi^2 permiten contrastar si es razonable pensar que la relación observada entre las variables puede ser simplemente atribuida al azar. En el nivel de significación influye, como en cualquier otra prueba estadística, no sólo la importancia o magnitud de la asociación, sino también el tamaño de la muestra y en ocasiones otros parámetros. Es posible obtener un resultado estadísticamente significativo con una débil asociación, si el tamaño de muestra es suficientemente grande, y viceversa, si la muestra es pequeña una asociación importante puede no llegar a ser estadísticamente significativa. Esto es algo que es de dominio común, y es universalmente aceptado en cualquier otra prueba estadística que nunca se debe presentar únicamente un valor de P, sino que éste debe acompañar a algún parámetro que exprese la magnitud del resultado, o mejor aún un intervalo de confianza para el efecto observado. Sin embargo esto, que es práctica habitual en el resto de pruebas estadísticas, no se lleva a cabo con las pruebas de asociación en tablas de contingencia, salvo que éstas sean 2x2, en cuyo caso se suele presentar como medida de la asociación alguna medida relativa como el odds ratio o el riesgo relativo, o bien una diferencia de proporciones.

¿A qué se debe que en tablas de contingencia de más de 2 filas o 2 columnas se indique casi siempre sólo el nivel de significación? Es debido probablemente a que no existe un único índice claro e intuitivo que permita cuantificar esa asociación cuando las variables que intervienen son de tipo nominal. En el enlace [1] se puede consultar una descripción de los coeficientes de asociación más empleados para variables cualitativas nominales. Entre ellos quizás el más utilizado es el denominado **V de Cramér**, cuyo valor puede ir desde 0 (no existe relación entre las variables) hasta 1 (asociación perfecta).

Vemos que en la tabla anterior el nivel de significación es 0.06 (no llegamos a rechazar la hipótesis de independencia al nivel 0.005) y que el coeficiente de asociación de Cramér es aproximadamente 0.3.

Hagamos ahora un pequeño ejercicio. Si multiplicamos por 2 todas frecuencias de la tabla 1, obtenemos esta otra tabla

	Fatiga=SI	Fatiga=NO	Total
1	2	10	12
1.5	12	8	20
2	76	38	114
2.5	30	16	46
3	48	16	64
4	18	2	20
Total	186	90	276

Tabla 2

cuyo valor del contraste de chi2 de Pearson es ahora 21, que para 5 grados de libertad, corresponde a un nivel de significación de 0.0008 (antes estábamos solo en 0.06). Sin embargo el coeficiente V de Crámer sigue siendo de 0.3: el grado de asociación que se manifiesta en esta tabla es el mismo que el de la tabla 1 (lógico tal y como la hemos construido); únicamente aporta mayor evidencia al ser una muestra con el doble de sujetos.

El coeficiente de asociación de Cramér se calcula muy fácilmente como:

$$V = \sqrt{\frac{\chi^2}{n(q-1)}}$$

donde n es el tamaño de la muestra y q el mínimo del nº de filas o del nº de columna.

Otro de los inconvenientes de fijarnos únicamente en el valor de P es que tampoco aporta ninguna información sobre la naturaleza de la asociación.

Si comparamos los valores de las frecuencias observadas y esperadas para cada celda, podemos analizar la estructura de esa asociación. Las frecuencia de cada celda pueden considerarse como recuentos que siguen una distribución de probabilidad de Poisson, cuya varianza es igual al valor esperado (media), por lo que las diferencias (f.observada − f.esperada) tienden a ser mayores en aquellas celdas en las que el valor de la frecuencia esperada es grande. Interesa por tanto analizar la magnitud relativa de esas diferencias, denominadas **residuo de Pearson**, que se calculan como:

$$\text{residuo}_{ij} = \frac{\text{f.obs}_{ij} - \text{f.esp}_{ij}}{\sqrt{\text{f.esp}_{ij}}}$$

O incluso mejor aún es calcular los denominados **residuos estandarizados de Pearson**, que tienen la interesante propiedad de distribuirse asintóticamente como una Normal(0,1) y que se calculan como:

$$\text{res.est.}_{ij} = \frac{\text{f.obs}_{ij} - \text{f.esp}_{ij}}{\sqrt{\text{f.esp}_{ij}(1-p_{i+})(1-p_{+j})}}$$

donde p_{i+} corresponde a la proporción de la fila i, y p_{j+} a la proporción de la columna j.

Puesto que éstos residuos estandarizados se distribuyen según una Normal(0,1), valores que excedan 2 o 3 en valor absoluto indican claramente que la frecuencia de esa celda no se ajusta a un modelo en el que se supone independencia entre las variables, y nos ayudan a analizar la estructura de la asociación.

Para la tabla 1 se obtienen los siguientes residuos estandarizados:

	Fatiga	
HY	**SI**	**NO**
1	−2,71	2,71
1.5	−0,52	0,52
2	−0,15	0,15
2.5	−0,24	0,24
3	1,05	−1,05
4	1,58	−1,58

Residuos estandarizados
Tabla 3

donde vemos que los mayores residuos se presentan en los extremos superiores e inferiores de la tabla.

En este caso, por tratarse de una tabla en la que una de las variables es binaria, también podíamos haber analizado la estructura de la asociación representando, por ejemplo, la proporción de cada categoría de la variable FATIGA para cada nivel de la escala HY:

	Fatiga	
HY	**SI**	**NO**
1	0,17	0,83
1.5	0,60	0,40
2	0.67	0,33
2.5	0,65	0,35
3	0,75	0,25
4	0,90	0,10
Total	0,67	0,33

Proprociones de FATIGA=SI,NO para cada estrato HY
Tabla 4

La proporción media de FATIGA=SI en nuestra muestra es de 0.67, proporción que sin embargo es muy inferior en el estrato HY=1 donde solo llega al 0.17, y mucho mayor en el estrato HY=4 donde es el 0.9.

Sin embargo en tablas con variables no binarias es donde el análisis de los residuos estandarizados puede ser de gran ayuda para investigar la naturaleza de la asociación.

En la siguiente tabla se representa la frecuencia de 4 pautas terapéuticas en 4 hospitales diferentes, con el fin de investigar si existe asociación.

	Hospital				
Terapia	**H1**	**H2**	**H3**	**H4**	**Total**
A	166	50	71	156	443
B	157	65	129	240	591
C	100	44	109	211	464
D	88	36	104	179	407
Total	511	195	413	786	1905

Tabla 5

Para esta tabla se obtiene un valor de chi2 = 46.3 ($p < 0.001$). Aunque el nivel de significación es alto, sin embargo el coeficiente de asociación de Cramér es pequeño 0.09.

Para comparar las frecuencias observadas con las esperables bajo la hipótesis de no asociación, representamos la tabla de residuos estandarizados, donde vemos fácilmente donde se producen las mayores discrepancias.

	H1	**H2**	**H3**	**H4**
A	5,77	0,83	−3,30	−2,95
B	−0,17	0,74	0,11	−0,39
C	−2,95	−0,64	1,09	2,12
D	−2,67	−1,04	2,14	1,26

Residuos estandarizados
Tabla 6

Como antes, además de esta tabla de residuos también puede ser bueno como complemento para ayudar a la interpretación, representar la proporción de cada terapia por hospital (esta vez se indican las proporciones en porcentaje %)

	H1	**H2**	**H3**	**H4**	**Total**
A	32,5	25,6	17,2	19,8	23,3
B	30,7	33,3	31,2	30,5	31,0
C	19,6	22,6	26,4	26,8	24,4
D	17,2	18,5	25,2	22,8	21,4

Distribución por hospitales de cada terapia
Tabla 7

Variables ordinales

Cuando una o las dos variables que intervienen en una tabla de contingencia son ordinales, es razonable pensar que si existe asociación ésta evolucione con ese orden. Se puede utilizar esa característica para efectuar contrastes más eficientes, para cuantificar mejor la asociación cuando ésta existe, y para ayudar a su interpretación.

En la tabla 1 la variable que se representa en las filas (escala HY) es precisamente una variable ordinal.

En el enlace [2] se puede consultar una relación de índices de asociación para variables ordinales. Uno de ellos aparece en los resultados de la tabla 1, denominado **gamma de Goodman−Kruskal**. El recorrido de valores posibles para gamma va de −1 a 1, y el valor de nuestro ejemplo de −0.32, indica que existe cierta asociación en el sentido de que al aumentar la categoría del HY aumenta la probabilidad de que el paciente esté clasificado como FATIGA=SI (por ser SI la primera categoría el coeficiente es negativo).

Cuando cruzamos una variable ordinal y una variable binaria disponemos de otras alternativas, sobre todo si a la variable ordinal se le puede asignar una puntuación numérica con cierto sentido, como es el caso de nuestro ejemplo donde precisamente el valor HY es una escala que nos proporciona un valor numérico. Podemos entonces tratar la otra variable, la binaria, como variable respuesta y analizar los datos mediante un **modelo de regresión**, que habitualmente será el **modelo logístico**.

Otra alternativa con una interpretación interesante consiste en utilizar un **contraste de Mann−Whitney**, descrito en un artículo anterior.

Empleando cualquiera de estas dos últimas alternativas (modelo de regresión y prueba de Mann−Whitney), en la tabla 1 comprobaríamos que el resultado sí es entonces estadísticamente significativo al nivel de 0.05, debido a que al considerar el orden en la variable HY hemos actuado con más eficiencia.

Otra posibilidad se basa en ajustar un modelo log−lineal para los datos en el que se tenga en cuenta esa posible asociación entre variables ordinales. Si en la hipótesis de independencia el modelo estima la frecuencia como:

$$\ln F_{ij} = \lambda + \lambda_i^f + \lambda_j^c$$

es decir que en el valor de la frecuencia de una celda incluye un valor medio, la fila y la columna.

Podemos considerar un modelo alternativo que considera la presencia de asociación lineal debida al valor x_i puntuación para la categoría de la fila, y_j puntuación para la categoría de la columna:

$$\ln F_{ij} = \lambda + \lambda_i^f + \lambda_j^c + \beta x_i y_j$$

Contrastando este último modelo frente al anterior podemos evaluar si la asociación representada por el término β es o no estadísticamente significativa.

** Agradezco a mi amigo el Dr. Pablo Martínez los datos de la tabla 1 y la pregunta que han motivado este artículo*

Referencias

- [1] Nominal Association: Phi, Contingency Coefficient, Tschuprow's T, Cramer's V, Lambda, Uncertainty Coefficient
- [2] Ordinal Association: Gamma, Kendall's tau−b and tau−c, Somers' d

G. David Garson
Professor of Public Administration
Editor, *Social Science Computer Review*
College of Humanities and Social Sciences
Box 8102, 206 1911 Building
North Carolina State University
Raleigh, North Carolina 27695

Indice de artículos

Principio de la página