Análisis de tablas de contingencia de más de 2 variables cualitativas

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

Abril 2003

Artículo en formato PDF

www.seh−lelha.org/stat1.htm

Introducción

Denominamos **variables cualitativas** a aquellas cuyo resultado es un valor o categoría de entre un conjunto finito de respuestas posibles. El *sexo*, el *estado civil* o el *grupo sanguíneo* son ejemplos de variables cualitativas. Cuando se analizan variables cualitativas es habitual representar en tablas las frecuencias de casos observados para cada una de las diferentes categorías de las variables, las cuales se denominan **tablas de contingencia**.

En la siguiente figura se presenta un ejemplo de tabla de contigencia para dos variables: en las filas se encuentra la variable *ESTUDIOS*, clasificada según tres categorías, y en las columnas representamos la variable *HTA*, según la clasificación propuesta en el documento The VI Report of the JNC on Prevention, Detection, Evaluation and Treatement of High Blood Pressure. Los datos corresponden a un conjunto de pacientes diabéticos.

Tabla 1

	HTA II a IV *	HTA I	Normal alta	Normal	Optima	Total
Sin estudios	30	117	94	49	11	301
1º grado	39	123	110	60	21	353
2º y 3º grado	11	35	58	39	24	167
Total	80	275	262	148	56	821

** Los niveles de HTA II a IV se han agrupado en una sola categoría.*

En este tipo de tablas habitualmente se desea conocer si existe **asociación** entre las dos variables, o si por el contrario se pueden considerar independientes. Dicho de otra forma: queremos saber si la proporción de casos para cada categoría de una de las variables es independiente del valor que toma la otra variable. En la tabla del ejemplo nos interesa saber si la proporción de sujetos en cada una de las categorías de la variable *HTA* es diferente según el nivel de estudios o si, por el contrario, se pueden considerar independientes.

El razonamiento para contrastar si existe o no asociación entre dos variables cualitativas se basa en calcular cuál serían los valores de frecuencia esperados para cada una de las celdas en el caso de que efectivamente las variables fuesen independientes, y compararlos con los valores realmente observados. Si no existe mucha diferencia entre ambos, no hay razones para dudar de que las variables sean independientes.

En el ejemplo, la proporción de pacientes con *HTA nivel I* en nuestra muestra es

$$275 / 821 = 33.5\%$$

Si las variables son independientes esta proporción debiera mantenerse (al menos de forma aproximada) en cada nivel de estudios. Así como tenemos 167 pacientes con *estudios de 2º o 3º grado*, el número de casos esperado con *HTA nivel I* es

167 x 0.335 = 55.9

mientras que el valor observado es sólo 35.

De forma general la frecuencia esperada para cada una de las celdas, cumpliéndose la hipótesis de independencia, se calcula multiplicando el total de la fila por el total de la columna correspondientes, y dividiéndolo por el tamaño global.

El contraste estadístico más utilizado para evaluar si las diferencias entre las frecuencias observadas y las esperadas pueden atribuirse al azar, bajo la hipótesis de independencia, es el denominado chi² de Pearson:

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

donde F_{ij} representa la frecuencia esperada para la celda situada en la fila *i* columna *j*, y f_{ij} representa la frecuencia efectivamente observada para esa celda. En la hipótesis de independencia este estadístico se distribuye de forma aproximada según una chi² con grados de libertad *(I−1)(J−1)*, siendo *I* el número de filas y *J* el número de columnas.

El estudio de la asociación entre dos variables cualitativas en ocasiones puede ser insuficiente, ya que la presencia de una tercera variable puede modificar las conclusiones respecto a esa asociación, e incluso puede interesar evaluar la influencia de más variables adicionales. En el ejemplo anterior si se calcula el valor del chi² obtenemos 35.6, que con 8 grados de libertad corresponde a un valor de probabilidad de 0.00002, lo que indica que los datos obtenidos están en clara contradicción con la hipótesis de independencia y debemos por lo tanto concluir, a partir de la evidencia de nuestros datos, que existe asociación entre el grado de *HTA* y el *nivel de estudios* de los pacientes.

Sin embargo, por las características sociales de nuestro país, sabemos que las personas de edad avanzada no tienen el mismo perfil educativo que las más jóvenes, siendo en general su nivel de estudios inferior. Si, por otro lado, la prevalencia de la *HTA* aumenta con la *edad*, pudiera ser que la asociación observada se explique porque en las categorías con nivel de estudios inferiores se encuentran más personas de edad avanzada, mientras que en las categorías con mayor nivel de estudios tengamos predominio de personas jóvenes. Por ello nos puede interesar en este caso incluir en nuestro análisis una tercera variable que recoja la *edad del paciente*, y que vamos a clasificar en los siguientes intervalos: Hasta 50 años, entre 50 y 64, entre 65 y 74, más de 74 años.

Puesto que ahora tenemos tres variables, *HTA, ESTUDIOS y EDAD*, vamos a ver cómo podemos analizar de forma conjunta la asociación entre tres o más variables cualitativas y para ello empezaremos con un poco de teoría. Estimado lector no se asuste porque encuentre en el texto fórmulas con letra grande, lo razón es para que se vean mejor en el navegador, y además son fórmulas muy sencillas, donde lo más complicado que encontramos es la presencia de algún logaritmo.

Modelos log−lineales para tablas de contingencia

Vamos a empezar nuestro razonamiento a partir de una tabla para dos variables con *I* filas y *J* columnas. Llamamos π_{ij} a la proporción total de sujetos clasificados como i en las filas y j en las columnas, y

π_{i+} a la proporción total de sujetos en la categoría i para la variable filas (olvidándonos de la otra

variable, como si no existiera, es decir juntando todas las columnas) y π_{+j} a la proporción total de sujetos en la categoría j para la variable columnas (juntando todas las filas). Hemos visto más arriba que bajo la hipótesis de independencia la proporción de cada celda se estima como:

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

Y por lo tanto para estimar la frecuencia F_{ij} multiplicaremos por el tamaño global N

$$F_{ij} = N\pi_{ij} = N\pi_{i+}\pi_{+j}$$

Si tomamos logaritmos obtenemos:

$$\ln F_{ij} = \ln N + \ln \pi_{i+} + \ln \pi_{+j}$$

Vamos a llamar *X* a la variable representada en las filas, *Y* a la variable correspondiente a las columnas. La fórmula anterior, que representa el modelo matemático para estimar la frecuencia de cada celda en la hipótesis de independencia, la reescribimos entonces como sigue:

$$\ln F_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

donde cada sumando se corresponde directamente con los de la fórmula anterior.

En este modelo podemos introducir un tercer término para considerar la presencia de asociación, y tendremos entonces un modelo en el que ya no se cumpliría la hipótesis de independencia:

$$\ln F_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

Por lo tanto la hipótesis de independencia es equivalente a plantear

$$\lambda_{ij}^{XY} = 0$$

Si estuviéramos analizando tres variables, añadimos una tercera con nombre *Z*, podemos generalizar el modelo anterior:

$$\ln F_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

Este tipo de modelos se conoce con el nombre de **modelos log−lineales**.

Antes de volver con nuestro ejemplo, en el que analizábamos las variables *hipertensión*, *nivel de estudios* y *edad*, vamos a plantear otro ejemplo también con 3 dimensiones, pero más sencillo en cuanto al número de categorías de cada variable, en la que se estudia la proporción de mujeres admitidas en una determinada Universidad frente a la proporción de hombres:

Tabla 2

	NO ADMITIDOS	**SI ADMITIDOS**	**% ADMITIDOS**	Total
HOMBRE	1493	1198	45	2691
MUJER	1278	557	30	1835
Total	2771	1755		4526

** Datos tomados del libro de Powers y Xie, correspondientes a un estudio realizado en la Universidad de California−Berkeley (Bickel et al. 1975; Freedman, Pisani, and Purves 1978)*

En base a estos datos se plantea una demanda a la citada Universidad acusándola de sexista en las pruebas de admisión, que parecen favorecer claramente a los hombres. Sin embargo los responsables de la Universidad presentaron los datos distribuidos por facultades (vamos a denominar a las diferentes facultades de forma genérica con las letras *A* hasta *F*):

Tabla 3

HOMBRES	**Total presentados**	**% Admitidos**
A	825	62
B	560	63
C	325	37
D	417	33
E	191	28
F	373	6
Total	2691	
MUJERES	**Total presentadas**	**% Admitidas**
A	108	82
B	25	68
C	593	34
D	375	35
E	393	24
F	341	7
Total	1835	

donde se puede ver que apenas hay diferencias en las tasas de admisión, salvo en la facultad A ¡donde el porcentaje de hombres admitidos es del 62% y el de mujeres es del 82%! Tenemos aquí un ejemplo claro de cómo la asociación entre dos variables cualitativas resulta ser espuria cuando se considera los valores de una tercera variable, situación que se conoce como Paradoja de Simpson y que también se puede dar en variables cuantitativas.

Para analizar este tipo de tablas multidimensionales utilizaremos los modelos log−lineales planteados más arriba. Para tres variables *X,Y,Z* podemos ajustar a nuestros datos diferentes modelos:

Tabla 4

Modelo **(X,Y,Z)**: todas las variables son mutuamente independientes, X Y son idependientes, X Z son independientes, Y Z son independientes, no existiendo asociación entre ellas, por lo que el modelo queda reducido a $$\ln F_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$
Modelo **(X,YZ)**. En este modelo sólo se considera la asociación YZ. X es independiente de las otras dos variables $$\ln F_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$$
Modelo **(XY,YZ)**. X es independiente de Z para cada valor de Y. $$\ln F_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$$
Modelo **(XY,YZ,XZ)**. Existe asociación dos a dos entre todas las variables, pero no se considera asociación conjunta entre las tres, de tal manera que la asociación entre dos de las variables es homogénea, no cambia, para cada nivel de la otra variable. $$\ln F_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$
Modelo **(XYZ)**. Si el modelo anterior no se ajusta bien a los datos quiere decir que hay que considerar la asociación de las tres variables, de tal manera que la asociación entre dos de ellas no es homogénea cuando cambia el nivel de la otra variable. $$\ln F_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

Selección del modelo

Procederemos a ajustar diferentes modelos a nuestros datos y ver cuál se adecua mejor a los valores observados.

Para contrastar dos modelos diferentes se utiliza el estadístico denominado **cociente de verosimilitud** (likelihood ratio), que se calcula como:

$$G^2 = 2\sum f \ln(f / F)$$

donde *f* es la frecuencia observada y *F* la frecuencia esperada según el modelo. Este estadístico se distribuye según una chi² en la hipótesis de que el modelo es correcto, con grados de libertad que dependen de los parámetros utilizados para ajustar el modelo. En la salida de los programa de estadística ingleses se suele presentar también como **Deviance** (**desviación respecto al modelo observado**, también conocido como

modelo saturado, porque en él se incluyen todos los términos de asociación posibles y se ajusta por tanto perfectamente a los datos observados). Valores elevados de G^2 reflejan un mal ajuste del modelo a los datos, lo que corresponderá a un valor de probabilidad bajo.

Ajustando diferentes modelos para nuestro ejemplo de admisión en la Universidad obtenemos los valores recogidos en la siguiente tabla:

Tabla 5

Modelo	**G^2**	**gl**	**P**
S,F,A	2097.7	16	0.0000
SF,A	877.1	11	0.0000
SF,AF	21.7	6	0.0014
SF,AF,SA	20.2	5	0.0011

S=Sexo, **F**=Facultad, **A**=Admitido

Como vemos el primer modelo, que supone independencia entre todas las variables, se ajusta muy mal a los datos obtenidos en nuestro estudio. En el segundo modelo, para el que la admisión A es independiente de las otras variables, que sí se consideran asociadas (Sexo y Facultad), el ajuste mejora considerablemente, y mucho más con el tercer modelo, aunque las diferencias de las frecuencias estimadas con respecto a las observadas todavía resultan estadísticamente significativas en el contraste G^2 frente al modelo saturado (p=0.0014).

Volvamos ahora sobre el primer ejemplo que planteamos al comienzo del artículo −pacientes diabéticos−, en el que se consideraban las variables *E: Edad, C: Nivel cultural e H: hipertensión*, y calculemos los estadísticos de ajuste correspondientes a diferentes modelos posibles

Tabla 6

Modelo	**G^2**	**gl**	**P**
E,C,H	232.2	50	0.0000
EC,H	108.2	44	0.0000
HE,CE	34.9	32	0.33

*Nota: Hemos llamado a la variable ESTUDIOS con la letra **C**, de Nivel Cultural, aunque no es lo mismo, porque la letra E la hemos tomado para la EDAD*

Vemos que el modelo **HE,CE** se ajusta bastante bien a nuestros datos (p=0.33), por lo que podemos considerar que el nivel de hipertensión en nuestra muestra NO se asocia con el nivel cultural en cada grupo de edad. De nuevo todo lo contrario de lo que se concluía si se consideraba únicamente las variables hipertensión y nivel cultural.

Indice de discrepancia

Volviendo otra vez sobre el ejemplo de Admisión en la Universidad, vemos en la tabla 5 que aunque el modelo *(SF,AF)* mejora considerablemente el ajuste frente al modelo de independencia de todas las variables *(S,F,A)*, sin embargo sigue siendo estadísticamente significativo el contraste con respecto al modelo saturado (frecuencias observadas). En este ejemplo estamos trabajando con una muestra bastante grade, de *4526* sujetos y en estos casos puede ocurrir que diferencias de escasa importancia práctica, sin embargo resulten estadísticamente significativas. Si representáramos los valores de frecuencias estimados por este modelo y los observados, veríamos que no hay grandes diferencias (no los presentamos aquí para no alargar la extensión de este artículo).

Buscando alguna otra alternativa para valorar el ajuste, podemos definir un **índice de discrepancia** (**dissimilarity index**)

$$ID = \sum_i |f_i - F_i| / 2N = \sum_i |p_i - \pi_i| / 2$$

que corresponde a la media de las diferencias entre las frecuencias observadas y las previstas por el modelo, en valor absoluto. Este índice puede tomar valores entre 0 y 1, correspondiendo los valores más pequeños a un mejor ajuste del modelo a los datos.

También podemos interpretarlo como la proporción de casos a los que hay que cambiar la clasificación para obtener un ajuste perfecto. Valores inferiores a 0.02 o 0.03 reflejan un buen ajuste.

En el ejemplo de admisión a la universidad, con el modelo *(SF,AF)*, en el que las variables *SEXO* y *ADMISION* son independientes para cada categoría de la variable *FACULTAD*, se obtiene un índice de discrepancia de 0.016; lo que quiere decir que con este modelo hay que cambiar la clasificación de menos del 2 % de los sujetos para obtener un ajuste perfecto; lo que ilustra claramente que en este ejemplo la significación estadística está detectando diferencias entre el modelo y los datos de poca importancia práctica, debido al gran tamaño de muestra del estudio (4526 sujetos).

Comparación entre dos modelos

Podemos comparar dos modelos log−lineales calculando la diferencia entre los valores de G^2 obtenido, lo que equivale al cociente de verosimilitud, y se distribuye aproximadamente como una chi^2 con grados de libertad igual a la diferencia entre los grados de libertad de los modelos. Así en el ejemplo de la Admisión en la Universidad, si queremos contrastar si existen diferencias significativas en el ajuste entre los modelos *(SF,AF)* y *(SF,AF,SA)*, con los datos de la tabla 5 calculamos:

$$G^2(SF,AF) - G^2(SF,AF,SA) = 21.7 - 20.2 = 1.5$$

que para una chi^2 con 6−5=1 gl corresponde a un nivel de probabilidad de 0.22, y por lo tanto podemos concluir que no existen diferencias significativas entre los ajustes logrados con ambos modelos.

Criterio Bayesiano de Información

Como se ha comentado en el párrafo anterior, la utilización únicamente del valor de G^2 como medida de bondad de ajuste favorecerá, cuando el tamaño de muestra sea grande, la elección de modelos complejos, con muchos términos de asociación y no nos permitirá distinguir entre una verdadera mejora del ajuste respecto de una mejora trivial. Además de la posibilidad de consultar el índice de discrepancia, también se ha propuesto utilizar el estadístico denominado **Criterio Bayesiano de Información**, que en las salidas de los programas se suele denominar **BIC** (**Bayesian Information Criterion**). La fórmula para su cálculo es la siguiente:

$$BIC = G^2 - gl \cdot \ln N$$

Aunque no vamos a profundizar en el razonamiento, se fundamenta en comparar la plausibilidad de dos modelos frente a simplemente comparar las diferencias absolutas del ajuste.

Cuando comparamos dos modelos un valor inferior del BIC indica un mejor modelo según ese criterio.

Si repetimos la tabla 5 para el ejemplo de Admisión en la Universidad añadiendo una columna con el BIC, tenemos:

Tabla 7

Modelo	G^2	gl	P	BIC
S,F,A	2097.7	16	0.0000	1958.0
SF,A	877.1	11	0.0000	779.5
SF,AF	21.7	6	0.0014	−29.4
SF,AF,SA	20.2	5	0.0011	−21.9

por lo que de acuerdo al criterio BIC nos decidiríamos por el modelo *(SF,AF)*, modelo que por otra parte es más fácil de interpretar.

Otras alternativas para analizar tablas de contigencia de más de dos variables

En ocasiones cuando se estudia la asociación entre variables cualitativas, una de ellas puede considerarse como **variable respuesta** y las otras como variables o factores explicativos de la respuesta. Los modelos log−lineales tratan todas las variables de forma simétrica, no distinguiendo entre variable respuesta y el resto de variables, por ello en el caso de que claramente se identifique una variable como respuesta puede ser más natural utilizar modelos logísticos, los cuales describen esa dependencia. Su utilización es más habitual sobre todo si la variable respuesta es **dicotómica** (dos categorías), ya que el modelo logístico nos permite cuantificar la asociación mediante los **odds ratio** correspondientes, que se pueden estimar directamente del modelo (aunque también puede ser estimados a partir del resultado del modelo log−lineal pero de forma un poco más compleja).

Para calcular el modelo logístico para los datos de Admisión en la Universidad en el que la variable *ADMISION* puede considerarse como variable respuesta y es además dicotómica, debemos previamente codificar la variable *FACULTAD* como variable "dummy".

En la siguiente tabla se indica el resultado obtenido al ajustar un modelo logístico en el que además se ha introducido un **término de interacción** entre las variables SEXO y FACULTAD (equivalente al concepto de **asociación** en el modelo log−lineal).

Tabla 8

Término	Coeficiente	Err.estándar	chi^2	P
Constante	0,4921	0,0717	47,044	0
SEXO	1,0521	0,2627	16,038	0,0000621
FACULTAD *			514,756	0,0000
FACULTAD 1	0,0416	0,1132		
FACULTAD 2	−1,0276	0,1355		
FACULTAD 3	−1,1961	0,1264		
FACULTAD 4	−1,4491	0,1768		
FACULTAD 5	−3,2619	0,2312		
SEXO,FACULTAD *			20,204	0,0011
SEXO,FACULTAD 1	−0,8321	0,5104		
SEXO,FACULTAD 2	−1,1770	0,2996		
SEXO,FACULTAD 3	−0,9701	0,3026		
SEXO,FACULTAD 4	−1,2523	0,3303		
SEXO,FACULTAD 5	−0,8632	0,4027		

A partir de este modelo, que tiene en cuenta la *FACULTAD* en la que se presenta la solicitud de admisión, el odds ratio favorece a las mujeres frente a los hombre con un valor de 2.86 (Int. confianza 95 % de 1.71 a

4.79), lo que contrasta con el valor de 0.54 (Int. confianza 95% de 0.48 a 0.62) obtenido sin tener en cuenta la facultad, que favorecía a los hombres.

Cuando la variable considerada como respuesta es **politómica** (más de dos categorías), como puede ser el caso de la HTA en el primer ejemplo, donde la variable respuesta sería la clasificación de nivel de hipertensión, una posible alternativa a utilizar es el modelo logístico para variables politómicas, sin embargo en estos casos a veces son más sencillos de interpretación los modelos log−lineales.

Cuando no hay claramente una variable respuesta o, por el contrario, son más de una las variables que pueden ser consideradas como respuesta, los modelos log−lineales son la alternativa adecuada para el análisis de tablas de contingencia multidimensionales.

En todo este artículo se han tratado las variables cualitativas como nominales, es decir que las respuestas son meras clasificaciones con un nombre o una etiqueta, no considerándose que exista ninguna relación de orden entre las distintas respuestas. Esto es así en variables como *SEXO* o *FACULTAD*, en uno de nuestros ejemplos, pero sin embargo en el ejemplo de la HTA sí que existe una relación ordinal en las variables *HIPERTENSION* y *NIVEL DE ESTUDIOS*, clasificación ordinal que puede ser tomada en cuenta a la hora de elaborar nuestro modelo, de tal forma que mejore el ajuste del mismo a los valores observados; pero esto será objeto de otro artículo.

Bibliografía seleccionada

- Agresti A. **An Introduction to Categorical Data Analysis**.
  New York: John Wiley&Sons, 1996.
- Powers DA, Xie Y. **Statistical Methods for Categorical Data Analysis**.
  San Diego: Academic Press, 2000

Enlaces de interés

- **LEM**
  Se trata de un excelente programa gratuito que permite todo tipo de análisis para variables cualitativas (analysis of categorical data), incluidos los modelos log−lineales.

  Enlace a la página de descarga (gratuita) del programa LEM, manual y ejemplos.

  Autor:

  - Dr. Jeroen Vermunt
    Department of Methodology
    Faculty of Social Sciences
    Tilburg University
    P.O. Box 90153 5000 LE Tilburg
    THE NETHERLANDS

Indice de artículos

Principio de la página