SUGERENCIAS METODOLOGICAS

Revista Investigación Educativa - Vol. 4 - n. 8 - 1986 (P.141-147)

ANÁLISIS DE TABLAS DE CONTINGENCIA MULTIDIMENSIONALES

por
Antoni Sans i Martín

En el campo de la investigación educativa son frecuentes las situaciones en las que la información de la que disponemos tiene, total o parcialmente, naturaleza cualitativa. Es corriente que tras laboriosas tareas y costosos procedimientos, especialmente en la aplicación de encuestas, se consiga disponer de información muy valiosa que normalmente es tratada con procedimientos de análisis de las frecuencias en tablas de contingencia bidimensionales con la técnica de X^2 de Pearson, con Y^2 razón de verosimilitud o técnicas parecidas, siguiendo como proceso último el control por terceras variables que se consideren por algún motivo especialmente importantes.

Afortunadamente en la actualidad (Goodman, L. A., 1884; Davis, J. A., 1985 y Bisquerra, R., 1988) podemos avanzar algo más en el análisis de los datos cuando disponemos de tablas de contingencia multidimensionales.

Para ver algún principio básico y comentar alguna de estas posibilidades, emplearemos una sencilla tabla de contingencia de tres dimensiones con valores hipotéticos:

Variables	*Etiquetas variables*	*Categorías*	*Etiquetas categorías*
TE	Tipo de estudios universitarios	TE1	Ciencias
		TE2	Letras
		TE3	Técnicos
DO	Orientación psicopedagógica	DO1	Con orientación
		DO2	Sin orientación
FE	Finalización de estudios univer.	FE1	Finalizados
		FE2	Abandonados

Tabla de frecuencias

TE	*DO*	*FE*	f_{ijk}
TE1	DO1	FE1	49
TE1	DO1	FE2	2
TE1	DO2	FE1	44
TE1	DO2	FE2	38
TE2	DO1	FE1	49
TE2	DO1	FE2	2
TE2	DO2	FE1	16
TE2	DO2	FE2	20
TE3	DO1	FE1	53
TE3	DO1	FE2	18
TE3	DO2	FE1	6
TE3	DO2	FE2	14

Marginales

— De una variable:

$$f_{0jk} = \sum_{i=1}^{I} f_{ijk}$$

TE	*DO*	*FE*
$f_{011} = 151$	$f_{101} = 93$	$f_{110} = 51$
$f_{012} = 22$	$f_{102} = 40$	$f_{120} = 82$
$f_{021} = 66$	$f_{201} = 65$	$f_{210} = 51$
$f_{022} = 72$	$f_{202} = 22$	$f_{220} = 36$
	$f_{301} = 71$	$f_{310} = 71$
	$f_{302} = 20$	$f_{320} = 20$

— De dos variables:

$$f_{00k} = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ijk}$$

TE-DO	*DO-FE*	*TE-FE*
$f_{001} = 217$	$f_{100} = 133$	$f_{010} = 173$
$f_{002} = 94$	$f_{200} = 87$	$f_{020} = 138$
	$f_{300} = 91$	

— «Gran total»:

$$f_{000} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} f_{ijk}$$

$$f_{000} = 311$$

De hecho es la suma de las frecuencias de todas las casillas o celdas.

1. Estudio de las relaciones

A partir de los datos disponibles, básicamente podemos realizar dos tipos de comparaciones:

1.1. *Comparación de las frecuencias de las casillas con los marginales*

En este caso lo que buscamos son pautas de asociación. En realidad lo que comparamos son proporciones o porcentajes. Supongamos que queremos estudiar la relación entre la variable DO (Orientación psicopedagógica) y FE (Finalización de estudios universitarios).

	FE1	*FE2*	
DO1	151	22	173
DO2	66	72	138
	217	94	311

Fijémonos que lo que aquí ocupa las celdas son marginales de la tabla de contingencia multidimensional. Así podemos comprobar que:

	FE1	*FE2*	
DO1	f_{011}	f_{012}	f_{010}
DO2	f_{021}	f_{022}	f_{020}
	f_{001}	f_{002}	f_{000}

Entre los sujetos que han disfrutado de servicios de orientación (DO1) encontramos que 151/217 = 0,70 han terminado sus estudios (FE1) y en cambio 122/94 = 0,23 no lo han hecho (FE2). La diferencia entre estas dos proporciones —llamada *d*— será una medida de asociación. Lo

mismo podemos hacer con la otra categoría de DO. Veámoslo en la siguiente tabla:

FE	*Porcentaje DO1*	*Porcentaje DO2*
FE1	151/217 = 0,70	66/217 = 0,30
FE2	22/94 = 0,23	72/94 = 0,77

$d_1 = 0{,}70 - 0{,}23 = 0{,}47$ $\quad d_2 = 0{,}30 - 0{,}77 = -0{,}47$

La diferencia de proporciones d toma el valor 0 cuando las dos variables son independientes y tiene un recorrido de + 1 a − 1 según la asociación sea positiva o negativa según planteo. Como puede comprobarse la d no se ve afectada por el cambio de orden (sólo en el signo), sin embargo no es simétrica. Comprobémoslo invirtiendo categorías por porcentajes:

DO	*Porcentaje FE1*	*Porcentaje FE2*
DO1	22/173 = 0,13	151/173 = 0,87
DO2	72/138 = 0,52	66/138 = 0,48

$d_3 = 0{,}13 - 0{,}52 = -0{,}39$ $\quad d_4 = 0{,}87 - 0{,}48 = 0{,}39$

Observemos que el valor de d varía según el orden de entrada de las variables en su cálculo, en este caso $|d_{1,2}| \neq |d_{3,4}|$ Cabe seguir la siguiente sugerencia: la variable independiente debe dar normalmente las categorías, mientras que se calculan las proporciones de esas categorías para una categoría de la variable dependiente (Sánchez, J. J., 1984). A partir de este sistema de diferencias entre proporciones pueden establecerse ecuaciones y ajustar modelos mediante la estimación de coeficientes a partir del cálculo de proporciones recurrentes estandarizadas.

1.2. *Comparación de las frecuencias de las celdas entre sí*

En este caso no obtenemos proporciones sino razones (odds). Por ejemplo, la razón de alumnos que terminan sus estudios universitarios y los que no los terminan será de 217/94 = 2,31, lo que nos permite decir que para cada individuo que no termina sus estudios habrá 2,31 que sí lo hagan. En este caso las llamaremos razones marginales.

También pueden encontrarse razones condicionales controlando por alguna categoría. Así, entre los estudiantes que terminan sus estudios universitarios la razón entre los que han tenido orientación y los que no la han tenido será 151/66 = 2,29, y entre los alumnos que no han terminado

sus estudios será 22/72 = 0,30. Puesto que 2,29 > 0,30 podemos concluir que no es lo mismo haber tenido orientación que no haberla tenido. Las variables DO y FE están asociadas.

Para conocer la fuerza y sentido de la asociación podemos dividir las dos razones condicionales calculando la razón de razones (odds ratio) 2,29/0,30 = 7,63. La razón de terminar estudios a no terminarlos es 7,63 superior en los que han tenido orientación.

Esta razón de razones será 1 cuando las variables sean independientes, y mayor o menor según la asociación sea positiva o negativa. El recorrido va de cero a infinito. Se trata de una escala desigual según el sentido, lo cual produce la sensación de distinta fuerza según sea el planteo. Por ejemplo, veamos las posibles razones de razones de DO-FE:

$$\frac{2,29}{0,30} = 7,6333333 \qquad \frac{0,30}{2,29} = 0,1310043$$

Una solución adoptada para resolver este problema ha sido transformar ambos números en sus logaritmos naturales de modo que se consigue que:

$$\log_e 7,6333333 = \quad 2,03252$$
$$\log_e 0,1310043 = -\ 2,03252$$

Ambos números son iguales, sólo cambia el signo. Esta medida de asociación y su transformación logarítmica es la base sobre la que se construyen los modelos lineales logarítmicos (Knoke, D., 1980; Cobo, E., 1986) que permitirán estudiar las diferencias en las frecuencias respecto a la equiprobabilidad y, en su caso, el estudio de los diferentes efectos de cada variable y de sus posibles asociaciones.

Esta medida permitirá plantear hipótesis respecto a cuáles son los efectos y cuál es su importancia en la explicación de los resultados obtenidos.

El punto de partida será la hipótesis de interacción que corresponde al modelo saturado e implica que la asociación entre dos variables cambia con cada categoría de la tercera. Para una tabla de tres dimensiones como la nuestra será:

$$\log_e F_{ijk} = \mu + \lambda_i^{TE} + \lambda_j^{DO} + \lambda_k^{FE} + \lambda_{ij}^{TE,\,DO} + \lambda_{ik}^{TE,\,FE} + \lambda_{jk}^{DO,\,FE} + \lambda_{ijk}^{TE,\,DO,\,FE}$$

Donde λ representa a los efectos de las dimensiones aisladas y sus asociaciones, y μ el efecto medio:

$$\mu = \frac{1}{I\,J\,K} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \log_e F_{ijk}$$

Existen varios procedimientos para conseguir el modelo más parsimonioso. El más utilizado es el llamado *screening* (Brown, M. B., 1976), que partiendo del modelo saturado en un primer paso incluye sólo las λ estandarizadas de valor $\pm$ 2 para proceder de modo parecido a la regresión eliminando aquellos parámetros que ajusten un modelo más parsimonioso y añadiendo aquellos que aporten una mejora significativa del ajuste. Debe tenerse en cuenta que la parsimonia puede aumentar si las variables son de naturaleza ordinal y se utilizan los procedimientos adecuados.

Cuando estamos previamente interesados por conocer los efectos de un conjunto de variables independientes sobre otra variable considerada dependiente se puede ajustar un modelo LOGIT, el cual considera todas las asociaciones e interacciones significativas que incluyan a la variable dependiente junto con la interacción entre las variables independientes. En este caso el modelo LOGIT saturado sería:

$$\log_e \frac{F_{ij1}}{F_{ij2}} = 2\,\lambda^{FE} + 2\,\lambda^{FE,\,DO} + 2\,\lambda^{FE,\,TE} + 2\,\lambda^{DO,\,TE,\,FE}$$

A partir de él se procede de modo similar que con el modelo general logarítmico lineal para encontrar la ecuación más parsimoniosa. Los modelos LOGIT, en el caso de que se disponga una teoría sustantiva sobre las relaciones causales, pueden ser el punto de partida para evaluar modelos causales con variables categóricas, aunque con algunas limitaciones en relación a «path analysis» o LISREL.

En la actualidad existen numerosos programas de ordenador que incluyen procedimientos de cálculo para estos modelos, entre ellos pueden citarse los siguientes: SPSSX (Statistical Package for Social Sciences) (Norussis, M. J., 1985), BMDP (Biomedical Computer Programs) (Snell, E. J., 1987), SAS (Statistical Analysis System), ECTA (Everyman's Contingency Table Analysis), LOGLIN (Log-linear probability model) y GLIM (Generalised Linear Modelling).

El desarrollo futuro de estas metodologías nos permite augurar un camino esperanzador en su aplicación a la investigación educativa, en la que a menudo nos encontramos en situaciones de naturaleza multivariable en las que, al menos en parte, las variables son de naturaleza cualitativa.

RESUMEN

En investigación educativa bastantes variables importantes son de naturaleza cualitativa. Esto origina la configuración de tablas de contingencia multidimensionales. Los métodos de análisis de datos de variables

nominales y ordinales serán muy últiles. Especialmente los modelos loglinear (jerárquicos y no jerárquicos).

Un caso especialmente interesante lo constituye el modelo logit.

ABSTRACT

In educational research enoughs important variables are qualitative nature. So that are origin multidimensional contingency tables. The methodologies of data analysis with categorical and ordinal variables are very useful. Specially the loglinear procedures (hierarchics and no hierarchics).

A very interesting procedure is the logit model.

BIBLIOGRAFÍA

BISQUERRA, R. (1988): *Introducción conceptual al análisis multivariable. Un enfoque informático con los paquetes SPSSX, BMDP, LISREL y SPAD*, PPU, Barcelona.

BROWN, M. B. (1976): «Screening effects in multidimensional contingency tables», *Aplications of Statistics*, 25:37-46.

COBO, E. (1986): *El análisis de tablas de contingencia. Introducción a los modelos log-lineales*, Edicions Universitat de Barcelona, Barcelona.

DAVIS, J. A. (1985): *The logic of Causal Order*, Sage, Beverly Hills, California.

GOODMAN, L. A. (1984): *The Analysis of Cross-Classified Data Having Ordered Categories*, Harvard University Press, Cambridge.

KNOKE, D. y BURKE, P. J. (1980): *Log-Linear Models*, Sage, Beverly Hills, California.

NORUSSIS, M. J. (1985): *Advanced Statistic Guide*, SPSS-X, McGraw-Hill Book Company, Nueva York.

SÁNCHEZ, J. J. (1984): *Análisis de tablas de contingencia: Modelos Lineales Logarítmicos*, en J. J. Sánchez Carrión (ed.), *Introducción a las Técnicas de Análisis Multivariable aplicadas a las Ciencias Sociales*, Centro de Investigaciones Sociológicas, Madrid, pp. 267-294.

SNELL, E. J. (1987): *Applied Statistics. A handbook om BMDP Analysis*, Chapman and Hall, Londres.

ANTONI SANS I MIRA
Departament M.I.D.E.
Universitat de Barcelona