

Análisis de datos cualitativos en la investigación sobre el rendimiento académico: resolubilidad ante casillas nulas mediante modelos Quasi-log-linear

por
Pedro Sánchez Algarra

Históricamente, en la mayor parte de los estudios sobre el rendimiento académico, se ha enfatizado el análisis de datos de medida, relegándose el tratamiento de los cualitativos a un mero análisis descriptivo. Cuando el estudio interesa ser llevado a cabo puntualmente (o en un seguimiento a partir de una sucesión más o menos prolongada de puntos de tiempo) y los datos de que se dispone son categóricos, el análisis «log-linear» es un útil instrumento que en los últimos años se ha expandido con fuerza, a pesar de que sus orígenes se situaran hacia 1960 debido, fundamentalmente, a los esfuerzos de Goodman, que culminaron casi dos décadas después (Goodman, L.A. 1978). De aquí que en aquellos casos en que se disponga de valoraciones cualitativas o informes acerca de diferentes aspectos a considerar sobre el rendimiento, basta tener en cuenta tantos criterios clasificatorios –y por tanto tantas categorizaciones o taxonomías dicotómicas o politómicas– como dimensiones vayan a ser consideradas, y en base a las cuales se partirá de una tabla de contingencia a resolver.

La construcción y análisis de tabulaciones cruzadas es una de las estrategias que últimamente se han extendido en la investigación sobre el rendimiento académico, y en este sentido basta recordar las ricas posibilidades existentes, por ejemplo, en el análisis de tareas cognitivas (Rodrigo, M.J. 1982), o en cualquier situación en que no sea suficiente una recogida de datos cuantitativos para captar la riqueza que subyace en una determinada estrategia que hay que valorar. Partiendo de tales premisas, la finalidad de estas líneas es la de aportar algunas sugerencias que agrandan sus posibilidades en el caso atípico de casillas nulas.

En efecto, existen una serie de restricciones que deben cumplirse satisfactoriamente (Kennedy, J.J. 1983), y entre ellas se halla la ocupación de todas las casillas de la tabla de contingencia. Debe distinguirse, en las casillas vacías, entre ceros de muestreo y ceros estructurales.

Los primeros (Fienberg, S.E. 1977) no son inherentes a la situación estudiada, y dependen de la muestra global, no suponiendo más que la ausencia de casos representativos en la tabla de clasificaciones cruzadas; si no es muy elevado el número de casillas vacías, se puede solventar mediante la adición de una pequeña cantidad (0,5) al cero existente, con lo cual se logra la necesaria viabilidad para el cálculo de las frecuencias esperadas. Cuando los ceros de muestreo son numerosos, sin embargo, pueden surgir inconvenientes, dado que se carecería de la necesaria información para inferir adecuadamente, la potencia estadística sería escasa, y se vulnerarían además las suposiciones que son básicas para un adecuado uso de la χ^2 -cuadrado (teniendo en cuenta que el análisis log-lineal implica una superación simultánea de χ^2 -cuadrado y ANOVA), cuya distribución, en muestras grandes, se aproxima a la ley multinomial.

Otro carácter presentan los ceros estructurales, que proceden de la ausencia de frecuencias en tablas en las que se sabía de antemano que era imposible la obtención de datos no nulos (Gilbert, 1981). La tabla resulta evidentemente incompleta, afectándose sus grados de libertad —que se reducen— y haciendo necesario su ajuste (Bishop, Y.M.M., Fienberg S.E. & Holland P.W., 1975), dando lugar a los modelos *quasi-log-linear* (Upton G.D.G., 1978), que estructuralmente son similares a los log-lineal, ya que generan frecuencias esperadas para la independencia de forma habitual, salvo en las filas o columnas que contienen uno o más ceros.

En muchos casos, es difícil que los propios investigadores reconozcan el hecho de la existencia de ceros estructurales, y con elevadísima frecuencia se produce el abandono del estudio en el análisis de datos, o bien se llega a resultados e interpretaciones inadecuadas. Y, en el caso en que origina menos trastornos, se tiende a modificar el sistema de categorías.

En tablas bidimensionales, que es el caso más frecuente, si S consiste en una serie de casillas en una ordenación I x J resultante después de eliminar las casillas vacías, si x_{ij} es la frecuencia observada en la casilla (i,j) y m_{ij} la correspondiente frecuencia esperada, consideramos que en las casillas que no se incluyen en la serie S es evidente que

$$x_{ij} = m_{ij} = 0, \quad (1)$$

de forma que podemos seguir utilizando la notación habitual para los marginales totales.

Así, si S se compone de todas las casillas excepto (1,1)

$$m_{1+} = \sum_{j=1}^J m_{1j} = \sum_{j=1}^J m_{ij} \quad (2)$$

representa el valor esperado para la primera fila de la tabla incompleta S.

Al utilizar los mismos modelos «log-linear» para ésta, y por analogía con el ANOVA, para las casillas (i,j) \in S, tenemos

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \quad (3)$$

donde

$$\sum_{i=1}^I u_{1(i)} = \sum_{j=1}^J u_{2(j)} = 0 \quad (4)$$

y

$$\sum_{i=1}^I \sigma_{ij} u_{12(ij)} = \sum_{j=1}^J \sigma_{ij} u_{12(ij)} = 0 \quad (5)$$

con

$$\sigma_{ij} = \begin{cases} 1 & \text{para } (i,j) \in S \\ 0 & \text{en los demás casos} \end{cases} \quad (6)$$

Los términos u_{12} de (5), que corresponden a casillas que no pertenecen a S son series iguales a una cantidad arbitraria finita, de forma que (5) está bien definido.

A partir de aquí, el modelo de cuasi-independencia se halla estableciendo

$$u_{12(ij)} = 0 \text{ para } (i,j) \notin S \quad (7)$$

de forma que

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \text{ para } (i,j) \in S \quad (8)$$

En otras palabras, las variables correspondientes a las filas y columnas son cuasi-independientes si podemos escribir $\{m_{ij}\}$ en la forma:

$$\begin{aligned} & \left\{ m_{ij} \right\} \text{ en la forma:} \\ m_{ij} &= \begin{cases} a_i b_j & \text{para } (i,j) \in S \\ 0 & \text{en los demás casos} \end{cases} \quad (9) \end{aligned}$$

La cuasi-independencia actúa, pues, como la independencia, y se aplica a las casillas no vacías de la tabla.

Por otra parte, la regla general para hallar los grados de libertad, se concreta (Fienberg S.E., 1981):

g.l. = número casillas – número parámetros ajustados (10)
y aquí es igualmente aplicable. Si hay e casillas eliminadas, entonces la serie S contiene $IJ - e$ casillas, y el número de parámetros ajustados son

$$I + J - 1 \text{ (1 por } u, I-1 \text{ por } u_{1(i)}, \text{ y } J-1 \text{ por } u_{2(j)})$$

resultando, en consecuencia,

$IJ - e - (I + J - 1) = (I - 1).(J - 1) - e$ grados de libertad

Es interesante considerar su repercusión práctica en estudios acerca del rendimiento académico, ya que en situaciones en donde una serie de casillas de la tabla son ocupadas por ceros es frecuente optar por modificar el sistema de categorías, con lo cual se pueden evitar tales ceros, pero, en cambio, se sacrifica capacidad discriminatoria entre tales categorías, hecho que puede llegar a ser especialmente preocupante si la escala es ordinal (Andrich, D. 1979), apareciendo, en cualquier caso, y como consecuencia, los efectos de errores de clasificar inadecuadamente (Fleiss, V.C. 1981).

Las posibilidades metodológicas de optimización son evidentes, y en el ámbito del rendimiento académico pueden llevarse a la práctica con relativa facilidad, por lo que la propuesta realizada de utilizar modelos quasi-log-linear ante la existencia de ceros estructurales es de esperar que produzca sus frutos.

REFERENCIAS BIBLIOGRÁFICAS

- ANDRICH, D. (1979) A model for contingency tables having an ordered response classification. *Biometrics*, 35, 403-415.
- BISHOP, Y.M.M.; FIENBERG, S.E. & HOLLAND, P.W. (1975) *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.
- FIENBERG, S.E. (1981) *The analysis of cross-classified categorical data*. Cambridge, Mass.: MIT Press.
- FLEISS, J.L. (1981) *Statistical methods for rates and proportions*. New York: Wiley & Sons.
- GOODMAN, L.A. (1978) *Analyzing qualitative/categorical data*. Cambridge, Mass.: Abt Books.
- KENNEDY, J.J. (1983) *Analyzing qualitative data. Introductory log-linear analysis for behavioral research*. New York: Praeger.
- RODRIGO, M.J. (1982) Las posibilidades del análisis de tareas como técnicas para el estudio de los procesos mentales. *Infancia y Aprendizaje*, 19-20, 159-173.
- UPTON, G.J.G. (1978) *The analysis of cross-tabulated data*. New York: Wiley & Sons.