Revista Investigación Educativa - N.º 21 - 1993 (P. 137-150)

PROBLEMAS FUNDAMENTALES DEL ANÁLISIS LOGARÍTMICO LINEAL (Y II): CASILLAS VACÍAS Y CASILLAS EXTREMAS

por
Ana Delia Correa Piñero
Dpto. Didáctica e Investigación Educativa. Universidad de La Laguna

RESUMEN

En este trabajo abordamos dos problemas que pueden presentarse en el análisis de tablas de contingencia multidimensionales: el tema de las casillas vacías —en su doble faceta de ceros muestrales y ceros estructurales— y el de las casillas y categorías extremas. Para ambos casos, se analizan sus repercusiones en el proceso y resultados del ajuste de un modelo logarítmico lineal y se ofrecen diversas soluciones, analizando sus ventajas e inconvenientes.

Descriptores: Análisis de modelos logarítmico lineales; casillas vacías; casillas y categorías «extremas»; análisis de residuales en tablas de contingencia.

ABSTRATC

Two problems that can appear in multi-way contingency tables analysis are treated in this paper. First one, empty cells in its double facet: random zeros and structural zeros. Second one, extreme (outliers) cells and strata. Both cases their effects into processe and outcomes of fitting a log-linear model are analized. Likewise different solutions are offered, analizing their advantages and disadvantages.

Key words: Log-Linear Models Analysis; empty cells; extreme (outliers) cells and strata; analysis of residuals in cross-classified tables.

I. INTRODUCCIÓN

Con este artículo terminamos la serie de «problemas fundamentales» que puede encontrar el investigador al analizar un conjunto de datos mediante el ajuste de modelos logarítmico lineales. Así, en CORREA (1991a) tratamos diversos problemas relacionados con el «colapsamiento» de variables (recombinar categorías, categorizar datos continuos, etc.). Otros problemas de interés, relacionados con el tamaño de la muestra, pueden encontrarse en CORREA (1991b). En esta ocasión nos centraremos en los problemas que supone para la fase de análisis la presencia de casillas vacías y casillas extremas. Este último tema también es conocido en la literatura como *análisis de residuales.*

Se presupone que el lector está familiarizado con el análisis logarítmico lineal en lo que respecta a la elaboración de tablas de contingencia multidimensionales, selección de modelos, ajuste de modelos y cálculo de parámetros y errores típicos. Pueden consultar una introducción al tema y otras referencias de interés en CORREA (1991c).

II. EL PROBLEMA DE LAS CASILLAS VACÍAS: CEROS MUESTRALES Y CEROS ESTRUCTURALES

Se dice que una casilla está vacía cuando tiene frecuencia cero. No hay ningún sujeto en una determinada combinación de categorías de las variables implicadas. Esto supone problemas, ya que la base del análisis logarítmico lineal está en el cálculo de razones (*ratios*) y la presencia de un cero en el denominador produciría una indeterminación. Este hecho puede producirse por dos motivos, que dan lugar a dos tipos de casillas vacías: los *ceros muestrales* y los *ceros estructurales.*

Cuando en la muestra no se obtiene ningún caso para determinadas combinaciones de categorías se producen casillas con frecuencia cero, pero sin que eso signifique que no existan esas combinaciones en la población. Simplemente, en la muestra obtenida no se presenta ningún caso. Cuanto más pequeña sea la muestra, aumenta la probabilidad de que alguna(s) casilla(s) quede(n) sin cubrir. Por el contrario, al crecer el tamaño de la muestra aumenta la probabilidad de que todas las casillas resulten ocupadas. A este tipo de casillas se les llaman *ceros muestrales o aleatorios* (GOLDSTEIN y DILLON, 1980).

Se han apuntado varias soluciones para este problema. GOODMAN (1970) propone añadir un pequeño valor constante (normalmente 0,5) a cada casilla antes de efectuar los cálculos. Éste es un procedimiento conservador que puede subestimar los parámetros y su significación: la constante añadida aumenta las frecuencias esperadas, lo cual reduce el valor de los estadísticos ji^2 y G^2 que son los que se utilizan para comprobar la calidad de ajuste del modelo (BROWN, 1983).

Ya que en los cálculos de las razones, como dijimos, la división entre 0 daría «indeterminado», otra solución sería definir, de forma arbitraria, que 0/0=0 (FIEN-

BERG, 1977). En ese caso debe tenerse precaución con dos aspectos: uno, la posible aparición de frecuencias esperadas muy pequeñas que ponen en entredicho la aplicabilidad de los estadístico de ajuste; el otro, si los cálculos se hacen mediante algún programa computarizado, asegurarse que esa solución es factible mediante las rutinas programadas o se corre el riesgo de que se «aborte» el análisis, como sucede, por ejemplo, al intentar invertir una matriz cuyo determinante es cero.

Otra posible solución, la más obvia pero no siempre posible por motivos diversos, es aumentar el tamaño de la muestra lo suficiente como para que se eliminen los ceros (KNOKE y BURKE, 1982). Se entiende que se vuelve a retroceder a la fase de obtención de datos hasta que desaparezca el problema.

Finalmente, queda una cuarta alternativa: colapsar o recombinar categorías, de manera que la casilla vacía quede «integrada» en otra casilla con frecuencias suficientes. Las posibles repercusiones del colapsamiento en el ajuste del modelo y en el valor de los parámetros se analizaron en el artículo anterior (CORREA, 1991a).

El segundo tipo de casilla vacía se produce cuando en la tabla hay ciertas categorías que no tienen referente empírico y, por tanto, tendrán frecuencia cero aunque se trabaje con toda la población. Son los *ceros lógicos*, llamados también *ceros estructurales,* ya que son consecuencia de la estructura del problema (FAGEN y MANKOVICH, 1980). Esta clase de ceros puede surgir por:

a) *Diseño del muestreo.* Es decir, que el investigador en su diseño del muestreo decida suprimir cierta casilla por razones teóricas. En realidad, este caso sí tendría referente empírico, pero sería teóricamente irrelevante.

b) *Secuencia ordinal de las variables.* Por ejemplo, en una tabla «Edad x Status Familiar», es poco probable encontrar la categoría «abuelos» en los primeros órdenes (intervalos) de la variable edad.

c) *Inconsistencia definicional.* Es el caso más neto de cero lógico. Por ejemplo, en una tabla «Sexo x Tipo de operación quirúrgica» es imposible encontrar referentes empíricos correspondientes a la casilla «mujer-próstata».

Por cualquiera de estas razones, pues, surgen las denominadas tablas de contingencia incompletas o truncadas. GOODMAN (1968) se ha interesado por este tipo de tablas, para el caso de dos y tres variables, formulando una técnica para su análisis conocida como prueba de cuasi-independencia. En una tabla con casillas vacías, la cuasi-independencia es una forma de «no asociación» entre las variables, pero referida sólo a la porción de la tabla donde no haya ceros. Es decir, es una forma de *independencia condicional* donde se modifican los grados de libertad restando el número de ceros estructurales existentes en la tabla. En una tabla IxJ, con Z casillas 0, los grados de libertad son: (I-1)(J-1)-Z. Así, a partir de las casillas disponibles, se obtienen estimaciones de máxima verosimilitud para las casillas vacías. Éstas son reemplazadas por dichos valores estimados. Luego, se calculan los valores esperados de todas las casillas como si la tabla fuese entera (véase BROWN, 1975). FIENBERG (1972) generaliza las ideas básicas a tablas mayores, formulando lo que él denomina modelos cuasi logarítmico lineales.

Por otra parte, SMITH (1973) señala la utilidad de las tablas incompletas para

abordar cierto tipo de problemas, por ejemplo: permiten prescindir de respuestas incorrectas cuando se desea analizar sólo las correctas, o de los «acuerdos» entre jueces cuando sólo interesa analizar los «desacuerdos» y similares.

Personalmente, en la práctica hemos encontrado un «tercer tipo» de casilla vacía, a medio camino entre el cero muestral y el estructural. En una encuesta a profesores de EGB sobre medios didácticos (CORREA y AREA, en prensa), utilizamos un cuestionario que incluía, entre otros, un ítem donde se pedía que indicaran la importancia que le concedían a una serie de elementos del libro de texto. Los elementos textuales en cuestión eran los siguientes:

1) Su relación con los Programas Oficiales.
2) El tipo de actividades que propone.
3) El lenguaje utilizado.
4) Los aspectos formales (colorido, tamaño, ilustraciones, etc.).
5) Su adecuación al nivel de conocimientos de los alumnos.
6) Lo innovador de su planteamiento.
7) La extensión con que trata los contenidos.
8) Su relación con el contexto.
9) El método de enseñanza que se desprende el texto o de su guía didáctica.

La importancia se reflejaba en una escala ordinal de 4 niveles (desde «ninguna» a «mucha»). Cada elemento textual, pues, representa una variable distinta, con categorías ordenadas. No resultó difícil encontrar sujetos que, para un elemento aislado, se repartiesen a lo largo de todo el continuo de valoración. El cruce de varios elementos (variables) produce tanto casillas con varias valoraciones simultáneas de «mucha importancia» como casillas donde se reflejan varias valoraciones simultáneas de «ninguna importancia». Pues bien, estas últimas casillas resultaron, en su mayoría, vacías. No se trataba de ceros estructurales, en sentido estricto. Siempre cabe pensar en la existencia de profesores/as en la población que puedan ubicarse en esas casillas, de forma que al aumentar el tamaño de la muestra encontrásemos algún «referente empírico» apropiado. Pero, debido a la formulación misma del contenido del ítem, pensamos que serían casos muy raros, de tal forma que ese necesario aumento de muestra debería ser muy considerable, lo cual hace que sea también difícil calificarlas de ceros muestrales estrictos. Los elementos textuales citados, innegablemente, son apreciativos. En cualquier caso, neutros; pero no peyorativos. De tal forma que la negación simultánea de la importancia de varios de ellos conjuntamente es un «caso raro». Y se va haciendo progresivamente más raro a medida que la tabla de contingencia incluya más elementos textuales. Es decir, es frecuente encontrar profesores/as que dan poca importancia a un elemento, considerado aisladamente (análisis univariado). Un poco menos frecuente, pero no raro, encontrar esa misma valoración referida a un par de elementos a la vez (tabla bivariada). Bastante raro, si la valoración hace referencia a tres elementos conjuntamente (tabla trivariada). Etc, etc.

Podría aducirse que este tipo de ceros sería mejor considerarlos como ceros

estructurales del tipo a), es decir, fruto del diseño del muestreo. Pero, en primer lugar, no se decidió a priori suprimir esas casillas. Y, en segundo lugar, en este caso no era teóricamente irrelevante, sino todo lo contrario, la constatación de su existencia.

En suma, en una aplicación concreta de esta técnica de análisis, podemos encontrar que «algo» en nuestras tablas de contingencia no se ajusta a los modelos standard. En ese caso, debería tomarse la decisión que se considere más apropiada, de acuerdo con las peculiaridades del estudio. Nosotros, en este caso y por las razones expuestas, decidimos recurrir a dos soluciones y posteriormente comparar los resultados. Una, tratar dichas casillas como ceros muestrales y seguir el criterio sugerido por Goodman de añadir una constante, en lugar de aumentar la muestra a la búsqueda, probablemente sin éxito, de referentes empíricos reticentes. La segunda solución fue combinar categorías adyacentes (véase artículo anterior sobre colapsamiento en esta misma revista) ya que las categorías de las variables eran ordenadas y su reagrupación, aunque disminuía el recorrido de la variable, no alteraba la ordenación subyacente. En la mayoría de los casos (tablas) fue preciso combinar ambos recursos. Los resultados que se ofrecen en la referencia citada (CORREA y AREA, en prensa) son precisamente estos últimos, ya que así obteníamos el ajuste con modelos más parsimoniosos.

III. EL PROBLEMA DE LAS CASILLAS EXTREMAS

En ocasiones, puede suceder que una frecuencia de casilla, incluso de una categoría entera, se aleje excesivamente de lo que cabría esperar a partir del modelo de referencia. Dicho de otra forma, la frecuencia esperada es excesivamente grande o excesivamente pequeña. Esto puede llevar a una impresión distorsionada de las relaciones entre variables, ya que «las casillas extremas pueden sesgar los valores esperados lo suficiente como para impedir que se hagan inferencias sobre el resto de casillas» (FAGEN y MANKOVICH, 1980:1.020). Es decir, la presencia de un valor extremo no sólo afecta la casilla donde se produce, sino a todas las casillas de la fila y columna a la que pertenece y, en último término, afecta la significación del modelo.

Es conveniente, pues, disponer de algún procedimiento para detectar casillas y/o niveles extremos, de la misma forma que otros procedimientos estadísticos disponen de sistemas para detectar «sujetos extremos», como la D^2 de Mahalanobis, el método de Comrey, el análisis de los residuos en la regresión, etc. (véase GAVIRIA, 1988). Esto podría explicar por qué en ocasiones se producen malos ajustes en modelos que, según todos los indicios eran muy prometedores. Estos indicios pueden venir dados, por ejemplo, por las pruebas previas de asociación parcial y marginal y su determinación de efectos «innecesarios», «necesarios» e «inciertos».

La base del análisis logarítmico lineal radica en la comparación de frecuencias observadas y esperadas. Una «anomalía» se detectaría entonces como una discrepancia, o residual, proporcionalmente grande entre ambos tipos de frecuencias. No

debe confundirse «casilla extrema» con «casilla de mayor frecuencia». La «extremosidad» viene dada por un mayor o menor alejamiento entre una frecuencia esperada y su correspondiente observada. Puede ser que, para parecerse a lo que se espera *bajo cierto modelo,* una frecuencia absoluta tenga que ser muy pequeña o puede que tenga que ser muy grande. Además, hemos dicho «proporcionalmente grande», es decir, que depende de los datos: no es lo mismo una diferencia de 100 cuando se comparan 12.000 y 11.900 que cuando se comparan 12 y 112. En este último caso, la diferencia es vital para el ajuste o desajuste del modelo.

Para detectar estos residuales se han propuesto diversas fórmulas, que dan lugar a otros tantos tipos de residuales (HABERMAN, 1973; UPTON, 1978). Como se trata de desviaciones a partir de las frecuencias esperadas, se les suele llamar así, *desviaciones*. Existen 3 tipos de desviaciones: de Freeman-Tukey, standarizadas y standarizadas ajustadas. Las dos primeras pueden ser aplicadas a tablas bi y multivariadas. La última, sólo a tablas de dos variables.

La desviación de Freeman-Tukey en cualquier casilla ij se define como:

$$d_{ij}(F\text{-}T) = \sqrt{(o_{ij})} + \sqrt{(o_{ij}+1)} - \sqrt{(4e_{ij}+1)} \qquad \text{(Ec.1)}$$

siendo *o* las frecuencias observadas en la casilla *ij* y *e* las frecuencias esperadas.

La interpretación de los valores obtenidos se basa en la similitud de estas desviaciones con las puntuaciones típicas. A partir de la distribución normal unitaria se determina el porcentaje correspondiente a esa puntuación (desviación), valorando así la magnitud del residual. No obstante, para poder hacer esta analogía con puntuaciones z, es preciso que los datos se hayan obtenido mediante una técnica de muestreo específica: la distribución Poisson.

Para una casilla ij cualquiera se define la desviación estandarizada o típica como:

$$d_{ij}(est) = (o_{ij} - e_{ij})/\sqrt{e_{ij}} \qquad \text{(Ec.2)}$$

Son los mismos elementos que intervienen en la conocida fórmula de ji^2, con la excepción de la raíz cuadrada. Para su análisis se recurre igualmente a la distribución normal.

Finalmente, las desviaciones típicas ajustadas son desviaciones estandarizadas, ajustadas de tal forma que tengan varianza 1 cuando los datos provengan de una distribución multinomial. El inconveniente, como decíamos, es que sólo se aplica a tablas bivariadas.

Su fórmula, dada la casilla ij, es:

$$d_{ij}(ajs) = \frac{o_{ij} - e_{ij}}{\sqrt{[e_{ij}(1-f_{i+}/N)(1-f_{+j}/N)]}} \qquad \text{(Ec.3)}$$

Los nuevos elementos, f_{i+} y f_{+j}, representan el total marginal de fila y columna, respectivamente. También se interpretan sus valores a partir de la tabla de la distribución normal.

Adicionalmente, habría otra medida de desviación: la diferencia entre las frecuencias observadas y las esperadas (o-e); este índice, sin embargo, es muy simple y no tiene base para una interpretación clara, como los anteriores.

Interpretación de los residuales

Al comienzo de este apartado comentábamos la importancia del análisis de casillas extremas para ver por qué un modelo hipotetizado no había ajustado, como se esperaba. Vamos a detallarlo a través de un ejemplo con los datos ficticios de una tabla tridimensional: SEXO x PERÍODO DE TIEMPO x CARRERA (véase Tabla 1). Todos los análisis se han efectuado mediante el programa 4F de BMDP.

TABLA 1
CARRERA x SEXO x PERÍODO DE TIEMPO (frecuencias observadas)

CARRERA		SEXO	
	PERÍODO	Hombre	Mujer
Ciencias	1960-70	69	12
	1971-80	124	34
Humanas	1960-70	85	113
	1971-80	68	116
Técnicas	1960-70	105	36
	1971-80	136	98

Inicialmente, aplicamos las pruebas de asociación parcial y marginal a los datos de la Tabla 1. Los resultados indican que las interacciones SC (SEXO y PERÍODO DE TIEMPO) y PC (PERÍODO DE TIEMPO y CARRERA) son efectos «definitivamente necesarios». La interacción SP, por su parte, resulta «incierta», ya que sólo es significativa en una de las pruebas. Sometemos a ajuste, por tanto, dos modelos: el SC,PC (que incluye sólo efectos «definitivamente necesarios») y el SC,PC,SP (que contiene los anteriores y, además, el efecto «incierto»). Los resultados del ajuste se muestran en la Tabla 2.

TABLA 2
AJUSTE DE MODELOS A LOS DATOS DE LA TABLA 1

MODELO	GL	G^2	PROB	JI^2	PROB
SC,PC	3	13,52	0,0036	13,21	0,0042
SC,PC,SP	2	2,49	0,2885	2,47	0,2907

Como se ve, el modelo SC,PC no presenta un buen ajuste: sus índices residuales (tanto ji^2 como G^2) son significativos; luego, el ajuste no lo es. En cambio, sí ajusta bien el modelo que incluye el efecto «incierto»: los residuales de SC,PC,SP no son significativos, por tanto, sí es significativo su ajuste. Sin embargo, el valor de G^2 *componente* correspondiente al efecto «incierto» (SP) es muy pequeño, lo que quiere decir que este efecto contribuye muy poco al ajuste del modelo SC,PC,SP. En concreto, su contribución o componente se determina mediante la sustracción de residuales de dos modelos que sólo se diferencien en el efecto de interés; sería: 13,52 - 2,49 = 11,03. La contribución de los otros dos efectos SC y PC, calculada de la misma forma y mediante los residuales de los modelos apropiados (que no se muestran en la Tabla 2), obtuvieron valores superiores: 120,2 y 33,3, respectivamente.

¿Por qué no ajustó bien el modelo SC,PC cuando todos los indicios hacían pensar lo contrario? Analicemos los residuales de ese modelo. Ya que se trata de una tabla de 3 variables no podemos usar las desviaciones estandarizadas ajustadas; las de Freeman-Tukey tampoco serían apropiadas, ya que los datos se corresponden con un esquema de muestreo multinomial (ningún marginal está fijado de antemano) y no un esquema Poisson; las diferencias o-e, dijimos que eran poco informativas. Así, pues, analizaremos mediante las desviaciones estandarizadas si se detecta la presencia de alguna casilla extrema, que haya contribuido al desajuste de un modelo tan prometedor. Las desviaciones estandarizadas para esos datos bajo el modelo SC,PC se muestran en la Tabla 3.

TABLA 3
DESVIACIONES ESTANDARIZADAS BAJO EL MODELO SC,PC

CARRERA		SEXO	
	PERÍODO	Hombre	Mujer
Ciencias	1960-70	0,4	–0,9
	1971-80	–0,3	0,7
Humanas	1960-70	0,6	–0,5
	1971-80	–0,7	0,5
Técnicas	1960-70	1,5	–2,0
	1971-80	–1,2	1,6

Numerando las casillas de arriba a abajo y de izquierda a derecha, las que muestran unas mayores desviaciones son la 9, 10, 11 y 12. Todas se ubican en el nivel «técnicas». Sobre todo la casilla 10 muestra una desviación considerable (–2,0). Una puntuación típica de –2,0 es un valor muy extremo. Se corresponde, en

la distribución normal, con un área de 0,0228. Es decir, que sólo un 2,28% de observaciones elegidas a partir de la distribución normal unitaria estaría *por azar* a esa distancia o mayor de la media 0. La probabilidad, pues, de que esta distancia se deba al azar es muy pequeña. Es una desviación significativa, que contribuye de forma más acusada que las otras al desajuste que mostró el modelo SC,PC.

El mismo cálculo para el modelo SC,PC,SP se muestra en la Tabla 4.

TABLA 4
DESVIACIONES ESTANDARIZADAS BAJO EL MODELO SC,PC,SP

CARRERA		SEXO	
	PERÍODO	Hombre	Mujer
Ciencias	1960-70	–0,0	0,0
	1971-80	0,0	–0,0
Humanas	1960-70	–0,5	0,5
	1971-80	0,6	–0,5
Técnicas	1960-70	0,5	–0,8
	1971-80	–0,4	0,5

En este caso las diferencias se reducen, sobre todo en las casillas 9, 11 y 12. En la 10 también se reduce (esto es lógico, ya que este modelo sí ajustó), pero, con todo, esta casilla sigue mostrando la mayor desviación de la tabla. Es decir, es la que menos contribuye al ajuste del modelo.

¿Qué conclusión sacar a partir de esto? Simplemente, que el nivel «técnicas» combinado con el resto de niveles del resto de variables, muestra un comportamiento muy distinto a los demás. Es un «caso muy raro», difícilmente explicable a partir del modelo SC,PC, aunque «menos raro» a la luz del modelo SC,PC,SP.

Esto también podría observarse con el cálculo de parámetros: los parámetros de menos valor (y menos significativos) para el modelo SC,PC,SP se obtendrán en aquellos términos donde esté implicado el nivel «técnicas». En concreto, una vez calculados, obtenemos que para el nivel «técnicas» al parámetro de interacción SEXO-CARRERA le corresponde un valor (lambda) de 0,033; y al de la interacción CARRERA-PERÍODO DE TIEMPO un valor de 0,077. Estas cifras suponen los menores valores de todo el conjunto de parámetros y, como se sabe, cuanto más cercanos a cero están los parámetros, menor es su importancia en el modelo de referencia.

A lo largo de este ejemplo, vemos como los resultados de las distintas fases del análisis van armonizando, complementándose, para dar una idea global del esquema que subyace a los datos: las pruebas de asociación parcial y marginal nos aconsejan

dos determinados modelos para su ajuste; probamos ambos modelos y uno de ellos ajusta y el otro no; sin embargo, el valor G^2 componente del término que diferencia a un modelo de otro es muy reducido, contribuye poco al ajuste; si esto es así, ¿por qué no ajustó también el otro modelo?; buscamos si existen casillas extremas que hayan podido impedirlo y, efectivamente, las encontramos; y coinciden, precisamente, con los parámetros de menor importancia en el modelo. El buen concierto de unos y otros resultados llega a producir un placer casi estético.

¿Qué decisión tomamos con estas casillas extremas y qué influencia puede tener en el análisis la decisión tomada? Ya que «técnicas» contribuye muy poco, parece lógico pensar que es muy probable que si se reduce la variable CARRERA a dos niveles, eliminando «técnicas», se mantenga el esquema de relaciones entre SEXO y CARRERA y PERÍODO DE TIEMPO y CARRERA. Vamos a comprobarlo.

Solución extrema para los datos extremos: eliminación de categorías

Supongamos que a la hora de obtener datos sólo hemos encuestado a hombres y mujeres, de los dos períodos, que estudian ciencias o humanas. Ningún técnico. La fila «técnicas» desaparecería de la Tabla 1 dando lugar a la Tabla 5.

TABLA 5
TABLA 1 SIN LA CATEGORÍA «TÉCNICAS» (FRECUENCIAS OBSERVADAS)

CARRERA		SEXO	
	PERÍODO	Hombre	Mujer
Ciencias	1960-70	69	12
	1971-80	124	34
Humanas	1960-70	85	113
	1971-80	68	116

La diferencia entre el N anterior (996) y el actual (621) no distorsiona lo que queremos averiguar. Al contrario, lo refuerza, ya que cuanto mayor sea N, aumenta la probabilidad de encontrar ajustes significativos, hasta con modelos complejos, como el presaturado, y viceversa. El ajuste del modelo SC,PC,SP para los nuevos datos es:

TABLA 6
AJUSTE DEL MODELO SC,PC,SP PARA LOS DATOS DE LA TABLA 5

MODELO	GL	G^2	PROB	ji^2	PROB
SC,PC,SP	1	0,24	0,6246	0,24	0,6265

No sólo se mantiene el ajuste que ya mostró la tabla original, sino que se mejora: los índices residuales (G^2 y ji^2) pasan de 2,49 y 2,47, respectivamente, a ser de 0,24 en ambas pruebas. Las desviaciones estandarizadas bajo este modelo se muestran en la Tabla 7.

TABLA 7
DESVIACIONES ESTANDARIZADAS DE LA TABLA 5 BAJO EL MODELO SC,PC,SP

CARRERA	PERÍODO	SEXO Hombre	SEXO Mujer
Ciencias	1960-70	0,1	–0,3
	1971-80	–0,1	0,2
Humanas	1960-70	–0,1	0,1
	1971-80	0,1	–0,1

Vemos que las desviaciones disminuyen considerablemente con respecto a las obtenidas para los datos originales. Eliminando «técnicas», el modelo que ya había ajustado en la tabla original mejora su ajuste. Veamos en la Tabla 8 qué sucede ahora con el modelo tan prometedor, que no llegó a ajustar.

TABLA 8
AJUSTE DEL MODELO SC,PC A LA TABLA 5

MODELO	GL	G^2	PROB.	ji^2	PROB.
SC,PC	2	3,02	0,2208	2,97	0,2270

En esta ocasión, el modelo sí ajusta. Y sus desviaciones estandarizadas muestran la desaparición de casillas extremas, tal como se muestra en la Tabla 9.

TABLA 9
DESVIACIONES ESTANDARIZADAS DE LA TABLA 5 BAJO EL MODELO SC,PC

CARRERA	PERÍODO	SEXO Hombre	SEXO Mujer
Ciencias	1960-70	0,4	–0,9
	1971-80	–0,3	0,7
Humanas	1960-70	0,6	–0,5
	1971-80	–0,7	0,5

A pesar de estos resultados, quizá parezca excesivamente riguroso suprimir de un plumazo una categoría entera. Al fin y al cabo, no todas las casillas donde intervenía el nivel «técnicas» eran igualmente extremas. Para el modelo SC,CP las casillas más extremas (9, 10, 11, 12) tenían desviaciones de 1,5, -2,0, -1,2 y 1,6, respectivamente. Quizá bastaría con eliminar la casilla 10.

La solución conciliadora: eliminación de casillas

Para aplicar esta solución menos drástica, es preciso dejar la tabla tal y como estaba, excepto que se anula una casilla, como si su frecuencia fuera cero. Esto supone obtener una tabla incompleta, en la cual se ajusta el modelo mediante una prueba de cuasi-independencia, como explicamos en el apartado de casillas vacías.

El programa 4F de BMDP puede, a partir de un modelo que no ha ajustado bien, ir identificando paso a paso las casillas más extremas (en cada paso, una casilla) y recalcular el ajuste eliminándolas por turno. Cesa el proceso cuando consigue un buen ajuste. El sistema, por tanto, no funciona con un modelo que ajuste bien desde el principio. Veamos en la Tabla 10 qué sucede al aplicar este procedimiento a los datos originales.

TABLA 10
AJUSTE DEL MODELO SC,PC CON ELIMINACIÓN AUTOMÁTICA DE CASILLAS EXTREMAS (DATOS TABLA 1)

PASO	MODELO X^2	SC,PC GL	PROB	DESVIACIÓN MÁXIMA	ENCONTRADA EN LA CASILLA
0	13,52	3	0,00363	-2,026	TÉCNICAS 1960-70 MUJER
1	3,02	2	0,22079		

En el paso 0 se ajusta el modelo con la tabla completa y se identifica la casilla con una desviación estandarizada máxima. Como ya sabíamos, es la casilla 10. Luego, paso 1, elimina esa casilla y calcula el ajuste de la tabla incompleta. Ahora los datos observados sí se ajustan al modelo y el proceso se detiene.

Efectivamente, suprimir el nivel «técnicas» entero, fue una decisión bastante drástica. Por ello, se aconseja utilizar primero este proceso paso a paso con casillas extremas y ver si así se consigue un buen ajuste. Si, después de todo, fuera preciso eliminar un nivel entero, se eliminaría en última instancia.

IV. CONSIDERACIONES FINALES

Hay una deducción, que quizá resulte obvia al lector, con respecto a las casillas «raras» (vacías o extremas) y los recursos más habituales para solventarlas. Tanto el colapsamiento como la eliminación de casillas o niveles extremos, favorecen el ajuste de modelos más parsimoniosos, menos complejos. En principio, esto parece deseable, pero podría alegarse que colapsando por aquí y eliminando por allá nos quedaremos con «la norma», cuando en ocasiones los casos «raros» pueden ser teóricamente interesantes, incluso deseables si se hipotetizan relaciones complejas, lo cual es muy probable en investigación educativa (en investigación social, en general).

En este punto, podemos citar la recomendación de GAVIRIA (1988) con respecto a los casos extremos: si se trata de investigación básica, con una muestra representativa, los casos extremos deben incluirse en el análisis, ya que son elementos representativos de grupos que, aunque poco numerosos, tendrían más probabilidad de aumentar a medida que se incrementase el tamaño de la muestra. Por el contrario, en investigación operativa, para tomar decisiones locales, sería preferible eliminar esos sujetos del análisis y considerarlos individualmente. No obstante, al margen de la utilidad de esta sugerencia, hay que recordar que en un análisis logarítmico lineal no siempre es aplicable la consideración de «casos (casillas) extremas» como «casos poco numerosos». Aquí una casilla es extrema no por su frecuencia absoluta, observada, sino por su discrepancia con la frecuencia esperada bajo cierto modelo. Así, (tenga una frecuencia alta o baja) puede ser extrema para ciertos modelos y no extrema para otros. Por tanto, la decisión está, sobre todo, en manos del investigador y del modelo que adopte. Con estos dos artículos sobre «problemas fundamentales» hemos querido contribuir a que las decisiones que tome al respecto en un análisis logarítmico lineal estén un poco mejor fundamentadas e informadas.

REFERENCIAS

BROWN, M. B. (1975): The asymptotic standard errors of some estimates of uncertainty in the two-way contingency table. *Psychometrika,* 40(3), 291-296.

BROWN, M. B. (1983): PF4. Two-way and Multi-way Frequency Tables - Measures of Association and the Log-Linear Model (Complete and Incomplete Tables). En W. J. Dixon (Ed.) *BMDP Statistical Software.* California: University of California Press.

CORREA, A. D. (1991a): Problemas fundamentales del análisis logarítmico lineal (I): el colapsamiento de variables y su influencia en el ajuste de modelos. *Revista de Investigación Educativa,* nº 18, 2º semestre.

CORREA, A. D. (1991b): Reglas prácticas en torno al tamaño de la muestra para el ajuste de modelos logarítmico lineales. *Qurriculum,* Nº Extra 1/2, 365-368.

CORREA, A. D. (1991c): Estudios multivariados con datos nominales: aportaciones del análisis logarítmico lineal. *Qurriculum,* nº 3, 35-52.

CORREA, A. D. y AREA, M.: ¿Qué opinan los profesores de EGB sobre el uso del libro de texto en las escuelas? *Qurriculum,* nº 4, en prensa.

FAGEN, R. M. & MANKOVICH, N. J. (1980): Two-act transitions, partitioned contigency tables, and the «significant cells» problem. *Animal Behaviour,* 28(4), 1.017-1.023.

FIENBERG, S. E. (1972): The analysis of incomplete multiway contingency tables. *Biometrics,* 28, 177-202.

FIENBERG, S. E. (1977): *The analysis of cross-classified categorical data.* Cambridge, Massachusetts: MIT Press.

GOLDSTEIN, M. & DILLON, W. R. (1980): A measure of separability and random zeros in statistical classification. *Multivariate Behavioral research,* 15(1), 57-71.

GAVIRIA, J. L. (1988): La detección de «outliers» en los análisis multivariantes. En I. Dendaluce (Coord.): *Aspectos metodológicos de la investigación educativa* (pp. 264-271). Madrid: Narcea.

GOODMAN, L. A. (1968): The analysis of cross-classified data. *Journal of the American Statistical Association,* 63, 1.091-1.131.

GOODMAN, L. A. (1970): The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal American Statistical,* 65, 226-256.

HABERMAN, S. J. (1973): The analysis of residual in cross-classified tables. *Biometrics,* 29, 205-220.

KNOKE, D. & BURKE, P. J. (1982): *Log-Linear Models.* Beverly Hills, CA: Sage Pubns.

SMITH, J. E. (1973): On tests of quasi-independence in psychological research. *Psychological Bulletin,* 80(4) 329-333.

UPTON, G. (1978): *The analysis of cross-tabulated data.* New York: Wiley.