El problema multivariante de la asociación con variables categóricas

F. Requena
Universidad de Extremadura

1.- INTRODUCCION

Con bastante frecuencia surge en la investigación el problema del estudio de la asociación o dependencia entre variables categóricas, ya sean variables de tipo cualitativo (atributos), que naturalmente vienen expresadas en categorías (sexo, estado civil, región de origen, ambiente, etc.) o variables esencialmente cuantitativas que son expresadas en categorías (edad -expresada en grupos de edad- nivel económico -expresada en varios niveles económicos- etc.). Para ello, disponemos de la información suministrada por un conjunto de datos obtenido mediante variables categóricas consideradas.

Evidentemente, el caso más simple lo tenemos cuando queremos estudiar la asociación entre dos variables categóricas (por ejemplo: Nivel económico y región de origen, ambiente -rural o urbano- y nivel cultural, etc.). Las notaremos por A (con categorías A_1.....A_I) y B (con categorías B_1.....B_J). La estructura de los datos viene representada por una tabla de doble entrada, con I filas y J columnas, que llamamos tabla de frecuencias observadas, tabla de contingencia o, simplemente, tabla IxJ (Tabla I), donde para la casilla ij, x_{ij} representa la frecuencia observada (número de casos) que corresponde a las categorías A_i de A y B_j de B. La tabla se completa con los totales marginales, x_{i+} (total de fila i) y x_{+j} (total de columna j) y el número total, $n = x_{++}$, de individuos en la muestra.

	B_1 ..... B_J	Total
A_1	x_{11} ..... x_{1J}	x_{1+}
	.....	
A_I	x_{I1} ..... x_{IJ}	x_{I+}
Total	x_{+1} ..... x_{+J}	$x_{++} = n$

Tabla I

Para el caso de tres variables, A, B, y C, la estructura de los datos viene expresada en la tabla II, donde el significado de las x_{ijk} y de los marginales es obvio. Igualmente, la generalización al caso de más de tres variables es inmediata.

	C_1			.....		C_K	
	B_1 ..... B_J				B_1 ..... B_J		
A_1	x_{11l} ..... x_{1J1}	x_{1+1}		A_1	x_{11K} ..... x_{1JK}	x_{1+K}
.	.....	.	.....	.	.....	.
A_I	x_{I11} ..... x_{IJ1}	x_{I+1}		A_I	x_{I1K} ..... x_{IJK}	x_{I+K}
	x_{+11} ..... x_{+J1}	x_{++1}			x_{+1K} ..... x_{+JK}	x_{++K}

Tabla II

De forma análoga a las tablas I y II, podemos escribir tablas para las probabilidades de pertenecer a cada casilla. Para el caso de dos variables tendríamos la tabla III, donde p_{ij} es la probabilidad de que un individuo pertenezca a las categorías A_i de A y B_j de B. La condición de independencia entre A y B es

$$p_{ij} = p_{i+} \cdot p_{+j} \qquad i = 1,.....,I \qquad j = 1,....., J$$

Si esto no se cumple, diremos que A y B son dependientes o están asociadas.

	B_1 ..... B_J	Total
A_1	p_{11} ..... p_{1J}	p_{1+}
.	.....	.
A_I	p_{I1} ..... p_{IJ}	p_{I+}
Total	p_{+1} ..... p_{+J}	1

Tabla III

	B_1 ..... B_J	Total
A_1	m_{11} ..... m_{1J}	m_{1+}
.	.....	.
A_I	m_{I1} ..... m_{IJ}	m_{I+}
Total	m_{+1} ..... m_{+J}	$m_{++} = n$

Tabla IV

Finalmente y dadas las p_{ij} anteriores, podemos hablar de la frecuencia o número de casos -del total n de casos de la muestra - que cabría esperar que se presentasen en cada casilla. Para la casilla ij tendremos $m_{ij} = n.p_{ij}$ como el número de casos que podemos esperar que se presenten con A_i y B_j. Así tendremos la tabla de frecuencias esperadas (Tabla IV). La condición de independencia en términos de frecuencia esperadas será

$$m_{ij} = \frac{m_{i+} \cdot m_{+j}}{n} \qquad (1)$$

La generalización, al caso de más de dos variables, de las probabilidades y de las frecuencias esperadas, así como de sus correspondientes tablas, es inmediata.

Otra cuestión de interés es el sistema de muestreo adoptado para obtener la tabla de frecuencias observadas. Aunque existen distintos sistemas o diseños de muestreo para obtener esta tabla, nosostros consideraremos el muestreo multinomial simple, ya que es el más habitual en el problema de asociación. Este consiste en una muestra aleatoria simple de tamaño n, cuyos individuos son clasificados respecto a las variables estudiadas, lo que nos proporciona la citada tabla de frecuencias.

Para el estudio del problema de asociación hemos de construir un modelo matemático adecuado para nuestros datos, que refleje la estructura de los mismos. Para ello, y en primer lugar, estudiaremos los modelos utilizados y su interpretación; en segundo lugar, abordaremos el problema de la estimación de estos modelos y, en tercer lugar, daremos algunas estrategias para la elección del modelo más adecuado. Finalmente, trataremos de dar una medida apropiada para el grado de asociación entre dos variables categóricas.

2.- MODELOS LOG-LINEAL

Aunque en algunas ocasiones se utilizan otros tipos de modelos, son los modelos log-lineal los que se construyen en la mayoría de las situaciones y son los que estudiaremos aquí. El modelo lo podemos construir, indistintamente, para las frecuencias observadas (x), las frecuencias esperadas (m) o las probabilidades (p) de cada casilla. Nosostros lo haremos para las frecuencias esperadas. Describiremos, en primer lugar, el modelo para el caso de dos variables y después generalizaremos al caso multivariante.

2.1.- Modelos log-lineal para dos variables (tabla IxJ)

El modelo log-lineal general para la tabla IV es

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \qquad i=1,...,I \quad j=1,...,J \qquad (2)$$

Lo que nos dice que el logaritmo de las frecuencias esperadas está compuesto de la suma de cuatro términos. El término u es una media global. El término $u_{1(i)}$ representa el efecto principal de la variable A y depende de i, esto es, cada categoría A_i tendrá un valor $u_{1(i)}$. Así, $u + u_{1(i)}$ es la medida correspondiente a la categoría A_i.

$$u + u_{1(i)} = \frac{1}{J} \sum_{j=1}^{J} \log m_{ij} \qquad (3)$$

Y podemos interpretar a $u_{1(i)}$ como la desviación (o diferencia) de la media de A_i respecto de la media global u. Si sumamos todas las desviaciones tendremos:

$$\sum_{i=1}^{I} u_{1(i)} = 0 \qquad (4)$$

Lo mismo decimos de término $u_{2(j)}$ para la variable B. Por último, el término $u_{12(ij)}$ representa la interacción o asociación entre ambas variables y depende de i y de j, esto es, a cada casilla ij de la tabla le corresponde un valor $u_{12(ij)}$ Igual que antes, también se cumple que

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

En el caso en que A y B sean independientes, lógicamente el témino de interacción $u_{12(ij)}$ se anulará y el modelo log-lineal apropiado será

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \qquad i=1,...,I \quad j=1,...,j \qquad (6)$$

Al modelo con todos los término, como el (2), le llamamos modelo saturado y a los que les falta algún término, como el (6), les llamamos modelos no saturados. El estudio de la asociación entre A y B se traduce, pues, en probar cuál de los dos modelos, el (2) o el (6), es el apropiado para nuestros datos.

Para el caso particular en que I=J=2 tendremos una tabla 2x2 (Tabla V). En ella podemos definir la razón de productos cruzados

	B_1	B_2	
A_1	m_{11}	m_{12}	m_{1+}
A_2	m_{21}	m_{22}	m_{2+}
	m_{+1}	m_{+2}	n

$$\alpha = \frac{m_{11} \cdot m_{22}}{m_{12} \cdot m_{21}} = \frac{m_{11}/m_{12}}{m_{21}/m_{22}}$$

Tabla V

que nos puede servir para medir el grado de asociación entre A y B, ya que se puede probar que la condición de independencia (1) es equivalente a la condición α=1 y mientras mayor sea α más nos apartamos de la independencia. Además, existe una relación entre α y el término de interacción del modelo $u_{12(ij)}$ que viene dada por la expresión

$$u_{12(11)} = -u_{12(12)} = -u_{12(21)} = u_{12(22)} = -1/4 - \log \alpha$$

lo que nos permite interpretar el término interacción (esto es, la asociación entre A y B) en función de α, el cual es fácilmente interpretable, según se deduce de su propia definición.

2.2.- Modelos log-lineal para más de dos variables

Para el caso de tres variables tendremos una tabla IxJxK como la tabla II y el modelo log-lineal saturado es

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

$$i=1,....,I \qquad j=1,....,J \qquad k=1,....,k \qquad (7)$$

cumpliéndose todas las condiciones del tipo (4) y (5) para cada término del modelo, esto es, la suma en los subíndices (i, j ó k) que le corresponda se anula. Los términos $u_{1(i)}$, $u_{2(j)}$ y $u_{3(k)}$ son los efectos principales de las variables A, B y C, respectivamente, y tienen la interpretación dada en la sección anterior. Los términos $u_{12(ij)}$, $u_{13(ik)}$ y $u_{23(jk)}$ corresponden a las interacciones entre cada par de variables. El término $u_{123(ijk)}$ representa una interacción conjunta entre las tres variables. Veamos el sentido de esta interacción. En la tabla II tenemos k tablas bidimensionales y podemos construir, por tanto, k modelos log-lineal, uno para cada tabla, con lo que tendríamos k términos de interacción entre A y B, que llamamos

$$v_{11(ij)}^{(k)} \qquad k = 1,.....,k$$

Si todos estos k términos son iguales, podemos decir que la variable C no influye en la interacción entre A y B, con lo que diríamos que no existe (sería nulo) en el modelo este término de interacción conjunta u123(ijk). En caso contrario, la interacción entre A y B depende de la variable C, con lo que decimos que existe una interacción conjunta entre A, B y C, expresada en el término u123(ijk).

Un modelo no saturado para tres variables sería aquel al que le faltase alguno de los términos que figuran en (7). A título ilustrativo, veamos la interpretación de algunos modelos no saturados que se pueden presentar en la práctica:

$$1º/\ \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} \qquad (8)$$

esta asociación para cada par de variables, siendo la asociación entre dos variables la misma para cualquier categoría de la tercera variable. Decimos, en este caso, que existe asociación parcial para cada par de variables.

$$2º/\quad \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)} \qquad (9)$$

Al ser $u_{12(ij)}=u_{123(ijk)}=0$, las variables A y B son independientes para cada nivel de C, pero cada una de ellas está asociada a la variable C. En otras palabras, A y B son condicionalmente independientes, dado cualquier categoría de C. Esto no implica necesariamente que A y B sean completamente independientes.

$$3º/\quad \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} \qquad (10)$$

Las variables B y C están asociadas, siendo esta asociación la misma para todas las categorías de A. Además, la variable A es completamente independiente de B y de C.

$$4º/\quad \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

Al no tener ningún término de interacción, este modelo representa la completa independencia entre las tres variables.

En una situación práctica hemos de encontrar el modelo log-lineal más apropiado a nuestros datos y los términos que posea dicho modelo nos pondrán de manifiesto las asociaciones existentes entre las variables estudiadas.

Por coherencia en la interpretación de los modelos suele admitirse el principio de jerarquización a la hora de buscar el modelo para nuestros datos. Este principio nos dice que si un modelo contiene un determinado término u, habrá de contener también a todos los términos cuyo conjunto de variables implicadas sea un subconjunto de las variables implicadas en el término u dado. Por ejemplo, si $u_{12(ij)}$ está presente en el modelo, dicho modelo habrá de contener los términos $u_{1(i)}$ y $u_{2(j)}$.

La generalización de los modelos log-lineal a cualquier número de variables es inmediata, aunque hay que tener en cuenta la dificultad práctica que conlleva un estudio con muchas variables, dado el gran número de términos que pueden estar presentes en el modelo y la complejidad a la hora de interpretar dichos términos, en especial los que involucran muchas variables. Con cada nueva variable que se incluya en el estudio se duplica el número de términos posibles del modelo. Así, si tenemos s variables en nuestro estudio, el número total de términos posibles en el modelo (modelo saturado) será de 2^s. Al número de variables que tenga un término u le llamamos orden, así, $u_{23(jk)}$ es de orden 2 y $u_{123(ijk)}$ es de orden 3. Para más detalles sobre estos modelos, véase Bishop, Fienberg and Holland (1980), Plackett (1981), Kennedy (1983) y Fienberg (1985).

Una cuestión de interés y que utilizamos más adelante es el número de parámetros independientes de que consta cada término de un modelo log-lineal. Está claro que el término u (media global) tiene un valor único y, por tanto, constituye un parámetro. Los términos de mayor orden tienen distintos valores al variar el subíndice, pero no todos estos valores son independientes. Así, por ejemplo, el término $u_{1(i)}$ lo constituyen I valores (ya que i=1,....I), pero como existe una restricción (4), en realidad serían sólo I-1 valores independientes, ya que el último valor, $u_{1(I)}$, vendría determinado por (4). Decimos, entonces que el término $u_{1(i)}$ tiene I-1 parámetros independientes. Igualmente se puede ver para los términos con dos variables (por ejemplo, $u_{12(ij)}$ tiene (I-1) (J-1) parámetros independientes), con tres variables (por ejemplo, $u_{123(ijk)}$ tiene (I-1) (J-1) (K-1) parámetros independientes), etc. Observemos que el número total de parámetros independientes en un modelo saturado coincide con el número de casillas de la tabla correspondiente.

3.- ESTIMACION DEL MODELO

El objetivo es ahora el siguiente: Dado un determinado modelo log-lineal jerarquizado, cómo estimar las frecuencias esperadas a partir de la información suministrada por los datos, esto es, a partir de la tabla de frecuencias observadas. Supondremos, aquí, que la tabla es completa, caso más habitual en la práctica. Para tablas incompletas, esto es, que tengan algunas casillas que por definición tengan siempre frecuencia observada 0, ver Bishop, Fienberg and Holland (1980).

3.1.- Notación

En primer lugar estableceremos claramente la notación a utilizar. Veámoslo primero para tres variables. Para la casilla ijk de la tabla, notamos x_{ijk}, m_{ijk} y $\hat{m}_{ijk}$ como la frecuencia observada, la frecuencia esperada y la frecuencia esperada estimada, respectivamente. Si en una tabla sumamos las casillas respecto a una variable obtendremos una nueva tabla en una dimensión menos. Por ejemplo, si en

la tabla II sumamos sobre todas las categorías de la variable C tendremos una tabla en dos dimensiones, IxJ, para las variables A y B. Cada casilla de esta nueva tabla será la suma de K casillas de la tabla original. A las frecuencias observadas de la nueva tabla les notaremos por x_{ij+}. Esta notación ya se ha utilizado en las tablas I a IV para expresar los marginales, indicando el subíndice + que se ha sumado sobre la variable que ocupaba este lugar. Lógicamente, se puede sumar en más de una variable y reducir, por tanto, en más de una dimensión la tabla, por ejemplo, x_{+j+} son los valores de una tabla en una dimensión, después de sumar en la primera y tercera variable de una tabla tridimensional. A todas estas tablas sumas les llamamos configuraciones de sumas de tablas o, sencillamente, configuraciones y las denotamos más simplificadamente por la letra C acompañada de subíndices que indican las variables y dimensión de la configuración. Por ejemplo, C12, C23 y C2 representan, respectivamente, a las tablas con valores x_{ij+}, x_{+jk} y x_{+j+}. A la tabla original x_{ijk} la representamos por la configuración C_{123}.

Esta notación se extiende fácilmente al caso general de un conjunto cualquiera de variables. Notamos simbólicamente por φ a dicho conjunto y por φ_1, φ_2,..... a determinados subconjuntos de variables de φ. Así, la notación para las frecuencias observadas, frecuencias esperadas y frecuencias esperadas estimadas será, respectivamente, x_φ , m_φ y $\hat{m}_\varphi$. Igualmente, dado un subconjunto cualquiera, φ_1, de variables de φ, x_{φ_1} serán los valores de la tabla suma o configuración obtenida de la tabla original x_φ al sumar en todas las variables menos en las φ_1 A esta configuración le denotaremos por C_{φ_1}

3.2.- Configuraciones suficientes

Consideremos el modelo (6) para dos variables y queremos estimar las frecuencias esperadas bajo este modelo, esto es, bajo el supuesto de que ambas variables son independientes. La estimación se llevará a cabo, lógicamente, a partir de la tabla I. Esta tabla contiene las frecuencias observadas, que constituyen una configuación C12 y además las marginales, que forman dos configuraciones de una sola dimensión, C1 y C2. Ahora bien, para estimar las frecuencias esperadas en este caso, ¿es necesaria toda la información de la tabla I o sería suficiente con conocer sólo las marginales? Se puede probar que es suficiente con conocer sólo las marginales, es decir, sólo necesitamos conocer las configuraciones C1 y C2. Diremos, pues, que las configuraciones C1 y C2 son suficientes para obtener la estimación de las frecuencias esperadas. En general, el problema es conseguir el conjunto de configuraciones que son suficientes para poder estimar las frecuencias esperadas bajo un determinado modelo. Esto es importante, ya que las propias estimaciones estarán en función de dichas configuraciones suficientes.

Vamos a dar una regla práctica general para obtener el conjunto de configuraciones suficientes. Partimos de un modelo log-lineal cuyo término y de mayor

orden es de orden t. En primer lugar, seleccionamos todos los términos u de orden t que figuren en el modelo. Después, examinamos en sucesivas etapas los términos de orden t-1, t-2,.... presentes en el modelo, seleccionando en cada etapa aquellos términos cuyo conjunto de variables implicadas no sea un subconjunto del conjunto de variables implicadas en algún término de orden mayor que haya sido ya seleccionado. Entonces, si el conjunto total de variables en nuestro estudio es φ y hemos seleccionado r términos u según el proceso anterior, siendo $\varphi_1,\ldots,\varphi_r$ los conjuntos de variables implicadas en los r términos seleccionados, el conjunto de configuraciones suficientes es C_φ ,...., C_φ. Para ilustrar el procedimiento consideremos el modelo (10). Por aplicación del método anterior seleccionaríamos los términos $u_{23(jk)}$ y $u_{1(i)}$ y, por tanto, las configuraciones suficientes serían C23 y C1, es decir, aquellas cuyos elementos son x_{+jk} y x_{i++}, *respectivamente.*

3.3.- Métodos de obtención de las estimaciones

En la práctica, se utilizan fundamentalmente dos métodos para la obtención de las frecuencias esperadas estimadas: Método de estimación directa y método proporcional iterativo.

El método directo, como su nombre indica, nos proporciona fórmulas explícitas con las que podemos obtener directamente las estimaciones de las frecuencias esperadas. Para que este método sea aplicable el modelo ha de cumplir unas condiciones determinadas que veremos más adelante. Sea un método que cumpla las condiciones anteriores, para un conjunto φ de variables estudiadas y cuyas configuracions suficientes son $C\varphi 1$, ...$C\varphi r$, que suponemos están colocadas en un orden tal que las $C\varphi i$ tales que su no contenga ninguna variable de las contenidas en las anteriores $\varphi_{i-1},\ldots,\varphi_1$, están colocadas al final. Entonces podemos llegar a la siguiente expresión general para obtener las estimaciones deseadas.

$$m' = \frac{x_{\varphi_1}\ldots\ldots x_{\varphi_r}}{x_{\beta_2}\ldots\ldots x_{\beta_r}}$$

donde $\beta_i = (\bigcup_{j=1}^{i-1} \varphi_1) \cap \varphi_i$ y si $\beta = \emptyset$ tomamos $x_{\beta_i} = n$

Por ejemplo, para el modelo (6) cuyas configuraciones suficientes son C_1 y C_2, las frecuencias esperadas estimadas se calcularían mediante

$$\hat{m}_{ij} = \frac{x_{i+} \cdot x_{+j}}{n} \qquad (11)$$

Y para el modelo (9) cuyas configuraciones suficientes son C_{13} y C_{23} se obtendrían mediante

$$\hat{m}_{ijk} = \frac{x_{i+k} \cdot x_{+jk}}{x_{++k}}$$

El método proporcional iterativo, en cambio, se puede aplicar para cualquier modelo log-lineal jerarquizado. Como tal procedimiento iterativo, partimos de unas estimaciones iniciales para las frecuencias esperadas y mediante sucesivas iteraciones vamos a llegar, después de adoptar una regla de parada, a nuestras estimaciones, en función únicamente de las configuraciones suficientes. Además, si el modelo permite la aplicación del método directo, por ambos métodos se llegarían a las mismas estimaciones.

Describimos el algoritmo del método para el caso general. Sean las variables en el estudio y tengamos un modelo con configuraciones suficientes $C\varphi 1,\ldots,$ $C\varphi r$. Partimos de unas estimaciones iniciales, $\hat{m}_{\varphi}^{(0)}$, para las frecuencias esperadas. El procedimiento consta de sucesivos ciclos y cada ciclo tiene r etapas, una para ajustar a cada una de las r configuraciones suficientes. En la etapa i (correspondiente a $C_{\varphi i}$) del ciclo s+1 calcularemos las estimaciones $\hat{m}_{\varphi}^{(rs+i)}$ en función de las $\hat{m}_{\varphi}^{(rs+i-1)}$ de la etapa anterior, según la expresión

$$\hat{m}_{\varphi}^{(rs+i)} = \hat{m}_{\varphi}^{(rs+i-1)} \cdot \frac{x_{\varphi_i}}{\hat{m}_{\varphi_i}^{(rs+i-1)}}$$

En la práctica, se recomienda tomar siempre como estimación inicial $m_{\varphi}^{(0)} = 1$. Con ello, el procedimiento siempre converge hacia las estimaciones requeridas, esto es, las estimaciones que vamos consiguiendo en cada etapa van siendo cada vez más precisas. En este sentido, es obvio que debemos buscar una regla que nos permita detener el proceso iterativo cuando hayamos alcanzado la precisión deseada. Como regla de parada podemos utilizar la siguiente: Detener el proceso al final del ciclo t+1 para el que se cumpla que para todas las casillas de la tabla

$$\left| \hat{m}_{\varphi}^{(tr+r)} - m_{\varphi}^{((tr)} \right| < \delta$$

donde δ es un valor pequeño (δ=0,1; δ=0,01;...) prefijado por nosotros en función de la precisión que deseemos. Mientras más pequeño sea d mayor número de ciclos necesitaremos realizar, pero las estimaciones que obtendremos al final serán más precisas.

A título ilustrativo, describimos el método para el caso de tres variables, utilizando el modelo (8) cuyas configuraciones suficientes son C12, C13 y C23. En cada ciclo tendremos tres etapas. Para el primer ciclo tendríamos, para todas las casilla ijk:

$$1^{a}\text{ etapa} \qquad \hat{m}^{(1)}_{ijk} = \hat{m}^{(0)}_{ijk} \cdot \frac{x_{ij+}}{\hat{m}^{(0)}_{ij+}}$$

$$2^{a}\text{ etapa} \qquad \hat{m}^{(2)}_{ijk} = \hat{m}^{(1)}_{ijk} \cdot \frac{x_{i+k}}{\hat{m}^{(1)}_{i+k}}$$

$$3^{a}\text{ etapa} \qquad \hat{m}^{(3)}_{ijk} = \hat{m}^{(2)}_{ijk} \cdot \frac{x_{+jk}}{\hat{m}^{(2)}_{+jk}}$$

La primera etapa del segundo ciclo partiría de las estimaciones $\hat{m}^{(3)}_{ijk}$. Así seguiríamos ciclo tras ciclo hasta detener el proceso al final del ciclo t+1 en el que se cumpla para toda casilla ijk.

$$\left| \hat{m}^{(3t+3)}_{ijk} - \hat{m}^{(3t)}_{ijk} \right| < \delta$$

Las estimaciones buscadas serían, pues, $\hat{m}^{(3t+3)}_{ijk}$.

Aunque el método proporcional iterativo es satisfactorio, no cabe duda que siempre que podamos aplicar el método directo lo aplicaremos. Por ello, conviene saber para qué modelos podemos aplicar el método directo. A continuación vamos a establecer una regla práctica para determinar si para un modelo, con sus correspondientes configuraciones suficientes, existe estimación directa. Sobre dichas configuraciones se llevan a cabo los siguientes pasos:

a) Reetiquetar como una nueva variable cualquier conjunto de variables que siempre aparezcan juntas.

b) Eliminar las variables que aparezcan en todas las configuraciones suficientes.

c) Eliminar cualquier variable que solamente aparezca en una configuración suficiente.

d) Eliminar las configuraciones suficientes que sean redundantes (aquéllas que sus variables sean un subconjunto de las variables de otra configuración suficiente).

e) Repetir los pasos a) al d) hasta que:

1) Se reduzca el número de configuraciones suficientes a sólo dos, en cuyo caso existe estimación directa.

2) No se puede reducir el número de configuraciones suficientes a sólo dos, en cuyo caso no existe estimación directa y tenemos que aplicar el método proporcional iterativo.

Los siguientes ejemplos ilustran la regla. En cada uno de ellos se parte de las configuraciones suficientes del modelo.

1º $C_{12}, C_{13}, C_{24} \rightarrow C_{12}, C_{13}, C_{2} \rightarrow C_{12}, C_{13}$

Existe estimación directa.

2º $C_{1235}, C_{1234}, C_{145} \xrightarrow{(Z=23)} C_{12'5}, C_{12'4}, C_{145} \rightarrow$

$\rightarrow C_{2'5}, C_{2'4}, C_{45}$. No existe estimación directa

4.- BONDAD DE AJUSTE Y ELECCION DE MODELOS

Entre otras varias cuestiones, nos interesa abordar ahora las siguientes:

a) Seleccionar un modelo log-lineal jerarquizado que describa adecuadamente a los datos. Este nos pondrá de manifiesto las relaciones o asociaciones existentes entre las variables. Cuando ajustemos el modelo, las frecuencias esperadas de las casillas nos proporcionarán una descripción suavizada de los datos, ya que son eliminadas las fluctuaciones aleatorias, aun manteniendo la estructura de los datos y las principales relaciones entre las variables.

b) Contrastar si dos variables específicas de nuestro estudio están asociadas o no y evaluar la magnitud de tal asociación. Y, en general, evaluar la magnitud de cualquier término de interacción entre variables.

Sea cual sea nuestro propósito, lo que parece evidente es que necesitamos una medida de la bondad de ajuste del modelo log-lineal considerado a los datos. Hay varias medidas con tal fin, pero la más conocida y utilizada es la de Chi-Cuadrado de Pearson, definida como

$$\chi^2 = \sum_{\varphi} \frac{(x_\varphi - \hat{m}_\varphi)^2}{\hat{m}_\varphi}$$

donde las $\hat{m}_\varphi$ son las frecuencias esperadas estimadas bajo el modelo del que se quiere medir su ajuste a los datos o frecuencias observadas x_φ. Para el caso más simple de dos variables, será

$$\chi^2 = \sum_{ij} \frac{(x_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \qquad (12)$$

Según se desprende de la definición χ^2, el ajuste del modelo a los datos será tanto mejor cuanto menor sea el valor de χ^2.

Ahora, una cuestión de fundamental importancia es determinar el número de grados de libertad que le corresponde al modelo que estamos considerando para la estimación de las frecuencias esperadas. Este número lo podemos obtener mediante la expresión V = Te - Tp, donde Te es número total de casillas de la tabla y Tp es el número total de parámetros independientes de todos los terminos del modelo. Por ejemplo, para el modelo (6) los grados de libertad serían

$$V = IJ - (1 + (I-1) + (J-1)) = (I-1)(J-1)$$

y para el modelo (9) serían

$$V = IJK - (1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) + (J-1)(K-1) = K(I-1)(J-1)$$

Esta regla se aplica sin problemas en los casos en que ninguna frecuencia esperada estimada sea 0, cosa que ocurre en la mayoría de las aplicaciones prácticas. En caso contrario, el cálculo de los grados de libertad es más complejo (Bishop, Fienberg and Holland (1980).

4.1.- Tests de hipótesis para contrastar una asociación

Si nuestro interés son cuestiones del tipo de las expresadas en el apartado b) anterior, habremos de realizar tests de hipótesis estadísticos, para llegar a la

conclusión de si la asociación entre las variables de interés es estadísticamente significativa, esto es, si podemos afirmar, bajo un nivel de error prefijado, que exista tal asociación. Para ello se habrá de obtener un valor experimental (estadístico de contraste) en función de las frecuencias observadas y de las esperadas estimadas y un valor teórico obtenido a partir de la distribución de probabilidad Chi-Cuadrado con un determinado número de grados de libertad, que vendrá especificado según la test que se esté realizando. De la comparación de ambos valores llegaremos a una conclusión sobre si tal asociación es significativa o no. (Para las nociones básicas sobre tests de hipótesis, ver cualquier texto sobre Estadística Inferencial).

Para probar si un término de interacción u que incluya un determinado número de variables (que representará la magnitud de la asociación conjunta entre esas variables) es estadísticamente significativo o no, ajustamos primeramente un modelo que incluya dicho término, pero que no incluya ningún término que contenga todas las variables del término que estamos probando, al que llamaremos modelo (a), y calcularemos la correspondiente medida de bondad de ajuste, que representaremos por $\chi^2(a)$. Después ajustaremos un segundo modelo, que sólo se diferencia del anterior en que no contenga el término de interacción que estamos probando y le llamaremos modelo (b). Calculamos, igualmente, su medida de bondad de ajuste $\chi^2(b)$. Entonces, el valor experimental del test lo calculamos como la diferencia $\chi^2(b) - \chi^2(a)$ y el valor teórico lo obtenemos de la distribución Chi-Cuadrado vb - va grados de libertad, donde vb y va son los grados de libertad que corresponde a los modelos (b) y (a), respectivamente. En general, el par de modelos (b) y (a) tal como lo hemos definido no es único. En este sentido y para realizar esta prueba de hipótesis, se recomienda tomar un par, de tal forma que uno de los modelos del par se ajuste bien a los datos.

Un caso particular del test anterior es la clásica prueba de Chi-Cuadrado para probar la asociación entre dos variables, a partir de la tabla I. En este caso el modelo (a) sería el modelo (2) con 0 grados de libertad y el modelo (b) sería el modelo (6) con (I-1) (J-1) grados de libertad. Además, en este caso podemos ver que $\chi^2(a)=0$, con lo que el valor experimental del test es el clásico estadístico Chi-Cuadrado (12), donde se sustituyen las mij por su valor para el modelo (6), dado en la expresión (11). El valor teórico se obtiene de la Chi-Cuadrado con (I-1)(J-1) grados de libertad.

Cuando tenemos una tabla multidimensional (más de dos variables) y queremos probar la significación estadística de la asociación entre dos variables dicotómicas (sólo tienen dos categorías), aunque se puede utilizar el procedimiento general anterior, es más cómodo y habitual en la práctica utilizar el test de Mantel-Haenszel (Mantel and Haenszel, 1959) y Mantel, 1963).

Por último, siempre podemos probar la significación estadística de un modelo concreto, de tal forma que si el test de hipótesis resulta significativo (para un determinado nivel preestablecido) concluiremos que el modelo no se ajusta bien a los datos (esto ocurrirá cuando su $\chi 2$ sea suficientemente grande). En caso contrario -test no significativo- admitiremos al modelo como adecuado para nuestros

datos (su $\chi 2$ será en este caso pequeña). La prueba se realiza tomando como valor experimental la $\chi 2$ calculada para ese modelo y como valor teórico el obtenido de la distribución Chi-Cuadrado con los grados de libertad que correspondan a dicho modelo.

4.2.- Elección del modelo

En principio el modelo a elegir habría de ser el que mejor se ajuste o describa a los datos, según un criterio establecido. El criterio comúnmente adoptado es el de la bondad del ajuste, dado por $\chi 2$. Así, habría que calcular la bondad de ajuste para todos los posibles modelos jerarquizados y elegir el de mejor ajuste. Pero esto es muy costoso, sobre todo si el número de variables no es pequeño (para cuatro variables tenemos 113 posibles modelos jerarquizados y conforme aumentamos el número de variables, el número de modelos se dispara).

Necesitamos, pues, una estrategia que nos permita llegar a seleccionar un modelo adecuado con un mínimo de cálculos. Se han propuesto varias estrategias, sin que pueda decirse que una es siempre mejor que las demás. Veremos, aquí, dos de ellas.

Estrategia 1

La motivación de esta estrategia es tratar de cubrir todos los posibles modelos. Para ello, en un primera etapa se ajusta un subconjunto de modelos, cada uno de ellos con términos de orden uniforme. Esto es, para s variables serían:

mod. 1. No posee término u de orden 2.
mod. 2. Posee todos los términos de orden 2, pero ninguno de orden 3.
.....
mod. (s-1). Posee todos los términos de orden 2-1 pero no el de orden s.

Examinamos la bondad de ajuste de estos modelos y nos quedamos con el par de modelos consecutivos, mod. r-1 y mod. r, tal que el mod. r-1 ajuste pobremente y el mod. r ajuste bien a los datos. Entonces el próximo paso es investigar los modelos intermedios entre los dos del par y se elegirá el más apropiado. Una forma de llevar a cabo esto es la conocida como selección adelantada. Se parte del mod. r-1 como modelo base y se le añade el término u de orden r que proporcione el mayor aumento en bondad de ajuste, siempre que el término u añadido sea estadísticamente significativo según vimos en 4.1. Este será nuestro nuevo modelo base y seguimos el proceso con los restantes términos u de orden r, hasta que ninguno de los posibles términos u de orden r a añadir sea estadísticamente significativo (ninguno de ellos aportaría un aumento significativo de la bondad de ajuste). El modelo conseguido de esta manera será el elegido.

Estrategia 2

Un modelo de asociación parcial es aquel al que sólo le falta un término de interacción de orden 2 y, lógicamente, todos los de orden superior que no puedan estar por el principio de jerarquización. Un modelo de este tipo tiene siempre sólo dos configuraciones suficientes, con lo que se puede aplicar el método de estimación directa y, por tanto, es fácil obtener su bondad de ajuste.

Así, en una primera etapa, podemos ajustar todos los modelos de asociación parcial y probar su significación estadística (como se vio en 4.1.). Naturalmente, si para uno de estos modelos nos sale significativo quiere decir que el único término de orden 2 que no posee debería figurar en nuestro modelo y lo contrario si nos sale no significativo. De esta manera, podemos determinar qué términos de orden 2 deben figurar realmente en nuestro modelo y los incluimos en él.

En una segunda etapa, inspeccionamos todos los posibles términos de orden 3 que, según el principio de jerarquización, puedan estar presentes en nuestro modelo e incluiremos en él todos los que sean estadísticamente significativos, según vimos en 4.1. Por este procedimiento construimos un modelo que será el que elegiremos para nuestros datos.

En cualquiera de las estrategias y como último paso, siempre podemos inspeccionar algunos modelos adyacentes al elegido (que se diferencia de él en un solo término), para ver si mejoran significativamente la bondad del ajuste.

En la práctica, el concepto de "mejor modelo" no es tan fácil de definir. Deberíamos tener en cuenta otros factores además de la bondad del ajuste. Si calculamos la bondad de ajuste para todos los posibles modelos y seleccionamos el de mejor ajuste, en la práctica, ¿es éste el "mejor modelo"? Puede ocurrir que este modelo tenga conjuntos de términos y que no sean fáciles de interpretar en la práctica y, en cambio, es probable que exista otro modelo cuya diferencia con el anterior no sea estadísticamente significativa (en el sentido de 4.1.) y sea mucho más fácil de interpretar. Parece razonable elegir a este último.

4.3.- Medidas de asociación

Tratamos, ahora, de dar una medida del grado de asociación entre dos variables categóricas que sean de interés. Para ello, partiremos de una tabla bidimensional, del tipo de la tabla I, para ambas variables y, por coherencia, habremos probado previamente que el término $u_{12(ij)}$ del modelo (2) es estadísticamente significativo. Se han dado bastantes medidas a tal fin. Según sea nuestro propósito y matización que queramos darle al concepto de asociación, elegiremos una u otra medida.

Si entendemos o interpretamos la asociación como el grado en el que se apartan de la independencia las variables, podemos utilizar como medida de asociación una función del grado de ajuste χ^2 del modelo de independencia (6) para

ambas variables. En este sentido, se han definido varias medidas, entre otras, el coeficiente de contingencia de Pearson

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}$$

que puede variar entre $\sqrt{(q-1)/q}$ $(q=\min\ (I, J))$. Mientras mayor sea C mayor es el grado de la asociación y mientras más próximo a 0 más cerca estaremos de la independencia entre las dos variables.

Una medida similar, pero normalizada, es la dada por Cramer,

$$V=\sqrt{\frac{\chi^2/n}{\min\ ((I-I), (J-1))}}$$

que varía entre 0 (caso de independencia) y 1 (caso de máxima asociación).

Por otra parte, si nos interesa estudiar la asociación con el objetivo de tratar de predecir una variable en fución de la otra, parece razonable matizar o entender el concepto de asociación como grado de información que nos proporciona una variable para predecir la otra. Podemos hacer la predicción bajo dos supuestos distintos: (a) ambas variables son independientes y (b) una variable es función de la otra. Naturalmente, al hacer la predicción la probabilidad de error en (a) es mayor que en (b). En esta situación, nos puede servir como medida de la asociación la reducción proporcional en el error, esto es,

$$\frac{\text{Prob. de error en (a) - Prob. de error en (b)}}{\text{Prob. de error en (a)}}$$

En este sentido, Goodman y Kruskal han definido la medida $\lambda_{B/A}$ para evaluar la reducción proporcional en el error cuando se predice B en función de A:

$$\lambda_{B/A}=\frac{\sum_{i=1}^{I} x_{im} - x_{+m}}{n - x_{+m}}$$

donde $x_{im}=\max\ (x_{i1}, \ldots, x_{ij})$ y $x_{+m}=\max\ (x_{+1}, \ldots, x_{+J})$

Esta medida varía entre 0 y 1, indicándonos el valor 0 que la variable A no nos proporciona ninguna información para predecir B y el valor 1 nos dice que la probabilidad de error al predecir B en función de A es nula.

También podemos entender la asociación, con fines de predicción, como el porcentaje de la variación de una variable que es explicada por la otra. Con este propósito y basado en el concepto de variación de Gini, podemos definir la siguiente medida de asociación, $\tau B/A$, que varía entre 0 y 1 y que si la multiplicamos por 100 nos dará el porcentaje de la variación de B que es debida a A.

$$\tau_{B/A} = \frac{\sum_j \frac{1}{x_{+j}} \cdot \sum_i x_{ij}^2 - \frac{1}{n} \sum_i x_{i+}^2}{n - \frac{1}{n} \sum_i x_{i+}^2}$$

Para el caso particular de tablas 2x2, además de la particularización de las anteriores, se han definido otras medidas. En función de χ^2 se ha definido el coeficiente

$$\emptyset = \sqrt{\frac{\chi^2}{n}} = \frac{x_{11} \cdot x_{22} - x_{12} \cdot x_{21}}{\sqrt{x_{1+} \cdot x_{2+} \cdot x_{+1} \cdot x_{+2}}}$$

que varía entre -1 y 1, tomando estos valores para cuando el grado de asociación es máximo y el valor 0 para el caso de independencia. Y en función de la razón de productos cruzados α se han definido varias medidas, una de ellas es la medida de asociación de Yule

$$Q = \frac{x_{11} \cdot x_{22} - x_{12} \cdot x_{21}}{x_{11} \cdot x_{22} + x_{12} \cdot x_{21}}$$

con el mismo rango de variabilidad y similar interpretación que Ø. Otros tipos de medidas de asociación han sido dadas por Altham (1970) y extensiones para el caso de más de dos variables categóricas han sido estudiadas por Davis (1971).

BIBLIOGRAFIA

ALTHAM, P.M.E. (1970): *The measurement of association of rows and columns for an rxs contingency table.* **J. Roy. Statist. Soc. Ser** B 32, 63-73.

BISHOP, Y.; FIENBERG, S. and HOLLAND, P. (1980): **Discrete multivariante analysis.** MIT Press.

DAVIS, J.A. (1971): **Elementary survey analysis.** Englewood Cliffs, N.J., Prentice- Hall.

FIENBERG, S.E. (1985): **The analysis of cross-classified categorical data.** MIT Press.

KENNEDY, J.J. (1983): **Analyzing qualitative data. Introductory log-linear analysis for behavioral research.** Praeger.

MANTEL N. (1963): *Chi-Square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure.* **J. Am. Stat. Assoc.**, 58, 690-700.

MANTEL, N. and HAENSZEL, W (1959): *Statistical aspects of the analysis of data from retrospective studies of disease.* **J. Natl. Cancer Inst.** 22, 719-48.

PLACKETT, R.L. (1981): *The analysis of categorical data.* **Griffin's Statis cal.** Monographs nº 35.