ESTADÍSTICA ESPAÑOLA
Vol. 44, Núm 149, 2002, págs. 29 a 46

Una propuesta de corrección de continuidad asimétrica para tablas de contingencia (2x2) con totales marginales fijos

por
JOSÉ MARÍA MONTERO LORENZO
Departamento de Economía y Empresa (Área de Estadística)
Universidad de Castilla-La Mancha

RESUMEN

En este trabajo se propone una corrección de continuidad asimétrica para tablas de contingencia (2x2) con totales marginales fijos que supera ampliamente las popularizadas por Yates, Mantel y Cochran. Se exponen las infraestimaciones y sobreestimaciones de la denominada "probabilidad exacta" a que conduce la corrección de continuidad de Yates, la corrección más utilizada que, al ser simétrica, aproxima mal la probabilidad exacta y, en determinadas situaciones, empeora la aproximación proporcionada por el estadístico ji-cuadrado no corregido de continuidad. La inclusión de la condición de asimetría en la generación de correcciones de continuidad proporciona, sin embargo, magníficas aproximaciones.

Palabras clave: tabla de contingencia (2x2), contraste de independencia, corrección de continuidad, corrección de Yates, test exacto de Fisher, probabilidad exacta, estadístico ji-cuadrado ajustado, corrección asimétrica.

Clasificación AMS: 62H17

1. INTRODUCCIÓN

La manera de proceder al contraste de independencia poblacional en una tabla de contingencia bifactorial no es otra que el cálculo de la probabilidad (bajo dicha hipótesis) de obtención de la estructura de frecuencias observada (tabla observada) y de todas aquéllas otras que evidencien al menos igual alejamiento de la hipótesis de independencia que la tabla observada (el alejamiento se entiende en la dirección marcada por la hipótesis alternativa). Una vez calculadas dichas probabilidades, se suman y esta suma se compara con el nivel de significación prefijado, con objeto de determinar si la estructura de frecuencias observada proporciona evidencia suficiente en contra de la hipótesis de independencia formulada.

A la probabilidad a la que acabamos de aludir, probabilidad que se comparará con el nivel de significación prefijado para decidir sobre el rechazo o no de la hipótesis de independencia, se le denominará en lo sucesivo, de forma abreviada, "probabilidad exacta".

Dada la laboriosidad que puede implicar el cálculo de la probabilidad exacta, se suele proporcionar una aproximación a la misma mediante una distribución Ji-cuadrado con un grado de libertad. El problema que surge es que la probabilidad exacta, calculada mediante una distribución discreta de probabilidad (hipergeométrica, binomial bivariante, multinomial, Poisson o binomial negativa bivariante, dependiendo del procedimiento de muestreo), se aproxima a través de una distribución continua de probabilidad (la distribución Ji-cuadrado)(1).

Tradicionalmente se ha considerado, en el caso de tablas de contingencia de orden (2x2), que dicha aproximación no es suficientemente buena si alguna de las estimaciones de las frecuencias esperadas, calculadas bajo la hipótesis de independencia, es inferior a 5, o si el total muestral es inferior a 20 (por ejemplo el paquete informático SPSS), y se ha señalado la necesidad de aplicar algún elemento corrector.

Las correcciones de continuidad surgen, pues, como intento de compensación de los desajustes que tienen lugar cuando la distribución de probabilidad de las frecuencias observadas, que es discreta, es aproximada por otra de carácter continuo. Se comete un error al calcular una determinada probabilidad, no mediante una distribución discreta, sino a través de su aproximación continua, y se pretende

(1) Véase RUIZ-MAYA PÉREZ, L.; MARTIN PLIEGO, F.J.; MONTERO LORENZO, J.M.; URIZ TOMÉ, P. (1995): "Análisis Estadístico de Encuestas: Datos Cualitativos". A.C., Madrid.

solventar dicho error "corrigiendo de la continuización realizada" o, simplemente, corrigiendo de continuidad.

Una vez vista la utilidad de las correcciones de continuidad surge la siguiente cuestión: ¿La manera de corregir de continuidad es independiente del diseño del experimento? o, por el contrario, ¿dependiendo de cuál sea el modelo la corrección de continuidad se lleva a cabo de una u otra manera?. La respuesta no es, ni mucho menos, obvia, pero las investigaciones llevadas a cabo en los últimos años abogan por diferentes correcciones de continuidad para diferentes diseños.

2. DISEÑO CON LOS TOTALES MARGINALES FIJOS: DEFICIENCIAS DE LA CORRECCIÓN DE YATES EN LA APROXIMACIÓN A LA PROBABILIDAD EXACTA EN TABLAS (2X2)

El diseño o procedimiento de muestreo en el que nos vamos a centrar es aquél en el que los totales marginales de ambos factores se consideran fijos, pues es el que contemplan los paquetes informáticos al uso para la realización de un "test exacto" que, en este caso, recibe el nombre de test exacto de Fisher.

En la literatura estadística existen célebres ejemplos de diseños experimentales para tablas de contingencia de orden (2x2) en los que los totales marginales de ambos factores se consideran fijos, entre los cuales el más destacado es el ya clásico de la dama inglesa que aseguraba saber discernir si en el té con leche se vertía primero el té o la leche(2).

En este diseño, la probabilidad exacta, es decir, la probabilidad de obtener, bajo la hipótesis nula de independencia, la tabla observada o aquellas otras con igual o mayor alejamiento (en cualquier dirección) de la hipótesis nula, se obtiene mediante la expresión(3)

$$P\left(\left|N_{ij}-\hat{E}_{ij}\right|\geq\left|n_{ij}-\hat{E}_{ij}\right|\right),\ \text{para cualquier } ij$$

donde n_{ij} es la frecuencia observada en la celda ij, $\hat{E}_{ij}$ es la estimación de la frecuencia esperada en dicha celda bajo la hipótesis de independencia y N_{ij} sigue, supuesta la hipótesis, una distribución de probabilidad hipergeométrica, $H(n;n_{.j};n_{i.})$,

(2) Véase FISHER, R.A. (1935). "The design of experiments". 8ª ed. 1966. Oliver and Boyd. Edinburgh.

(3) Véase RUIZ-MAYA PÉREZ, L.;MARTIN PLIEGO, F.J.; MONTERO LORENZO, J.M.; URIZ TOMÉ, P. (1995): "Análisis Estadístico de Encuestas: Datos Cualitativos". A.C., Madrid.

distribución de probabilidad discreta, por estar fijados los totales marginales de los dos factores involucrados en la tabla bifactorial.

La anterior probabilidad, si se dan las condiciones apropiadas, se puede aproximar a través de(4)

$$P\left[\chi_1^2 \geq \frac{(n-1)(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}\right]$$

donde, como es sabido, la distribución Ji-cuadrado con un grado de libertad es una distribución de probabilidad continua.

Si bien al aproximar la anteriormente denominada "probabilidad exacta" a través de una distribución continua se simplifica mucho el contraste, no es menos cierto que se está cometiendo un "error de continuidad", es decir, un error debido a la aproximación de una distribución de probabilidad discreta mediante otra continua. Para corregir ese "error de continuidad", Yates propuso una corrección que se opera en el estadístico χ^2_{ajd} (5)

$$\chi^2_{ajd} = \frac{(n-1)(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} = \frac{(n-1)}{n}\sum_i\sum_j\frac{(n_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$$

y que consiste en restar 0,5 a las desviaciones positivas de las frecuencias observadas (n_{ij}) respecto de las estimaciones de las esperadas bajo la hipótesis de independencia ($\hat{E}_{ij}$), y sumar 0,5 en caso de que dichas desviaciones sean negativas, y ello siempre antes de elevar al cuadrado las anteriores desviaciones. En otros términos, la corrección de continuidad de Yates consiste en restar 0,5 al valor absoluto de las diferencias entre n_{ij} y $\hat{E}_{ij}$, obteniéndose una aproximación a la "probabilidad exacta", ya corregida de continuidad, mediante

(4) Ibidem.

(5) En realidad se opera en el estadístico χ^2, pero se expone sobre el estadístico χ^2_{ajd} por ser este último más apropiado para llevar a cabo el contraste de la hipótesis de independencia en este diseño. La demostración de la equivalencia de las dos expresiones expuestas del estadístico χ^2_{ajd} puede verse en RUIZ-MAYA PEREZ, L.;MARTIN PLIEGO, F.J.; MONTERO LORENZO, J.M.; URIZ TOME, P; LOPEZ ORTEGA, J. (1990): "Metodología Estadística para el Análisis de Datos Cualitativos". C.I.S., Madrid.

$$P\left[\ \chi_1^2 \geq \frac{(n-1)}{n}\sum_i\sum_j\frac{\left(\left|\, n_{ij} - \hat{E}_{ij}\right| - 0{,}5\,\right)^2}{\hat{E}_{ij}}\ \right]$$

o bien a través de

$$P\left[\ \chi_1^2 \geq \frac{(n-1)\,(\,|\,n_{11}n_{22} - n_{12}n_{21}\,| - 0{,}5\,n)^2}{n_{1.}\,n_{2.}\,n_{.1}\,n_{.2}}\ \right]$$

En general, cualquiera que sea ij, la probabilidad de obtener la tabla observada o aquellas otras que se alejen de la hipótesis de independencia tanto o más que ella viene dada por

$$P\left[\left|N_{ij} - \hat{E}_{ij}\right| \geq \left|n_{ij} - \hat{E}_{ij}\right|\right] =$$

$$P\left[N_{ij} - \hat{E}_{ij} \leq -\left|n_{ij} - \hat{E}_{ij}\right|\right] + P\left[N_{ij} - \hat{E}_{ij} \geq \left|n_{ij} - \hat{E}_{ij}\right|\right] =$$

$$P\left[N_{ij} \leq \hat{E}_{ij} - \left|n_{ij} - \hat{E}_{ij}\right|\right] + P\left[N_{ij} \geq \hat{E}_{ij} + \left|n_{ij} - \hat{E}_{ij}\right|\right]$$

y si $n_{ij} \leq \hat{E}_{ij}$

$$P\left[N_{ij} \leq \hat{E}_{ij} - \left|n_{ij} - \hat{E}_{ij}\right|\right] + P\left[N_{ij} \geq \hat{E}_{ij} + \left|n_{ij} - \hat{E}_{ij}\right|\right] =$$

$$P\left[N_{ij} \leq \hat{E}_{ij} + \left(n_{ij} - \hat{E}_{ij}\right)\right] + P\left[N_{ij} \geq \hat{E}_{ij} - \left(n_{ij} - \hat{E}_{ij}\right)\right] =$$

$$P\left[N_{ij} \leq n_{ij}\right] + P\left[N_{ij} \geq 2\hat{E}_{ij} - n_{ij}\right]$$

y si $n_{ij} \geq \hat{E}_{ij}$

$$P\left[N_{ij} \leq \hat{E}_{ij} - \left|n_{ij} - \hat{E}_{ij}\right|\right] + P\left[N_{ij} \geq \hat{E}_{ij} + \left|n_{ij} - \hat{E}_{ij}\right|\right] =$$

$$P\left[N_{ij} \leq \hat{E}_{ij} - \left(n_{ij} - \hat{E}_{ij}\right)\right] + P\left[N_{ij} \geq \hat{E}_{ij} + \left(n_{ij} - \hat{E}_{ij}\right)\right] =$$

$$P\left[N_{ij} \leq 2\hat{E}_{ij} - n_{ij}\right] + P\left[N_{ij} \geq n_{ij}\right]$$

La inclusión de la corrección de continuidad de Yates en las expresiones anteriores llevaría a

$$P\left[\left|N_{ij}-\hat{E}_{ij}\right| \geq \left|n_{ij}-\hat{E}_{ij}\right|-\frac{1}{2}\right] =$$

$$P\left[N_{ij}'-\hat{E}_{ij} \leq -\left|n_{ij}-\hat{E}_{ij}\right|+\frac{1}{2}\right]+P\left[N_{ij}-\hat{E}_{ij} \geq \left|n_{ij}-\hat{E}_{ij}\right|-\frac{1}{2}\right] =$$

$$P\left[N_{ij} \leq \hat{E}_{ij}-\left|n_{ij}-\hat{E}_{ij}\right|+\frac{1}{2}\right]+P\left[N_{ij} \geq \hat{E}_{ij}+\left|n_{ij}-\hat{E}_{ij}\right|-\frac{1}{2}\right]$$

de tal forma que:

A) Si $n_{ij} \leq \hat{E}_{ij}$, entonces

$$P\left[| N_{ij}-\hat{E}_{ij} | \geq | n_{ij}-\hat{E}_{ij} | -\frac{1}{2}\right]=P\left[N_{ij} \leq n_{ij}+\frac{1}{2}\right]+P\left[N_{ij} \geq 2\hat{E}_{ij}-n_{ij}-\frac{1}{2}\right]$$

B) Si $n_{ij} \geq \hat{E}_{ij}$, tendríamos

$$P\left[| N_{ij}-\hat{E}_{ij} | \geq | n_{ij}-\hat{E}_{ij} | -\frac{1}{2}\right]=P\left[N_{ij} \leq 2\hat{E}_{ij}-n_{ij}+\frac{1}{2}\right]+P\left[N_{ij} \geq n_{ij}-\frac{1}{2}\right]$$

donde la distribución de probabilidad de N_{ij} bajo la hipótesis de independencia, hipergeométrica, se puede aproximar, si se dan las condiciones, mediante una ley normal

$$N_{ij} \underset{aprox}{-} N\left(\frac{n_{i.}n_{.j}}{n}; \sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}\right)$$

Denominando θ a la parte entera de $2\hat{E}_{ij}$ - n_{ij}, se puede establecer la siguiente casuística(6):

---

(6) En dicha casuística podríamos haber separado el caso en que $2\hat{E}_{ij}-n_{ij}=\theta+\frac{1}{2}$, pero hemos preferido no hacerlo por no extendernos en demasía. Dicho caso será tratado como caso particular de las situaciones A1), A2), B1) y B2).

Caso A.1)

$$n_{ij} \leq \hat{E}_{ij}$$

y además

$$\theta < 2\hat{E}_{ij} - n_{ij} \leq \theta + \frac{1}{2}$$

se aproxima la probabilidad exacta mediante el estadístico χ^2_{ajd} , o lo que es igual, mediante

$$P\left[N_{ij} \leq n_{ij}\right] + P\left[N_{ij} \geq 2\hat{E}_{ij} - n_{ij}\right] \quad \text{ya que } n_{ij} \leq \hat{E}_{ij}$$

con

$$N_{ij} \underset{aprox}{\longrightarrow} N\left(\frac{n_{i.}n_{.j}}{n}; \sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}\right)$$

tienen lugar dos distorsiones(7):

1) Infraestimación:

$$\frac{1}{2}P\left[N_{ij} = n_{ij}\right]$$

2) Sobreestimación

$$\left[\left(\theta + \frac{1}{2}\right) - \left(2\hat{E}_{ij} - n_{ij}\right)\right]P\left[N_{ij} = \theta\right]$$

entendiéndose por "infraestimación" la aproximación a la probabilidad exacta por defecto y por "sobreestimación" la aproximación por exceso. Incluyendo la corrección de continuidad de Yates, es decir, aproximando la probabilidad exacta mediante

(7) La segunda distorsión, es decir la sobreestimación, es nula en el caso en que $2\hat{E}_{ij} - n_{ij} = \theta + \frac{1}{2}$

$$P\left[N_{ij} \leq n_{ij} + \frac{1}{2}\right] + P\left[N_{ij} \geq 2\hat{E}_{ij} - n_{ij} - \frac{1}{2}\right] \text{ ya que } n_{ij} \leq \hat{E}_{ij}$$

1) Se corrige la infraestimación

2) Aumenta la sobreestimación hasta:

$$\left[(\theta + 1) - \left(2\hat{E}_{ij} - n_{ij}\right)\right] P\left[N_{ij} = \theta\right]$$

Como ilustración del caso A.1., y a modo de ejemplo, considérense la tabla y el gráfico que se presentan a continuación.

Tabla 1

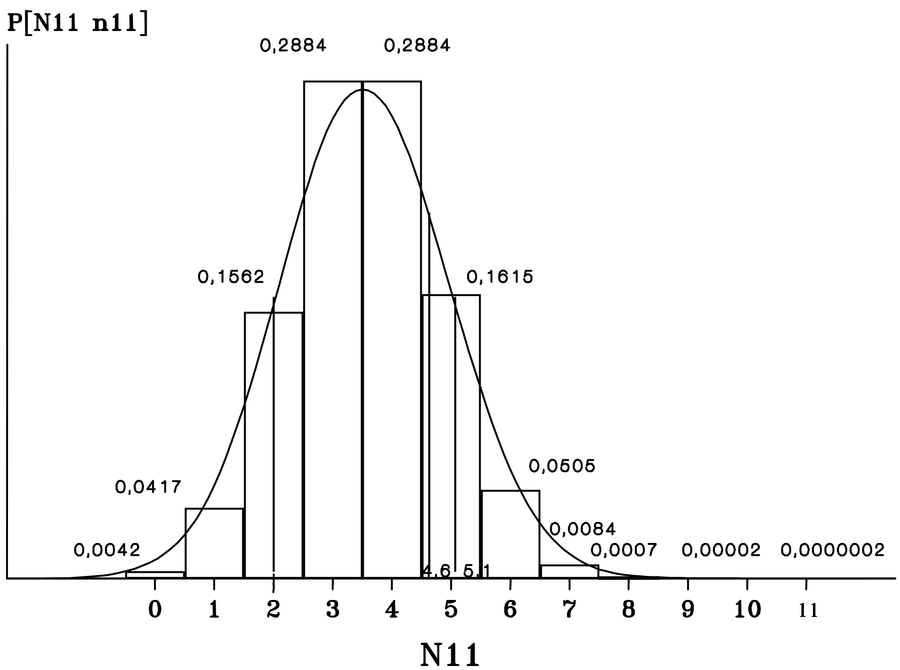
	Factor A		
Factor B	Nivel 1	Nivel 2	
Nivel 1	2	8	10
Nivel 2	9	12	21
	11	20	31

donde

$$\hat{E}_{11} = 3{,}5483871 \ ; \ n_{11} \leq \hat{E}_{11} \ ; \ \theta < 2\hat{E}_{11} - n_{11} < \theta + \frac{1}{2}$$

Gráfico 1

$P(N_{11}=n_{11})$



Con la celda (1;1) de referencia, la probabilidad exacta viene dada, gráficamente, por el área rectangular correspondiente a los valores de N_{11} menores o iguales que 2 o mayores o iguales que 5,1; es decir, el área rectangular correspondiente a los valores: 0, 1, 2, 6, 7, 8, 9, 10. La aproximación normal (o mediante el estadístico χ^2_{ajd}) viene dada por el área bajo la normal a la izquierda de 2 y a la derecha de 5,1, perdiéndose en la aproximación la mitad de la probabilidad de que N_{11} tome el valor 2 e incorporándose un 40% de la probabilidad de que N_{11} tome el valor 5. Cuando se incluye la corrección de continuidad de Yates, se incorpora de nuevo la mitad de la probabilidad de que N_{11} tome el valor 2, pero se añade, adicionalmente, un 50% de la probabilidad de que N_{11} tome el valor 5.

Análogamente se establecen las demás situaciones.

Caso A.2)

$$n_{ij} \le \hat{E}_{ij}$$

y además

$$\theta + \frac{1}{2} \le 2\hat{E}_{ij} - n_{ij} < \theta + 1$$

Aproximando la probabilidad exacta mediante el estadístico χ^2_{ajd} se produce una doble infraestimación(8):

1) Infraestimación:

$$\frac{1}{2} P\left[N_{ij} = n_{ij}\right]$$

2) Infraestimación:

$$\left[\left(2\hat{E}_{ij} - n_{ij}\right) - \left(\theta + \frac{1}{2}\right)\right] P\left[N_{ij} = \theta + 1\right]$$

y con la corrección de continuidad de Yates

1) Se corrige la infraestimación evaluada en:

$$\frac{1}{2} P\left[N_{ij} = n_{ij}\right]$$

2) Se incurre en sobreestimación:

$$\left[(\theta + 1) - \left(2\hat{E}_{ij} - n_{ij}\right)\right] \quad P\left[N_{ij} = \theta\right]$$

Caso A.3)

$$n_{ij} \le \hat{E}_{ij}$$

(8) La segunda es nula cuando $2\hat{E}_{ij} - n_{ij} = \theta + \frac{1}{2}$

y además

$$2\hat{E}_{ij} - n_{ij} = \theta$$

La aproximación de la probabilidad exacta utilizando el estadistico nos lleva, al igual que en el caso A.2, a una doble infraestimación:

1) Infraestimación:

$$\frac{1}{2} P\left[N_{ij} = n_{ij}\right]$$

2) Infraestimación:

$$\frac{1}{2} P\left[N_{ij} = 2\hat{E}_{ij} - n_{ij}\right] = \frac{1}{2} P\left[N_{ij} = \theta\right]$$

que se corrige incluyendo la corrección de continuidad de Yates.

Caso B.1)

$$n_{ij} \geq \hat{E}_{ij}$$

y además

$$\theta < 2\hat{E}_{ij} - n_{ij} \leq \theta + \frac{1}{2}$$

Aproximando la probabilidad exacta mediante el estadístico χ^2_{ajd}, o lo que es igual, a través de

$$P\left[N_{ij} \leq 2\hat{E}_{ij} - n_{ij}\right] + P\left[N_{ij} \geq n_{ij}\right] \text{ ya que } n_{ij} \geq \hat{E}_{ij}$$

con

$$N_{ij} \underset{aprox}{\longrightarrow} N\left(\frac{n_{i.}n_{.j}}{n}; \sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}\right)$$

se produce una doble infraestimación(9):

1) Infraestimación:

$$\frac{1}{2}P\left[N_{ij}=n_{ij}\right]$$

2) Infraestimación:

$$\left[\left(\theta+\frac{1}{2}\right)-\left(2\hat{E}_{ij}-n_{ij}\right)\right]P\left[N_{ij}=\theta\right]$$

y si se incluye la corrección de continuidad de Yates, es decir, aproximando la probabilidad exacta mediante

$$P\left[N_{ij}\leq 2\hat{E}_{ij}-n_{ij}+\frac{1}{2}\right]+P\left[N_{ij}\geq n_{ij}-\frac{1}{2}\right] \quad \text{ya que } n_{ij}\geq\hat{E}_{ij}$$

se tiene que

1) Se corrige la infraestimación evaluada en

$$\frac{1}{2}P\left[N_{ij}=n_{ij}\right]$$

2) Se incurre en sobreestimación

$$\left[\left(2\hat{E}_{ij}-n_{ij}\right)-\theta\right]\ P\left[N_{ij}=\theta+1\right]$$

Caso B.2)

$$n_{ij}\geq\hat{E}_{ij}$$

y además

$$\theta+\frac{1}{2}\leq 2\hat{E}_{ij}-n_{ij}<\theta+1$$

(9) La segunda nula cuando $2\hat{E}_{ij}-n_{ij}=\theta+\frac{1}{2}$

La utilización del estadístico χ^2_{ajd} para aproximar la probabilidad exacta provoca,

1) Infraestimación

$$\frac{1}{2} P\left[N_{ij} = n_{ij}\right]$$

2) Sobreestimación

$$\left[\left(2\hat{E}_{ij} - n_{ij}\right) - \left(\theta + \frac{1}{2}\right)\right] P\left[N_{ij} = \theta + 1\right]$$

si bien, como viene siendo habitual, la segunda distorsión, en este caso la sobreestimación, se anula en caso de que $2\hat{E}_{ij} - n_{ij} = \theta + \frac{1}{2}$.

Incluyendo la corrección de continuidad de Yates se tiene que

1) Se corrige la infraestimación

2) Se incrementa la sobreestimación hasta

$$\left[\left(2\hat{E}_{ij} - n_{ij}\right) - \theta\right] P\left[N_{ij} = \theta + 1\right]$$

Caso B.3)

$$n_{ij} \geq \hat{E}_{ij}$$

y además

$$2\hat{E}_{ij} - n_{ij} = \theta$$

La aproximación de la probabilidad exacta utilizando el estadístico χ^2_{ajd} nos lleva, al igual que en el caso B.1, a una doble infraestimación:

1) Infraestimación

$$\frac{1}{2} P\left[N_{ij} = n_{ij}\right]$$

2) Infraestimación

$$\frac{1}{2} P \left[N_{ij} = 2\hat{E}_{ij} - n_{ij} \right] = \frac{1}{2} P \left[N_{ij} = \theta \right]$$

que se corrige incluyendo la corrección de continuidad de Yates.

Como puede apreciarse, salvo en los casos A.3 y B.3, en los que las estimaciones de las frecuencias esperadas bajo la hipótesis de independencia son múltiplos de 0,5,(10) la utilización de la corrección de continuidad de Yates no conduce a buenas aproximaciones de la probabilidad exacta.

3. DISEÑO CON LOS TOTALES MARGINALES FIJOS: UNA ALTERNATIVA A LA CORRECCIÓN DE CONTINUIDAD DE YATES EN LA APROXIMACIÓN DE LA PROBABILIDAD EXACTA EN TABLAS (2X2)

Una propuesta para corregir las sobreestimaciones o infraestimaciones de la probabilidad exacta que tienen lugar cuando se aplica la corrección de continuidad de Yates consiste en aproximar la misma de forma asimétrica.

Se propone, en concreto, el siguiente procedimiento:

A) Si $n_{ij} \leq \hat{E}_{ij}$

$$P_{exacta} \cong P\left(N_{ij} \leq n_{ij} + \frac{1}{2} \right) + P\left(N_{ij} \geq 2\hat{E}_{ij} - n_{ij} + \Delta \right)$$

B) Si $n_{ij} \geq \hat{E}_{ij}$

$$P_{exacta} \cong P\left(N_{ij} \leq 2\hat{E}_{ij} - n_{ij} + \Delta \right) + P\left(N_{ij} \geq n_{ij} - \frac{1}{2} \right)$$

siendo

$$N_{ij} \rightarrow_{aprox} N\left(\frac{n_{i.} n_{.j}}{n} ; \sqrt{\frac{n_{1.} n_{2.} n_{.1} n_{.2}}{n^2 (n-1)}} \right)$$

(10) Sólo en este caso se verifica $2\hat{E}_{ij} - n_{ij} = \theta$

O bien

A) Si $n_{ij} \leq \hat{E}_{ij}$

$$P_{exacta} \cong P\left(\xi^* \leq \frac{\hat{D}_{ij} + \frac{1}{2}}{\sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}}\right) + P\left(\xi^* \geq \frac{\Delta - \hat{D}_{ij}}{\sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}}\right)$$

B) Si $n_{ij} \geq \hat{E}_{ij}$

$$P_{exacta} \cong \left(\xi^* \leq \frac{\Delta - \hat{D}_{ij}}{\sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}}\right) + P\left(\xi^* \geq \frac{\hat{D}_{ij} - \frac{1}{2}}{\sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}}\right)$$

siendo $\hat{D}_{ij} = n_{ij} - \hat{E}_{ij}$, ξ^* una normal estándar y Δ una cantidad que se calcula como

$$\Delta = \theta + \frac{1}{2} - 2\hat{E}_{ij} + n_{ij}$$

En lo que a la anterior forma de proceder se refiere, es necesario establecer las siguientes salvedades:

En el caso A:

$$\text{Si } 2\hat{E}_{ij} - n_{ij} > \min(n_{1.}n_{2.}n_{.1}n_{.2}) \text{ entonces } P\left(\xi^* \geq \frac{\Delta - \hat{D}_{ij}}{\sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}}\right) = 0$$

En el caso B:

$$\text{Si } 2\hat{E}_{ij} - n_{ij} < 0 \text{ entonces } P\left(\xi^* \leq \frac{\Delta - \hat{D}_{ij}}{\sqrt{\frac{n_{1.}n_{2.}n_{.1}n_{.2}}{n^2(n-1)}}}\right) = 0$$

Existen otras correcciones de continuidad que pretenden superar las deficiencias de la de Yates en las tablas bifactoriales con los totales marginales fijos. Entre ellas cabe destacar dos formuladas por Cochran (correcciones I y II) y otra elaborada por de Mantel(11).

A modo de ilustración, aunque puede comprobarse para cualquier tabla, supóngase de nuevo la tabla (2x2) con totales marginales fijos anteriormente considerada, donde

$$\hat{E}_{11} = 3{,}5483871 \text{ y } n_{11} \leq \hat{E}_{11}$$

De acuerdo con lo anteriormente expuesto, las tablas que se alejan de la hipótesis de independencia tanto o más que la observada son, además de ésta, aquéllas que verifican

$$N_{11} \leq n_{11} \quad N_{11} \geq 2\hat{E}_{11} - n_{11}$$

En el caso que nos ocupa, aquéllas con $N_{ij} \leq 2$ ó $N_{ij} \geq 7{,}0967742$, es decir las tablas T_0, T_1, T_2, T_6, T_7, T_8, T_9 y T_{10}, siendo T_i la tabla con N_{11}=i.

La probabilidad exacta (suma de las probabilidades de las tablas especificadas) se cifra en 0,2617(12). La aproximaciones de la probabilidad exacta computadas han sido las siguientes:

a) Mediante el estadístico χ^2_{ajd}: 0,2224 (Infraestimación: 0,0393).

b) Con la corrección de continuidad de Yates: 0,4040 (Sobreestimación: 0,1423).

c) Con la corrección I de Cochran: 0,2360 (Infraestimación: 0,0257).

d) Con la corrección II de Cochran: 0,2360 (Infraestimación: 0,0257).

e) Con la corrección de Mantel: 0,2302 (Infraestimación:0,0315).

f) Con la alternativa expuesta: 0,2638 (Sobreestimación: 0,0021).

---

(11) Su desarrollo puede verse en RUIZ-MAYA, L; MARTIN PLIEGO, F.J.; MONTERO LORENZO, J.M.; URIZ TOME, P. (1995) "Analisis Estadistico de Encuestas: Datos Cualitativos", A.C., Madrid.

(12) La probabilidad de la tabla T_i, con los totales marginales fijos, y bajo el supuesto de independencia, se calcula como

$$P_i = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

pudiéndose apreciar en el caso expuesto -aunque sería generalizable a cualquier otra estructura de frecuencias- la bondad de la alternativa propuesta en la aproximación a la probabilidad exacta.

REFERENCIAS

AGRESTI, A. (1990). «Categorical Data Analysis». John Wiley. New York.

COCHRAN W. G. (1942). «The 2x2 Correction for Continuity». *Iowa State College Journal of Science*, 16, 421-436.

COX, D.R. (1970). «The continuity correction». *Biometrika*, 57, 217-219.

FISHER, R.A. (1935). «The design of experiments». 8ª ed. 1966. Oliver and Boyd. Edinburgh.

HABER, M. (1980). «A Comparison of Some Continuity Corrections for the Chi-Squared Test on 2x2 Tables». *Journal of the American Statistical Association*, Vol 75, 371, 510-515.

HABERMAN, S. (1988). «A Warning on the Use of Chi-Squared Statistics With Frecuency Tables With Small Expected Cell Counts». *Journal of the American Statistical Association*, Vol 83, 402, 555-560.

MANTEL, N. (1974). «Some Reasons for Not Using The Yates Continuity Correction on 2x2 Contingency Tables - Comment and a Suggestion». *Journal of the American Statistical Association*, 69, 378-380.

MANTEL, N. (1976). «The Continuity Correction». *The American Statistician*, 30, 103-104.

MANTEL, N. - Greenhouse, S. (1968). «What is the Continuity Correction?». *The American Statistician*, 22 nº5, 27-30.

PLACKETT, R.L. (1964). «The continuity correction on 2x2 tables». *Biometrika*, 64, 37-42.

RUIZ-MAYA, L.; MARTÍN PLIEGO, J.; MONTERO LORENZO, J.M.; URIZ, P. (1990). «Metodología Estadística para el análisis de datos cualitativos». C.I.S. Madrid.

RUIZ-MAYA, L. - MARTÍN PLIEGO, J. - MONTERO LORENZO, J.M. - URIZ, P. (1995). « Análisis Estadístico de Encuestas: Datos Cualitativos». A.C, Madrid.

UPTON, G.J.G. (1992). «Fisher's Exact Test». *Journal of the Royal Statistic Society*, Ser. A, Part 3, 395-402.

YATES, F. (1934). «Contingency Tables Involving Small Numbers and the X^2 Test». *Journal of the Royal Statistical Society*, Ser. B, Supp. Vol.1, 217-235.

AN ASYMMETRIC ONTINUITY CORRECTION FOR 2X2 CONTINGENCY TABLES WITH FIXED MARGINAL TOTALS

SUMMARY

In this study it is proposed an asymmetric continuity correction applicable to (2X2) contingency tables with fixed marginal totals, which brodaly improves the well-known corrections from Yates, Mantel and Cochran. At that point we have formulated the underestimations and overestimations of the so-called "exact probability" obtained by applying the Yates continuity correction. In spite of the fact that this correction is the most commonly applied, it is simmetric and therefore it does not perfectly approximate the exact probability. What is more, in some cases it worsens the approximation provided by the Chi-square statistic without continuity correction. On these grounds, the inclusion of the asymmetry condition in the fieds of continuity corrections leads to magnificient approximations.

Key words: (2X2) contingency table, independence test, continuity correction, Fisher's exact probability, adjusted chi-square statistic, asymmetric correction.

AMS Classification: 62H17