

# **PROBLEMÁTICA Y VALORACIÓN DE LOS CONTRASTES SUSTITUTIVOS DEL TEST EXACTO DE FISHER: UN PROCEDIMIENTO ALTERNATIVO.**

Jose M<sup>a</sup> Montero Lorenzo

Universidad de Castilla-La Mancha

## **ABSTRACT**

El contraste de independencia en tablas bifactoriales de dos niveles por factor puede plantearse bajo distintos procedimientos de muestreo, pero suele llevarse a cabo mediante el denominado "test exacto de Fisher". No obstante, en la práctica, en muchas ocasiones, dicho contraste se realiza mediante el estadístico Ji-cuadrado incorporando la corrección de continuidad de Yates. Este trabajo plantea algunas expresiones operativas que facilitan la realización del test exacto de Fisher, por lo que no se justifica, salvo en casos extremos, la utilización de procedimientos aproximados. Además, pone de manifiesto la casuística de errores cometidos al aproximar la probabilidad exacta a través del estadístico Ji-cuadrado ajustado, incluyendo o no la corrección de continuidad de Yates que es la que está programada en la mayoría de los paquetes informáticos, lo que reforzará aún más la tesis relativa a la no realización de procedimientos aproximados. Por último, en caso de querer llevarlos a cabo, se presenta un procedimiento sencillo que proporciona aproximaciones a la probabilidad exacta notablemente mejores a las del contraste Ji-cuadrado con la corrección de continuidad de Yates.

## **PROBLEMÁTICA Y VALORACIÓN DE LOS CONTRASTES SUSTITUTIVOS DEL TEST EXACTO DE FISHER: UN PROCEDIMIENTO ALTERNATIVO.**

### **1. Introducción.**

La forma clásica de proceder al contraste de independencia poblacional en una tabla de contingencia bifactorial -en general, si bien en adelante consideraremos la situación de dos niveles por factor- no es otra que el cálculo de la probabilidad (bajo dicha hipótesis) de obtener 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 si ambos factores presentan dos niveles. *El problema que surge es que la probabilidad exacta, calculada mediante una distribución discreta de probabilidad (hipergeométrica, binomial bivalente, multinomial, Poisson y binomial negativa bivalente, dependiendo del procedimiento de muestreo), se aproxima a través de una distribución continua de probabilidad (la distribución Ji-cuadrado).* (NOTA 1)

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 solventar dicho error "corrigiendo de la continuización realizada" o, simplemente, corrigiendo de continuidad. No obstante, 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.

El contraste de independencia en una tabla bifactorial con dos niveles por factor se suele llevar a cabo, sea cual sea el procedimiento de muestreo, con el condicionante de que los totales marginales de ambos factores sean fijos, y el procedimiento más popular entre los investigadores de las Ciencias Sociales en general, y de la Economía en particular, probablemente mediatizados por los paquetes informáticos de contenido estadístico al uso y la literatura tradicional sobre esta materia, consiste en llevar a cabo un test exacto de Fisher 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, y considerar que la aproximación Ji-cuadrado que incluye la corrección de continuidad de Yates es aceptable en las demás condiciones. (NOTA 2)

El objetivo de este trabajo consistirá en demostrar que la supuesta laboriosidad de la realización del test exacto de Fisher no es tal puesto que el cálculo de las probabilidades de las tablas cuyo alejamiento de la hipótesis de independencia es igual o superior al de la tabla observada puede llevarse a cabo de una forma muy simple, por lo que no se justifica, salvo en casos extremos, la realización de procedimientos aproximados. Se pondrá de manifiesto, además, la casuística de errores cometidos al aproximar la probabilidad exacta a través del estadístico Ji-cuadrado ajustado, incluyendo o no la corrección de continuidad de Yates que es la que está programada en la mayoría de los paquetes informáticos, lo que reforzará aún más la tesis relativa a la no realización de procedimientos aproximados. Por último, en caso de querer llevarlos a cabo, se presenta un procedimiento muy sencillo que proporciona aproximaciones a la probabilidad exacta notablemente mejores a las del contraste Ji-cuadrado con la corrección de continuidad de Yates.

**2. Diseño con los totales marginales fijos: Expresiones operativas para el test exacto de Fisher y deficiencias de los tests Ji-cuadrado y Ji-cuadrado con corrección de continuidad de Yates en la aproximación de la probabilidad exacta.**

Como ya se avanzó, el diseño o procedimiento de muestreo en el que nos vamos a centrar, para la contrastación de la hipótesis de independencia en una tabla (2x2) 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 recibe el nombre de test exacto de Fisher.

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

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

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,  $H(n; n_j; n_i)$ , (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

$$P \left[ \chi^2_1 \geq \frac{(n-1) (n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.} n_{2.} n_{.1} n_{.2}} \right]$$

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  $X^2_{ajd}$  (NOTA 3)

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

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

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

O bien a través de

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

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

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

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

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

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

$$\begin{aligned}
& P\left[N_{ij} \leq \hat{E}_{ij} - |n_{ij} - \hat{E}_{ij}|\right] + P\left[N_{ij} \geq \hat{E}_{ij} + |n_{ij} - \hat{E}_{ij}|\right] = \\
& P\left[N_{ij} \leq \hat{E}_{ij} + (n_{ij} - \hat{E}_{ij})\right] + P\left[N_{ij} \geq \hat{E}_{ij} - (n_{ij} - \hat{E}_{ij})\right] = \\
& P\left[N_{ij} \leq n_{ij}\right] + P\left[N_{ij} \geq 2\hat{E}_{ij} - n_{ij}\right]
\end{aligned}$$

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

$$\begin{aligned}
& P\left[N_{ij} \leq \hat{E}_{ij} - |n_{ij} - \hat{E}_{ij}|\right] + P\left[N_{ij} \geq \hat{E}_{ij} + |n_{ij} - \hat{E}_{ij}|\right] = \\
& P\left[N_{ij} \leq \hat{E}_{ij} - (n_{ij} - \hat{E}_{ij})\right] + P\left[N_{ij} \geq \hat{E}_{ij} + (n_{ij} - \hat{E}_{ij})\right] = \\
& P\left[N_{ij} \leq 2\hat{E}_{ij} - n_{ij}\right] + P\left[N_{ij} \geq n_{ij}\right]
\end{aligned}$$

La realización del test exácto de Fisher no plantea dificultad alguna pues las expresiones expuestas nos especifican claramente las tablas que se deben calificar de "tan raras o más" que la observada. Una vez seleccionadas estas tablas, el cálculo de su probabilidad de ocurrencia no presenta dificultad analítica alguna ya que se puede calcular la probabilidad de una de ellas, la i-ésima,

$$P(i) = \frac{n_{1.}! n_{2.}! n_{.1}! n_{.2}!}{n! n_{11}! n_{12}! n_{21}! n_{22}!}$$

y la de las demás se obtienen a partir de ella mediante las expresiones

$$\begin{aligned}
P(i+k) &= \frac{\prod_{l=1}^{k-1} (n_{12} - l)(n_{21} - l)}{\prod_{l=1}^k (n_{11} + l)(n_{22} + l)} P(i) \\
P(i-k) &= \frac{\prod_{l=1}^{k-1} (n_{11} - l)(n_{22} - l)}{\prod_{l=1}^k (n_{12} + l)(n_{21} + l)} P(i)
\end{aligned}$$

siendo  $n_{11}$ ,  $n_{12}$ ,  $n_{21}$  y  $n_{22}$  las frecuencias observadas en la i-ésima tabla, y siendo las tablas i, i+k e i-k,

aquéllas en la que  $N_{ij}=i$ ,  $N_{ij}=i+k$  y  $N_{ij}=i-k$  respectivamente, cualquiera que sea  $ij$ .

En consecuencia, ni la detección de las tablas que conforman la región crítica, ni el cálculo de sus probabilidades conlleva complicación o laboriosidad alguna por lo que no existen razones suficientes, salvo casos extremos, para los procedimientos aproximados.

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

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

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

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

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

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

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

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

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

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

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

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

$$\text{Caso A.1) } n_{ij} \leq \hat{E}_{ij} \text{ y adem s } \mathbf{q} < 2\hat{E}_{ij} - n_{ij} < \mathbf{q} + \frac{1}{2}$$

Si se aproxima la probabilidad exacta mediante el estadístico  $X^2_{\text{ajd}}$ , o lo que es igual, mediante

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

o a través de

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

con

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

tienen lugar dos distorsiones:

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

$$2) \text{ Sobreestimación: } \left[ \left( \mathbf{q} + \frac{1}{2} \right) - (2\hat{E}_{ij} - n_{ij}) \right] P \left[ N_{ij} = \mathbf{q} \right]$$

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

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

ocurre que



1) Se corrige la infraestimación

2) Aumenta la sobreestimación hasta:  $\left[ (q+1) - (2\hat{E}_{ij} - n_{ij}) \right] P[N_{ij} = q]$

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

TABLA 1

donde

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

#### GRAFICO

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 5,09; es decir, el área rectangular correspondiente a los valores: 0, 1, 2, 6, 7, 8, 9, 10. La aproximación normal (o mediante el estadístico  $X^2_{ajd}$ ) 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

$$\text{Caso A.2)} \quad n_{ij} \leq \hat{E}_{ij} \quad \text{y adem s} \quad q + \frac{1}{2} < 2\hat{E}_{ij} - n_{ij} < q + 1$$

Aproximando la probabilidad exacta mediante el estadístico  $X^2_{ajd}$  se produce una doble infraestimación:

$$1) \text{ Infraestimación: } \frac{1}{2} P [N_{ij} = n_{ij}]$$

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

y con la corrección de continuidad de Yates

$$\text{Caso A.3) } n_{ij} \leq \hat{E}_{ij} \text{ y adem s } 2 \hat{E}_{ij} - n_{ij} = q \text{ Se corrige la infraestimación evaluada en } \frac{1}{2} P [N_{ij} = n_{ij}]$$

$$2) \text{ Se incurre en sobreestimación: } \left[ (q + 1) - (2 \hat{E}_{ij} - n_{ij}) \right] P [N_{ij} = q]$$

La aproximación de la probabilidad exacta utilizando el estadístico  $X^2_{\text{ajd}}$  nos lleva, al igual que en el caso A.2, a una doble infraestimación:

$$1) \text{ Infraestimación: } \frac{1}{2} P [N_{ij} = n_{ij}]$$

$$2) \text{ Infraestimación: } \frac{1}{2} P [N_{ij} = 2 \hat{E}_{ij} - n_{ij}] = \frac{1}{2} P [N_{ij} = q]$$

que se corrige incluyendo la corrección de continuidad de Yates.

$$\text{Caso B.1) } n_{ij} \geq \hat{E}_{ij} \text{ y adem s } q < 2 \hat{E}_{ij} - n_{ij} < q + \frac{1}{2}$$

Aproximando la probabilidad exacta mediante el estadístico  $X^2_{\text{ajd}}$  se produce una doble infraestimación:

$$1) \text{ Infraestimación: } \frac{1}{2} P [N_{ij} = n_{ij}]$$

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

y se incluye la corrección de continuidad de Yates se tiene que

Caso B.2)  $n_{ij} \geq \hat{E}_{ij}$  Se corrige la infraestimación evaluada en  $\frac{1}{2} P [N_{ij} = n_{ij}]$

2) Se incurre en sobreestimación:  $\left[ \left( 2 \hat{E}_{ij} - n_{ij} \right) - q \right] P [N_{ij} = q + 1]$

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

1) Infraestimación:  $\frac{1}{2} P [N_{ij} = n_{ij}]$

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

mientras que incluyendo la corrección de continuidad de Yates se tiene

Caso B.3)  $n_{ij} \geq \hat{E}_{ij}$  y además  $2\hat{E}_{ij}$  corrige la infraestimación

2) Se incrementa la sobreestimación hasta:  $\left[ \left( 2 \hat{E}_{ij} - n_{ij} \right) - q \right] P [N_{ij} = q + 1]$

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

1) Infraestimación:  $\frac{1}{2} P [N_{ij} = n_{ij}]$

2) Infraestimación:  $\frac{1}{2} P [N_{ij} = 2 \hat{E}_{ij} - n_{ij}] = \frac{1}{2} P [N_{ij} = q]$

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 sean múltiplos de 0,5, (NOTA 4) 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 el siguiente procedimiento:

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

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

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

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

siendo

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

O bien

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

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

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

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

siendo  $\xi^*$  una normal estandar y  $\Delta$  una cantidad que se calcula como

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

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

En el caso A:

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

En el caso B:

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

A modo de ilustración supóngase, de nuevo, la tabla (2x2) con totales marginales fijos anteriormente considerada

TABLA 1

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

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

En el caso que nos ocupa, aquéllas con  $N_{ij} \leq 2$  ó  $N_{ij} \geq 7,0967742$ , es decir las tablas T<sub>0</sub>, T<sub>1</sub>, T<sub>2</sub>, T<sub>6</sub>, T<sub>7</sub>, T<sub>8</sub>, T<sub>9</sub> y T<sub>10</sub>, siendo T<sub>i</sub> la tabla con  $N_{11}=i$ .

La probabilidad exacta (suma de las probabilidades de las tablas especificadas) se cifra en 0,2617.

La aproximaciones de la probabilidad exacta computadas han sido las siguientes:

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

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

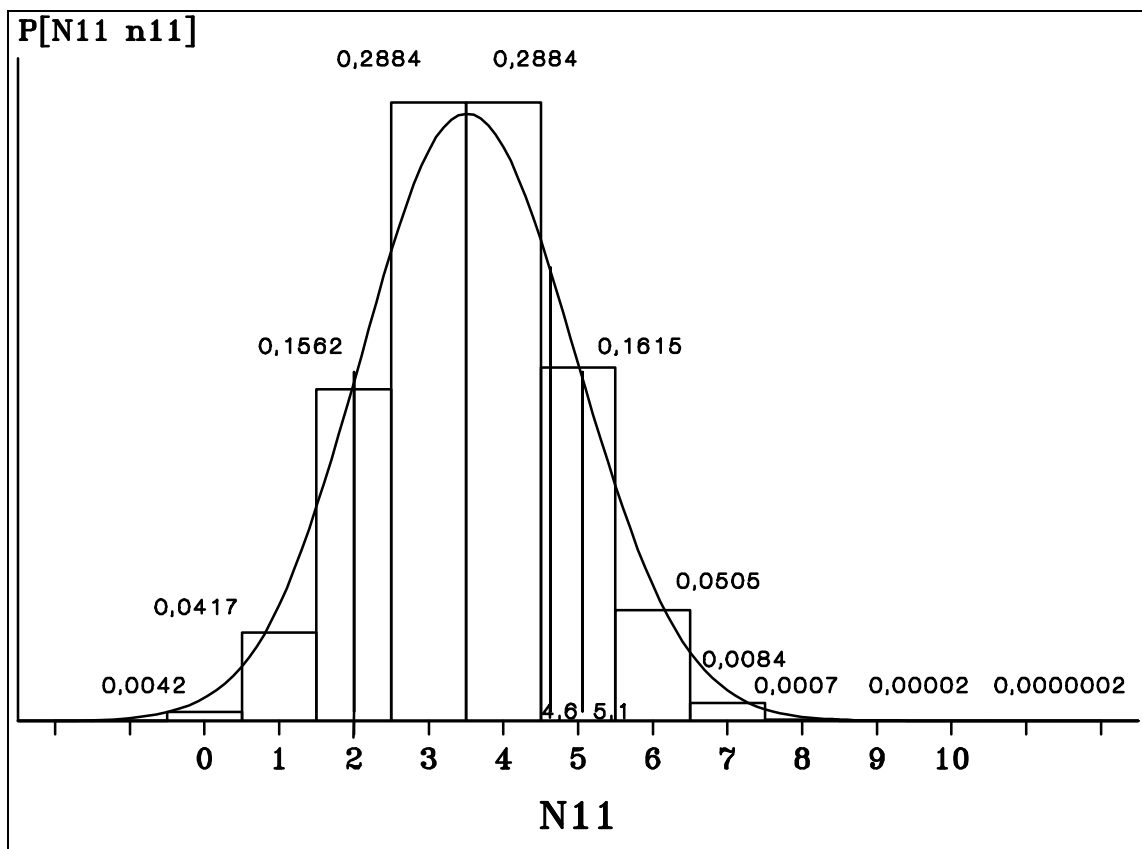
c) Con la alternativa expuesta: 0,2638 (Sobreestimación: 0,0021).

pudiéndose apreciar, en el caso expuesto, la bondad de la alternativa propuesta en la aproximación a la probabilidad exacta.

TABLA 1

		FACTOR A		
		NIVEL 1	NIVEL 2	
FACTOR B	NIVEL 1	2	8	10
	NIVEL 2	9	12	21
		11	20	31

GRAFICO



## NOTAS A PIE DE PAGINA

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

(2) Algunas de las versiones más recientes de los paquetes informáticos de contenido estadístico - por ejemplo la versión 6.1.2 del popular SPSS- ya superan esta limitación.

(3) En realidad se opera en el estadístico  $X^2$ , pero se expone sobre el estadístico  $X^2_{ajd}$  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  $X^2_{ajd}$  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.

(4) Sólo en este caso se verifica

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



## BIBLIOGRAFIA.

**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 Frequency 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, J.; MONTERO, J.M.; URIZ, P.; LÓPEZ, J.** (1990). *Metodología Estadística para el análisis de datos cualitativos*. C.I.S. Madrid.

**RUIZ-MAYA, L.; MARTÍN, J.; MONTERO, 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.

Jose M<sup>a</sup> Montero Lorenzo

Universidad de Castilla La Mancha



