LES CORRECCIONS DE CONTINUÏTAT EN DISTRIBUCIONS BINOMIAL I POISSON, I LA CORRECCIÓ DE YATES EN EL TEST KHI-QUADRAT EN TAULES DE CONTINGÈNCIA 2×2

M.S. NIKULIN*
C.M. CUADRAS†

Aquest article desenvolupa i comenta diverses correccions de continuïtat a les aproximacions normal i khi-quadrat d'algunes distribucions discretes.

INTRODUCCIÓ

La correcció de continuïtat en les aproximacions a les distribucions binomial, Poisson i khi-quadrat relacionada amb les taules de contingència 2×2, és un tema sovint no ben conegut per professors, usuaris i estudiants d'estadística. Per exemple, la correcció de continuïtat de Yates tendeix a donar un estadístic khi-quadrat baix i aleshores el test és conservador (tendència a acceptar la hipòtesi nul·la tot i no sent certa).

En aquest article es proporciona una justificació de la correcció de continuïtat en les distribucions esmentades, que és il·lustrada amb exemples.

1. CORRECCIÓ DE CONTINUÏTAT PER A LA DISTRIBUCIÓ BINOMIAL

Sigui ν una variable aleatòria i considerem la hipòtesi H_0 afirmant que ν segueix una distribució binomial $B(n,p)$ amb paràmetres n i p, $(0 < p < 1)$. Aleshores sota H_0 tenim

*M.S. Nikulin. *Mathematiques Stochastiques*, Universite Bordeaux 2, France. Steklov Mathematical Institute, St.Petersburg, Russia.

† C.M. Cuadras. Departament d'Estadística. Universitat de Barcelona. Diagonal, 645. 08028, Barcelona.

$$E\nu = np \quad \text{i} \quad \mathbf{Var}\ \nu = np(1-p).$$

Com és ben sabut, pel teorema de Laplace-de Moivre tenim

$$\lim_{n\to\infty} \boldsymbol{P}\left\{ \frac{\nu - np}{\sqrt{np(1-p)}} \leq x | H_0 \right\} = \Phi(x), \tag{1}$$

on Φ és la distribució normal N(0,1). És a dir, sota H_0 la variable aleatòria ν és asimptòticament normal amb paràmetres np i $np(1-p)$. A més, el teorema Laplace-de Moivre ens diu que per a qualsevol p, $0 < p < 1$, essent $n \to \infty$

$$\boldsymbol{P}\{\nu \leq x | H_0\} = \Phi\left(\frac{x - np + c}{\sqrt{np(1-p)}} \right) + O\left(\frac{1}{\sqrt{n}} \right), \tag{2}$$

on c és un número real que sovint és 0. En la pràctica estadística hom agafa c igual a 0.5 i aleshores es parla de la *correcció de continuïtat*. És fàcil justificar l'elecció de $c = 0.5$ tenint en compte que la variable aleatòria $n - \nu$ segueix la distribució binomial $B(n, 1-p)$, i atès que per a qualsevol $x = 0, 1, ..., n$

$$\boldsymbol{P}\{\nu \leq x\} + \boldsymbol{P}\{\nu \geq x+1\} = 1$$

tenim

$$\boldsymbol{P}\{\nu \leq x\} + \boldsymbol{P}\{n - \nu \leq n - x - 1\} = 1.$$

Aleshores de (2) obtenim que per $n \to \infty$

$$\Phi\left(\frac{x - np + c}{\sqrt{np(1-p)}} \right) + \Phi\left(-\frac{x - np + (1-c)}{\sqrt{np(1-p)}} \right) = 1 + O\left(\frac{1}{\sqrt{n}} \right). \tag{3}$$

Però sabem que

$$\Phi(z) + \Phi(-z) \equiv 1, \qquad z \in \mathbb{R}; \tag{4}$$

llavors, si $c = 0.5$, la suma a (3) val exactament 1. Així, amb la correcció de continuïtat $c = 0.5$, de (2) tenim que per a $n \to \infty$

$$\boldsymbol{P}\{\nu \leq m | H_0\} = \Phi\left(\frac{m - np + 0.5}{\sqrt{np(1-p)}} \right) + O\left(\frac{1}{\sqrt{n}} \right), \tag{5}$$

$$\begin{aligned} \boldsymbol{P}\{\nu \geq M | H_0\} &= 1 - \boldsymbol{P}\{\nu \leq M-1 | H_0\} = \\ &\Phi\left(-\frac{M - 0.5 - np}{\sqrt{np(1-p)}} \right) + O\left(\frac{1}{\sqrt{n}} \right). \end{aligned} \tag{6}$$

on $0 < m, M \leq n$. De (5) i (6) se segueix que si volem tenir un criteri estadístic per contrastar el test d'hipòtesi H_0 amb nivell de significació $\approx \alpha$ $(0 < \alpha < 0.5)$, aleshores cal rebutjar H_0 si

$$\Phi\left(\frac{\nu + 0.5 - np}{\sqrt{np(1-p)}}\right) \leq \frac{\alpha}{2} \quad \text{ó} \quad \Phi\left(-\frac{\nu - 0.5 - np}{\sqrt{np(1-p)}}\right) \leq \frac{\alpha}{2}, \tag{7}$$

on ν és l'observació de la variable. Això significa que rebutgem H_0 si es presenta un dels dos esdeveniments:

$$\frac{\nu - np}{\sqrt{np(1-p)}} \leq \Psi\left(\frac{\alpha}{2}\right) - \frac{1}{2\sqrt{np(1-p)}}, \tag{8}$$

$$\frac{\nu - np}{\sqrt{np(1-p)}} \geq -\Psi\left(\frac{\alpha}{2}\right) + \frac{1}{2\sqrt{np(1-p)}}, \tag{9}$$

essent $\Psi(x)$ la funció inversa de $\Phi(x)$. De (8),(9) tenim que H_0 ha d'esser rebutjada si

$$X_n^2 = \frac{(\nu - np)^2}{np(1-p)} \geq \left[\Psi\left(1 - \frac{\alpha}{2}\right) + \frac{1}{2\sqrt{np(1-p)}}\right]^2, \tag{10}$$

perquè $\Psi(z) + \Psi(1-z) \equiv 0, z \in (0,1)$, on X_n^2 segueix asimptòticament la distribució khi-quadrat de Pearson amb un grau de llibertat. Així si escollim el valor crític

$$c_\alpha = \left[\Psi\left(1 - \frac{\alpha}{2}\right) + \frac{1}{2\sqrt{np(1-p)}}\right]^2, \tag{11}$$

obtenim un criteri khi-quadrat per contrastar H_0 amb un nivell aproximat de significació α. El segon terme que hi ha dins (11) és aleshores l'anomenada *correcció de Yates*, que és una conseqüència natural d'introduir la correcció de continuïtat a (5). Cal remarcar aquí que sovint a la literatura estadística la correcció de continuïtat de Yates és utilitzada incorrectament.

Exemple 1

Suposem que tenim un generador de nombres aleatoris $x_1, x_2, \ldots, x_n, \ldots$ que són considerats (hipòtesi $\boldsymbol{H}_0$) com a realitzacions de variables aleatòries independents $X_1, X_2, \ldots, X_n, \ldots$, amb distribució uniforme discreta sobre el conjunt $S = \{0, 1, 2, \ldots, 9\}$, és a dir,

$$\boldsymbol{P}\{X_1 = i | \boldsymbol{H}_0\} = 0.1, \qquad i \in S. \tag{12}$$

Ara suposem que tenim una mostra $\mathbb{X} = (X_1, X_2, \ldots, X_n)^T$ de grandària n=10000, produïda per l'esmentat generador aleatori, i volem contrastar $\boldsymbol{H}_0$ si la mostra $\mathbb{X}$ és acceptable, enfront de l'alternativa que no és acceptable com a provinent d'una distribució uniforme discreta (12). Si a la mostra els nombres x_i no excedint 4 foren observats només 4901 vegades, a quin nivell de significació hem de rebutjar $\boldsymbol{H}_0$?

Solució

Sigui $\mu_n = \text{freq}\,\{X_i \leq 4\}$. De les observacions tenim que

$$\frac{\mu_n}{n} = \frac{4901}{10000} = 0.4901,$$

que és bastant pròxim a 0.5. Consegüentment, si la nostra suposició (hipòtesi $\boldsymbol{H}_0$) és correcte aleshores μ_n té una distribució binomial B(n,p) amb paràmetres $n = 10000$, $p = 0.5$, i, sota $\boldsymbol{H}_0$,

$$\boldsymbol{E}\mu_n = np = 5000 \quad \text{i} \quad \textbf{Var}\ \mu_n = np(1-p) = 2500. \tag{13}$$

Així, per a qualsevol $x = 1, 2, \ldots$, pel teorema de Laplace-de Moivre tenim que (amb la correcció de continuïtat 0.5)

$$\begin{aligned} \boldsymbol{P}\{|\mu_n - np| \leq x | \boldsymbol{H}_0\} &= \boldsymbol{P}\left\{\frac{n}{2} - x \leq \mu_n \leq \frac{n}{2} + x\right\} \approx \\ &\approx \Phi\left(\frac{0.5n + x + 0.5 - 0.5n}{\sqrt{n \cdot 0.5 \cdot 0.5}}\right) - \Phi\left(\frac{0.5n - x - 0.5 - 0.5n}{\sqrt{n \cdot 0.5 \cdot 0.5}}\right) = \\ &= 2\Phi\left(\frac{2x+1}{\sqrt{n}}\right) - 1. \end{aligned} \tag{14}$$

Sigui α el nivell de significació, $0 < \alpha < 0.5$, del test amb regió crítica

$$\mathcal{K}_x = \left\{\left|\mu_n - \frac{n}{2}\right| > x\right\}. \tag{15}$$

En aquest cas, a un donat valor crític x correspon el nivell de significació

$$\alpha \approx 2 - 2\Phi\left(\frac{2x+1}{\sqrt{n}}\right), \qquad (n = 10000) \tag{16}$$

com es dedueix de (14). En particular, si $x = 98$, aleshores tenim

$$\alpha \approx 2 - 2\Phi\left(\frac{2 \times 98 + 1}{\sqrt{n}}\right) - 1 = 2 - 2\Phi(1.97) = 0.049.$$

Fent ara *inferència estadística*, d'acord amb el test estadístic basat en la regió crítica

$$\mathcal{K}_{98} = \{|\mu_n - 5000| > 98\},$$

la hipòtesi $\boldsymbol{H}_0$ *és rebutjada* amb un nivell de significació $\alpha \approx 0.05$.

Considerem ara una situació diferent. Suposem que $\mu_n = \text{freq}\,\{x_i \leq 4\} = 4999$. Aleshores

$$\frac{\mu_n}{n} = \frac{4999}{10000} = 0.4999$$

que és molt pròxim a 0.5, tan pròxim que aquest generador és "sospitós". Seguint Rao (1989), que ens diu que un ajust molt bo entre observacions i teoria postulada pot ser indici de manipulació (R. A. Fisher va concloure això amb els experiments de G. Mendel), considerem la regió crítica

$$\mathcal{K}'_x = \left\{\left|\mu_n - \frac{n}{2}\right| \leq x\right\}.$$

Ara tenim que a cada x correspon un nivell de significació

$$\alpha \approx 2\Phi\left(\frac{2x+1}{\sqrt{n}}\right) - 1 \qquad (n = 10000).$$

Si $x = 1$ tenim, anàlogament

$$\alpha \approx 2\Phi\left(\frac{3}{\sqrt{n}}\right) - 1 = 2\Phi(0.03) - 1 = 0.024.$$

Fent *inferència estadística* basada en la regió crítica

$$\mathcal{K}'_1 = \{|\mu_n - 5000| \leq 1\}.$$

La hipòtesi H_0 és rebutjada amb un nivell de significació $\alpha \approx 0.025$.

Notem la diferència entre les dues regions $\mathcal{K}_x, \mathcal{K}'_x$. La primera serveix per rebutjar H_0 perquè és massa lluny del resultat esperat. La segona serveix també per rebutjar H_0, però ara per ser-hi massa a prop. També succeeix en ciències experimentals l'anomenat "falsament de segon ordre", que consisteix a obtenir estadístics khi-quadrat que indiquin ser ni massa a prop ni massa lluny del model postulat. Però hi ha tests estadístics que permeten detectar aquest falsament (Rao, 1989, p. 48).

Per a més detalls sobre la correcció de continuïtat hom pot veure, per exemple, Cox (1970), Mantel (1976), Mantel i Greenhouse (1968), Bolshev i Smirnov (1968), Martín i Luna (1990), Huber i Nikulin (1991).

2. CORRECCIÓ DE CONTINUÏTAT PER A LA DISTRIBUCIÓ DE POISSON

Sigui γ_m una variable aleatòria gamma amb m graus de llibertat, és a dir, podem escriure per a qualsevol $\lambda > 0$:

$$\begin{aligned} \boldsymbol{P}\{\gamma_m \geq \lambda\} &= \frac{1}{\Gamma(m)}\int_\lambda^\infty x^{m-1}\mathrm{e}^{-x}\mathrm{d}x = \frac{1}{\Gamma(m+1)}\int_\lambda^\infty \mathrm{e}^{-x}\mathrm{d}(x^m) = \\ &= -\frac{\lambda^m}{\Gamma(m+1)}\mathrm{e}^{-\lambda} + \frac{1}{\Gamma(m+1)}\int_\lambda^\infty x^m\mathrm{e}^{-x}\mathrm{d}x = \\ &= -\frac{\lambda^m}{\Gamma(m+1)}\mathrm{e}^{-\lambda} + \boldsymbol{P}\{\gamma_{m+1} \geq \lambda\}. \end{aligned} \tag{17}$$

Per tant tenim

$$\begin{aligned} \boldsymbol{P}\{\gamma_{m+1} \geq \lambda\} &= \boldsymbol{P}\{\gamma_m \geq \lambda\} + \frac{\lambda^m}{\Gamma(m+1)}\mathrm{e}^{-\lambda} = \\ &= \boldsymbol{P}\{\gamma_m \geq \lambda\} + \boldsymbol{P}\{Z = m|\lambda\}, \end{aligned} \tag{18}$$

on Z és una variable aleatòria que segueix una distribució de Poisson amb paràmetre λ:

$$\boldsymbol{P}\{Z = m|\lambda\} = \frac{\lambda^m}{m!}\mathrm{e}^{-\lambda}, \quad m = 0, 1, 2, \ldots$$

Així hem provat que per a qualsevol $m = 0, 1, 2, \ldots$

$$\boldsymbol{P}\{Z \leq m|\lambda\} = \sum_{k=0}^{m}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} = \boldsymbol{P}\{\gamma_{m+1} \geq \lambda\}. \tag{19}$$

D'altra banda és ben conegut que

$$2\gamma_m = \chi^2_{2m}, \tag{20}$$

i per tant de (19) i (20) se segueix que

$$\boldsymbol{P}\{Z \leq m|\lambda\} = \boldsymbol{P}\{\gamma_{m+1} \geq \lambda\} = \boldsymbol{P}\{\chi^2_{2m+2} \geq 2\lambda\}. \tag{21}$$

Aquesta fórmula ens mostra les relacions existents entre les distribucions Poisson, gamma i khi-quadrat.

Recordem ara que la mitjana i variància de la variable aleatòria χ^2_n són

$$\boldsymbol{E}\chi^2_n = n \quad \text{i} \quad \mathbf{Var}\ \chi^2_n = 2n.$$

Pel teorema central del límit tenim que per a valors grans de n

$$P\left\{\frac{\chi_n^2 - n}{\sqrt{2n}} \leq x\right\} = \Phi(x) + O\left(\frac{1}{\sqrt{n}}\right), \tag{22}$$

així és que per a $n \to \infty$

$$P\{\chi_n^2 \leq x\} = \Phi\left\{\frac{x-n}{\sqrt{2n}}\right\} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right). \tag{23}$$

Una altra aproximació normal per a la distribució khi-quadrat fou proposada per R. A. Fisher, segons la qual per a valors grans de n

$$P\left\{\sqrt{2\chi_n^2} - \sqrt{2n-1} \leq x\right\} = \Phi(x) + O\left(\frac{1}{\sqrt{n}}\right). \tag{24}$$

A fi de provar (24) notem que

$$\begin{aligned} \sqrt{2\chi_n^2} - \sqrt{2n-1} &= \frac{2\chi_n^2 - 2n + 1}{\sqrt{2\chi_n^2} + \sqrt{2n-1}} = \\ &= \frac{\frac{1}{\sqrt{2n}}(\chi_n^2 - n - 0.5)}{\frac{1}{2}\left(\sqrt{\frac{1}{n}\chi_n^2} + \sqrt{1 - \frac{1}{2n}}\right)}. \end{aligned} \tag{25}$$

Si $n \to \infty$, aleshores

$$\frac{1}{n}\chi_n^2 \to 1 \quad \textit{(en probabilitat)}, \quad \sqrt{1 - \frac{1}{2n}} \to 1 \tag{26}$$

i així de (22),(25),(26) i el teorema de Slutsky es dedueix que per a qualsevol x fixat tenim (si $n \to \infty$) que efectivament

$$P\left\{\sqrt{2\chi_n^2} - \sqrt{2n-1} \leq x\right\} = \Phi(x) + O\left(\frac{1}{\sqrt{n}}\right).$$

Fent ús d'aquestes relacions asimptòtiques aconseguim una altra aproximació normal amb la *correcció de continuïtat.*

Suposem que λ és gran. Aleshores de (21) obtenim

$$\begin{aligned} P\{Z \leq m|\lambda\} &= P\{\chi_{2m+2}^2 \geq 2\lambda\} = 1 - P\{\chi_{2m+2}^2 \leq 2\lambda\} = \\ &= 1 - P\left\{\frac{\chi_{2m+2}^2 - (2m+2)}{\sqrt{4m+4}} \leq \frac{2\lambda - 2m - 2}{\sqrt{4m+4}}\right\} \cong \\ &\cong 1 - \Phi\left(\frac{\lambda - m - 1}{\sqrt{m+1}}\right) = \\ &= \Phi\left(\frac{m+1-\lambda}{\sqrt{m+1}}\right). \end{aligned} \tag{27}$$

És interessant de notar que si fem ús de l'aproximació de Fisher (24) aleshores tindrem

$$\begin{aligned} \boldsymbol{P}\{Z \leq m|\lambda\} &= \boldsymbol{P}\{\chi^2_{2m+2} \geq 2\lambda\} = 1 - \boldsymbol{P}\{\chi^2_{2m+2} \leq 2\lambda\} = \\ &\cong 1 - \Phi\left(\sqrt{4\lambda} - \sqrt{4m+3}\right) = \Phi\left(\sqrt{4m+3} - \sqrt{4\lambda}\right) = \\ &= \Phi\left(\sqrt{4(m+0.5)+1} - 2\sqrt{\lambda}\right). \end{aligned} \tag{28}$$

El nombre 0.5 a (28) pot ésser considerat com la *correcció de continuïtat* de l'aproximació normal per a la distribució de Poisson. Aquesta fórmula implica que per a grans valors de λ ($\lambda \geq 25$) l'estadístic

$$\sqrt{4Z+3} - 2\sqrt{\lambda}$$

segueix aproximadament la distribució normal N(0,1). Aquesta aproximació és molt útil a la pràctica per contrastar la hipòtesi $\boldsymbol{H}_0$ que diu que els elements Z_i de la mostra $\mathbb{Z} = (Z_1, \ldots, Z_n)^T$ segueixen la mateixa distribució de Poisson amb paràmetre λ, quan tot Z_i és gran.

D'altra banda, com que la mitjana i la variància de la Poisson és λ, és fàcil obtenir, si λ és gran, una altra aproximació normal estàndard de la distribució de Poisson,

$$\boldsymbol{P}\{Z \leq m|\lambda\} \cong \Phi\left(\frac{m - \lambda + 0.5}{\sqrt{\lambda}}\right). \tag{29}$$

Exemple 2

Un comptador assimilat a un *procés de Poisson* registra 150 pulsacions durant la primera hora, 117 pulsacions durant la segona hora. Podem suposar que la ràtio de pulsacions per unitat de temps és constant (hipòtesi $\boldsymbol{H}_0$)?

Si $\boldsymbol{H}_0$ és certa, els valors observats 150 i 117 poden ésser interpretats com a realitzacions de dues variables aleatòries independents Z_1 i Z_2 seguint la mateixa distribució de Poisson amb paràmetre λ, on el valor de λ és desconegut. Com que $\boldsymbol{H}_0$ implica que les variables aleatòries

$$Y_1 = \sqrt{4Z_1+3} - 2\sqrt{\lambda} \qquad Y_2 = \sqrt{4Z_2+3} - 2\sqrt{\lambda}$$

estan aproximadament distribuïdes segons la llei N(0,1) i són independents, perquè Z_1 i Z_2 ho són. Per tant si la hipòtesi $\boldsymbol{H}_0$ és certa, aleshores, com que $(Y_1 - Y_2)/\sqrt{2}$ és aproximadament N(0,1),

$$X^2 = \frac{1}{2}\left(\sqrt{4Z_1+3} - \sqrt{4Z_2+3}\right)^2, \tag{30}$$

està aproximadament distribuït segons la llei khi-quadrat amb un grau de llibertat, és a dir,

$$P\{X^2 \geq x|H_0\} \cong P\{\chi_1^2 \geq x\}.$$

Les taules de la distribució khi-quadrat proporcionen el valor crític

$$c_\alpha = \chi^2_{1,0.05} = 3.841$$

pel nivell de significació $\alpha = 0.05$, és a dir

$$P\{\chi_1^2 \geq 3.841\} = 0.05.$$

Ara, fent ús dels valors observats Z_1 =150 i Z_2 = 117, podem calcular el valor X^2 de l'estadístic (30) de contrast:

$$X^2 = 0.5\left(\sqrt{4 \times 150 + 3} - \sqrt{4 \times 117 + 3}\right)^2 = 4.071. \tag{31}$$

Com que

$$X^2 = 4.071 > c_\alpha = \chi^2_{1,0.05} = 3.841,$$

concloem que la H_0 (*ràtio de pulsació constant*) ha de ser rebutjada pel test khi-quadrat al nivell de significació $\alpha = 0.05$.

3. SOBRE LA CORRECCIÓ DE CONTINUÏTAT EN TAULES 2×2

Seguint Bolshev i Smirnov (1968), considerem un exemple de Cramer (1946, p.444-445.), vegeu també Fleiss (1981). Suposem (la hipòtesi H_0) que N elements són dividits aleatòriament en dos grups de n i $N-n$ elements i suposem que en el conjunt total dels N elements n'hi ha M posseint alguna propietat Y, i els restants elements no la tenen. La divisió resultant pot ser expressada en la forma d'una taula 2×2 , on μ és el nombre d'elements amb la propietat Y, pertanyent al primer grup:

	amb la prop. Y	sense la prop. Y	total
Mostra del 1r grup	μ	$n-\mu$	n
Mostra del 2n grup	$M-\mu$	$N-n-M+\mu$	$N-n$
Total	M	$N-M$	N

Si la divisió dels elements en els dos grups és realment aleatòria i no depèn de la possessió de la propietat Y, en altres paraules, si $\boldsymbol{H}_0$ és certa

$$(32)\ \boldsymbol{E}\mu = \boldsymbol{E}\{\mu|\boldsymbol{H}_0\} = \frac{nM}{N} \text{ i } \mathbf{Var}\ \mu = \mathbf{Var}\ \{\mu|\boldsymbol{H}_0\} = \frac{nM(N-n)(N-M)}{N^2(N-1)}.$$

Si N és suficientment gran i cap dels $n, M, N-n, N-M$ són gaire petits, aleshores, sota $\boldsymbol{H}_0$, la variable normalitzada

$$(33) \qquad \frac{\mu - \boldsymbol{E}\mu}{\sqrt{\mathbf{Var}\ \mu}}$$

és aproximadament normal N(0,1), i sota $\boldsymbol{H}_0$ per a N gran el quadrat de la variable normalitzada

$$(34) \qquad X^2 = \frac{(\mu - \boldsymbol{E}\mu)^2}{\mathbf{Var}\ \mu}$$

està distribuïda aproximadament com una χ_1^2. L'estadístic X^2 és sovint utilitzat com un criteri per contrastar la hipòtesi $\boldsymbol{H}_0$ que diu que els N elements estan dividits aleatòriament en dos grups. Per a N petit està recomanat que, en lloc de X^2, s'utilitzi l'estadístic

$$(35) \qquad Z^2 = \left(|X| - \frac{1}{2\sqrt{\mathbf{Var}\ \mu}}\right)^2$$

i seguir la regla següent:

si $P(Z^2;1) \leq \alpha$, aleshores rebutgem la hipòtesi $\boldsymbol{H}_0$;
si $P(Z^2;1) \geq \alpha$, aleshores hom pot concloure que els valors observats no contradiuen la hipòtesi $\boldsymbol{H}_0$, essent $P(x;1) = \boldsymbol{P}\{\chi_1^2 \geq x\}$.

A la fórmula (34) el terme

$$\frac{1}{2\sqrt{\mathbf{Var}\ \mu}}$$

és l'anomenada *correcció de continuïtat* per l'aproximació a la funció de distribució $P(x;1)$.

Hi ha taules de valors crítics exactes de l'estadístic $M - \mu$ (criteri de Fisher) per a $n = 3(1)20$ i $N - n = 2(1)n$ (vegeu Pearson i Hartley (1956)), Latscha (1953)). Si $N \geq 20$ aleshores la regió crítica dels tests unilaterals aproximats estan definits (amb l'òbvia correcció de continuïtat) per les desigualtats ($0 < \alpha < 0.5$):

$$(36) \qquad \begin{aligned} \frac{\mu - \boldsymbol{E}\mu}{\sqrt{\mathbf{Var}\ \mu}} &\geq \Psi\left(1 - \frac{\alpha}{2}\right) + \frac{1}{2\sqrt{\mathbf{Var}\ \mu}} \\ \text{i} \\ \frac{\mu - \boldsymbol{E}\mu}{\sqrt{\mathbf{Var}\ \mu}} &\leq -\Psi\left(1 - \frac{\alpha}{2}\right) - \frac{1}{2\sqrt{\mathbf{Var}\ \mu}}, \end{aligned}$$

on $\Psi(x) = \Phi^{-1}(x)$. Aquesta correcció difereix de la correcció de continuïtat proposada per Yates (1934), que no hauria d'esser utilitzada.

De (35) se segueix que la regió crítica del test bilateral aproximat amb nivell de significació α ve donada per la desigualtat

$$X^2 = \frac{(\mu - E\mu)^2}{\mathbf{Var}\ \mu} \geq \left[\Psi\left(1 - \frac{\alpha}{2}\right) + \frac{1}{2\sqrt{\mathbf{Var}\ \mu}}\right]^2 . \tag{37}$$

Ara podem veure que les fórmules aproximades per valors crítics del test exacte de Fisher, basat en (36), ens porten a resultats que són diferents dels basats en la fórmula (34). De fet, segons (34) l'estadístic Z^2 està distribuït (sota $\boldsymbol{H}_0$) aproximadament com χ^2_1, i de (36) segueix que X^2 està distribuïda (sota $\boldsymbol{H}_0$) com

$$\left(\xi + \frac{1}{2\sqrt{\mathbf{Var}\ \mu}}\right)^2 ,$$

on ξ és la variable aleatòria N(0,1). En altres paraules, l'estadístic X^2 està distribuït (sota $\boldsymbol{H}_0$) aproximadament segons la llei khi-quadrat no central $\chi^2_1(\lambda)$ amb paràmetre de no-centralitat $\lambda = (4\mathbf{Var}\ \mu)^{-1}$. Si

$$\mathbf{Var}\ \mu = \frac{nM(N-n)(N-m)}{N^2(N-1)} \to \infty$$

quan $n \to \infty$, aleshores les distribucions de X^2 i Z^2 són asimptòticament iguals.

Més detalls sobre la correcció de continuïtat de Yates es poden trobar a Cochran (1942), (1952), Conover (1974), Mantel (1974), Martín i Luna (1990), Mirvaliev i Nikulin (1992), Nikulin i Greenwood (1990), Plackett (1964), Yates (1934).

AGRAÏMENTS

El primer autor està molt agraït als Professors S. Kotz (USA), A. I. Dale (South Africa) i T. Smith (Canada) per les observacions tan útils i l'encoratjament que li han proporcionat durant la preparació d'aquest article.

REFERENCIES

[1] **Bolshev, L.N.** and **Smirnov, N.V.** (1968). *Tables of Mathematical Statistics.* Nauka, Moscow.

[2] **Cochran, W.G.** (1942). "The 2×2 correction for continuity". *Iowa State College Journal of Science*, **16**, 421–436.

[3] **Cochran, W.G.** (1952). "The χ^2 test of goodness of fit". *Annals of Mathematical Statistics*, **23**, 315–345.

[4] **Conower, W.J.** (1974). "Some reasons for not using the Yates continuity correction on 2×2 contingency tables". *JASA*, **69**, 374–376.

[5] **Cox, D.R.** (1970). "The continuity correction". *Biometrika*, **57**, 217–219.

[6] **Cramer, H.** (1946). *Mathematical Methods in Statistics.* Princeton Univ. Press, Princeton, NJ.

[7] **Fleiss, J.L.** (1981). *Statistical Methods for Rates and Proportions.* Wiley & Sons, New-York.

[8] **Haber, M.** (1980). "A comparison of some continuity corrections for the khi-quadrat test on 2×2 tables". *JASA*, **75**, #**371**, 510–515.

[9] **Huber, C.** and **Nikulin, M.** (1991). *Transformations des variables aléatoires. Application au choix et à la réduction d'un model statistique.* Le Rapport technique du Laboratoire de la Statistique médicale, l'Université Paris 5, 220p.

[10] **Latscha, R.** (1953). "Significance test in 2×2 contingency table". *Biometrika*, **40**, 74–86.

[11] **Mantel, N.** (1974). "Some reasons for not using the Yates continuity correction on 2×2 contingency tables — Comment and suggestion". *JASA*, **69**, 378–380.

[12] **Mantel, N.** (1976). "The continuity correction". *The American Statistician*, **30**, 103–104.

[13] **Mantel, N.** and **Greenhouse S.W.** (1968). "What is the continuity correction?". *The American Statistician*, **22**, #**5**, 27–30.

[14] **Martín, A. y Luna, J. de D.** (1990). *Bioestadística para las Ciencias de la Salud.* Ed. Norma, Madrid. Tercera edició.

[15] **Mirvaliev M.** and **Nikulin M.** (1992). "Goodness-of-fit tests of the chi-square type". *Industrial Laboratory* (Plenum Publishing Corporation), pp. 280–291.

[16] **Nikulin M.** and **Greenwood P.E.** (1990). "A Guide to Chi-Square Testing". *Technical report* #**94**, Department of Statistics, University of British Columbia, Vancouver, Canada.

[17] **Nikulin M.** (1991). *Some recent results on chi-squared tests.* Queen's papers in Pure and Applied Mathematics, #**86**, Queen's University, Kingston, Canada.

[18] **Pearson, E.S.** and **Hartley, H.O.** (1958). *Biometric Tables for Statisticians.* Vol. 1. Cambridge University Press.

[19] **Plackett, R.L.** (1964). "The continuity correction in 2×2 tables". *Biometrika*, **51**, 139–167.

[20] **Rao, C.R.** (1989). *Statistics and Truth.* Int. Co -op Pub. House, Fairland, Maryland.

[21] **Yates, F.** (1934). "Contingency tables involving small numbers and the χ^2 test". *JRSS*, Ser.**B**, Suppl.**1**, 217–235.