REVUE DE STATISTIQUE APPLIQUÉE

E. MORICE

Emploi des tables de la loi de F pour le calcul de l'intervalle de confiance du paramètre p d'une loi binomiale

Revue de statistique appliquée, tome 25, n° 2 (1977), p. 33-38

<http://www.numdam.org/item?id=RSA_1977__25_2_33_0>

EMPLOI DES TABLES DE LA LOI DE F POUR LE CALCUL DE L'INTERVALLE DE CONFIANCE DU PARAMÈTRE p D'UNE LOI BINOMIALE

E. MORICE

La relation entre la loi binomiale et la loi de F :

$$\alpha = \sum_{x=0}^{k} C_n^x p^x (1-p)^{n-x} = 1 - \Pr\left[F < \frac{n-k}{k+1}\,\frac{p}{1-p}\right], \tag{1}$$

$$\nu_1 = 2(k+1) \quad , \quad \nu_2 = 2(n-k)$$

permet de calculer de manière précise les bornes de l'intervalle de confiance bilatéral, $(p_i < p < p_s)$, au niveau de confiance $1 - 2\alpha$:

$$p_i = \frac{k}{k + (n-k+1) F_{1-\alpha}} \quad ; \quad \nu_1 = 2(n-k+1) \quad , \quad \nu_2 = 2k \tag{2}$$

$$p_s = \frac{(k+1) F_{1-\alpha}}{n-k+(k+1) F_{1-\alpha}} \quad ; \quad \nu_1 = 2(k+1) \quad , \quad \nu_2 = 2(n-k) , \tag{3}$$

à condition de posséder des tables suffisamment détaillées de la loi de $F_{1-\alpha}(\nu_1, \nu_2)$.

Les tables calculées par M. Merrington and C.M. Thompson [1], et publiées – soit partiellement, soit après avoir été complétées – dans de nombreux ouvrages donnent les valeurs de $F_{1-\alpha}(\nu_1, \nu_2)$ pour un nombre limité de valeurs de α, ν_1 et ν_2.

Leur emploi peut donc impliquer, dans certains cas, des calculs d'interpolation à l'aide de formules données quelquefois avec les tables.

Ces formules d'interpolation sont du type :

$$\log_{10} F_{1-\alpha}(\nu_1, \nu_2) = \frac{a}{[2\nu_1\nu_2/(\nu_1+\nu_2)] - b} - c\,\frac{\nu_2 - \nu_1}{\nu_1\nu_2} ,$$

dans lesquelles a, b, c sont des constantes données qui varient avec α. Ces calculs, relativement longs, semblent, pour bien des tables d'usage courant, pouvoir être remplacés par des calculs beaucoup plus élémentaires, permettant le calcul de p_i et p_s avec une précision pratiquement suffisante.

Mots-clés : **Loi binomiale. Intervalle de confiance.**

Si l'on utilise, par exemple, les tables statistiques de Hald, [2], on constate que, pour les valeurs assez généralement utilisées de $1 - \alpha$:

$$0{,}95 - 0{,}975 - 0{,}99 - 0{,}995 \quad ,$$

chacune des tables est partagée en 4 zones :

$$\begin{array}{ll}
A : (1 \leqslant \nu_1 \leqslant 30 & , \; 1 \leqslant \nu_2 \leqslant 50) \\
B : (30 < \nu_1 < \infty & , \; 1 \leqslant \nu_2 \leqslant 50) \\
C : (1 \leqslant \nu_1 \leqslant 30 & , \; 50 \leqslant \nu_2 < \infty) \\
D : (30 < \nu_1 < \infty & , \; 50 < \nu_2 < \infty)
\end{array}$$

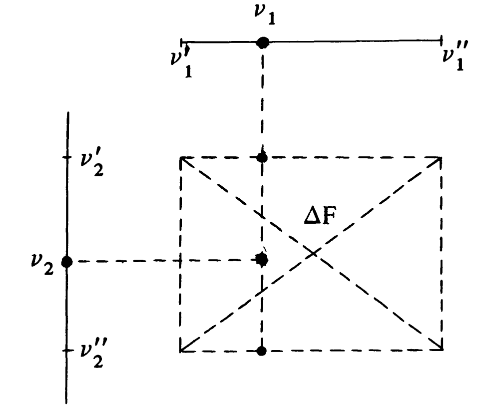
Dans la région A, il n'y a pas de problème d'interpolation pour le calcul de p_i ou p_s : toutes les valeurs paires de ν_1 et ν_2 y figurant.

Dans la région B, le problème d'interpolation ne se posera que dans les lignes où les valeurs successives de F diffèrent au maximum de $\Delta F = 0{,}04$ pour $\alpha = 0{,}05$, jusqu'à $\Delta F = 0{,}06$ pour $\alpha = 0{,}005$.

Dans la région C, le problème d'interpolation ne se posera que dans les colonnes où les valeurs successives de F diffèrent au maximum de 0,02 pour $\alpha = 0{,}05$ à 0,06 pour $\alpha = 0{,}005$.

Dans la région D, le problème d'interpolation se posera à l'intérieur des quadrilatères $(\nu'_1 \, \nu''_1 \, , \nu'_2 \, \nu''_2)$ définis par des couples de valeurs consécutives de ν_1 et ν_2.

Dans ces quadrilatères, la différence maximale ΔF entre les valeurs de F correspondant à deux sommets opposés d'un tel quadrilatère variera de 0,09 pour $\alpha = 0{,}05$ à 0,15 pour $\alpha = 0{,}005$.



De plus, si l'on examine ces divers éléments de la surface $Z = F(\nu_1 , \nu_2)$, on constate que l'on a très sensiblement :

$$\frac{1}{2} \left[F(\nu'_1 , \nu'_2) + F(\nu''_1 , \nu''_2) \right] = \frac{1}{2} \left[F(\nu''_1 , \nu'_2) + F(\nu'_1 , \nu''_2) \right]$$

à 0,05 près, ce qui montre que les quadrilatères élémentaires en lesquels les tables sont partagées, sont très voisins de rectangles plans, et justifie l'emploi de l'interpolation linéaire pour le calcul de $F(\nu_1, \nu_2)$.

Mais cette interpolation exige – dans un de ces rectangles – un triple calcul :

Interpolation pour ν_1 entre ν_1' et ν_1'' pour ν_2' donné

Interpolation pour ν_1 entre ν_1' et ν_1'' pour ν_2'' donné

et enfin, interpolation pour ν_2 entre les deux valeurs de F ainsi calculées.

Ces calculs élémentaires, mais fastidieux, peuvent être pratiquement simplifiés avec une précision suffisante pour le calcul de p_i et p_s.

Des relations (2) et (3) on déduit pour une erreur dF sur l'estimation de $F(\nu_1, \nu_2)$:

$$dp_i = - \frac{k(n-k+1)}{[k+(n-k+1)F]^2} dF \tag{4}$$

$$dp_s = \frac{(k+1)(n-k)}{[n-k+(k+1)F]^2} dF \tag{5}$$

Si l'on ne tient pas compte du fait que F est fonction de k et $(n-k+1)$ ou de $(k+1)$ et $(n-k)$, on obtient aisément, en tenant compte du fait que $F > 1$, une valeur majorante de la valeur de l'extremum de dp, (dp_i ou dp_s) :

$$|dp| < \frac{1/4\,(n+1)^2}{(n+1)^2} |dF| = 0{,}25\ |dF|$$

Il en résulte que, si pour un couple (ν_1, ν_2), on prend comme valeur de F, inconnue, la valeur moyenne des moyennes correspondant aux extrémités des diagonales et pour valeur de dF la valeur $dF = 1/2\,\Delta F$ fournie par la diagonale $[(\nu_1'\,\nu_2'), (\nu_1''\nu_2'')]$, on aura pour toute l'étendue des tables considérées :

$$|dp| \leqslant 0{,}25 \times 0{,}5 \times 0{,}15$$

$$|dp| \leqslant 0{,}019\ .$$

Mais si, de plus on tient compte des liaisons entre $F_{1-\alpha}$ et (n , k), c'est-à-dire si on étudie les variations de dp_i et dp_s en fonction de α, n et k, on constate que :

Les valeurs maximales de dp restent inférieures à 0,012 pour $\alpha = 0{,}005$ et à 0,009 pour $\alpha = 0{,}05$.

D'où la règle pratique :

Pour toutes valeurs de n et k, telles que ν_1 et ν_2 se trouvent dans un intervalle (linéaire ou rectangulaire) de la table, on utilisera, avec ces valeurs de n et k, les formules (2) et (3) en prenant comme valeur de F la valeur centrale de cet intervalle.

Le résultat sera obtenu avec une erreur généralement inférieure à 0,01.

Exemple : $n = 160$, $k = 54$, $1 - \alpha = 0{,}975$.

Cet exemple, choisi dans une région de la table où

$$dF = \frac{1}{2}(1{,}39 - 1{,}28) = 0{,}055$$

est grand et pour lequel l'approximation normale est très bonne, montre que la table de la loi de F donne une approximation comparable à celle de la loi normale avec des calculs numériques beaucoup plus simples.

$$\begin{array}{cccc}
 & 100 & 110 & 200 \\
200 & 1{,}39 & & 1{,}32 \\
212 & & 1{,}3375 & \\
300 & 1{,}36 & & 1{,}28
\end{array}$$

Emploi de la table F :

$$p_{sF} \simeq \frac{55 \times 1{,}3375}{106 + 55 \times 1{,}3375} = 0{,}409$$

Emploi de l'approximation normale :

$$\frac{54 + 0{,}5}{160} - p_s = 1{,}96 \sqrt{\frac{p_s(1 - p_s)}{160}}$$

d'où

$$p_{sN} \simeq 0{,}417$$

soit une différence 0,008 pratiquement négligeable.

L'interpolation dans la table $F_{0,975}$ pour $\nu_1 = 110$, $\nu_2 = 212$ donnerait :

$$\log_{10} F_{0,975}(110, 212) = \frac{1{,}7023}{\sqrt{h} - 1{,}14} - 0{,}846 \frac{212 - 110}{110 \times 212}$$

avec :

$$h = \frac{2 \times 110 \times 212}{110 + 212} = 144{,}8447$$

$$\log_{10} F_{0,975}(110, 212) = 0{,}1383 \quad , \quad F = 1{,}375,$$

valeur qui donnerait $p_{sF} = \dfrac{55 \times 1{,}375}{106 \times 55 \times 1{,}375} = 0{,}416$

C'est beaucoup de calculs, pour une précision supplémentaire pratiquement sans intérêt.

On aurait pu d'ailleurs remarquer dans le cas présent que le point (ν_1 , ν_2) étant voisin du point (100, 200) il suffisait pratiquement de prendre F = 1,39 qui donne

$$p_{sF} = \frac{55 \times 1{,}39}{106 + 55 \times 1{,}39} = 0{,}419$$

Remarque

On s'est attaché, dans ce qui précède, à l'étude d'une limite supérieure de l'erreur dp.

Cette erreur est une fonction compliquée de n, k et F(n , k). On notera que, pour les petites valeurs de $k/n = \hat{p}$ elle est particulièrement négligeable.

Par exemple, pour n = 100, k = 5, on trouverait dans un test bilatéral à $1 - \alpha = 0{,}95$ pour

$$p_i = 0{,}016 \quad , \quad \hat{p} = 0{,}05 \quad , \quad p_s = 0{,}112$$

$$dp_s < 0{,}001 \quad , \quad dp_i < 0{,}0001 \, .$$

(voir tableau ci-après).

Valeurs approchées des erreurs commises sur p_i et p_s, lorsque l'on utilise comme valeur de $F_{0,995}(\nu_1, \nu_2)$ la valeur centrale de l'intervalle tabulaire comprenant le point (ν_1, ν_2) : Table de Hald.

$$dp_s = \frac{(k+1)(n-k)}{[n-k+(k+1)F]^2} dF \quad \nu_1 = 2(k+1) \quad , \quad \nu_2 = 2(n-k)$$

$$dp_i = \frac{k}{[k+(n-k+1)F]^2} dF \quad \nu_1 = 2(n-k+1) \quad , \quad \nu_2 = 2k$$

dF = demi intervalle tabulaire (ligne, colonne ou diagonale)
dp_s : ligne supérieure ; dp_i : ligne inférieure.
Pour dp_s, si $k < 15$ et $n - k < 25$, il n'y a pas d'interpolation.
Pour dp_i, si $n - k < 15$ et $k < 25$, il n'y a pas d'interpolation.
(Les erreurs dp diminuent avec le niveau de confiance $1 - \alpha$).

k / n – k		15	30	50	100	250
15		0,002 0,002	0,002 0,002	0,002 0,007	0,0005 0,005	0,0002 0,002
30		0,006 0,002	0,005 0,005	0,006 0,008	0,003 0,003	0,001 0,001
50	< 0,001	0,006 0,001	0,008 0,006	0,010 0,010	0,008 0,012	0,005 0,004
100	< 0,001	0,005 0,0005	0,008 0,003	0,012 0,007	0,011 0,011	0,008 0,011
250	< 0,001	0,003 0,0002	0,004 0,002	0,009 0,004	0,011 0,008	0,012 0,012

BIBLIOGRAPHIE

[1] M. MERRINGTON & C. THOMPSON. – *Biometrika,* Vol. 33 (1943), p. 73-88. *Biometrika, Tables* Tome I, 1969, Table 18.

[2] A. HALD. – Statistical Tables and Formulas. *J. Wiley and Sons.*