NORMAL APPROXIMATION TO THE DISTRIBUTION OF TWO INDEPENDENT BINOMIALS, CONDITIONAL ON FIXED SUM

BY J. HANNAN[1] AND W. HARKNESS[2,3]

Michigan State University

In 2×2 contingency tables $[n_{ij}]$ resulting from $n_{..}$ independent trials with probability matrix $[p_{ij}]$, n_{11} and n_{21} are, conditional on fixed $n_{1\cdot}$, independent binomials with parameters $(n_{1\cdot}, p_{11}/p_{1\cdot})$ and $(n_{2\cdot}, p_{21}/p_{2\cdot})$. Further conditioned on fixed $n_{\cdot 1}$, their distribution is that of our title and a normal approximation has been suggested by Patnaik [6]. That this approximation is based on an erroneous limiting distribution, unless $|[p_{ij}]| = 0$ (where it reduces to the classical normal approximation to the hypergeometric), was promptly pointed out by Stevens [8], who also asserted that the conditional mean and variance of n_{11} are approximated by the normal parameters suggested by the theorem to follow.

The distribution of independent binomials with parameters (n_i, p_i), $0 < p_i < 1$, conditional on fixed sum c in $\{0, 1, \cdots, n_1 + n_2\}$ is

$$(1) \quad f(k_1) = \prod_{i=1}^{2} b(k_i; n_i, p_i) \Big/ \sum_{k_1+k_2=c} \prod_{i=1}^{2} b(k_i; n_i, p_i) \quad \text{on} \quad k_1 + k_2 = c$$

(notation not otherwise defined is that of Chapters VI and VII of Feller [3]). Noting that f depends on the p_i only through $\lambda = p_1 q_2 / q_1 p_2$, we prepare (1) for simultaneous normal approximation to the separate binomial probabilities by replacing (p_1, p_2) by the unique (P_1, P_2), $0 < P_i < 1$, satisfying

$$(2) \qquad P_1 Q_2 = \lambda Q_1 P_2, \qquad N_1 P_1 + N_2 P_2 = c + 1,$$

where, here and throughout, $N_i = n_i + 1$.

THEOREM. *With* $H_i = (N_i P_i Q_i)^{-\frac{1}{2}}$, $H^2 = H_1^2 + H_2^2$ *and* $X_k = H(k - N_1 P_1 + \frac{1}{2})$,

$$(a) \qquad f(k) \sim H\phi(X_k) \qquad \text{as } H, HX_k^3 \to 0,$$

$$(b) \qquad \sum_{\alpha}^{\beta} f(k) \sim \Phi(X_{\beta+\frac{1}{2}}) - \Phi(X_{\alpha-\frac{1}{2}}) \qquad \text{as } H, HX_\alpha^3, HX_\beta^3 \to 0,$$

$$(c) \qquad \sum_{\alpha}^{c} f(k) \lesssim X_\alpha^{-1}\phi(X_\alpha) \qquad \text{as } HX_\alpha \to 0+,$$

$$(d) \qquad \sum_{\alpha}^{c} f(k) \sim 1 - \Phi(X_{\alpha-\frac{1}{2}}) \qquad \text{as } H, HX_\alpha^3 \to 0.$$

Received May 9, 1963.

[1] Supported in part by National Science Foundation Grant No. G-5011.

[2] Supported by Office of Naval Research Grant No. Nonr-2410(00).

[3] Now at Pennsylvania State University.

REMARK 1. By (a) we mean that $f(k)/H\phi(X_k)$, admittedly not a function of H and HX_k^3, will be in any pre-assigned neighborhood of 1 provided (H, HX_k^3) is in a sufficiently small Euclidean neighborhood of (0, 0); by (c) we mean the corresponding assertion for $X_\alpha \sum_\alpha^c f(k)/\phi(X_\alpha)$, a neighborhood of $\{r \mid 0 < r \leqq 1\}$ and a right hand neighborhood of 0.

REMARK 2. n_1, n_2, c, and $n_1 + n_2 - c \to \infty$ as $H \to 0$ and, if λ and n_1/n_2 are bounded away from 0 and ∞, the converse follows from (2).

REMARK 3. The distribution function convergence as $H \to 0$, which is implied by (b), will, after the preparation (2), also follow from the very general theorems of Steck ([7], Section 2).

PROOF OF THE THEOREM. Let f^* be defined by

$$f^*(k_1) = (2\pi)^{\frac{1}{2}} H \prod_1^2 H_i^{-1} b(k_i\,;n_i\,,P_i) \quad \text{on} \quad k_1 + k_2 = c, \tag{3}$$

and let (a*)–(d*) denote the propositions (a)–(d) with f replaced by f^*. Noting that $f(k) = f^*(k)/\sum_0^c f^*(k)$ and that (d*) implies (by two applications) that $\sum_0^c f^*(k) \sim 1$ as $H \to 0$, it follows that (a*)–(d*) imply (a), (b), (d). Since (c) is trivially true for H bounded away from 0, it too is so implied.

Noting that $k_2 - N_2P_2 + \frac{1}{2} = -(k_1 - N_1P_1 + \frac{1}{2})$, (a*) follows on combining the two applications of the superior normal approximation to the binomial ([3]; VII, problems 19–21 (the additional condition, $H_i \to 0$, eliminates the need for fixed P_i)),

$$b(k_i\,;n_i\,,P_i) \sim H_i\phi(H_iX_{k_1}/H) \quad \text{as} \quad H_i\,,\, H_i(H_iX_{k_1}/H)^3 \to 0. \tag{4}$$

As in the binomial case, since X_k^3 is in $[X_\alpha^3, X_\beta^3]$ for k in $[\alpha, \beta]$ it follows from (a*) that

$$\sum_\alpha^\beta f^*(k) \sim \sum_\alpha^\beta H\phi(X_k) \quad \text{as} \quad H, HX_\alpha^3, HX_\beta^3 \to 0. \tag{5}$$

(b*) follows from (5) since[4], for $0 < h$ and $-\infty < x < \infty$,

$$e^{-h^2/24} < \Phi]_{x-h/2}^{x+h/2}/h\phi(x) < h^{-1}\int_{x-h/2}^{x+h/2} e^{x(x-t)}\,dt < e^{h^2x^2/24} \tag{6}$$

by Jensen's inequality and the elementary inequalities, $x^2 - t^2 \leqq 2x(x-t)$ and $u^{-1}\sinh u < \exp(u^2/6)$, and hence

$$e^{-H^2/24} < \Phi]_{X_{\alpha-\frac{1}{2}}}^{X_{\beta+\frac{1}{2}}} \Big/ \sum_\alpha^\beta H\phi(X_k) < \exp(H^2 \max[X_\alpha^2, X_\beta^2]/24). \tag{7}$$

(c*) is an asymptotic version of the analogue of the binomial tail bound ([3]; VI (3.5)). Letting $t_i = H_i^2/H^2$ and $V_k = HX_k$,

[4] For $0 < h < 1$ and $|xh| < 1.4$, Feller [1; Lemma 1] (cf. Nicholson [5]) has given much tighter bounds. (6) is a substitute for (7.6.3) of Feller [2] and an alternative to VII(2.15-17) of Feller [3].

$$
(8)\quad \frac{f^*(k+1)}{f^*(k)} = \frac{(N_1-k-1)P_1}{(k+1)Q_1}\,\frac{(c-k)Q_2}{(N_2-c+k)P_2} < \frac{1-t_1P_1V_k}{1+t_1Q_1V_k}\,\frac{1-t_2Q_2V_k}{1+t_2P_2V_k}
$$

and, hence, if $X_\alpha \geqq 0$, $\sum_\alpha^c f^*(k)$ is bounded by a geometric series,

$$
(9)\quad \sum_\alpha^c f^*(k) < \frac{f^*(\alpha)}{1-f^*(\alpha+1)/f^*(\alpha)} < f^*(\alpha)\,\frac{(1+t_1Q_1V_\alpha)(1+t_2P_2V_\alpha)}{V_\alpha + t_1t_2(P_2-P_1)V_\alpha^2}.
$$

(c^*) then follows from (a^*) and (9).

(d^*) is a slightly strengthened analogue of the "large deviation" theorem ([3]; VII(5.1)). Abbreviating $X_{\alpha-\frac{1}{2}}$ by a, $X_{\beta+\frac{1}{2}}$ by b and taking β so that, as $H \to 0$, $Hb^3 \to 0$ and b, $b^2 - (a^+)^2 \to \infty$ (for example, $\beta =$ the least integer with $b \geqq \max\{a + \log a^+, H^{-\frac{1}{4}}\}$), VII (6.1) of [3] insures $b^{-1}\phi(b) \sim \Phi]_b^\infty$ and $\Phi]_b^\infty/\Phi]_a^\infty \to 0$ as $H \to 0$. Hence (d^*) follows from the implication of (b^*) and a slight weakening of (c^*),

$$
(10)\qquad \Phi]_a^b \lesssim \sum_\alpha^c f^*(k) \lesssim \Phi]_a^b + b^{-1}\phi(b) \quad \text{as} \quad H, HX_\alpha^3 \to 0,
$$

and the proof of the theorem is complete.

REFERENCES

[1] Feller, W. (1945). On the normal approximation to the binomial distribution. *Ann. Math. Statist.* **16** 319–329.

[2] Feller, William (1950). *An Introduction to Probability Theory and Its Applications*, **1**. Wiley, New York.

[3] Feller, William (1957). *An Introduction to Probability Theory and Its Applications*, **1** (2nd ed.). Wiley, New York.

[4] Harkness, William Leonard (1959). An investigation of the power function for the test of independence in 2 × 2 contingency tables. Ph.D. Thesis, Michigan State Univ. Library.

[5] Nicholson, W. L. (1956). On the normal approximation to the hypergeometric distribution. *Ann. Math. Statist.* **27** 471–483.

[6] Patnaik, P. B. (1948). The power function of the test for the difference between two proportions in a 2 × 2 table. *Biometrika* **35** 157–175.

[7] Steck, George P. (1957). Limit theorems for conditional distributions. *Univ. California Publ. Statist.* **2** 237–284.

[8] Stevens, W. L. (1951). Mean and variance of an entry in a contingency table. *Biometrika* **38** 468–470.