BIOMETRICS 34, 483–486
September, 1978

An Improved Approximate Formula for Calculating Sample Sizes for Comparing Two Binomial Distributions

J. T. CASAGRANDE and M. C. PIKE

Department of Community and Family Medicine, University of Southern California School of Medicine, 2025 Zonal Avenue, Los Angeles, California 90033, U.S.A.

P. G. SMITH

D.H.S.S. Cancer Epidemiology and Clinical Trials Unit, Oxford University, Oxford, England

Summary

An improved approximate formula is derived for determination of the sample sizes required for detecting the difference between two binomial parameters with specified type I and type II errors.

1. Introduction

If x and y are each binomially distributed with index n and parameters p_1 and p_2 respectively, then the comparison of these two binomial distributions is usually displayed as a 2×2 table and Fisher's "exact" (one-sided) test may be used to test the null hypothesis, $p_1 = p_2$, against the alternative hypothesis that $p_1 > p_2$.

The exact test is based on arguing conditionally on the observed number of "successes", i.e., $x + y$, see Yates (1934) and Fisher (1935). The distribution of x with $x + y = m$ fixed depends on p_1 and p_2 only through the odds ratio $\theta = (p_1q_2)/(q_1p_2)$, where $q_i = 1 - p_i$. The conditional distribution is $\Pr(x|\theta) \equiv C(n,x)\, C(n,y)\, \theta^x / \Sigma_i C(n,i) C(n,m-i)\theta^i$, where i takes the values $L = \max(0, m - n)$ to $U = \min(n, m)$.

Let x_c be the critical value of x for the exact test of $p_1 > p_2$ (i.e., $\theta > 1$) against the null hypothesis $\theta = 1$ with type I error α, so that

$$\sum_{i=x_c}^{U} \Pr(i|\theta = 1) \le \alpha \quad \text{and} \quad \sum_{i=x_c-1}^{U} \Pr(i|\theta = 1) > \alpha.$$

The conditional power is thus $\beta(\theta|m) = \Sigma_{i=x_c} \Pr(i|\theta)$ and the expected power is $\beta(\theta, p_1) = \Sigma_m\, \beta(\theta|m\Pr(m)$, where m takes the values 1 to $2n - 1$, and

$$\Pr(m) = \sum_{i=L}^{U} C(n,i)\, p_1{}^i\, q_1{}^{n-i}\, C(n,m-i)\, p_2{}^{m-i}\, q_2{}^{n-m+i}.$$

To find the minimum n to achieve a power of 100β percent an iterative procedure is required. This involves very extensive calculations and numerous approximations have thus been suggested. The two most commonly employed are:

(1) The "arcsin formula" as given, for example, in Cochran and Cox (1957),

Key Words: Sample size; 2×2 table.

$$n = (z_{1-\alpha} + z_\beta)^2/[2(\arcsin \sqrt{p_1} - \arcsin \sqrt{p_2})^2], \tag{1}$$

where $\Phi(z_\gamma) = \gamma$ and Φ is the cumulative normal distribution.

(2) The "uncorrected χ^2 formula" as given, for example, in Fleiss (1973),

$$n = [z_{1-\alpha} \sqrt{(2\bar{p}\bar{q})} + z_\beta \sqrt{(p_1q_1 + p_2q_2)}]^2/(p_1 - p_2)^2, \tag{2}$$

where $\bar{p} = (p_1 + p_2)/2$, $\bar{q} = 1 - \bar{p}$.

In general formulae (1) and (2) give very similar answers, which are too low (Haseman 1978). A "corrected χ^2 method" given by Kramer and Greenhouse (1959) is

$$n = A[1 + \sqrt{\{1 + 8\,(p_1 - p_2)/A\}}]^2/[4\,(p_1 - p_2)^2], \tag{3}$$

where $A = [z_{1-\alpha} \sqrt{(2\bar{p}\bar{q})} + z_\beta \sqrt{(p_1q_1 + p_2q_2)}]^2$. It can be seen that equation (3) is equation (2) with a correction factor. Unlike equations (1) and (2), equation (3) is conservative.

Table 1 shows the exact values compared to the values obtained by the three approximations for several values of p_1 and p_2. Inspection of this table shows that an average of the corrected and uncorrected χ^2 approximations would give a very close estimate of the exact values. This led us to reassess the derivation of the Kramer-Greenhouse corrected χ^2 method.

2. *Derivation of Corrected* χ^2

Consider the statistic $d = 0.5(x - y)$. The usual χ^2 test with Yates' correction may be written

$$X^2 = (d - 0.5)^2/V(d), \tag{4}$$

where the variance of d (on the null hypothesis) is $V(d) = 0.5n\,[(x + y)/2n][1 - (x + y)/2n]$.

Now simplify (4) by replacing $x + y$ by its expectation, and then solve it for d^* with X^2 set to the specified type I error, i.e., $z_{1-\alpha}{}^2 = (d^* - 0.5)^2/(0.5n\bar{p}\bar{q})$,

thus

$$d^* = 0.5 + z_{1-\alpha} \sqrt{(0.5n\bar{p}\bar{q})}. \tag{5}$$

If this n and d^* are to lead to a test with power β then $\Pr(d \geq d^*) = \beta$. But d has expected value $0.5n(p_1 - p_2)$ and variance $0.25(p_1q_1 + p_2q_2)$, so that $\Pr(d \geq d^*) \doteq 1 - \Phi(w)$, where Φ is the cumulative normal distribution function, $w = [d^* - 0.5n(p_1 - p_2) - c]/\sqrt{[0.25n(p_1q_1 + p_2q_2)]}$, and c is a correction factor to allow for the discreteness in the distribution of d. Thus we obtain a second equation for d^* to satisfy, i.e.,

$$-z_\beta = [d^* - 0.5n(p_1 - p_2) - c]/\sqrt{[0.25n(p_1q_1 + p_2q_2)]}.$$

TABLE 1

Comparison of Exact Values with Those of Three Approximations with a 90% Chance of Declaring a Significant Result on a One-Sided Test at the 5% Level

	$p_1 = 0.40$ $p_2 = 0.25$	0.50 0.25	0.60 0.50	0.80 0.50
Exact method	178	71	445	47
Arcsin Approximation	165	63	420	41
Uncorrected χ^2	166	63	423	42
Kramer-Greenhouse χ^2	191	78	462	54

TABLE 2

Comparison of 'Exact' Values with Those Obtained by an Improved Approximate Formula for a Power of 90% and a One-Sided Test at the 5% Level

	$\delta = p_1 - p_2$ where $p_1 > p_2$													
p_2	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
0.05	504	165	89	57	42	33	25	21	18	15	13	11	10	9
	513	172	95	63	46	35	28	23	20	17	14	12	11	9
0.10	782	232	119	74	52	39	31	25	20	17	15	12	11	10
	787	237	121	77	54	41	32	26	21	18	15	13	11	10
0.15	1024	289	142	87	60	45	34	27	23	18	16	12	11	10
	1027	292	144	89	61	45	35	28	23	19	16	13	11	10
0.20	1231	338	162	97	65	47	36	30	23	18	16	12	11	10
	1233	339	163	98	66	48	37	29	23	19	16	14	11	10
0.25	1402	377	178	106	71	51	36	31	24	18	16	12	11	9
	1404	378	179	106	71	51	38	30	24	19	16	13	11	9
0.30	1538	408	190	111	72	53	40	31	24	18	16	12	10	—
	1541	408	190	111	73	52	39	30	24	19	16	13	11	—
0.35	1640	428	200	115	72	53	40	31	23	18	15	11	—	—
	1644	429	198	114	75	53	39	30	23	19	15	12	—	—
0.40	1710	445	202	116	72	53	36	30	23	17	13	—	—	—
	1713	442	202	115	75	52	38	29	23	18	14	—	—	—
0.45	1746	446	202	115	72	51	36	27	20	15	—	—	—	—
	1747	447	202	114	73	51	37	28	21	17	—	—	—	—
0.50	1746	445	200	111	71	47	34	25	18	—	—	—	—	—
	1747	442	198	111	71	48	35	26	20	—	—	—	—	—

Upper figure = Exact Value

Thus

$$d^* = 0.5n(p_1 - p_2) + c - z_\beta \sqrt{[0.25(np_1q_1 + np_2q_2)]}. \tag{6}$$

To solve for n we set (5) = (6). This leads to

$$n = A[1 + \sqrt{\{1 + 4(1 - 2c)(p_1 - p_2)/A\}}]^2/[4(p_1 - p_2)^2]. \tag{7}$$

Setting $c = -0.5$ in (7), one obtains the Kramer-Greenhouse formula (equation 3) and setting $c = 0.5$, one obtains the uncorrected formula (2). For $c = 0$ *a* value for n roughly midway between (2) and (3) is obtained.

The choice of $c = 0$ would appear to be correct since the equations for d^* are not integral and the correction factor has already been employed in its calculation (equation 5). The choice $c = -0.5$ will clearly lead to a conservative test. We therefore suggest as an improved χ^2 formula

$$n = A[1 + \sqrt{\{1 + 4(p_1 - p_2)/A\}}]^2/[4(p_1 - p_2)^2]. \tag{8}$$

We have tested this approximation over a wide range of values of p_1 and p_2. Table 2 compares the exact values with those obtained with the improved approximation at the one-sided significance level $\alpha = 0.05$ for $\beta = 0.90$. The values of p_1 and p_2 were chosen to correspond to those of Cochran and Cox (1957). Formula (8) is clearly an excellent approximation and effectively eliminates the necessity of ever requiring the exact values.

Isographs calculated using Formula (1) were recently presented by Feigl (1978) with the statement that care should be exercised in their use due to the known inaccuracy of Formula (1); recomputed isographs using Formula (8) could be used without such a precaution.

Acknowledgments

This research is supported by Contract N01-CP-53500 and Grant P01-CA-17054 from the National Cancer Institute, National Institutes of Health.

Résumé

Une formulation d'approximation améliorée est déterminée pour les tailles d'échantillons nécessaires à la sélection des différences entre deux paramètres binomiaux avec des erreurs de type I et II spécifiés.

References

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, John Wiley and Sons, Inc., New York.

Feigl, P. (1978). A graphical aid for determining sample size when comparing two independent proportions. *Biometrics 34*, 111–122.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, Series A 98*, 39–54.

Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions*, John Wiley and Sons, Inc., New York.

Haseman, J. K. (1978). Exact sample sizes for use with the Fisher-Irwin Test for 2 $\times$ 2 tables. *Biometrics 34*, 106–109.

Kramer, M. and Greenhouse, S. W. (1959). Determination of sample size and selection of cases. In *Psychopharmacology: Problems in Evaluation*. J. O. Cole and R. W. Gerard (eds.), National Academy of Sciences, National Research Council, Washington, D. C., Publication 583, 356–371.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society, Supplement 1*, 217–235.

Received November 1977; *Revised March* 1978