Statistica Sinica 3(1993), 189-196

# A NEW APPROXIMATION TO THE DISTRIBUTION OF PEARSON'S CHI-SQUARE

Charles S. Davis

*University of Iowa*

*Abstract:* A new approximation to the distribution of Pearson's chi-square statistic for testing independence in two-way contingency tables is described. Using the exact first three moments of the test statistic under the conditional permutation distribution, the distribution is approximated by that of $(aW)^k$, where $W$ has a chi-square distribution. In an extensive comparison of the new generalized chi-square procedure and several other tests of independence with the "exact" conditional test, the new method consistently yields estimated $p$-values which agree closely with the exact results.

*Key words and phrases:* Generalized gamma approximation, Pearson's chi-square statistic, test of independence, two-way contingency table.

## 1. Introduction

When a sample of $N$ observations is classified with respect to two qualitative variables, the resulting frequencies are often displayed in an $r \times c$ contingency table, in which case $n_{ij}$ is the observed count in the $i$th row and $j$th column. Let $n_{i\cdot}$ and $n_{\cdot j}$ denote the row and column marginal totals, respectively, and let $E_{ij} = n_{i\cdot}n_{\cdot j}/N$. In testing the independence of the two qualitative variables, Pearson's statistic

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

is commonly used. If the expected counts are not "too small", $X^2$ has an approximate chi-square distribution with $(r-1)(c-1)$ degrees of freedom $(\chi^2_{(r-1)(c-1)})$.

For contingency tables with many small expectations, Cochran (1954) fitted a normal approximation to the distribution of $X^2$, using the exact mean and variance given by Haldane (1940). Alternative gamma and lognormal two-moment approximations were studied by Nass (1959) and Lawal and Upton (1984). Lewis et al. (1984) derived the exact third central moment of $X^2$ and studied a three-moment location-shifted gamma approximation. The third moment was independently derived by Mielke and Berry (1985), who proposed a Pearson type III

approximation. Both approaches are equivalent to a location-shifted chi-square approximation $a + b\chi_p^2$, where $a$, $b$ and $p$ are chosen to match the first three moments of $X^2$. In empirical studies reported by Lewis et al. (1984) and Berry and Mielke (1988), the three-moment $\chi^2$ approximation resulted in significance levels closer to the nominal levels than the other tests considered.

A generalized chi-square approximation to the distribution of $X^2$ is proposed in Section 2. The distribution is approximated by that of $(aW)^k$, where $W$ has a $\chi_p^2$ distribution and the parameters $a$, $k$ and $p$ are chosen to match the first three moments of $X^2$. A simple iterative procedure for determining the moment estimators of the parameters is described. Like the location-shifted chi-square approximation, this method uses the first three moments of $X^2$ and thus may yield more accurate results than the asymptotic $\chi^2_{(r-1)(c-1)}$ distribution or a two-moment approximation. An important advantage over the location-shifted approximation is that the generalized chi-square approximation has the same range as that of $X^2(0, \infty)$.

Section 3 summarizes an extensive numerical comparison of the generalized chi-square approximation and several other approximate tests with the "exact" conditional test. In contrast with published empirical studies emphasizing comparisons between nominal and actual significance levels under the null hypothesis, differences between approximate and exact $p$-values are studied. Over a large number of configurations of two-way tables with dimensions ranging from $2 \times 3$ to $4 \times 5$, the new approximation consistently yields estimated $p$-values which agree closely with the exact results.

## 2. A New Approximation to the Distribution of $X^2$

Let $Y = (aW)^k$, where $W$ has a $\chi_p^2$ distribution. Note that $Y$ has the generalized gamma distribution (Stacy (1962), Johnson and Kotz (1970, p.197)). The $t$th moment about the origin of $Y$ is

$$\mu_t'(Y) = \frac{(2a)^{kt}\Gamma(kt+v)}{\Gamma(v)},$$

where $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$ and $v = p/2$. For a given $r \times c$ table, the distribution of $X^2$ can be approximated by that of $Y$, where $a$, $k$ and $p$ are chosen to match the exact first three moments of $X^2$. The case $a = k = 1$, $p = (r-1)(c-1)$ corresponds to the usual asymptotic approximation.

In solving for $a$, $k$ and $p$, it is convenient to define

$$\begin{aligned} M_1(Y) &= \frac{\mu_2'(Y)}{\{\mu_1'(Y)\}^2} = \frac{\Gamma(v)\Gamma(2k+v)}{\{\Gamma(k+v)\}^2}, \\ M_2(Y) &= \frac{\mu_3'(Y)}{\{\mu_1'(Y)\}^3} = \frac{\{\Gamma(v)\}^2\Gamma(3k+v)}{\{\Gamma(k+v)\}^3}. \end{aligned} \tag{1}$$

Now let $\mu$, $\mu_2'$ and $\mu_3'$ denote the exact first three moments about the origin of $X^2$ and let $M_1(X^2) = \mu_2'/\mu^2$ and $M_2(X^2) = \mu_3'/\mu^3$. Given values of $k$ and $p$ satisfying (1) with $M_1(X^2)$ and $M_2(X^2)$ replacing $M_1(Y)$ and $M_2(Y)$, the value of $a$ can be found by equating the means of $X^2$ and $Y$.

This method was studied by Davis (unpublished M.Sc. Thesis, McMaster University (1975)) and used by Solomon and Stephens (1977, 1980, 1983) in approximating positive random variables with known first three moments. However, explicit solutions for $k$ and $p$ were not provided due to the complexity of (1). Instead, values of $M_1(Y)$ and $M_2(Y)$ were tabulated over an extensive grid of $(k,p)$ values. For given $(M_1, M_2)$ of the random variable to be approximated, the corresponding $(k,p)$ values were found by inverse double interpolation.

Explicit solutions can be found by rewriting equation (1) as

$$\begin{aligned} \log M_1(Y) &= \log\Gamma(v) + \log\Gamma(2k+v) - 2\log\Gamma(k+v), \\ \log M_2(Y) &= 2\log\Gamma(v) + \log\Gamma(3k+v) - 3\log\Gamma(k+v), \end{aligned} \tag{2}$$

where $\log\Gamma(x)$ is the natural logarithm of the gamma function. The partial derivatives of (2) with respect to $k$ and $v$ are in terms of the $\psi$ function (Abramowitz and Stegun (1965, p.258)). Since fast and accurate algorithms for computing $\log\Gamma(x)$ and $\psi(x)$ are available (MacLeod (1989), Bernardo (1976)), Equation (2) can easily be solved using Newton's method and starting values of $k_0 = 1$ and $v_0 = \frac{1}{2}(r-1)(c-1)$. If the calculated value of $X^2$ for a given $r \times c$ table is $x$, we then approximate the tail probability $\mathrm{pr}(X^2 > x)$ by $\mathrm{pr}(W > (x/a)^{1/k})$, which can be calculated for nonintegral degrees of freedom using the algorithm of Shea (1988).

## 3. Numerical Comparisons

Most studies of the empirical properties of Pearson's $X^2$ and other tests of independence have assessed the accuracy of significance levels under the null hypothesis. Due to recent developments enabling the computation of exact conditional tail probabilities for $r \times c$ tables once beyond the range of computational feasibility, more extensive evaluations were carried out.

The basic design of the empirical study involved the following factors:

1. eight table dimensions $(2\times3, 2\times4, 2\times5, 3\times3, 3\times4, 3\times5, 4\times4, 4\times5)$;
2. three average expected cell frequencies (3, 6, 9);
3. three row marginal total patterns (ratios of largest marginal total to smallest marginal total of 1, 2 and 5, with a common value for the remaining margins);
4. three column marginal total patterns (same as for row marginal totals).

These define a reasonably comprehensive set of contingency table configurations covering most practical settings involving small and moderately-small expectations. Tests of independence in $2 \times 2$ tables are not considered, since this case has been studied extensively, e.g., Haber (1980).

For each combination of the design factors, multiple random contingency tables were generated using the algorithm of Patefield (1981); random numbers from the uniform $(0, 1)$ distribution were obtained using the Wichmann and Hill (1982) generator. Exact two-tailed $p$-values were calculated using the algorithm of Mehta and Patel (1986). The sampling scheme involved the generation of tables over a wide range of exact $p$-values; further details are available from the author. In all, 5532 unique tables were used in the study.

For each table generated, the exact $p$-value was compared with the approximate $p$-value resulting from each of the following nine statistics:

A. Pearson's $X^2$;
B. Likelihood ratio (LR) chi-square statistic: $2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log(n_{ij}/E_{ij})$;
C. Cressie-Read (1984) power divergence statistic with $\lambda = 2/3$;
D. Three-moment generalized chi-square approximation;
E. Location-shifted chi-square approximation (Lewis et al. (1984), Mielke and Berry (1985));
F. William's (1976) corrected LR statistic (with correction term calculated using expected values);
G. William's (1976) corrected LR statistic (using the minimum value of the correction term);
H. Gart's (1966, p.170, Equation 5.8) modified LR statistic;
I. Gart's (1966, p.169, Equation 5.6) simpler modified LR statistic.

These methods include the classical statistics (A, B), the Cressie-Read power divergence statistic (C), the three-moment approximations (D, E) and "corrected" LR tests (F–I). Two-moment approximations were not included since these have been shown to be less accurate than approximation E (Lewis et al. (1984), Berry and Mielke (1988)). Significance levels for all statistics except D and E were calculated using the asymptotic $\chi^2_{(r-1)(c-1)}$ distribution.

Table 1 displays average $(\sum d/5532)$ and root mean square $(\sum d^2/5532)^{1/2}$ errors of the nine approximate tests over all 5532 tables, where $d$ denotes approximate $p$-value − exact $p$-value. In addition, separate tabulations are given for exact $p$-values in the range 0–0.15 (2448 tables), 0.15–0.85 (1707 tables) and 0.85–1 (1377 tables). Overall, the generalized chi-square approximation (D) has the smallest average and root mean square errors. This approximation has the smallest average error when the exact $p$-value is in either tail and is nearly as

accurate as the location-shifted chi-square approximation (E) when the exact $p$-value is in the range 0.15–0.85.

Table 1. Average and root mean square error of differences between $p$-values (approximate – exact)

| | | | Exact $p$-value range | | | | | |
|---|---|---|---|---|---|---|---|---|
| | All tables ($n = 5532$) | | $< 0.15$ ($n = 2448$) | | 0.15-0.85 ($n = 1707$) | | $> 0.85$ ($n = 1377$) | |
| Method* | Mean | RMSE | Mean | RMSE | Mean | RMSE | Mean | RMSE |
| A | –0.0139 | 0.040 | 0.0012 | 0.020 | –0.0313 | 0.058 | –0.0193 | 0.038 |
| B | –0.0351 | 0.062 | –0.0170 | 0.026 | –0.0635 | 0.092 | –0.0322 | 0.061 |
| C | –0.0169 | 0.039 | –0.0016 | 0.015 | –0.0353 | 0.058 | –0.0213 | 0.041 |
| D | –0.0077 | 0.035 | –0.0004 | 0.020 | –0.0180 | 0.051 | –0.0080 | 0.029 |
| E | –0.0080 | 0.035 | –0.0007 | 0.021 | –0.0175 | 0.051 | –0.0092 | 0.029 |
| F | –0.0116 | 0.042 | –0.0007 | 0.020 | –0.0232 | 0.061 | –0.0167 | 0.043 |
| G | –0.0189 | 0.047 | –0.0065 | 0.019 | –0.0352 | 0.070 | –0.0209 | 0.049 |
| H | 0.0509 | 0.078 | 0.0580 | 0.079 | 0.0747 | 0.101 | 0.0088 | 0.028 |
| I | 0.0538 | 0.083 | 0.0621 | 0.085 | 0.0779 | 0.106 | 0.0090 | 0.028 |

*See text for definition of the approximations

Method E also provides accurate approximations over the wide range of tables considered here, but does less well in the lower tail of the distribution. Based on average approximation errors, William's (1976) corrected LR test (F) appears to be the third best overall method, followed by Pearson's $X^2$ (A). William's (1976) simpler corrected test (G) and the Cressie-Read (1984) statistic (C) are consistently less accurate than methods A, D, E and F. While the Gart (1966) approximations (H, I) do well in the lower tail, they are noticeably less accurate otherwise. The poor performance of the LR statistic (B) agrees with the results of other studies (e.g., Larntz (1978), McCullagh (1986)). The same general conclusions follow based on the magnitudes of the root mean square errors, although the absolute rankings of the approximate methods vary somewhat.

For each table size and exact $p$-value range, Table 2 displays the average approximation errors for the three tests with smallest absolute average errors. With four exceptions, the generalized chi-square method (D) was always one of the top three approximations. For $2 \times 5$, $3 \times 4$ and $4 \times 4$ tables with exact $p$-values less than 0.15, the average errors for approximation D were $-0.0012$, 0.0021 and 0.0033, respectively. Similarly, the average error for $2 \times 3$ tables with exact $p$-values in the range (0.15, 0.85) was $-0.0468$. In each case, these average errors are comparable to those for the third best approximation.

Table 2. Average differences ($\times 10^4$) between approximate and exact $p$-values for the three best approximations at each table size and exact $p$-value range

| Table size | All tables | Exact $p$-value range: $< 0.15$ | 0.15-0.85 | $> 0.85$ |
|---|---|---|---|---|
| $2 \times 3$ | H 68 | F −40 | H 22 | I −535 |
| | I 88 | A −41 | I 43 | H −536 |
| | D −249 | D −46 | E −445 | D −615 |
| | ($n = 384$) | ($n = 213$) | ($n = 132$) | ($n = 39$) |
| $2 \times 4$ | D −107 | A −9 | E −183 | I −231 |
| | E −110 | F −18 | D −193 | H −233 |
| | F −144 | D −20 | F −302 | D −339 |
| | ($n = 693$) | ($n = 409$) | ($n = 209$) | ($n = 75$) |
| $2 \times 5$ | D −83 | F 3 | E −170 | I −70 |
| | E −86 | A 9 | D −176 | H −71 |
| | F −106 | C −10 | F −265 | D −182 |
| | ($n = 843$) | ($n = 480$) | ($n = 243$) | ($n = 120$) |
| $3 \times 3$ | D −58 | D −5 | E −116 | H 7 |
| | E −60 | E −6 | D −119 | I 9 |
| | F −80 | A 7 | F −184 | D −121 |
| | ($n = 848$) | ($n = 454$) | ($n = 254$) | ($n = 140$) |
| $3 \times 4$ | D −53 | C −4 | F −156 | D −39 |
| | E −54 | F −9 | E −162 | E −43 |
| | F −94 | E 18 | D −166 | H 133 |
| | ($n = 1160$) | ($n = 481$) | ($n = 357$) | ($n = 322$) |
| $3 \times 5$ | D −46 | E 4 | E −113 | D −24 |
| | E −47 | D 5 | D −115 | E −26 |
| | F −102 | C −12 | F −162 | F −119 |
| | ($n = 806$) | ($n = 239$) | ($n = 271$) | ($n = 296$) |
| $4 \times 4$ | D −27 | C 1 | E −133 | E 2 |
| | E −28 | F −11 | D −135 | D 3 |
| | F −82 | E 32 | F −162 | F −74 |
| | ($n = 623$) | ($n = 165$) | ($n = 173$) | ($n = 285$) |
| $4 \times 5$ | D −127 | C −7 | E −274 | D −37 |
| | E −128 | E 23 | D −276 | E −39 |
| | F −210 | D 24 | F −337 | F −132 |
| | ($n = 175$) | ($n = 7$) | ($n = 68$) | ($n = 100$) |

## 4. Discussion

Relative to the other methods studied, the generalized chi-square approximation (D) consistently resulted in estimated $p$-values which agree more closely with the exact results. Both this method and its closest competitor, the location-shifted chi-square approximation (E), are best suited for machine computation. Although the improvement over the location-shifted chi-square method is relatively modest, there is also little additional cost in obtaining a better approximation.

Based on the commonly used criteria, the sample sizes and configurations used in this study involve "small" expected counts. Of the 155 marginal total configurations, 121 (78%) had 25% or more of the expected cell frequencies which were less than 5. Even in this sparse data setting, it was possible to obtain relatively accurate approximations to the exact $p$-value. Although the Mehta and Patel (1986) algorithm has greatly extended the capability for exact tests, the extensive computing time requirements still necessitate the use of approximate methods for tables with at least three rows and three columns.

## Acknowledgement

This research was supported by Grant CA39065 from the National Cancer Institute. The helpful comments of the referees and editor are gratefully acknowledged.

## References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions.* National Bureau of Standards, Washington, D.C.

Bernardo, J. M. (1976). Algorithm AS 103: Psi (digamma) function. *Appl. Statist.* **25**, 315–317.

Berry, K. J. and Mielke, P. W. (1988). Monte Carlo comparisons of the asymptotic chi-square and likelihood-ratio tests with the nonasymptotic chi-square test for sparse $r \times c$ tables. *Psycho. Bull.* **103**, 256–264.

Cochran, W. G. (1954). Some methods for strengthening the common $\chi^2$ tests. *Biometrics* **10**, 417–451.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser.B* **46**, 440–464.

Gart, J. J. (1966). Alternative analyses of contingency tables. *J. Roy. Statist. Soc. Ser.B* **28**, 164–179.

Haber, M. (1980). A comparison of some continuity corrections for the chi-squared test on $2 \times 2$ tables. *J. Amer. Statist. Assoc.* **75**, 510–515.

Haldane, J. B. S. (1940). The mean and variance of $\chi^2$, when used as a test of homogeneity, when expectations are small. *Biometrika* **31**, 346–355.

Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Multivariate Distributions–1.* John Wiley, New York.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *J. Amer. Statist. Assoc.* **73**, 253–263.

Lawal, H. B. and Upton, G. J. G. (1984). On the use of $X^2$ as a test of independence in contingency tables with small cell expectations. *Austral. J. Statist.* **26**, 75–85.

Lewis, T., Saunders, I. W. and Westcott, M. (1984). The moments of the Pearson chi-squared statistic and the minimum expected value in two-way tables. *Biometrika* **71**, 515–522. Correction (1989), **76**, 407.

MacLeod, A. J. (1989). Algorithm AS 245: A robust and reliable algorithm for the logarithm of the gamma function. *Appl. Statist.* **38**, 397–402.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *J. Amer. Statist. Assoc.* **81**, 104–107.

Mehta, C. R. and Patel, N. R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Amer. Statist. Assoc.* **78**, 427–434.

Mehta, C. R. and Patel, N. R. (1986). Algorithm 643—FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Trans. Math. Software* **12**, 155–161.

Mielke, P. W. and Berry, K. J. (1985). Non-asymptotic inferences based on the chi-square statistic for $r$ by $c$ contingency tables. *J. Statist. Plann. Infer.* **12**, 41–45.

Nass, C. A. G. (1959). The $\chi^2$ test for small expectations in contingency tables, with special reference to accidents and absenteeism. *Biometrika* **46**, 365–385.

Patefield, W. M. (1981). Algorithm AS 159: An efficient method of generating random $R \times C$ tables with given row and column totals. *Appl. Statist.* **30**, 91–97.

Shea, B, L. (1988). Algorithm AS 239: Chi-squared and incomplete gamma integral. *Appl. Statist.* **37**, 466–473.

Solomon, H. and Stephens, M. A. (1977). Distribution of a sum of weighted chi-square variables. *J. Amer. Statist. Assoc.* **72**, 881–885.

Solomon, H. and Stephens, M. A. (1980). Approximations to densities in geometric probability. *J. Appl. Probab.* **17**, 145–153.

Solomon, H. and Stephens, M. A. (1983). An approximation to the distribution of the sample variance. *Canad. J. Statist.* **11**, 149–154.

Stacy, E. W. (1962). A generalization of the gamma distribution. *Ann. Math. Statist.* **33**, 1187–1192.

Wichmann, B. A. and Hill, I. D. (1982). Algorithm AS 183: An efficient and portable pseudo-random number generator. *Appl. Statist.* **31**, 188–190.

Williams, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika* **63**, 33–37.

Department of Preventive Medicine, University of Iowa, Iowa City, Iowa 52242, U.S.A.

(Received December 1990; accepted May 1992)