Jour. Ind. Soc. Ag. Statistics
44 (2), 1992 : 160–170

Small Sample Comparisons for the Generalised Chi-Square Statistics

K. Aruna Rao and B.N. Nagnur*
Mangalore University, Mangalagangothri
(Received : November, 1991)

Summary

The adequacy of the Chi-square approximation to the null distribution of the members of the generalised Chi-square statistics under the simple hypothesis has been examined. Expressions for the first two moments of the statistics are obtained up to the order of $O(n^{-1})$. Two corrections have been considered to improve the Chi-square approximation. Numerical accuracy of the approximations have been checked through simulation. Bartlett adjusted statistic of a member of this family emerges as a competitor for Pearson Chi-square statistic.

Key words : Multinomial distribution; Generalised Chi-square Statistics; Null distribution; Equiprobable null hypothesis; level of significance.

Introduction

Pearson Chi-square statistic is wideely used to test the hypothesis that the observed frequencies $\mathbf{n} = (n_1, \ldots, n_k)$ in k classes are according to a multinomial distribution with specified cell probabilities $(\pi_1, \ldots, \pi_k)$. Cressie and Read [4] unified the theory of the 'Chi-square' goodness-of-fit tests by considering a class of goodness-of-fit statistics. This class includes the well known statistics, that is, Pearson Chi-square, likelihood Ratio, the Freeman-Tukey statistic, and Neyman's modified Chi-square statistic. Another class of goodness-of-fit statistics abailable in the literature, is the Generalised Chi-square statistics (Taylor [15]).

Given the sample $n = (n_1, \ldots, n_k)$ from the multinomial distribution with cell probabilities $\pi_1, \ldots, \pi_k$ where $\pi_i = \pi_i(\theta_1, \ldots, \theta_s)$ the family of generalised Chi-square statistics is defined by

*Karnatak University, Dharwad

$$D_n\left\{\mathbf{p}, \Pi(\tilde{\theta})\right\} = n\sum \frac{\left[h(p_i) - h\left\{\pi_i(\tilde{\theta})\right\}\right]^2}{\pi_i(\tilde{\theta})\left[h'\left\{\pi_i(\tilde{\theta})\right\}\right]^2} \qquad (1.1)$$

Where

$P_i = \frac{n_i}{n}$, $i = 1, \ldots, k$ and $\tilde{\theta}$ is any Best Asymptotically normal (BAN) estimator of θ, and $h(x)$ is a monotone function of x for $0<x<1$. For all practical purposes, it is convenient to consider the family when $h(x) = x^{\eta}$, $\eta \in R$ and (1.1) takes the form

$$D_n^{(\eta)}\left\{\mathbf{p}, \Pi(\tilde{\theta})\right\} = n\sum \frac{\left[p_i^{\eta} - \pi_i^{\eta}(\tilde{\theta})\right]^2}{\eta^2 \pi_i(\tilde{\theta})^{2\eta-1}} \qquad (1.2)$$

It may be noted that $D_n^{(\eta)}$ is a sum of weighted squared differences, between the observed frequencies raised to a power and the expected frequencies raised to the same power. Also, $D_n^{(\eta)}$ is approximately equal to the power divergence statistics (Read and Cressie [11]. Some particular cases of $D_n^{(\eta)}$ are given in Table 1.

Table 1. Members of $D_n^{(\eta)}\left\{\mathbf{p}, \underline{\pi}(\theta)\right\}$ for selected values of η.

η	$D_n^{(\eta)}\left\{\mathbf{p}, \underline{\pi}(\theta)\right\}$	Statistic
0	$n\sum \pi_i(\theta)\left[\log p_i - \log \pi_i(\theta)\right]^2$	
1/2	$4n\sum\left[\sqrt{p_i} - \left\{\pi_i(\theta)\right\}^{1/2}\right]^2$	Freeman-Tukey
2/3	$\frac{9n}{4}\sum \frac{\left[p_i^{2/3} - \pi_i^{2/3}(\theta)\right]^2}{\pi_i^{1/3}(\theta)}$	
1	$n\sum \frac{\left[p_i - \pi_i(\theta)\right]^2}{\pi_i(\theta)}$	Pearson Chi-square

It is worth to point out here that the well known statistics namely likelihood Ratio Statistic and Neyman's modified Chi- square statistic are not members of this family.

Using Taylor's expansion, under the null hypothesis

$H_0 : \pi_i = \pi_i(\theta)$, $i = 1, \ldots, k$, unspecified, as well as contiguous alternative hypothesis $H_1 : \pi_i = \pi_i(\theta) + \frac{C_i}{\sqrt{n}}$, $i = 1, \ldots, k$ with $\sum C_i = 0$,

$$D_n^{(\eta)}(\mathbf{p}, \underline{\pi}(\theta)) = n\sum \frac{[p_i - \pi_i(\theta)]^2}{\pi_i(\theta)} + o_p(1) \qquad (1.3)$$

Thus all the members of $D_n^{(\eta)}(\mathbf{p}, \underline{\pi}(\theta))$ have the same asymptotic distribution both under H_0 and H_1. Under the null hypothesis it is that of a central Chi-square random variable with k–s–1 degrees of freedom (d.f.) and under the contiguous alternative hypothesis that of a non-central Chi-square random variable with d.f. k–s–1 and the same non-centrality parameter. Some properties of $D_n^{(\eta)}(\mathbf{p}, \underline{\pi}(\theta))$ have been studied by Sutrick [14] and Rao [9].

For the Pearson statistic, the large sample approximation under H_0 is quite accurate for moderate and small sample sizes especially when the cells are equiprobable (Yarnold[17]). The approximation is markedly less accurate for other members of the power divergence family (Larntz [7], Read [10]). The purpose of the paper is to study the small sample behaviour of this family, when the hypothesis is simple. For this purpose, more accurate approximations to the first and second moments of the generalised Chi-square statistics are obtained so as to check the adequacy of the chi-square approximation to the asymptotic distribution of $D_n^{(\eta)}(\mathbf{p}, \underline{\pi}_0)$ where $\underline{\pi} = \underline{\pi}_0 = (\pi_{01}, \ldots, \pi_{0k})'$ and possibly to improve the approximation (Cox and Hinkley [3], Read [10]).

2. *Moments under the Simple Null Hypothesis*

Under the simple null hypothesis $\underline{\pi} = \underline{\pi}_0$ [illegible]

Correction 1 : $$D_n^{(\eta)'} = D_n^{(\eta)}\left[1+\frac{b(\eta)}{n}\right]^{-1} \quad (3.1)$$

and

Correction 2 : $$D_n^{(\eta)''} = \frac{D_n^{(\eta)} - d_n}{\sqrt{c_n}} \quad (3.2)$$

Where $b_n = \frac{a_n}{k-1}$ and a_n and b_n are the co efficients of $\frac{1}{n}$ in $E\left[D_n^{(\eta)}\right]$ and $V\left[D_n^{(\eta)}\right]$ respectively and $C_n = 1+\frac{b_n}{2\,\eta(k-1)}$ and $d_n = (k-1)(1-\sqrt{C_n})+\frac{a_n}{n}$. The expected value of $D_n^{(\eta)'}$ is (k–1) to $O(n^{-2})$. The statistic $D_n^{(\eta)''}$ has mean k–1 and variance 2(k–1) up to the order of $O(n^{-2})$. As the correction is based on the first two moments, $D_n^{(\eta)''}$ is expected to provide better approximation to the distribution.

4. *Finite Sample comparisons*

Although it is of interest to compare the exact distribution of $D_n^{(\eta)}$ in the entire range of the variate, in practical situations of testing, we are concerned with only the tail probability. In this empirical study, we restrict to the finite sample comparison of the attained level of the test for different members of $D_n^{(\eta)}$. Consider mainly the symmetric null hypothesis $\pi_i = \frac{1}{k}$, i=1, . . ., k. There are various studies indicating that equiprobable class intervals produce the most sensitive tests. They are locally most powerful and unbiased (Kendall and Stuart [6], Cohen and Sackrowitz [2], Spruill [13]).

The attained level of the test is obtained by simulation. n pseudo-random numbers are generated in the unit-interval (0,1). They are then inserted into the appropriate one of the k groups by using the limits $0, \pi_1, \pi_1+\pi_2, \ldots, \sum_{i=1}^{k-1}\pi_i, 1$.

Next the values of a_n, b_n, C_n and d_n computed. The values of $D_n^{(\eta)}$, $D_n^{(\eta)'}$ and $D_n^{(\eta)''}$ are then calculated and compared to the

appropriate tabled alpha point of the central Chi-square distribution. The sampling process is repeated 1,000 times for each of the combination of n, k and η and the estimate of the actual attained level of significance is the proportion of times that the value of $D_n^{(\eta)}$, $D_n^{(\eta)^*}$ and $D_n^{(\eta)^{**}}$ is in the rejection region. Also from 1000 samples, the values of α for the simulated distribution of $D_n^{(\eta)}$, $D_n^{(\eta)^*}$ and $D_n^{(\eta)^{**}}$ are obtained. The values of n, k, η and α considered for the study of the equiprobable null hypothesis are as follows:

n = 10, 15, 25, 50, 100

k = 3, 4, 8, 10

n = –2, –1, –0.5, 0, 0.5, 2/3, 1, 2, 5.

α = 0.10. 0.05. 0.01

For k = 4, 8 and values of η and α as mentioned above; in addition to the equiprobable null hypothesis, two other null hypothesis are also considered. They are

(i) $\left(\frac{3}{8}+(2k)^{-1}, \frac{1}{8}+(2k)^{-1}, \ldots, (2k)^{-1}\right)$

(ii) $0.9\left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)+0.1\left(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0\right)$

The values of η=2/3 is included because minimization of $D_n^{(\eta)}$ produces estimators which has got same second order efficiency as maximum likelihood estimator (Nagnur and Hegde [8]). We present the results for n=15 and k=4, 8 for the equiprobable null hypothesis in tables 3 and 4. The results for others are similar to those that are presented.

Based on the empirical study, the following observations are made:

(1) For $\eta < 0$, the attained level of significance is considerably larger than the nominal level of significance. The two statistics $D_n^{(\eta)^*}$ and $D_n^{(\eta)^{**}}$ improve the approximations but still the attained level of significance is higher than the nominal level. [See Table 3 and 4]. This is due to the fact that for $\eta < 0$, for all

Table 3. Estimates of the attained levels of significance under symmetric null hypothesis for k=4, n=15 and different values of η along with the α% point of the simulated distribution of $D_n^{(\eta)}$, $D_n^{(\eta)^*}$ and $D_n^{(\eta)^{**}}$.

Value of η	α = 0.01			α = 0.10			α% point of the simulated distribution					
							α = 0.01			α = 0.10		
	D_1	D_2	D_3	D_1	D2	D3	D_1	D2	D3	D1	D2	D3
–2	0.404	0.319	0.319	0.648	0.319	0.319	∞	∞	∞	166.55	38.28	40.08
–1	0.319	0.319	0.132	0.404	0.319	0.319	∞	∞	∞	32.28	12.74	11.66
–0.5	0.319	0.063	0.062	0.337	0.319	0.319	∞	∞	∞	16.76	8.88	7.96
0	0.063	0.052	0.052	0.320	0.146	0.132	∞	∞	∞	9.80	6.92	6.36
0.5	0.055	0.052	0.052	0.147	0.101	0.087	19.85	17.71	17.02	7.02	6.27	6.11
2/3	0.026	0.026	0.026	0.128	0.101	0.101	14.05	13.24	13.21	6.86	6.47	6.46
1	0.015	0.015	0.015	0.105	0.105	0.105	11.40	11.40	11.69	6.60	6.60	6.73
2	0.081	0.023	0.021	0.233	0.092	0.092	22.65	17.65	14.37	7.64	5.96	5.29
5	0.570	0.125	0.233	0.570	0.233	0.233	927.67	146.47	169.64	72.23	11.41	12.76
α% point of the central Chi-square distribution for 3 d.f.							α = 0.01 11.34			α = 0.10 6.25		

$D_1 = D_n^{(\eta)}$, $D_2 = D_n^{(\eta)^*}$ $D_3 = D_n^{(\eta)^{**}}$.

Table 4. Estimates of the attained levels of significance under symmetric null hypothesis for k=8, n=15 and different values of η along with the α% point of the simulated distribution of $D_n^{(\eta)}$, $D_n^{(\eta)^*}$ and $D_n^{(\eta)^{**}}$.

Value of η	α = 0.01			α = 0.10			α% point of the simulated distribution					
							α = 0.01			α = 0.10		
	D_1	D_2	D_3	D_1	D2	D3	D_1	D2	D3	D1	D2	D3
–2	0.764	0.764	0.764	0.864	0.764	0.764	∞	∞	∞	∞	∞	∞
–1	0.764	0.764	0.764	0.764	0.764	0.764	∞	∞	∞	∞	∞	∞
–0.5	0.764	0.764	0.764	0.764	0.764	0.764	∞	∞	∞	∞	∞	∞
0	0.764	0.764	0.764	0.764	0.764	0.764	∞	∞	∞	∞	∞	∞
0.5	0.213	0.057	0.049	0.415	0.314	0.308	29.32	23.69	22.47	20.70	16.73	16.02
2/3	0.029	0.010	0.010	0.223	0.156	0.156	20.78	18.67	18.60	14.86	13.35	13.32
1	0.012	0.012	0.012	0.014	0.014	0.014	18.60	18.60	19.01	12.20	12.20	12.38
2	0.256	0.066	0.062	0.345	0.144	0.114	79.91	45.67	34.78	24.85	14.20	12.17
5	0.853	0.345	0.345	0.973	0.345	0.714	39339.97	2850.72	4552.60	1483.83	107.52	167.68
α% point of the central Chi-square distribution for 7 d.f.							α = 0.01 18.48			α = 0.10 12.02		

$D_1 = D_n^{(\eta)}$, $D_2 = D_n^{(\eta)^*}$ $D_3 = D_n^{(\eta)^{**}}$.

partitions with zero observations in one or more cells, $D_n^{(\eta)}$ is infinite, and the exact critical region will always contain all such partitions. Since $D_n^{(\eta)'}$ and $D_n^{(\eta)''}$ are infinite whenever $D_n^{(\eta)}$ is infinite, the approximations are not suitable. Therefore, for $\eta < 0$ and for moderate sample size, the use of $D_n^{(\eta)}$ is not appropriate.

(2) For values of η somewhere between 0.5 to 1.5, the exact level of significance is very close to nominal level of significance. For these values of η the Chi-square approximation is adequate to the asymptotic distribution of $D_n^{(\eta)}$.

(3) The Chi-square approximation is quite satisfactory for the Pearson Chi-square statistic. Similar conclusion is by Good *et. al.* [5], Chapman [1], Larntz [7], and Read [10].

(4) The two approximations produce significance levels that are quite close to the nominal levels in the range (0.5 to 2.0). Among the two approximations, the second approximation produce significance levels much closer to the nominal level than the first one. Thus Freeman-Tukey Chi-square statistic can be used with the second approximation.

(5) For large positive values of η, the Chi-square approximation is not satisfactory as indicated by the attained level of significance and the simulated estimates of the α% values of the test statistics. The two approximations fail to improve the situation.
(See for η=5 in Tables 3 and 4).

(6) η=2/3 seems to be a competitor for Pearson Chi-square statistic. For values fo k from 3 to 10 and n < 50, the distribution of $D_n^{(\eta)'}$ can be well approximated by a Chi-square distribution and the second correction does not improve the situation. For values of n≥100, the distribution of $D_n^{(\eta)}$ can well be approximated by Chi-square distribution and no correction is required.

REFERENCES

[1] Chapman, J.W., 1976. A comparsion of the x^2,-2 log R, and the multinomal probability criteria for significance testing when expected frequencies are small. *J. Amer.Statist. Assoc.*, **71**, 854-863.

[2] Cohen, A. and Sackrowitz, H.B., 1975. Unbiasedness of the Chi-squaree, likelihood ratio, and other goodness-of-fit tests for the equal cell case. *Ann. Statist.*, **3**, 959=964.

[3] Cox, D.R. and Hinkley, D.V., 1974. Theoretical Statistics. London; Chapman and Hall.

[4] Cressie, N and Read, T.R.C, 1984. Multinomial goodness-of-fit tests. *J. Royal Statist. Soc. B*, **46**, 440-464.

[5] Good, I.J., Gover, T.N. and Mitchell, C.J., 1970. Exact distribution for χ^2 and for the likelihood ratio statistic for the equiprobable multinomial distribution. *J. Amer. Statist. Assoc.*, **65**, 267-283.

[6] Kendall, M. and Stuart, A., 1973. The Advanced Theory of Statistics, Vol.2. London: Charles, Griffin and Co. Ltd.

[7] Larntz, K, 1978. Small sample comparisons of exact levels of Chi-square goodness-of-fit statistics. *J. Amer. Statist. Assoc.*, **73**, 253-263.

[8] Nagnur, B.N. and Hegde, M.S., 1988. On second order efficiency of minimum discripancy and minimum distance estimator for the multinomial distribution. *Cal. Statist. Assoc. Bull.*, **37**, 27-28.

[9] Rao, K.A., 1989. Some results on generalised Chi-square type statistics. Unpublised Ph. D. Thesis submitted to Karanataka University, Dharwad.

[10] Read, T.R.C., 1984. Small sample comparisons for the power divergence goodness-of-fit statistics. *J. Amer. Statist. Assoc.*, **79**, 929-936.

[11] Read, T.R.C. and Cressie, N., 1988. Goodness-of-fit Statistics for discrete multivariate data. New York: Springer-Verlag.

[12] Smith, P.J., Rae D.S. Manderscheid, R.W. and Silbergeld, S, 1981. Approximating the moments and distribution of likelihood ratio statistic for multinomial goodness-of-fit. *J. Amer. Statist. Assoc.*, **76**, 737-740.

[13] Spruill, M.C., 1977. Equally likely intervals in the Chi-square test. *Sankhya, A*, **39**, 299-302.

[14] Sutrick, K.H., 1986. Asymptotic power comparisions of the Chi-square and likelihood ratio tests. *Ann. Inst. Statist. Math.*, **38**, 503-511.

[15] Taylor, W.F., 1953. Distance functions and regular best asymptotically normal estimators. *Ann. Math. Statist.*, **24**, 85-92.

[16] Willianms, D.A., 1976. Improved likelihood ratio tests for complete contingency tables. *Biometrika*, **63**, 33-37.

[17] Yarnold, J.K., 1972. Asymptotic approximation for the Probability that a sum of lattice random rectors lies in a convex set. *Ann. Math. Statist.*, **43**, 1566-1580.