SMALL SAMPLE COMPARISONS OF CHI-SQUARE STATISTICS

Kinley Larntz, University of Minnesota

Several statistics are commonly used to judge the goodness-of-fit for counted data models. In this paper, two of these statistics will be compared with respect to their samll sample properties under the null hypothesis. The usual chi-square statistic (Pearson statistic) is defined by

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed - Expected})^2}{\text{Expected}} \qquad (1)$$

A suggested alternative statistic that has some asymptotically optimal properties is the likelihood-ratio statistic

$$G^2 = 2 \sum_{\text{all cells}} \text{Observed} \log_e\left(\frac{\text{Observed}}{\text{Expected}}\right) \qquad (2)$$

Many statisticians prefer the use of one or the other of these statistics, although among everyday users the Pearson statistic is far more popular. Also, some statisticians follow the practice of reporting both statistics (see for example, Goodman [1973]), but little guidance is available concerning the occurrence of large discrepancies between the two statistics.

THE MODEL

Comparisons between the statistics are made for a particular parametric model that arises naturally in a group helping situation. Individuals or groups are given the opportunity to help another individual in distress. The degree of help is graded I, II, or III: I for not helping, III for actively helping, and II for an intermediate action. Further details can be found in Fienberg and Larntz [1971] or Staub [1970]. Similar models are also used in component testing problems (see Easterling and Prairie [1971]).

Data were gathered for individuals and groups of size two. Let p_1, p_2, and p_3 be the probabilities of observing an individual with help graded I, II, and III, respectively. Then if the individuals in a group act independently and if only the higher grade of help is scored, p_1^2, $p_2^2 + 2p_1p_2$, and $p_3^2 + 2p_1p_3 + 2p_2p_3$ are the respective probabilities of observing I, II, and III for groups of size two.

Suppose N_1 individuals and N_2 groups are tested. Under the above assumptions, (n_{11}, n_{21}, n_{31}) follows a multinomial distribution with probability vector (p_1, p_2, p_3), and (n_{12}, n_{22}, n_{32}) follows a multinomial distribution with probability vector (g_1, g_2, g_3) where

$$\begin{aligned} g_1 &= p_1^2 \\ g_2 &= p_2^2 + 2p_1p_2 \\ g_3 &= p_3^2 + 2p_1p_3 + 2p_2p_3 \ . \end{aligned} \qquad (3)$$

For this case the unique maximum likelihood estimates for (p_1, p_2, p_3) can be written down directly as

$$\begin{aligned} p_1 &= (-n_{31} + \sqrt{n_{31}^2 + 4ac}\,)/2a \\ p_2 &= r\hat{p}_1 \\ p_3 &= 1 - (1+r)\hat{p}_1 \end{aligned} \qquad (4)$$

where

$$r = \frac{n_{21} - 2n_{11} - 4n_{12} + \sqrt{s}}{2(n_{11} + 2n_{12})}$$

$$s = (2n_{11} + 4n_{12} - n_{21})^2 + 8(n_{21} + n_{22})(n_{11} + 2n_{12}) \qquad (5)$$

$$\begin{aligned} a = (1 + r)[(n_{11} &+ 2n_{12})(1 + r) \\ &+ (n_{31} + 2n_{32}) + 2n_{22}(1 + r)/(2 + r)] \ , \end{aligned} \qquad (6)$$

and

$$c = n_{11} + 2n_{12} + 2n_{22}/(2 + r) \ .$$

The selection of this model for making comparisons between the likelihood-ratio and Pearson chi-squares provides several advantages:

(a) The model depends on two parameters, p_1 and p_2, and thus the goodness-of-fit test for the null hypothesis involves the estimation of these parameters. Comparisons are therefore made for a composite null hypothesis.

(b) Since the maximum likelihood estimates can be written down in closed form, iteration is not necessary for finding the estimates. This is important when considering the feasibility of doing large amounts of computation.

(c) Examining (3), note that the probability of Help Grade I for groups is p_1^2. When p_1 is small, p_1^2 is quite small. Thus the selection of this model allows for comparisons of very skew multinomials, which means comparisons can be made for small as well as moderate minimum cell expectations. Previous studies (Cochran [1952], Yarnold [1970]) have indicated that, for small expected values, the Pearson statistic does not follow the chi-square distribution well, while some suggestion has been indicated (cf. Bliss [1967]) that the likelihood-ratio statistic would be better in such situations.

SMALL SAMPLE PROPOERTIES UNDER THE NULL HYPOTHESIS

Under the null hypothesis the goodness-of-fit statistics, X^2 and G^2, have asymptotic chi-square distributions with 2 degrees of freedom. However, for small samples the chi-square approximation in many cases does not agree well with the actual distribution. Several studies (Cochran [195], Fisher [1958], Roscoe and Byars [1971], Yarnold [1970]) have given conflicting points of view as to at what point the approximation is "reasonable" for the Pearson chi-square statistic. Standard rules specify that the minimum cell expectation should be 5, with possibly a few smaller. The emphasis here will not be on finding such a rule, but in comparing the likelihood-ratio and Pearson statistics with regard to the approximation. In other words, we ask for small samples, which of the two statistics is better approximated by the asymptotic chi-square distribution?

The initial task in this study of the small sample properties is to determine the distribution of the statistics G^2 and X^2 when the null hypothesis holds. Several methods are available to handle such a problem. The principal method used here was that of enumeration. The number of possible outcomes of two trinomials with sample sizes N_I and N_G is given by

$$\text{Outcomes} = \binom{N_I + 2}{2} \binom{N_G + 2}{2} \qquad (8)$$

For $N_I = N_G = 8$, the number of possible outcomes is 2025. Thus, for a given value of (p_1,p_2,p_3), N_I, and N_G, the distribution of G^2 and X^2 were determined by computer.

One question that arises in the use of this method is how to deal with zero cell counts and zero expected values. The maximum likelihood estimates were extended by continuity to provided well-defined procedures. In the same manner, when a cell had zero expected value, it contributed zero to the chi-square statistic.

Figure A gives a contour plot of the mean of G^2 for $N_I = N_G = 8$. Barycentric coordinates were chosen to represent the 3 probabilities. Each corner of the triangle represents one of the probability vectors (1, 0, 0), (0, 1, 0), and (0, 0, 1), while a general point in the triangle corresponds to the probability vector (p_1,p_2,p_3). Figure B gives a similar plot for X^2. The asymptotic mean for both statistics is, of course, 2.0. The mean of G^2 overshoots that value for a large set of (p_1,p_2,p_3). The peak value is approximately 2.51. In viewing Figure B, it can be seen that the mean of the Pearson statistic is a smoother function of (p_1,p_2,p_3) than the mean of G^2. For a large set of (p_1,p_2,p_3), the mean of X^2 is close to 2.0. The peak value is approximately 2.12. Thus, considering the mean only, the Pearson statistic appears better.

Another method of comparison is to check the agreement of the actual small sample percentage points with the corresponding asymptotic values. Results analogous to the case of the mean hold here. Namely, the likelihood-ratio tends to overshoot the corresponding large sample value while the Pearson statistic tends to be closer to the asymptotic value for a large range of (p_1,p_2,p_3).

Several questions concerning the likelihood-ratio statistic arise from these results. First, is it still possible that the optimality properties of the likelihood-ratio statistic carry over in spite of the poor characteristics of its null distribution? This will be the subject of another paper comparing the powers of the statistics. Second, can the statistics be easily adjusted to remove some of its poor behavior? And third, exactly how does the likelihood-ratio behave as the "small" sample size increases? An attempt at answering the last question will be given below.

The question of adjusting the likelihood-ratio statistic poses large difficulties. A simple-minded correction for the mean yielded mixed results, partly due to a problem of overcorrection with respect to size. Other corrections involving more moments or quantiles may be possible, but practical use would require a simple multiplicative or additive correction, such as those given in Bartlett [1947] and Box [1949].

PROPERTIES OF THE LIKELIHOOD_RATIO CHI-SQUARE STATISTIC

The asymptotic distribution of G^2 for the model considered here is that of a chi-square variate with 2 degrees of freedom. Figure C gives a graph of the mean values of G^2 for (.6, .2, .2). Figure C is indicative of what happens to the mean as the sample size changes. It begins below its asymptotic value, rises to a peak, and descends to the correct value. The true sizes follow a similar pattern. Because of the discreteness of the distribution, the rise and descent may be slightly rocky, but the general pattern remains the same.

The sample size at which the peak is reached varies considerably depending on the probability vector (p_1, p_2, p_3). Some evidence has been given that the minimum cell expectation governs the closeness of the small sample distribution to asymptotic theory for several chi-square problems (see for example, Cochran [1952], Cramer [1946], Odoroff [1970], Yarnold [1970]). In the problem at hand, small expected values are found for small values of p_1 (since the first cell for pairs has probability p_1^2) and for very small values of p_2 and p_3. Evidence from this study indicates that the larger minimum cell expectation cases are closer to the behavior predicted by the asymptotic theory.

POWER CHARACTERISTICS

In order to compare the power functions of the test statistics, it was necessary to adjust for the level of significance differences between X^2 and G^2. Let the adjusted level be defined as

$$\text{Adj. level } (z) = \sup_{(p_1, p_2, p_3)} P(\text{statistic} \geq z) \quad (9)$$

Thus for a given alternative $(p_1, p_2, p_3; g_1, g_2, g_3)$, the power of X^2 or G^2 can be computed as a function of the adjusted level.

Many methods can be used to compare the power functions of the statistics. One interesting comparison can be made by means of the median significance level (Joiner [1969]. For a particular alternative, let

$$\text{M.S.L.} = \text{Adjusted Level } (z_M) \quad (10)$$

where z_M is the median of the statistic under the alternative distribution. In comparing two statistics, the one with the lower median significance level would be considered better. For this and several other methods of comparison, it was found that the Pearson statistic was more powerful than the likelihood-ratio for most alternatives.

The stochastic limit ratio (defined below) gives a method of determining the alternatives where the likelihood-ratio was more powerful than the Pearson. For an alternative $p_a = (p_1, p_2, p_3; g_1, g_2, g_3)$, let $G^2(p_a)$ be the value of the likelihood-ratio chi-square calculated using $n_{11} = p_1$, $n_{12} = p_2$, $n_{13} = p_3$, $n_{21} = g_1$, $n_{22} = g_2$, and $n_{23} = g_3$. Similarly, let $X^2(p_a)$ be defined. Then for an alternative define the stochastic limit ratio as

$$\text{S.L.R.}(p_a) = \frac{G^2(p_a)}{X^2(p_a)} \quad (11)$$

When S.L.R. is large (71.05), the likelihood-ratio statistic appears more powerful based on small samples; whereas, with S.L.R. >1, the Pearson statistic is better. The differentiation in the middle range (1 - 1.05) is not clear with some cases going to likelihhod-ratio and some to Pearson. However, for large areas of the alternative parameter space, S.L.R. < 1.

CONCLUSIONS

For one special model with a composite null hypothesis, the samll sample distributions of two chi-square statistics were examined. Using as criterion the closeness of small sample dsitribution to the asymptotic chi-square approximation, the Pearson chi-square statistic is by far the more desirable. The likelihood-ratio statistic has an expected value in excess of the nominal and yields far too many rejections under the null distribution.

It was also noted that the expected value and level of significance for the likelihood-ratio statistic displayed a consistent regularity in which the mean and level rose to a peak and then declined toward the asymptotic value as the sample size increased.

Power comparisons also indicated the desirability of using the Pearson statistic over the likelihood-ratio -- at least when proper adjustments are made for the differing levels of significance.

REFERENCES

Bartlett, M. S., "Multivariate Analysis," Journal of the Royal Statistical Society, Series B, 9 (1947), 176-197.

Bliss, C. I., Statistics in Biology, Volume I, New York: McGraw-Hill Book Co., 1967.

Box, G. E. P., "A General Distribution Theory for a Class of Likelihood Criteria," Biometrika, 36 (1949), 317-346.

Cochran, W. G., "The χ^2 Test of Goodness of Fit," Annals of Mathematical Statistics, 23 (1952), 315-346.

Cramer, H., Mathematical Methods of Statistics, Princeton, N. J.: Princeton University Press, 1945.

Easterling, R. G. and Prairie R. R., "Combining Component and System Information," Technometrics, 13 (1971), 271-280.

Fienberg, S. E. and Larntz, K., "Some Models for Individual-Group Comparisons and Group Behavior," Psychometrika, 36 (1971), 349-367.

Fisher, R. A., Statistical Methods for Research Workers, 13th ed., New York: Hafner Publishing Co., 1958.

Goodman, L. A., "Guided and Unguided Methods for Selecting Models for a Set of T Multidimensional Contingency Tables," Journal of the American Statistical Association, 68 (1973), 165-175.

Joiner, B. L., "The Median Significance Level and Other Small Sample Measures of Test Efficacy," Journal of the American Statistical Association, 64 (1969), 971-985.

Odoroff, C. L., "A Comparison of Minimum Logit Chi-Square Estimation and Maximum Likelihood Estimation in 2 x 2 x 2 and 3 x 2 x 2 Contingency Tables: Tests for Interactions," Journal of the American Statistical Association, 65 (1970) 1617-1631.

Roscoe, J. T. and Byars, J. A., "Sample Size Restraints Commonly Imposed on the Use of the Chi-Square Statistic," Journal of the American Statistical Association, 66 (1971), 755-759.

Staub, E., "A Child in Distress: The Influence of Age and Number of Witnesses on Children's Attempts to Help," Journal of Personality and Social Psychology, 14 (1970), 130-140.

Yarnold, J. K., "The Minimum Expectation in X^2 Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution," Journal of the American Statistical Association, 65 (1970), 864-886.

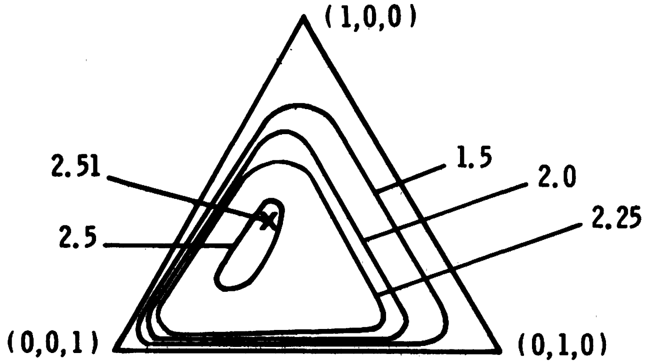

Fig. A

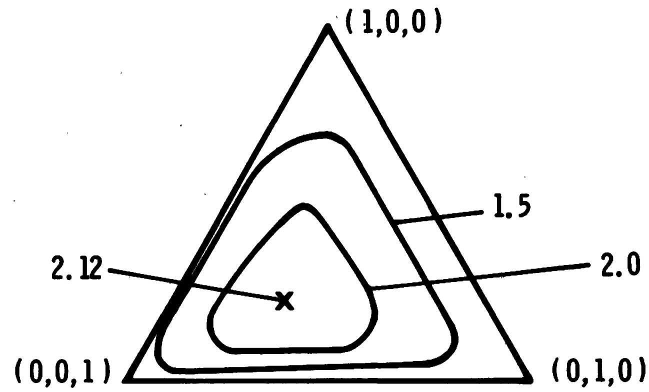

Fig. B

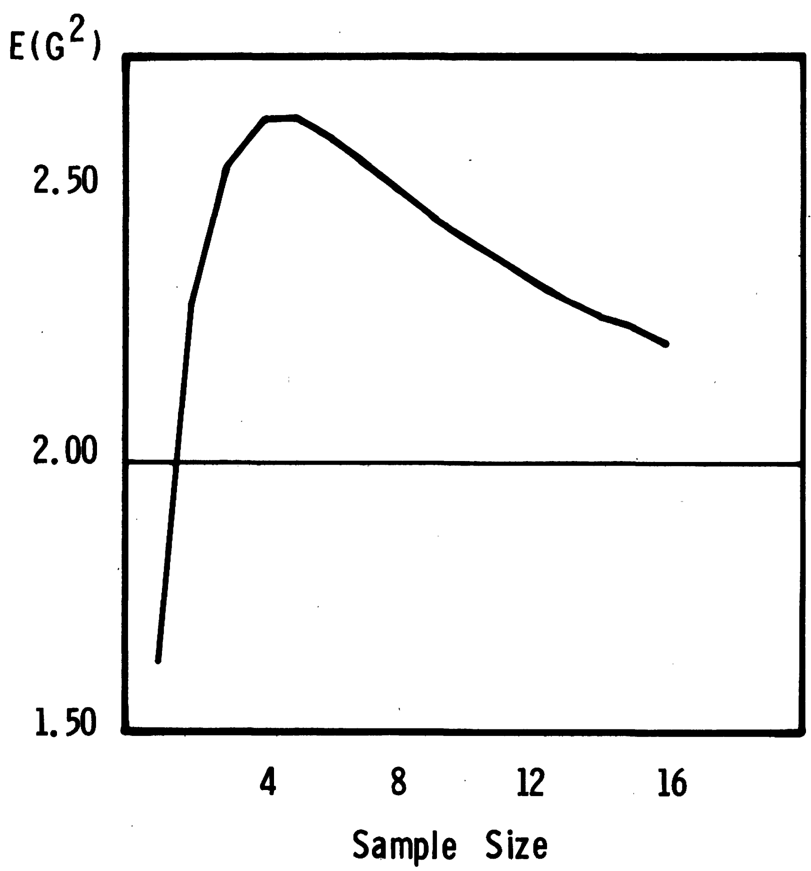

Fig. C