Psychological Bulletin
1983, Vol. 94, No. 1, 166-176

The Use and Misuse of Chi-Square: Lewis and Burke Revisited

Kevin L. Delucchi
Department of Education University of California, Berkeley

The proper use of Karl Pearson's chi-square for the analysis of contingency tables is reviewed. The 1949 article by Lewis and Burke, who cited nine sources of error in the use of chi-square, is updated. Research on the application of the chi-square statistic is examined and supplementary and alternative approaches are discussed. Emphasis is placed on techniques that are of use to the practicing researcher who often deals with qualitative ordered and unordered data.

In 1949 the landmark article by Lewis and Burke entitled "The Use and Misuse of the Chi-Square Test" appeared in the *Psychological Bulletin.* The purpose of the article was to counteract the improper use of this statistic by researchers in the behavioral sciences. The paper addressed nine major sources of error, cited examples from the literature to illustrate these points, and caused a stir among practicing researchers. The Lewis and Burke paper was followed by several responses (Edwards, 1950; Pastore, 1950; Peters, 1950) and a rejoinder by Lewis and Burke (1950).

Since then, a great deal of research has been conducted on the chi-square procedure and several methods have been developed to handle some of the problems cited by Lewis and Burke. This article is a review of that literature. It is an attempt to address the problems listed by Lewis and Burke in light of current knowledge and to form recommendations regarding the use and misuse of the chi-square test.

The Use and Misuse of Chi-Square

Lewis and Burke centered their 1949 article around nine principle sources of error they found in their review of published research. Those nine sources are:

1. Lack of independence among single events or measures
2. Small theoretical frequencies
3. Neglect of frequencies of non-occurrence
4. Failure to equalize the sum of the observed frequencies and the sum of the theoretical frequencies
5. Indeterminant theoretical frequencies
6. Incorrect or questionable categorizing
7. Use of nonfrequency data
8. Incorrect determination of the number of degrees of freedom
9. Incorrect computations.

The errors they cited still occasionally appear, and, as a consequence, their article should be required reading by anyone who intends to use the chi-square statistic. There is little to add to Lewis and Burke's discussion with respect to most of these issues. The major exception centers on the question of the minimal size of expected cell frequencies. A number of statisticians have addressed this point in the years since publication of the Lewis and Burke paper, and the following section will summarize their work.

Small Theoretical Frequencies

Lewis and Burke (1949) called the use of expected frequencies that are too small the most common weakness in the use of chi-square (p. 460). They took the position that expected values of 5 were probably too low and stated a preference for a minimum expected value of 10, with 5 as the absolute lowest limit. Lewis and Burke cited as examples two published studies that used chi-square tests with expected values below 10.

This article is based on a paper presented to the Annual Convention of the American Educational Research Association held in Los Angeles in April of 1981. The author gratefully acknowledges the encouragement and helpful comments of Leonard A. Marascuilo, Patricia Busk, and Jerome L. Myers on earlier drafts.

Requests for reprints should be sent to Kevin L. Delucchi, Department of Education, Tolman Hall, University of California, Berkeley, California 94704.

It appears today that their position, a popular one among researchers, may be overly conservative.

This problem has been examined from the perspectives of two different applications. In testing goodness-of-fit hypotheses, the categories are chosen arbitrarily, permitting control over the size of the expected values by choice of the category sizes. In contrast, the categories of contingency tables used for testing association hypotheses are relatively limited, and one is forced to increase the expected values by increasing the sample size and/or collapsing rows and/or columns. Research taken from the perspective of this latter case will be considered first.

Tests of association hypotheses in contingency tables. Recommendations with respect to minimum expected cell frequencies have included recommended minimum values of 1 (Jeffreys, 1961; Kempthorne, 1966; Slakter, 1965), 5 (Fisher, 1938), 10 (Cramer, 1946), and 20 (Kendall, 1952). Wise (1963) recommended small (i.e., less than five) but equal expected frequencies over the case where a few expected values are small and the remaining frequencies are well above most criteria.

Cochran (1952) suggested that chi-square may be applied if no more than 20% of the cells have expected values between one and five. Good, Grover, and Mitchell (1970) concluded that if the expected values are equal, they may be as low as 1/3 (p. 275). This apparent robust nature of the procedure is also supported by Lewontin and Felsenstein (1965), who used Monte Carlo methods to examine $2 \times N$ tables with fixed marginals. With small expected values in each cell and degrees of freedom (*df*) greater than 5, they concluded that the test tends to be conservative. Even the occurrences of expected values below one generally do not invalidate the procedure. Bradely, Bradely, McGrath, and Cutcomb (1979) conducted a series of sampling experiments to examine the Type I error rates of chi-square in the presence of small expected values in tables as large as 4×4. Their results offer strong support for the robustness of the statistic in meeting preassigned Type I error rates. Additional support comes from Camilli and Hopkins' (1978) study of chi-square in 2×2 tables; they found expected values as low as one or two were acceptable when the total sample size was greater than 20.

Testing goodness-of-fit hypotheses. In testing goodness-of-fit hypotheses, Kendall and Stuart (1969), following suggestions by Mann and Wald (1942) and Gumbel (1943), recommended that one choose the boundaries of categories so that each has an expected frequency equal to the reciprocal of the number of categories. They prefer a minimum value of five categories. Slakter (1965, 1966), Good (1961), and Wise (1963) have all found that in testing goodness-of-fit, expected values may be as low as one or two for an alpha of .05 when the expected values are equal. For unequal expected values or an alpha of .01, the expected frequencies should be at least four or larger.

In an article based on his dissertation, Yarnold (1970) numerically examined the accuracy of the approximation of the chi-square goodness-of-fit statistic. He proposed that "If the number of classes, s, is three or more, and if r denotes the number of expectations less than five, then the minimum expectation may be as small as $5r/s$" (p. 865). He concluded that "the upper one and five percentage points of the χ^2 approximation can be used with much smaller expectations than previously considered possible" (p. 882).

After considering earlier work, Roscoe and Byars (1971) recommended that for the goodness-of-fit statistic with more than one degree of freedom, one should be concerned with the *average* expected value. In the case of equal expected cell frequencies, they suggested an average value of 2 or more for an alpha equal to .05 and 4 or more for an alpha equal to .01. In the nonuniform case, they recommend average expected values of 6 and 10, respectively. They urged the use of this average expected value rule in the test for independence as well, even when the sample sizes are not equal.

As Horn (1977) has noted, this average expected value rule is in agreement with Slakter's (1965, 1966) suggestion that what may be most important is the average of the expected frequencies. Horn also noted that this subsumes Cochran's rule that 20% of

the expected frequencies should be greater than one.

Summarizing this work on minimum expected values for both association and goodness-of-fit hypotheses, it seems that, as a general rule, the chi-square statistic may be properly used in cases where the expected values are much lower than previously considered permissible. In the presence of small expected values, the statistic is quite robust with respect to controlling Type I error rate, especially under the following conditions: (a) the total N is at least five times the number of cells, (b) the average expected value is five or more, (c) the expected values tend toward homogeneity, and (d) the distribution of the margins is not skewed. Additional references on this matter that may be of interest to readers can be found in Hutchinson (1979).

For most applications, Cochran's rule, which states that all expected values be greater than one and not more than 20% be less than five, offers a fair balance between practicality and precision. An alternative to consider, especially in the case of small tables, is the computation of an exact test. For reference to the exact test the reader is referred to Agresti and Wackerly (1977) and Baker (1977). Berkson (1978) and Kempthorne (1979) present a debate over the use of the exact test in 2×2 tables that the interested researcher should examine.

Power considerations. An important point that is easily overlooked regards the effect of small expected values on the power of the chi-square test. Overall (1980) has examined the effect of low expected frequencies in one row or column of a 2×2 design on the power of the chi-square statistic. This most often results from the analysis of infrequently occurring events. Setting $(1 - \alpha) = .70$ as a minimally acceptable level, Overall concluded that when expected values are quite low, the power of the chi-square test drops to a level that produces a statistic that, in his view, is almost useless.

Correction for Continuity

Lewis and Burke presented the Yates correction for continuity, noting that it is justified only in the case of a 2×2 table. Questions have arisen regarding the appropriateness of the use of a correction for continuity.

Since categorical variables are discrete and the chi-square distribution is continuous, a correction to improve the approximation can be made. The most well known was proposed by Yates (1934) and is made by adding or subtracting ½ to each observed frequency so as to move the observed value closer to the expected value. Thus it becomes more difficult to reject the hypothesis being tested. Symbolically, the corrected chi-square is written as

$$x_c^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{[(x_{ij} + \tfrac{1}{2}) - \epsilon(x_{ij})]^2}{\epsilon(x_{ij})}. \quad (1)$$

The analytical derivation of the correction expressed in Equation 1 is given by Cox (1970).

The disagreement over the use of this correction is based not on its theoretical grounding but on its applicability. Plackett (1964), confirming empirical results of Pearson (1947), argued that the correction is inappropriate if the data come from independent binomial samples. Grizzle (1967) extended Plackett's results to the general case and concluded that the correction is so conservative it is rendered useless for practical purposes.

The consensus of several investigators (Camilli & Hopkins, 1978; Conover, 1974a, 1974b; Mantel, 1974; Mantel & Greenhouse, 1968; Miettinen, 1974; Starmer, Grizzle, & Sen, 1974) seems to be that the correction for continuity becomes overly conservative when either or both of the marginals in a table are random. As this is often the case in social science research, it would appear that the use of the correction should not be given the blanket recommendation that often accompanies it. If strong conservatism is desired and/or the marginal totals in the contingency table being analyzed are fixed values, then the Yates correction should be applied. In all other cases, however, one must be cautious in its use because the correction for continuity will produce very conservative probability estimates.

Misclassification

One issue of categorical analysis that has received little attention in social science research is the effect of misclassification on the power and Type I error rate of the chi-square

test. The majority of relevant literature is found in biostatistics (e.g., Mote & Anderson, 1965). One notable exception to this is an article by Katz and McSweeney (1979), who discuss the effects of classification error on the significance level and power of the test for equality of proportions.

They develop and discuss a correction procedure based on estimates of the probability of false negatives and false positives. As Katz and McSweeney note, the detrimental effects of misclassification can be marked, including a loss in power. This is especially true when one of the proportions being tested is small and the probability of misclassification is not equivalent across groups. Any researcher who suspects the presence of misclassified data points should consult the Katz and McSweeney (1979) article and the references they cite. The key to using their procedure, and its major drawback, is the need for estimates of the rate of misclassification that may often be unobtainable.

Supplementary and Alternative Procedures

Log- and Logit-Linear Models

A major drawback to the use of Pearson's chi-square lies in the fact that it does not readily extend to the analysis of multidimensional contingency tables. An alternative approach, which readily extends to higher order tables, is the use of log-linear and logit-linear models. Many articles and texts are now available for these procedures, including the works of Bishop, Fienberg, and Holland (1975), Goodman (1978), Haberman (1978), and Fienberg (1980). These procedures are implemented through several packaged computer programs including SAS FUNCAT, Goodman's ECTA, BMDP 4F, Nelder's GLIM, and Bock's MULTIQUAL, which are familiar to many researchers.

Although most applicable for analyzing multidimensional tables, it should be pointed out that these models can be used on two-dimensional tables as well. It is not difficult to argue that log-linear models will eventually supersede the use of Pearson's chi-square in the future because of their similarity to analysis of variance (ANOVA) procedures and their extension to higher order tables. Discussion of this methodology, however, is beyond both the scope and focus of this article. The following topics of partitioning, the use of G^2, and the analysis of ordered categories are often cited as particular instances where log- and logit-linear models may be preferred.

Log-Likelihood Ratio

An alternative procedure to calculating Pearson's chi-square to test a hypothesis concerning a multinomial is the use of the likelihood ratio statistic. It is a maximum likelihood estimate labeled G^2 and defined as

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij} \log_e \left(\frac{x_{ij}}{e_{ij}} \right). \qquad (2)$$

In their text on discrete multivariant analysis, Bishop, Fienberg, and Holland (1975) used log-linear models, as opposed to additive models, for contingency table analysis. As a summary statistic, they stated a preference for maximum likelihood estimators (MLEs) on theoretical grounds. Additionally, practical reasons for the use of this procedure were given:

1. Ease of computation for linear models.
2. MLEs satisfy certain marginal constraints they call intuitive.
3. "The method of maximum likelihood can be applied directly to multinomial data with several observed cell values of zero, and almost always produces non-zero estimates for such cells (an extremely valuable property in small samples)" (p. 58).

They further state,

> MLEs necessarily give minimum values of G^2, it is appropriate to use G^2 as a summary statistic . . . although the reader will observe that, in those samples where we compute both X^2 and G^2, the difference in numerical value of the two is seldom large enough to be of practical importance. (p. 126)

There are cases where the likelihood-ratio statistic may be preferred over chi-square. Such may occur when some expected values are quite small or when the contingency table contains a structural zero.

Several investigators have compared X^2 and G^2 in a variety of research situations. Chapman (1976) provides an overview of much of this research, including the work of Neyman and Pearson (1931), Cochran (1936),

Fisher (1950), Good, Grover, and Mitchell (1970), and West and Kempthorne (1972). From these comparisons, neither of the two procedures emerges a clear favorite. When one method is better in some respect than the other, it seems to result from a particular configuration of sample size, number of categories, expected values, and the alternative hypothesis. An exception to the general equivalence of these two statistics can be found in the literature on partitioning of contingency tables, which is discussed next.

Partitioning

At about the same time that Lewis and Burke were writing, the first extensive work on the partitioning of an $I \times J$ contingency table into components was being conducted by Lancaster (1949, 1950, 1951), who demonstrated that a general term of a multinomial can be reduced to a series of binomial terms, each with one degree of freedom. Irwin (1949) presented a formula for exact partitioning, which was simplified algebraically by Kimbal (1954) for easier computation. Ten years later Kastenbaum (1960) generalized the partitioning procedure to handle cases where some of the desired partitions contained more than one degree of freedom. Castellan (1965) reviewed these partitioning procedures and argued for their use in place of constructing a series of 2×2 tables, based on the following two points.

First, in setting up the full contingency table, it is assumed that the marginal totals represent the population values. It is more likely that the marginals for any 2×2 table, taken from the full table, will not adequately reflect those population values. Instead, they will reflect a population different from other populations generated from the same table. There will be as many populations represented as there are 2×2 tables produced.

Second, following the procedure Castellan presented, the 2×2 tables produce statistics that sum to the chi-square value for the original table. This independence of tables produces uncorrelated chi-squares and thus allows for more meaningful interpretation.

Bresnahan and Shapiro (1966) examined methods for partitioning, including the methods for determining possible partitions. They concluded that all forms of a partitioning follow three basic rules: (a) each cell appears alone once and only once, (b) the same combination of cells appears only once, and (c) the dividing lines of a partition do not hold for other partitions. Following these rules, additional partitioning schemes may be employed. A general equation for the chi-square is derived that may be applied to any table that results from partitioning. The equation for an $I \times J$ table is written as follows:

$$\chi_{(l-1)(m-1)}^2 = \sum_{i=1}^{l} \sum_{j=1}^{m} \left(\frac{x_{ij}^2}{e_{ij}} \right) - \sum_{i=1}^{l} \left(\frac{o_{i\cdot}}{e_{i\cdot}} \right) - \sum_{j=1}^{m} \left(\frac{o_{\cdot j}}{e_{\cdot j}} \right) + \frac{O}{E}, \quad (3)$$

where

l = the number of rows in the partitioned table;
m = the number of columns in the partitioned table;
e_{ij} = the expected value for cell ij calculated from the original table;
x_{ij} = the observed frequency in cell ij

$$e_{i\cdot} = \sum_{j=1}^{m} e_{ij};$$

$$e_{\cdot j} = \sum_{i=1}^{l} e_{ij};$$

$$o_{i\cdot} = \sum_{j=1}^{m} x_{ij};$$

$$o_{\cdot j} = \sum_{i=1}^{l} x_{ij};$$

$$O = \sum_{i=1}^{l} \sum_{j=1}^{m} x_{ij};$$

$$E = \sum_{i=1}^{l} \sum_{j=1}^{m} e_{ij};$$

Bresnahan and Shapiro (1966) advocated the use of this formula in cases where some cells have low expected values. Instead of pooling data or discarding it to raise the low expected values, one can calculate a chi-square based on the table configuration that contains adequate expected values. The value of the chi-square for the partitioned table will

be the contribution of that part of the table to the chi-square for the entire table. For the special case of the analysis of $2 \times k$ tables, Brunden (1972) proposed the use of rank sums as suggested by Steel (1960) and extended by Dunn (1964) instead of partitioning.

Shaffer (1973) has taken exception to the use of these methods of partitioning, claiming that they do not actually test the questions of interest. For example, a 2×4 table may be partitioned into three separate tests, each with one degree of freedom. Shaffer demonstrated that to test the first of the three resulting hypotheses actually entails testing that all three partitions do not contain significant differences against the alternate hypothesis that the first partition is significant and that the other two are not. This results from the fact that the data from the entire table enter the calculation for a portion of the table in the determination of the expected values. She therefore contends that the data from the entire table should not enter into a partition, since the test produced is not the statistic desired.

On the basis of this argument, Shaffer (1973) proposed the use of the likelihood ratio statistic. Though it does not partition exactly, its use overcomes the problem of testing "inappropriate" hypotheses. Shaffer has noted that although there is no evidence for the superiority of one method over another, Pearson's method has historical priority and a greater ease of computation.

Regardless of which method one uses, partitioning increases the amount of information one is able to glean from the data. If the partitions are orthogonal to one another, the information rendered from each partition does not overlap with any other. However, Shaffer's paper presents an interesting dilemma.

If one requires a test of a partition, independent of the structure of the rest of the partitions, then one must use the log-likelihood ratio as she proposed. Although the likelihood ratios for each partition do not sum to the ratio for the complete table, this may not always be a problem. Often, only one partition is meaningful and/or accounts for much of the total variation. In such cases, the choice between the use of the log-likelihood statistic and chi-square rests on the alternate hypothesis that is of interest. If one wishes to test a single partition for homogeneity against the hypothesis that it is not homogeneous and the rest of the partitions are, then chi-square is appropriate. If the test is to be done independently of the structure of the rest of the table, then the log-likelihood ratio is the method of choice. Other applications of the log-likelihood ratio will be discussed in the next section.

Several procedures that supplement or provide an alternative to partitioning are available. Graphical analysis is discussed and exemplified by Boardman (1977), Cohen (1980), Cox and Laugh (1967), Fienberg (1969), and Snee (1974). One version of graphical analysis, based on Brown's work (1974, 1976), is implemented by BMDP's 2F procedure (Dixon & Brown, 1977). Other alternative procedures are comparisons of individual proportions and testing hypotheses about order. These will be considered next.

Comparison of Individual Proportions

The chi-square procedure, as Berkson noted in 1938, is an omnibus test. In the case of a test for homogeneity among K groups classified by J levels of the dependent variable A, the hypothesis under test is that

$$H_o: \begin{bmatrix} P(A_1|G_1) \\ P(A_2|G_1) \\ \vdots \\ P(A_J|G_1) \end{bmatrix} = \begin{bmatrix} P(A_1|G_2) \\ P(A_2|G_2) \\ \vdots \\ P(A_J|G_2) \end{bmatrix}$$

$$= \cdots = \begin{bmatrix} P(A_1|G_K) \\ P(A_1|G_K) \\ \vdots \\ P(A_J|G_K) \end{bmatrix} = \begin{bmatrix} P(A_1) \\ P(A_2) \\ \vdots \\ P(A_J) \end{bmatrix}$$

against the alternative that H_o is false. If the hypothesis is rejected, one would like to be able to find the contrasts among the proportions that are significantly different from zero.

This may be accomplished by a well-known procedure that allows one to construct simultaneous confidence intervals for all contrasts of the proportions in the design, across groups, while maintaining the specified Type I error probability. The method is

an extension of Scheffé's (1953) theorem, which is used for the construction of contrasts in the analysis of variance.

If a linear contrast in the population proportions in a contingency table is denoted as ψ, the sample estimate is $\hat{\psi}$ and is defined as

$$\hat{\psi} = \Sigma a_k \hat{p}_k, \tag{4}$$

where $\hat{p}_k$ is the proportion in Group k and $\Sigma a_k = 0$. The limiting probability is $(1 - \alpha)$ that, for all contrasts,

$$\hat{\psi} - SE_{\hat{\psi}}\sqrt{\chi_{K-1;1-\alpha}^2} < \psi < \hat{\psi} + SE_{\hat{\psi}}\sqrt{\chi_{K-1;1-\alpha}^2}, \tag{5}$$

where

$$SE_{\hat{\psi}}^2 = \Sigma a_k^2 \frac{(\hat{p}_k \hat{q}_k)}{(n_k)}, \quad \hat{q}_k = 1 - \hat{p}_k, \tag{6}$$

and $\sqrt{\chi_{K-1;1-\alpha}^2}$ is the $(1 - \alpha)$th percent value from the chi-square distribution with $K - 1$ degrees of freedom. Some of the earlier work with this procedure may be found in Gart (1962), Gold (1963), and Goodman (1964).

The only drawback to this post hoc procedure is its lack of power relative to a planned set of contrasts. A generally more powerful procedure results from the use of a Bonferroni type critical value where the Type I error probability is spread over just the contrasts of interest. Such a value may be found in the table given by Dunn (1961). The value $\sqrt{\chi_{K-1;1-\alpha}^2}$ in the confidence interval is replaced by the value taken from Dunn's table based on Q, which equals the number of planned contrasts and the degrees of freedom, which equals infinity.

Analysis of Ordered Categories

In spite of its usefulness, there are conditions under which the use of Pearson's chi-square, although appropriate, is not the optimum procedure. Such a situation occurs when the categories forming a table have a natural ordering. The value of the statistic expressed in Equation 4 will not be altered if the rows and/or columns in a table are permuted. However, if ordering of the rows or columns exists, their order cannot meaningfully be changed. This is information that chi-square is not sensitive to. Instead, the researcher may choose among several alternatives.

If both rows and columns contain a natural ordering, two methods are available. The first is a procedure taken from Maxwell (1961) as modified by Marascuilo and McSweeney (1977). It is used to test for a monotonic trend in the responses across categories.

The first step is to quantify the categories using any arbitrary numbering system. As the method is independent of the numbers chosen, both Maxwell and Marascuilo and McSweeney recommend numbers that simplify the calculations such as the linear coefficients in a table of orthogonal polynomials. These coefficients are then applied to the marginal frequencies, the $Y_{i.}$ and $Y_{.j}$, to produce the sums and sums of squares for use in calculating a slope coefficient by the usual formula,

$$\hat{\beta} = \frac{N(\Sigma\Sigma Y_i Y_j - (\Sigma Y_i)(\Sigma Y_j)}{N(\Sigma Y_i^2) - \Sigma(Y_j^2)}. \tag{7}$$

Under the assumption that $\beta = 0$, the standard error of $\hat{\beta}$ is calculated as

$$SE_{\hat{\beta}}^2 = \frac{S_{Y_j}^2}{(N-1)S_{Y_i}^2}. \tag{8}$$

Then the hypothesis of no linear trend may be tested by

$$X^2 = \frac{\hat{\beta}^2}{SE_{\hat{\beta}}^2} \sim \chi_{\nu-1}^2. \tag{9}$$

A second procedure for examining tables with ordered marginal categories involves the use of Kendall's (1970) rank tau, corrected for ties. If the observed tau is statistically significant, the hypothesis of no association is rejected. In addition, the statistic itself is a measure of association or array of the data. Further comments are contained in the section of measures of association.

When one of the two variables defining a table is ordered, Kruskal and Wallis's (1952) nonparametric one-way analysis-of-variance procedure may be utilized to test for equality of distributions. This procedure is described by Marascuilo and Dagenais (1982). Consider the case of an $I \times K$ contingency table, where the dimension I is defined by mutually exclusive ordered categories. The Kruskal-Wallis statistic is based on a simultaneous comparison of the sum of the ranks for the K groups. To apply the statistic in the case

of an $I \times K$ table the frequencies within a category along dimension I are considered to be tied and, therefore, are all assigned a mid-rank value. One then sums the ranks across I, within Group k, to obtain the summed ranks used in calculating the statistic.

Comparison of Two Independent Chi-Squares

Situations may occur in which one may want to test the equality of two independent chi-square values. Knepp and Entwisle (1969) have presented, in tabular form, the 1% and 5% critical values for this comparison for degrees of freedom = 1 to 100. They also provide a normal approximation calculated as

$$z = \frac{\frac{1}{2}X_1^2 - \frac{1}{2}X_2^2}{\sqrt{\nu}}, \qquad (10)$$

where X_1^2 and X_2^2 are two independent sample chi-square values, each with ν degrees of freedom. The statistic z is approximately distributed as a unit normal variable.

D'Agostino and Rosman (1971) have offered another simple normal approximation for comparing two chi-square values in the form of

$$z = \frac{\sqrt{X_1^2} - \sqrt{X_2^2}}{\sqrt{1 - \frac{1}{4\nu}}}. \qquad (11)$$

This approximation was tested by Monte Carlo methods and found to be quite good for cases with degrees of freedom greater than 2. For one degree of freedom the researcher must use Knepp and Entwistle's tabled values, which are 2.19 for $\alpha = .05$ and 3.66 for $\alpha = .01$. D'Agostino and Rosman also note that for dfs greater than 20, the denominator in Equation 11 makes little difference and

$$z = \sqrt{X_1^2} - \sqrt{X_2^2} \qquad (12)$$

may be used in place of Equation 10.

One should note that these procedures should be used cautiously for at least two reasons. It is possible for very different configurations within two tables to produce the same chi-square values. It is also possible to obtain different values of chi-square from tables with identical internal patterns if the sample sizes differ between tables.

Measures of Association

The value of a chi-square statistic is difficult to evaluate as it is both a function of the truth of the hypothesis under test and a function of sample size. To double the size of a sample, barring sample-to-sample fluctuations, will double the size of the associated chi-square. To compensate for this, the data analyst should always calculate an appropriate measure of association so as to allow for judging the practical, that is, the meaningful significance of the findings.

If the data are generated from a single sample, then the proper test is one of independence and a measure of association is the mean square contingency coefficient. Designated as ϕ^2, it's sample estimate is calculated as

$$\hat{\phi}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{x_{ij}^2}{x_{i.}x_{.j}} - 1. \qquad (13)$$

It can be shown that the maximum value that $\hat{\phi}^2$ can attain is $\phi^2_{\max}$ = the minimum of $I - 1$ or $J - 1$. To correct for this compute

$$\hat{\phi}^{2\prime} = \frac{\hat{\phi}^2}{\hat{\phi}^2_{\max}}, \qquad (14)$$

which is referred to as Cramer's measure of association (Cramer, 1946).

In the case of a table generated from K samples, the proper measure of association is given by the work of Light and Margolin (1971) as a ratio of the sum of squares between the K groups over the total sum of squares. Their measure, R^2_{LM}, is tested for significance by a chi-square statistic calculated as $X^2 = (N - 1)(I - 1)R^2_{LM}$, which is tested at $df = (I - 1)(K - 1)$. Light and Margolin have shown that their statistic tends to be larger in some situations, and therefore more powerful, than the ordinary chi-square in the analysis of a K-group design.

When the frequencies of the K groups are cross-classified by a dependent variable that is ordered, Serlin, Carr, and Marascuilo (1982) have proposed a measure that is the ratio of the calculated test statistic to the maximum the statistic can reach. Their measure ranges from zero to unity, and it is interpreted just as eta-squared is in the parametric ANOVA.

If both variables are ordered, one is presented with a variety of choices including the

standard product-moment correlation coefficient (Kendall & Stuart, 1969), tau (Kendall, 1970), and gamma (Goodman & Kruskal, 1954, 1959, 1963). Comparison of these methods is given by Gans and Robertson (1981) and Cesa (1982). Tau is generally recommended as it approaches the normal distribution faster than Spearman's rho (Kendall, 1970) and is not inflated by the exclusion of tied values as gamma is.

In the case of a 2×2 table, the well-known measure of association based on chi-square is phi and is calculated as

$$\hat{\phi}^2 = \frac{X^2}{N}. \tag{15}$$

If Kendall's tau is calculated for the same table, then it will be seen that $\phi = \tau$. An alternative to the use of phi is to employ the odds ratio (Fienberg, 1980).

For a 2×2 table the categories defining the table may be labeled as A, $\bar{A}$, B, and $\bar{B}$. The probability of observing B, given the presence of A, can be expressed as

$$\frac{P(B|A)}{P(\bar{B}|A)} \tag{16}$$

Alternately, the probability of observing B, given the absence of A, is

$$\frac{P(B|\bar{A})}{P(\bar{B}|\bar{A})} \tag{17}$$

A simple measure of association, apparently first proposed by Cornfield (1951), is the ratio of these two odds. In the sample, the measure is calculated as

$$\hat{\gamma} = \frac{x_{11}x_{22}}{x_{12}x_{21}}, \tag{18}$$

with a standard error estimated as

$$SE_{\hat{\gamma}} = \sqrt{x_{11} + x_{22} + x_{12} + x_{21}}. \tag{19}$$

A useful discussion of this measure including additional references may be found in Fleiss (1973). The choice between the two coefficients, tau and phi, for the 2×2 table is not clear cut and the reader is referred to Fleiss for further discussion.

Summary

In summary, a few points bear repetition. Under certain conditions, expected cell frequencies less than five do not substantially alter the Type I error rate of the chi-square statistic. The decrease in power that accompanies these small expected values, though, should encourage one to use large sample sizes.

The debate over the use of the Yates correction for continuity is unresolved. There is general agreement, however, that the correction often results in a most conservative test when the margins in a table are generated from random variables.

There are a number of supplementary and alternative approaches to the use of Pearson's chi-square that the researcher should know. Often the questions one asks of data may be more directly or more efficiently answered by planned contrasts of proportions, partitioning of the total chi-square, or the use of log-linear models. A useful paper on this subject was written by Cochran (1954). He presented methods for dealing with some specific contingency table designs and probability distributions. In addition to the previously mentioned recommendations regarding minimum expected values, he discussed testing goodness-of-fit hypotheses in different distributions, degrees of freedom in $2 \times N$ tables, and combining 2×2 tables.

References

Agresti, A., & Wackerly, D. Some exact conditional tests of independence for $R \times C$ cross-classification tables. *Psychometrika*, 1977, *42*, 111–125.

Baker, R. J. Algorithm AS112. Exact distributions derived from two-way tables. *Applied Statistics*, 1977, *26*, 199–206.

Berkson, J. Some difficulties in interpretation of the chi-square test. *Journal of the American Statistical Association*, 1938, *33*, 526–536.

Berkson, J. In dispraise of the exact test. *Journal of Statistical Planning and Inference*, 1978, *2*, 27–42.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass.: MIT Press, 1975.

Boardman, T. J. Graphical contribution to the X^2 statistic for two-way contingency tables. *Communications in Statistics: Theory and Methods*, 1977, *A6*, 1437–1451.

Bradely, D. R., Bradely, T. D., McGrath, S. G., & Cutcomb, S. D. Type I error rate of the chi-square test of independence in $R \times C$ tables that have small expected frequencies. *Psychological Bulletin*, 1979, *86*, 1290–1297.

Bresnahan, J. L., & Shapiro, M. M. A general equation and technique for the exact partitioning of chi-square contingency tables. *Psychological Bulletin*, 1966, *66*, 252–262.

Brown, M. B. The identification of sources of significance in two-way contingency tables. *Applied Statistics*, 1974, *23*, 405–413.

Brown, M. B. Screening effects in multidimensional contingency tables. *Applied Statistics,* 1976, *25,* 37–46.

Brunden, M. N. The analysis of non-independent 2 × 2 tables from 2 × *c* tables using rank sums. *Biometrics,* 1972, *28,* 603–607.

Camilli, G., & Hopkins, K. D. Applicability of chi-square to 2 × 2 contingency table with small expected cell frequencies. *Psychological Bulletin,* 1978, *85,* 163–167.

Castellan, J. N. Jr. On the partitioning of contingency tables. *Psychological Bulletin,* 1965, *64,* 330–338.

Cesa, T. *Comparisons among methods of analysis for ordered contingency tables in psychology and education.* Unpublished dissertation, University of California, Berkeley, 1982.

Chapman, J. A. W. A comparison of the χ^2, $-2 \log R$, and multinomial probability criteria for significance tests when expected frequencies are small. *Journal of the American Statistical Association,* 1976, *71,* 854–863.

Cochran, W. G. The χ^2 distribution for the binomial and Poisson Series with small expectations. *Annals of Eugenics,* 1936, *2,* 207–217.

Cochran, W. G. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics,* 1952, *23,* 315–345.

Cochran, W. G. Some methods for strengthening the common χ^2 tests. *Biometrics,* 1954, *10,* 417–451.

Cohen, A. On the graphical display of the significant components in two-way contingency tables. *Communications in Statistics: Theory and Methods,* 1980, *A9,* 1025–1041.

Conover, W. J. Rejoinder. *Journal of the American Statistical Association,* 1974, *69,* 382. (a)

Conover, W. J. Some reasons for not using the Yates continuity correction on 2 × 2 contingency tables. *Journal of the American Statistical Association,* 1974, *69,* 374–382. (b)

Cornfield, J. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute,* 1951, *11,* 1269–1275.

Cox, D. R. The continuity correction. *Biometrika,* 1970, *57,* 217–219.

Cox, D. R., & Laugh, E. A note on the graphical analysis of multidimensional contingency tables. *Technometrics,* 1967, *9,* 481–488.

Cramer, H. *Mathematical Methods of Statistics.* Princeton, N.J.: Princeton University Press, 1946.

D'Agostino, R. B., & Rosman, B. A normal approximation for testing the equality of two independent chi-square values. *Psychometrika,* 1971, *36,* 251–252.

Dixon, W. J., & Brown, M. B. (Eds.). *BMDP-77: Biomedical computer programs P-series.* Berkeley: University of California Press, 1977.

Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association,* 1961, *56,* 52–64.

Dunn, O. J. Multiple comparisons using rank sums. *Technometrics,* 1964, *3,* 241–252.

Edwards, A. E. On "The use and misuse of the chi-square test": The case of the 2 × 2 contingency table. *Psychological Bulletin,* 1950, *47,* 341–346.

Fienberg, S. E. Preliminary graphical analysis and quasi-independence for two-way contingency tables. *Applied Statistics,* 1969, *18,* 153–168.

Fienberg, S. E. *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, Mass.: MIT Press, 1980.

Fisher, R. A. *Statistical methods for research workers* (7th ed.). London: Oliver and Boyd, 1938.

Fisher, R. A. The significance of deviations from expectations in a Poisson series. *Biometrics,* 1950, *6,* 17–24.

Fleiss, J. L. *Statistical methods for rates and proportions.* New York: Wiley, 1973.

Gans, L., & Robertson, C. A. The behavior of estimated measures of association in small and moderate sample sizes for 2 × 3 tables. *Communications in Statistics: Theory and Methods,* 1981, *A10,* 1673–1686.

Gart, J. J. Approximate confidence limits for the relative risk. *Journal of the Royal Statistical Society, Series B,* 1962, *24,* 454–463.

Gold, R. Z. Tests auxiliary to χ^2 tests in a markov chain. *Annals of Mathematical Statistics,* 1963, *34,* 56–74.

Good, I. J. The multivariate saddlepoint method and chi-squared for the multinominal distribution. *Annals of Mathematical Statistics,* 1961, *32,* 535–548.

Good, I. J., Grover, T. N., & Mitchell, G. J. Exact distributions for χ^2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution. *Journal of the American Statistical Association,* 1970, *65,* 267–283.

Goodman, L. A. Simultaneous confidence intervals for cross-products ratios in contingency tables. *Journal of the Royal Statistical Society, Series B,* 1964, *26,* 86–102.

Goodman, L. A. *Analyzing qualitative/categorical data.* Cambridge, Mass.: Abt Books, 1978.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. *Journal of the American Statistical Association,* 1954, *49,* 732–764.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. II: Further discussion and references. *Journal of the American Statistical Association.* 1959, *54,* 123–163.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. III: Approximate sampling theory. *Journal of the American Statistical Association,* 1963, *58,* 310–364.

Grizzle, J. E. Continuity correction in the chi-square test for 2 × 2 tables. *American Statistician,* 1967, *21*(4), 28–32.

Gumbel, E. J. On the reliability of the classical chi-square test. *Annals of Mathematical Statistics,* 1943, *14,* 253–263.

Haberman, S. J. *Analysis of qualitative data. Volume 1: Introductory topics.* New York: Academic Press, 1978.

Horn, S. D. Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics,* 1977, *33,* 237–248.

Hutchinson, T. P. The validity of the chi-squared test when expected frequencies are small: A list of recent research references. *Communications in Statistics: Theory and Methods,* 1979, *A8,* 327–335.

Irwin, J. O. A note on the subdivision of χ^2 into components. *Biometrika,* 1949, *36,* 130–134.

Jeffreys, H. *Theory of Probability* (3rd ed.). Oxford: Clarendon Press, 1961.

Kastenbaum, M. A. A note on the additive partitioning of chi-square in contingency tables. *Biometrics,* 1960, *16,* 416–422.

Katz, B. M., & McSweeney, M. Misclassification errors

and data analysis. *Journal of Experimental Education,* 1979, *47,* 331–338.

Kempthorne, O. The classical problem of inference: Goodness of fit. In J. Neyman (Ed.), *Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1966.

Kempthorne, O. In dispraise of the exact test: Reactions. *Journal of Statistical Planning and Inference,* 1979, *3,* 199–213.

Kendall, M. G. *The advanced theory of statistics* (Vol. 1, 5th ed.). London: Griffin, 1952.

Kendall, M. G. *Rank correlation methods* (4th ed.). London: Griffin, 1970.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 3, 3rd ed.). London: Griffin, 1969.

Kimball, A. W. Short cut formulas for the exact partitioning of χ^2 in contingency tables. *Biometrics,* 1954, *10,* 452–458.

Knepp, D. L., & Entwisle, D. R. Testing significance of differences between two chi-squares. *Psychometrika,* 1969, *34,* 331–333.

Kruskal, W. H., & Wallis, W. A. Use of rank in one-criterion variance analysis. *Journal of the American Statistical Association.* 1952, *47,* 401–412.

Lancaster, H. O. The derivation and partition of χ^2 in certain discrete distributions. *Biometrika,* 1949, *36,* 117–129.

Lancaster, H. O. The exact partitioning of χ^2 and its application to the problem of pooling of small expectations. *Biometrika,* 1950, *37,* 267–270.

Lancaster, H. O. Complex contingency tables treated by the partition of χ^2. *Journal of the Royal Statistical Society; Series B,* 1951, *13,* 242–249.

Lewis, D., & Burke, C. J. The use and misuse of the chi-square test. *Psychological Bulletin,* 1949, *46,* 433–489.

Lewis, D., & Burke, C. J. Further discussion of the use and misuse of the chi-square test. *Psychological Bulletin,* 1950, *47,* 347–355.

Lewontin, R. C., & Felsenstein, J. The robustness of homogeneity tests in $2 \times n$ tables. *Biometrics,* 1965, *21,* 19–33.

Light, R. J., & Margolin, B. H. An analysis of variance for categorical data. *Journal of the American Statistical Association,* 1971, *66,* 534–544.

Mann, H. B., & Wald, A. On the choice of the number of intervals in the application of the chi-square test. *Annals of Mathematical Statistics,* 1942, *13,* 306–317.

Mantel, N. Comment and a suggestion. *Journal of the American Statistical Association,* 1974, *69,* 378–380.

Mantel, N., & Greenhouse, S. W. What is the continuity correction? *The American Statistician,* 1968, *22*(5), 27–30.

Marascuilo, L. A., & Dagenais, F. Planned and post hoc comparisons for tests of homogeneity where the dependent variable is categorical and ordered. *Educational and Psychological Measurement,* 1982, *42,* 777–781.

Marascuilo, L. A., & McSweeny, M. *Nonparametric and Distribution-Free Methods for the Social Sciences.* Monterey, Calif.: Brooks/Cole, 1977.

Maxwell, A. E. *Analysing qualitative data.* London: Methuen, 1961.

Miettinen, O. S. Comment. *Journal of the American Statistical Association,* 1974, *69,* 380–382.

Mote, V. L., & Anderson, R. L. An investigation of the effect of misclassification on the properties of χ^2-tests in the analysis of categorical data. *Biometrika,* 1965, *52,* 95–109.

Neyman, J., & Pearson, E. S. Further notes on the χ^2 distribution. *Biometrika,* 1931, *22,* 298–305.

Overall, J. E. Power of chi-square tests for 2×2 contingency tables with small expected frequencies. *Psychological Bulletin,* 1980, *87,* 132–135.

Pastore, N. Some comments on "the use and misuse of the chi-square test." *Psychological Bulletin,* 1950, *47,* 338–340.

Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine,* July, 1900, pp. 157–175. In E. S. Pearson (Ed.), *Karl Pearson's Early Statistical Papers.* Cambridge: Cambridge at the University Press, 1947.

Peters, C. C. The misuse of chi-square: A reply to Lewis and Burke. *Psychological Bulletin,* 1950, *47,* 331–337.

Plackett, R. L. The continuity correction for 2×2 tables. *Biometrika,* 1964, *51,* 327–337.

Roscoe, J. T., & Byars, J. A. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association,* 1971, *66,* 755–759.

Scheffé, H. A method for judging all contrasts in the analysis of variance. *Biometrika,* 1953, *40,* 87–104.

Serlin, R. C., Carr, J. C., & Marascuilo, L. A. A measure of association for selected nonparametric procedures. *Psychological Bulletin,* 1982, *92,* 786–790.

Shaffer, J. P. Testing specific hypotheses in contingency tables: Chi-square partitioning and other methods. *Psychological Reports,* 1973, *33*(2), 343–348.

Slakter, M. J. A comparison of the Pearson chi-square and Kolmogorov goodness of fit tests with respect to validity. *Journal of the American Statistical Association,* 1965, *60,* 854–858.

Slakter, M. J. Comparative validity of the chi-square and two modified chi-square goodness of fit tests for small but equal expected frequencies. *Biometrika,* 1966, *53,* 619–622.

Snee, R. D. Graphical display of two-way contingency tables. *American Statistician,* 1974, *28,* 9–12.

Starmer, C. F., Grizzle, J. E., & Sen, P. K. Comment. *Journal of the American Statistical Association,* 1974, *69,* 376–378.

Steel, R. D. G. A rank sum for comparing all pairs of treatments. *Technometrics,* 1960, *2,* 197–208.

West, E. N., & Kempthorne, O. A comparison of the chi-square and likelihood ratio tests for composite alternatives. *Journal of Statistical Computation and Simulation,* 1972, *1,* 1–33.

Wise, M. E. Multinomial probabilities and the X^2 and χ^2 distributions. *Biometrika,* 1963, *50,* 145–154.

Yarnold, J. K. The minimum expectation in χ^2 goodness-of-fit tests and the accuracy of approximation for the null distribution. *Journal of the American Statistical Association,* 1970, *65,* 864–886.

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

Received November 8, 1982
Revision received February 11, 1983 ■