Nerlove, M., & Press, S. J. (1976). *Multivariate log-linear probability models for the analysis of qualitative data.* Discussion paper no. 1, Center for Statistics and Probability, Northwestern University.

Plackett, R. L. (1974). *The analysis of categorical data.* London: Griffin.

Rao, C. R. (1965). *Linear statistical inference and its applications.* New York: Wiley.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika, 40,* 105–110.

Williams, O. D., & Grizzle, J. E. (1972). Contingency tables having ordered response categories. *Journal of the American Statistical Association, 67,* 55–63.

10 On the Use and Misuse of Chi-Square

Kevin L. Delucchi
Developmental Studies Center, San Ramon, California

One of the most useful tools available to any data analyst—especially one who deals with social science data—is Pearson's statistic known as chi-square. Its usefulness stems primarily from the fact that much of the data collected by social scientists is categorical in nature—whether ordered or unordered. Not only are variables such as sex, school, ethnicity, and experimental group categorical, but one can argue that many other measures are best, that is, conservatively, analyzed by being treated as categorical variables. This would include, for example, the ubiquitous Likert-type item often found in questionnaires and other measures.

As well as being applicable in many common analysis situations, the chi-square statistic is also quite widely known, relatively easy to compute, and available on most computer packages of statistical software. Like the good-natured nextdoor neighbor who always lends a hand without complaining, however, the chi-square statistic is easy to take for granted and easy to misuse.

The title of this chapter comes from a 1949 landmark article by Lewis and Burke entitled "The Use and Misuse of the Chi-Square Test," which appeared in *Psychological Bulletin.* The purpose of their article was to counteract the improper use of this statistic by researchers in the behavioral sciences. It 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, use of the chi-square statistic among social scientists has increased, a great deal of research has been conducted on its behavior under a variety of conditions, and several methods have been developed to handle some

of the problems cited by Lewis and Burke. Several years ago I reviewed developments since Lewis and Burke's original paper (Delucchi, 1983). In this chapter, I provide a further update, reviewing those common errors, providing examples of some of them, and discussing supplementary and complementary procedures for the analysis of data commonly analyzed with Pearson's chi-square statistic.

THE USE OF CHI-SQUARE

To begin, let me remind the reader that there is a distinction between Pearson's chi-square statistic (Pearson, 1900) and the chi-square distribution. The former is a number calculated from data, which is compared to the latter, a family of theoretical distributions defined by their degrees of freedom. Unless stated otherwise, the phrase *chi-square* refers here to the computed statistic, symbolized by X^2, as opposed to the Greek letter chi (χ), which is used to denote the distribution.

As originally proposed by Pearson, the statistic is based on comparing the observed frequencies in a contingency table with those frequencies that would be expected under the hypothesis of no association when testing for independence between two variables in the single sample model, or with those expected under the hypothesis of homogeneity of distributions in the multiple sample model.

Table 10.1 illustrates the case of a 2×3 contingency table.[1] In this example we have the responses of 79 teachers from two groups of schools. Teachers at one group of schools are involved in an educational intervention, whereas those in the other school are serving as a control group. As part of an effort to determine the effects of the intervention on the teachers' perceptions of school climate, the teachers filled out a questionnaire that included a section asking them to indicate how typical a series of descriptions were of their school. The item used in Table 10.1 read, "The principal determines the educational program and philosophy." Their responses were classified into one of three categories: not typical, somewhat typical, and typical. We wish to know if there is evidence that teachers from the two groups view this aspect of school climate differently. Pearson's statistic is defined as:

$$X^2_\nu = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{[(f_{ij} - E(f_{ij})]^2}{E(f_{ij})} \qquad (1)$$

[1]These data, as are most of the other examples in this chapter, were collected as part of the evaluation of the Child Development Project (CDP), a multiyear demonstration program that is attempting to promote the prosocial development of elementary-age children. Interested readers are referred to Solomon, Watson, Delucchi, Schaps, and Battistich (1988) and Watson, Solomon, Battistich, Schaps, and Solomon (1989) for additional information.

TABLE 10.1
Contingency Table of School by Principal's Perceived Role

SCHOOL by PDETER

Principal determines the educational program and philosophy

	Count	not typical 1	somewhat 2	typical 3	Row Total
SCHOOL					
Control	1	7	14	24	45
					57.0
Program	2	12	16	6	34
					43.0
	Column Total	19	30	30	79
		24.1	38.0	38.0	100.0

Chi-Square	*Value*	*DF*	*Significance*
Pearson	10.92937	2	.00423
Likelihood Ratio	11.49290	2	.00319

Minimum Expected Frequency—8.177

Where:

I = number of rows
J = number of columns
ν = degrees of freedom
$= (I - 1)(J - 1)$
f_{ij} = observed frequency in ith row, jth column
$E(f_{ij})$ = expected value of the observed frequency
$= \dfrac{(F_{i.})(F_{.j})}{F_{..}}$

Computing the expected values gives us the following:

$$X^2_{(2-1)(3-1)} = \frac{[(7 - 10.8)^2 + (14 - 17.1)^2 + (24 - 17.1)^2 + (12 - 8.2)^2 + (16 - 12.9)^2 + (6 - 12.9)^2]}{10.8 + 17.1 + 17.1 + 8.2 + 12.9 + 12.9}$$

$$= \frac{863.42}{79}$$

$$= 10.9, \qquad (2)$$

which we then compare to a tabled value for $\alpha = .05$ from the chi-square distribution with two degrees of freedom, $X^2 = 5.99$. Our computed value is greater than the tabled value, so we have evidence to reject the hypothesis under study, which is the null hypothesis of no group differences.

The 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, in the order Lewis and Burke listed them, 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; and
9. incorrect computations.

Two of these errors, (8) incorrect determination of the number of degrees of freedom and (9) incorrect computations, are largely obsolete thanks to the widespread use of computer packages of statistical software. Nevertheless, it does no harm to remind the reader that errors in computation, program coding, data entry, and so forth are easy to make. A very good habit to acquire is to doubt your results and check the integrity of your data all the way back to the original raw data file. This is especially important in small-sample data sets where each data point carries substantial weight in the final results.

The seventh error, the use of nonfrequency data, is also an error that is not often encountered in the current research literature. This is probably the result of greater familiarity with chi-square among practitioners and journal reviewers. Suffice it to note the data entered into Equation 1 must be frequencies, not percentages, means, or any number that is not a count.

Lewis and Burke cited the first error in their list, lack of independence among single events or measures, as the error they found most frequently in their brief review of articles that used the chi-square statistic to analyze data. This is also probably the most likely cause of the fifth source of misuse, indeterminant theoretical frequencies, which they noted, "commonly arises from a lack of independence between measures" (p. 478).

It is interesting to note that one of the examples Lewis and Burke used to illustrate this error, in which a set of die are thrown repeatedly, is actually a poor, if not incorrect, example (Pastore, 1950).

One of the basic assumptions under which the statistic is derived is that of independence of the data. Just as one should not compute a two-sample t test on matched-pair data, so also one must not compute Pearson's chi-square on dependent measures. This is true regardless of what produces the interdependence; repeated measurement of the same person, sample matching, or correlation inherent in the subjects themselves such as data from spouses, siblings, parent and child combinations, and so on. The proper statistic for correlated data is McNemars's measure in the 2 × 2 table, and either Stuart's or Bowker's test in the K × K table (Marascuilo & McSweeny, 1977).

Small Theoretical Frequencies

Lewis and Burke (1949) labeled the second error in their list, the use of expected frequencies that are too small, as 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. As examples they cited two published studies that used chi-square tests with expected values below 10. It appears today that their position, a popular one among researchers, may be overly conservative.

This problem of small expected values 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 category sizes. In contrast, the categories of contingency tables used for testing association hypotheses are relatively fixed, 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 are considered first.

Tests of Association Hypotheses in Contingency Tables. Based on Monte Carlo and empirical studies, recommendations with respect to minimum expected cell frequencies in testing hypotheses of association have included recommended minimum values of 1 (Jeffreys, 1961; Kempthorne, 1966; Slakter, 1965), 5 (Fisher, 1938), 10 (Cramer, 1946), and 20 (Kendall, 1952). Cochran (1952) first proposed the oft-cited rule-of-thumb that chi-square may be applied if no more than 20% of the cells have expected values between one and five. Wise (1963) suggested that small (i.e., less than five) but equal expected frequencies were preferable to unequal frequencies where a few expected values are small, and the remaining frequencies are well above most criteria. Good, Grover, and Mitchell (1970) concluded that if the expected values are equal, they may be as low as 0.33 (p. 275).

This view of the statistic as being robust with respect to minimum expected values is also supported by the findings of 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 greater than five, they concluded that the test tends to be conservative. Even the occurrence of expected values below one generally does 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 that 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 preferred a minimum value of five categories. Slakter (1965, 1966), Good (1961), and Wise (1963) found that in testing goodness of fit, expected values may be as low as one or two for an alpha of .05 when expected values are equal. For unequal expected values or an alpha of .01, the expected frequencies should be at least four.

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 X^2 approximation can be used with much smaller expectations than previously considered possible" (p. 882).

After considering earlier work, Roscoe and Byars (1971) concluded that one should be concerned primarily with the average expected value when considering the goodness-of-fit statistic with more than one degree of freedom. 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) 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 and also 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, 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 or sparse tables, is the computation of an exact test (Agresti & Wackerly, 1977; Baker, 1977; Mehat & Patel, 1980; Mehat and Patel, 1983; Mehat, Patel, & Gray, 1985). In recent years, these procedures have become more accessible due to the availability of increased computer power and efficient algorithms. A comprehensive implementation can be found in the *Statxact* software (Cytel Software, 1991). In spite of its name, however, the use of an "exact test" is not without controversy. As is discussed in a later section, debate still continues over the appropriate use of both exact tests and continuity corrections. Berkson (1978), Kempthorne (1979), Upton (1982) and D'Agostino, Chase, and Belanger (1988) offered the opposition to its use in 2×2 tables.

Power Considerations. An important point that is easily overlooked concerns the effect of small expected values on the power of the chi-square test. Overall (1980) 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 because low power means the inability to detect an existing difference.

Specific advice as to the selection of sample size is difficult to provide as the requirements and standards of researchers vary. In general, following Cochran's rule will provide sufficient power in most cases. Tables for computing power in the use of chi-square are given in Cohen (1988, chap. 7). The point here is to remind the reader that Type II error rates go up as sample size goes down.

Neglect of Frequencies of Nonoccurrence

Omitting frequencies of nonoccurrence from contingency tables is a surprisingly easy error to make, and examples can still occasionally be found. Consider, for example, the case of some of the early work on the detection of item bias. In

1979, Scheuneman proposed a method of detecting potentially biased items in an otherwise unbiased test analogous to the more demanding item-response theory approach by categorizing test-takers based on the total test score to equate for ability differences. To test an item for evidence of bias against some subgroup based on, say, sex or ethnicity of the test-taker, she proposed classifying each person in the sample on three dimensions: their group membership, total test score, and whether or not they passed the item.

But the contingency table she formed for calculating chi-square on was not this three-dimensional table, but rather a two-dimensional table defined by total-score grouping and group membership—counting only the numbers of each group that passed the item in question. As noted by several critics (Baker, 1981; Marascuilo & Slaughter, 1981), the resulting statistic is not distributed as chi-square because she neglected to count the frequency of the group members who did not pass the item. For an example of the statistically correct approach the reader is referred to Zwick and Ericikan (1989).

By neglecting the frequencies of nonoccurrence one usually commits the fourth error, failure to equalize the sum of the observed frequencies and the sum of the theoretical frequencies. Although relatively rare, this will result directly from the error discussed earlier—neglecting the frequency of nonoccurrence. One quick check of the validity of a contingency table for computing chi-square is to see if the sum of the observed frequencies is equal to the sum of the expected. If they are not equal, something is wrong.

Incorrect or Questionable Categorizing

This problem, more an issue of methodology than of mathematical statistics, is found in situations where the data need to be categorized in some arbitrary form in the absence of naturally occurring categories such as group membership. The distribution of frequencies within a set of categories is at the heart of the statistic, so the selection of those categories obviously will have a great deal of influence on the obtained value. The conservative data analyst will define categories before collecting data (preferably as a result of collecting and analyzing pilot data). The categories should be mutually exclusive so that each outcome belongs in one, and only one, category, and they must be as well-defined as possible so that there is no question about what constitutes membership in a given category. The categories themselves should cover the full range of possible responses, yet not be so narrowly defined that the resulting frequencies produce very low expected values.

While on this subject of classification, a comment on the matter of misclassification is appropriate. 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. Most of the relevant literature

is found in the biostatistics literature (e.g., Mote & Anderson, 1965). One exception to this in the area of educational research is an article by Katz and McSweeny (1979), who discussed the effects of classification errors on the significance level and power of the test for equality or proportions. They developed and discussed a correction procedure based on estimates of the probability of false negatives and false positives and noted that the detrimental effects of misclassification can be marked, including a loss in power. This problem is especially likely to occur 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 should consult the Katz and McSweeny (1979) article and the references they cited. The key to using their procedure, and its major drawback, is the need for estimates of the rate of misclassification that often may be unobtainable.

Correction for Continuity

As part of their discussion on the proper use of the chi-square statistic, Lewis and Burke presented the Yates correction for continuity, noting that it is justified only in the case of a 2×2 table. Since the time of their writing, questions have arisen regarding the appropriateness of the use of a correction for continuity.

Categorical variables are discrete and the chi-square distribution is continuous, thus a correction to improve the approximation can be made. The most well-known correction was proposed by Yates (1934) and is formed by adding or subtracting $\frac{1}{2}$ 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, X_c^2, is written as

$$X_c^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left[\left(f_{ij} + \frac{1}{2} - E(f_{ij})\right]\right]^2}{E(f_{ij})} \tag{3}$$

The analytical derivation of the correction expressed in Equation 3 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 as to render it useless for practical purposes.

The consensus of several investigators (Camilli & Hopkins, 1978; Conover, 1974a, 1974b; D'Agostino, Chase, & Belanger, 1988; Mantel, 1974; Mantel & Greenhouse, 1968; Miettinin, 1974; Starmer, Grizzle, & Sen, 1974; Upton,

1982) is 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, the use of the correction should not be given the blanket recommendation that often accompanies it.

These critics are not without critics of their own. In a paper read to the Royal Statistical Society, Yates (1984, followed by comments from several noteworthy statisticians) held that the correction for continuity is misunderstood due to strict adherence to Neyman–Pearson critical levels, the use of strict nominal levels and a refusal by investigators to accept his arguments for conditioning on the marginals. In any event, as a couple of the discussants following Yates noted, even the simple 2×2 table contains a great deal of potential information and the analysis of even such a simple case cannot be taken lightly.

So the debate continues after 50 years. 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.

Having reviewed common sources of misuse, let us move on to supplementary and alternative procedures that can aid in the exploration of data appropriate to a chi-square–based analysis.

SUPPLEMENTARY AND ALTERNATIVE PROCEDURES

Whereas a properly executed chi-square statistic may well be a thing of beauty to behold—at least to some of us—in many ways it is only the simplest of forms of statistical analysis. There are at least three major shortcomings to its use: (a) it is an omnibus test, (b) it does not necessarily utilize all of the information available in a contingency table such as the ordering of categories, and (c) its significance level is partly a function of sample size. So by itself a significant chi-square statistic may not provide all of the information contained in the table. The researcher should keep in mind several procedures that supplement or serve as an alternative to a chi-square test. A comprehensive treatment of these and other methods may be found in Agresti (1990).

One way to understand why a contingency table produces a statistically significant test statistic is to examine the cell entries expressed as more than just counts. Table 10.2 is a table produced by SPSSX from the data shown in Table 10.1. The difference in these two tables results from the information requested of the software.

In addition to cell counts, Table 10.2 displays the cell information in terms of each cell's expected value, its count as a percentage of the row, column, and

TABLE 10.2
Expanded Display of Table 10.1

SCHOOL by PDETER

Principal determines the educational program and philosophy

		not typical 1	somewhat 2	typical 3	Row Total
SCHOOL	Count				
	Exp Val				
	Row Pct				
	Col Pct				
	Tot Pct				
	Residual				
	Std Res				
	Adj Res				
Control	1	7	14	24	45
		10.8	17.1	17.1	57.0%
		15.6%	31.1%	53.3%	
		36.8%	46.7%	80.0%	
		8.9%	17.7%	30.4%	
		−3.8	−3.1	6.9	
		−1.2	−.7	1.7	
		−2.0	−1.4	3.2	
Program	2	12	16	6	34
		8.2	12.9	12.9	43.0%
		35.3%	47.1%	17.6%	
		63.2%	53.3%	20.0%	
		15.2%	20.3%	7.6%	
		3.8	3.1	−6.9	
		1.3	.9	−1.9	
		2.0	1.4	−3.2	
	Column Total	19 24.1	30 38.0%	30 38.0%	79 100.0%

Chi-Square	*Value*	*DF*	*Significance*
Pearson	10.92937	2	.00423
Likelihood Ratio	11.49290	2	.00319

Minimum Expected Frequency—8.177

total N, and as a residual from the expected value in "raw," Studentized, and adjusted forms. Note that the largest residuals are found in the column marked "typical" where 53.3% (24 out of 45) Control teachers chose this response versus 17.6% (6 out of 34) of the Program teachers. By re-expressing the cell entries in each of these forms the data analyst may begin to see more of the information contained in the table that the basic cell counts alone cannot provide.

Comparison of Individual Proportions

The chi-square procedure, as Berkson noted in 1938, is an omnibus test.[2] 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 expressed as

$$H_0: \begin{bmatrix} P(A_1|G_1) \\ P(A_2|G_1) \\ \vdots \\ P(A_J|G_1) \end{bmatrix} = \begin{bmatrix} P(A_1|G_2) \\ P(A_2|G_2) \\ \vdots \\ P(A_J|G_2) \end{bmatrix} = \ldots = \begin{bmatrix} P(A_1|G_K) \\ P(A_2|G_K) \\ \vdots \\ P(A_J|G_K) \end{bmatrix} = \begin{bmatrix} P(A_1) \\ P(A_2) \\ \vdots \\ P(A_J) \end{bmatrix} \quad (4)$$

against the alternative that H_0 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 Scheffe's (1953) theorum, which is used for the construction of contrasts in the analysis of variance. Scheffe's work was extended by Dunn (1961) and applied to qualitative variables by Goodman (1964) in the 1960s.

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

$$\hat{\Psi} = \Sigma a_k \hat{p}_k, \quad (5)$$

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,

$$\hat{\Psi} - SE_\Psi \sqrt{\chi^2_{k-1;1-\alpha}} < \psi < \hat{\Psi} + SE_\Psi \sqrt{\chi^2_{k-1;1-\alpha}} \quad (6)$$

where

$$SE^2_{\hat{\psi}} = \sum a_k^2 \frac{\hat{p}_k \hat{q}_k}{n_k}, \; \hat{q}_k = 1 - \hat{p}_k \quad (7)$$

and $\sqrt{\chi^2}$ 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).

Table 10.3 contains an example of such a contrast. Here, the proportion of teachers from each group who chose "very typical" as their answer are compared.

[2]It is intriguing that in spite continuing criticism of omnibus tests as not providing specific answers to research questions, they are still widely used. See Rosnow and Rosenthal (1989) for further discussion including their rule of thumb which states that whenever we use a chi-square or F test with greater than one degree of freedom, we have probably tested a question in which we are not interested.

TABLE 10.3
Computing a Confidence Interval for the Difference Between Proportions

$$\hat{\psi} = (1)\frac{24}{45} + (-1)\frac{6}{34}$$
$$= .5333 - .1765$$
$$= .3568$$
$$SE^2_{\hat{\psi}} = (1)^2 \frac{(.5333)(.4667)}{45} + (-1)^2 \frac{(.1765)(.8235)}{34}$$
$$= \frac{.2489}{45} + \frac{1453}{34}$$
$$= .0055 + .00427$$
$$= .0098$$
$$SE^2_{\hat{\psi}} = \sqrt{SE^2_{\hat{\psi}}} = \sqrt{.0098} = .09902$$
$$.3568 - \sqrt{.0098}\sqrt{5.99} < \psi < .3568 + \sqrt{.0098}\sqrt{5.99}$$
$$.114 < \psi < .599$$

The only drawback to this post hoc application 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 originally by Dunn (1961) and included in many tests (cf. Marascuilo & Serlin, 1988). The value $\sqrt{\chi^2}$ 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.

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 the sample size. To double the size of a sample, barring sample-to-sample fluctuations, will double the size of the computed chi-square statistic. To compensate for this, the data analyst should always calculate an appropriate measure of association in order to assess the practical, that is, the meaningful significance of the findings.

Bishop, Fienberg, and Holland (1975, chap. 11) provided an overview of various measures of association for two-dimensional tables. They made an important point when they noted that the issue today is not to develop an appropriate measure of association for a given problem, but rather "to choose wisely from among the variety of existing measures" (p. 373). For example, SPSSX and BMDP both offer over 12 measures of association to choose from. Table 10.4 is a

TABLE 10.4
Measures of Association for the Data of Table 10.1

Statistic	*Value*	*ASE1*	*T value*	*Approximate Significance*
Phi	.37195			.00423 *1
Cramer's V	.37195			.00423 *1
Contingency Coefficient	.34862			.00423 *1
Lambda :				
symmetric	.20482	.11844	1.61109	
with SCHOOL dependent	.20588	.18347	1.00639	
with PDETER dependent	.20408	.11224	1.64993	
Goodman & Kruskal Tau :				
with SCHOOL dependent	.13835	.07330		.00454 *2
with PDETER dependent	.07190	.03956		.00367 *2
Uncertainty Coefficient :				
symmetric	.08259	.04599	1.79391	.00319 *3
with SCHOOL dependent	.10643	.05921	1.79391	.00319 *3
with PDETER dependent	.06747	.03762	1.79391	.00319 *3
Kendall's Tau-b	−.34075	.09591	−3.53772	
Kendall's Tau-c	−.38584	.10906	−3.53772	
Gamma	−.55844	.13738	−3.53772	
Somers's D :				
symmetric	−.33725	.09492	−3.53772	
with SCHOOL dependent	−.29510	.08352	−3.53772	
with PDETER dependent	−.39346	.11052	−3.53772	
Pearson's R	−.35403	.10173	−3.32169	.00069
Spearman Correlation	−.36041	.10141	−3.39041	.00055
Eta :				
with SCHOOL dependent	.37195			
with PDETER dependent	.35403			

*1 Pearson chi-square probability
*2 Based on chi-square approximation
*3 Likelihood ratio chi-square probability

copy of the measures of association produced by SPSSX for the example in Table 10.1.

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, its sample estimate is calculated as

$$\hat{\phi}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{f_{ij}^2}{f_{i\cdot} - f_{\cdot j}} - 1. \tag{8}$$

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

$$\hat{\phi}^{2'} = \frac{\hat{\phi}^2}{\phi^2_{\max}}, \tag{9}$$

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

If both variables are ordered, one is presented with a variety of choices including the standard product-moment correlation coefficient (Kendall & Stuart, 1969), tau-a and tau-b (Kendall, 1970; Kendall & Stuart, 1979), Goodman and Kruskal's tau, 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.

When the frequencies of the K groups are cross-classified by a dependent variable that is ordered, Serlin, Carr, and Marascuilo (1982) 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 η^2 is in the parametric analysis of variance (ANOVA). For Table 10.2, $\eta = .37$.

In the case of a 2×2 table, the well-known measure of association based on χ^2 is ϕ^2 and is calculated as

$$\hat{\phi}^2 = \frac{X^2}{N}. \tag{10}$$

If Kendall's tau is calculated for the same table, 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, not-A, B, and not-B. The probability of observing B, given the presence of A, can be expressed as

$$\frac{P(B|A)}{P(\bar{B}|\bar{A})}. \tag{11}$$

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

$$\frac{P(B|\bar{A})}{P(\bar{B}|\bar{A})}. \tag{12}$$

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

$$\hat{\gamma} = \frac{f_{11} f_{22}}{f_{12} f_{21}} \tag{13}$$

with a standard error estimated as

$$SE_{\hat{\gamma}} = \sqrt{\frac{1}{f_{11}} + \frac{1}{f_{22}} + \frac{1}{f_{12}} + \frac{1}{f_{21}}}. \tag{14}$$

A useful discussion of this measure, which is widely used in bio-medical research, 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.

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 5 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 to which chi-square is not sensitive. 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 McSweeny (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 McSweeny recommended 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 $Y_{i\cdot}$ and $Y_{\cdot j}$, to produce the sums and sums of squares for use in calculating a slope coefficient by the usual equation:

$$\hat{\beta} = \frac{N(\Sigma\Sigma Y_i Y_j - (\Sigma Y_i)(\Sigma Y_j))}{N(\Sigma Y_i^2) - (\Sigma Y_j^2)}. \tag{15}$$

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

$$SE_{\beta} = \frac{S_{Y_j}^2}{N - 1(S_{Y_j}^2)}. \tag{16}$$

Then the hypothesis of no linear trend may be tested by

$$X^2 = \frac{\hat{\beta}^2}{SE^2\hat{\beta}^2} \sim \chi^2_{\nu-1}. \tag{17}$$

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, as discussed in the previous section.

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 $I \times J$ 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 assigned a midrank value. One then sums the ranks across I, within Group k, to obtain the summed ranks used in calculating the statistic.

Log- and Logit-Linear Models

This versatile statistic of Pearson's can also be extended to three-dimensional tables as well (Agresti, 1990; Fienberg, 1980). Given the expected frequencies derived from a model, one computes the statistic as shown in Equation 1. The degrees of freedom are computed as the number of cells in the table minus the number of parameters fitted. As Fienberg (p. 40) noted, Equation 1 is asymptotically equivalent to G^2 which is -2 times the log of the likelihood ratio statistic. The choice between these two statistics is discussed in the next section.

The derivation of the expected values in multidimensional tables are, of course, at the heart of log-linear and logit-liner models. Many articles and texts are now available for these procedures, including the works of Bishop et al. (1975), Goodman (1978), Haberman (1978), and Fienberg (1980). These procedures are implemented through several packaged computer programs including LOGLINEAR in SPSSX, SAS CATMOD, 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 likely that log-linear models will eventually supersede the use of Pearson's chi-square in the future because of their similarity to ANOVA procedures and their extension to higher-order tables. Discussion of this methodology, however, is beyond the scope of this chapter.

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

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \log_e \frac{f_{ij}}{E(f_{ij})}. \tag{18}$$

In their text on discrete multivariate analysis, Bishop et al. (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 called 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 stated, "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 the 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) provided an overview of much of this research, including the work of Neyman and Pearson (1931), Cochran (1936), Fisher (1950), Good et al. (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 following the next section.

Comparison of Two Independent Chi-Squares

It is conceivable that situations may occur in which one may want to test the equality of two independent chi-square values. One direct method to accomplish this would be to compute the same measure of association for each table and visually compare their values. If a test is required, Knepp and Entwisle (1969) presented, in tabular form, the 1% and 5% critical values for this comparison for degrees of freedom that equal 1 to 100. They also provided a normal approximation calculated as

$$Z = \frac{\frac{1}{2}(X_1^2) - \frac{1}{2}(X_2^2)}{\sqrt{\nu}}, \tag{19}$$

where X_{21} and X_{22} 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) offered another simple normal approximation for comparing two chi-square value in the form of

$$\frac{\sqrt{x_1^2} - \sqrt{x_2^2}}{\sqrt{1 - \frac{1}{4\nu}}}. \tag{20}$$

This approximation was tested by Monte Carlo methods and found to be quite good for cases with degrees of freedom greater than two. With 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 noted that for *df*'s greater than 20, the denominator in Equation 20 makes little difference and

$$\sqrt{X_1^2} - \sqrt{X_2^2} \tag{21}$$

may be used in place of Equation 19.

The same question that produced the data in Table 10.1 was asked of 68 teachers from two groups in a different school district. Pearson's chi-square for this second sample equaled 5.106 compared to a value of 10.929 in Table 10.1. With only two degrees of freedom we can use Equation 19 to obtain a z statistic of 2.05, leading us to conclude that the two sample statistics are different from each other. In other words, the lack of homogeneity between groups is not the same for our two samples.

As noted by Serlin (personal communication, 1990) one should be able to extend this same approach to tables with different degrees of freedom. Using the relatively accurate cube-root approximation one should be able to compute a z statistic as

$$Z = \frac{\sqrt[3]{\frac{X_1^2}{\nu_1}} - \sqrt[3]{\frac{X_2^2}{\nu_2}}}{\sqrt{\frac{2}{9\nu_1} + \frac{2}{9\nu_2}}}. \tag{22}$$

Although this approximation is quite good for even two or three degrees of freedom, this is still a large-sample approximation.

One should note that these procedures should be used with extreme caution

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 chi-square values from tables with identical internal patterns if the sample sizes differ between tables.

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. This work along with the work of Irwin (1949), Kimbal (1954), Kastenbaun (1960), Castellan (1965), and Bresnahan and Shapiro (1966) allows one to decompose a contingency table into a set of smaller tables whose individual chi-square statistics sum to the total chi-square.

The partitioning of contingency tables is not often seen in the literature, however, for two primary reasons. First, log-linear analysis, the examination of residuals, and the use of contrasts permit one to examine the sources of variation as easily. Second, Shaffer (1973) demonstrated that to test one partition for statistical significance is actually to test the hypothesis that no partition is significant against the alternative that one is significant and the remaining partitions are not. The interested reader is referred to the references cited earlier.

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, 1983).

CONCLUSIONS

Ninety years after its original development, Pearson's chi-square statistic remains a useful and powerful tool in our attempts to account for variation in data. Its ready availability makes for widespread use while research into its various properties and over its appropriate applications continues. In addition to reminding the researcher to pay heed to all of the usual issues and warnings applicable to any inferential statistic, such as being aware of its assumptions and what precise hypothesis it tests, 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 an overly 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 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.

ACKNOWLEDGMENTS

This chapter was written while the author was a research associate at the Developmental Studies Center. He is now senior statistician at the Treatment Research Unit, University of California, San Francisco.

The author would like to acknowledge and thank Drs. Patricia Busk and Ron Serlin for helpful discussions, Drs. Daniel Solomon, Victor Battistich, and an anonymous reviewer for many helpful suggestions and the late Dr. Leonard Marascuilo who suggested this topic to the author many years ago.

REFERENCES

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

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

Baker, R. J. (1977). Algorithm AS 112. Exact distributions derived from two-way tables. *Applied Statistics, 26*, 199–206.

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18*, 59–62.

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

Berkson, J. (1978). In dispraise of exact tests. *Journal of Statistical Planning and Inference, 2*, 27–42.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

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

Bradely, D. R., Bradely, T. D., McGrath, S. G., & Cutcomb, S. D. (1979). Type I error rate of the chi-square test of independence in R × C tables that have small expected frequencies. *Psychological Bulletin, 86*, 1920–1927.

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

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

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

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

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

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

Chapman, J. A. W. A. (1976). A comparison of the chi-square, −2 Log R, and multinomial probability criteria for significance tests when expected frequencies are small. *Journal of the American Statistical Association, 71,* 854–863.

Cochran, W. G. (1936). The chi-square distribution for the binomial and Poisson series with small expectations. *Annals of Eugenics, 2,* 207–217.

Cochran, W. G. (1952). The chi-square test of goodness-of-fit. *Annals of Mathematical Statistics, 23,* 315–345.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics, 10,* 417–451.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

Cox, D. R. (1970). The continuity correction. *Biometrics, 57,* 217–219.

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

Cramer, H. (1946). *Mathematical methods of statistics.* Princeton, NJ: Princeton University Press.

CYTEL Software. (1991). *User manual, version 2.* Cambridge, MA: Author.

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

D'Agostino, R. B., Chase, W., & Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistician, 42*(2), 198–203.

Delucchi, K. L. (1983). The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin, 94,* 166–176.

Dixon, W. J. (Ed.) (1983). *BMDP statistical software.* Berkeley: University of California Press.

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

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

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

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

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

Fisher, R. A. (1950). The significance of deviations from expectations in a Poisson series. *Biometrics, 6,* 17–34.

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

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

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

Gold, R. Z. (1963). Tests auxiliary of chi-square tests in a markov chain. *Annals of Mathematical Statistics, 34,* 56–74.

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

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

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

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

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

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

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

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

Gumbel, E. J. (1943). On the reliability of the classical chi-square test. *Annals of Mathematical Statistics, 14,* 255–263.

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

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

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

Irwin, J. O. (1949). A note on the subdivision of chi-square into components. *Biometrics, 36,* 130–134.

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

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

Katz, B. M., & McSweeney, M. (1979). Misclassification errors and data analysis. *Journal of Experimental Education, 47,* 331–338.

Kempthorne, O. (1966). The classical problem of inference: Goodness-of-fit. In J. Neyman (Ed.), *Fifth Berkeley symposium on mathematical statistics and probability* (pp. 235–249). Berkeley: University of California Press.

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

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

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

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

Kendall, M. G., & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 2, 4th ed.). London: Griffin.

Kimball, A. W. (1954). Short cut formulas for the exact partitioning of chi-square in contingency tables. *Biometrics, 10,* 452–458.

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

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

Lancaster, H. O. (1949). The derivation and partition of chi-square in certain discrete distributions. *Biometrika, 36,* 117–129.

Lancaster, H. O. (1950). The exact partitioning of chi-square and its application to the problem of pooling of small expectations. *Biometrika, 37,* 267–270.

Lancaster, H. O. (1951). Complex contingency tables treated by the partition of chi-square. *Journal of the Royal Statistical Society, Series B, 13,* 242–249.

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

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

Lewontin, R. C., & Felsenstein, J. (1965). The robustness of homogeneity tests in 2-×-n tables. *Biometrics, 21,* 19–33.

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

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

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

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

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences.* New York: W. H. Freeman.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on chi-square statistics. *Journal of Educational Measurement, 18,* 229–248.

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

Mehat, C. R., & Patel, N. R. (1980). A network algorithm for the exact treatment of the 2-×-K contingency table. *Communication in Statistics: Simmulation and Computation, B9,* 649–664.

Mehat, C. R., & Patel, N. R. (1983). A network algorithm for performing Fisher's exact test in r × c contingency tables. *Journal of the American Statistical Association, 78,* 427–434.

Mehat, C. R., Patel, N. R., & Gray, R. (1985). On computing an exact confidence interval for the common odds ratio in several 2-×-2 contingency tables. *Journal of the American Statistical Association, 80,* 969–973.

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

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

Neyman, J., & Pearson, E. S. (1931). Further notes on the chi-square distribution. *Biometrika, 22,* 298–305.

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

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

Pearson, K. (1900). 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, 157–175. In E. S. Pearson (Ed.), (1947). *Karl Pearson's early statistical papers.* Cambridge: Cambridge University Press.

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

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

Roscoe, J. T., & Byars, J. A. (1971). 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, 66,* 755–759.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

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

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143–152.

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

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

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

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

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

Solomon, D., Watson, M. S., Delucchi, K. L., Schaps, E., & Battistich, V. (1988). Enhancing children's prosocial behavior in the classroom. *American Educational Research Journal, 25,* 527–554.

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

Upton, G. J. G. (1982). A comparison of alternative tests for the 2-×-2 comparative trial. *Journal of the Royal Statistical Society, Series B, 145,* 86–105.

Watson, M., Solomon, D., Battistich, V., Schaps, E., & Solomon, J. (1989). The child development project: Combining traditional and developmental approaches to values education. In L. P. Nucci (Ed.), *Moral development and character education* (pp. 51–92). Berkeley: McCutchan.

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

Wise, M. E. (1963). Multinomial probabilities and the X^2 and chi-square distributions. *Biometrika, 50,* 145–154.

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

A HANDBOOK FOR DATA ANALYSIS IN THE BEHAVIORAL SCIENCES: Statistical Issues

Edited by

Gideon Keren
Free University of Amsterdam

Charles Lewis
Educational Testing Service

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1993 Hillsdale, New Jersey Hove & London