Method Note

American Journal of Evaluation
33(3) 448-458

DOI: 10.1177/1098214011426594
http://aje.sagepub.com

The Chi-Square Test: Often Used and More Often Misinterpreted

Todd Michael Franke[1], Timothy Ho[2], and Christina A. Christie[3]

Abstract

The examination of cross-classified category data is common in evaluation and research, with Karl Pearson's family of chi-square tests representing one of the most utilized statistical analyses for answering questions about the association or difference between categorical variables. Unfortunately, these tests are also among the more commonly misinterpreted statistical tests in the field. The problem is not that researchers and evaluators misapply the results of chi-square tests, but rather they tend to over interpret or incorrectly interpret the results, leading to statements that may have limited or no statistical support based on the analyses preformed.

This paper attempts to clarify any confusion about the uses and interpretations of the family of chi-square tests developed by Pearson, focusing primarily on the chi-square tests of independence and homogeneity of variance (identity of distributions). A brief survey of the recent evaluation literature is presented to illustrate the prevalence of the chi-square test and to offer examples of how these tests are misinterpreted. While the omnibus form of all three tests in the Karl Pearson family of chi-square tests—independence, homogeneity, and goodness-of-fit,—use essentially the same formula, each of these three tests is, in fact, distinct with specific hypotheses, sampling approaches, interpretations, and options following rejection of the null hypothesis. Finally, a little known option, the use and interpretation of post hoc comparisons based on Goodman's procedure (Goodman, 1963) following the rejection of the chi-square test of homogeneity, is described in detail.

Keywords

chi-square test, quantitative methods, methods use, using chi-square test

[1] Department of Social Welfare, Meyer and Rene Luskin School of Public Affairs, University of California, Los Angeles, CA, USA
[2] Department of Education, Graduate School of Education and Information Sciences, University of California, Los Angeles, CA, USA
[3] Department of Education, Social Research Methods Division, Graduate School of Education and Information Sciences, University of California, Los Angeles, CA, USA

Corresponding Author:
Todd Michael Franke, Department of Social Welfare, Meyer and Rene Luskin School of Public Affairs, University of California, Box 951656, Los Angeles, CA, 90095, USA
Email: tfranke@ucla.edu

Karl Pearson initially developed the chi-square test in 1900 and applied it to test the goodness of fit for frequency curves. Later, in 1904, he extended it to contingency tables to test for independence between rows and columns (Stigler, 1999). Since then, the Pearson family of chi-square tests has become one of the most common sets of statistical analyses in evaluation and social science research. Unfortunately, these tests are also among the more commonly misinterpreted statistical tests in the field. The problem is not that researchers and evaluators misapply the results of chi-square tests, but rather they tend to over interpret or incorrectly interpret the results, leading them to make statements that may have limited or no statistical support based on the analyses preformed.

In this article, we will attempt to clarify any confusion about the uses and interpretations of the family of chi-square tests developed by Pearson, focusing primarily on the chi-square tests of independence and homogeneity of variance (identity of distributions). First, the family of chi-square statistics will be presented, including distinguishing features of and appropriate uses for each specific test. Next, a brief survey of the recent evaluation literature will be presented to illustrate the prevalence of the chi-square test and to offer examples of how these tests are misinterpreted. Finally, a little known option, the use of post hoc comparisons based on Goodman's procedure (Goodman, 1963) following the rejection of the chi-square test of homogeneity, will be described.

The Karl Pearson Family of Chi-Square Tests

The chi-square test is computationally simple. It is used to examine independence across two categorical variables or to assess how well a sample fits the distribution of a known population (goodness of fit). The chi-square tests in the Karl Pearson family are not to be confused with others such as the Yates chi-square test (correction for continuity), the Mantel–Haenszel chi-square or the Maxwell–Stuart tests of correlated proportions. Each of these has its own applications, though they all utilize the chi-square distribution as the reference distribution. In fact, many tests that assess model fit use the chi-square distribution as the reference distribution. For example, many covariance structure analyses, including factor analysis and structural equation modeling, assess model fit by comparing the sample covariances to those derived from the model. Again, while they are based on the same chi-square distribution, these tests are similar to the Karl Pearson family of tests only in that they compare an observed set of data to what is expected.

The omnibus form of all three tests in the Karl Pearson family of chi-square tests—goodness of fit, independence, homogeneity—use essentially the same formula. Each of these three tests is, in fact, distinct with specific hypotheses, interpretations, and options following rejection of the null hypothesis. The formula for computing the test statistic is as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where n is the number of cells in the table. The obtained test statistic is compared against a critical value from the chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom.

The main difference across each of the three chi-square tests relates to the appropriate situations for which each should be used. The chi-square goodness of fit test is used when a sample is compared on a variable of interest against a population with known parameters. For example, a goodness of fit test might be applied on a survey sample to compare whether the ethnicity or income of the survey respondents is consistent with the known demographic makeup of the geographic locale from which the sample was drawn. The null and alternative hypotheses are:

Hypothesis$_0$: The data follow a specified distribution.
Hypothesis$_A$: The data do not follow the specified distribution.

The interpretation upon rejection is that the sample differs significantly from the population on the variable of interest.

The chi-square test of independence determines whether two categorical variables in a single sample are independent from or associated with each other. For example, a survey might be administered to 1,000 participants who each respond with their hair color and favorite ice cream flavor. The test would then be used to determine whether hair color and ice cream preference are independent of each other. The null and alternative hypotheses are as follows:

$Hypothesis_0$: The variables of interest are independent.
$Hypothesis_A$: The variables of interest are associated.

A significant test rejecting the null hypothesis would suggest that within the sample, one variable of interest is associated with a second variable of interest.

Finally, the chi-square test of homogeneity is used to determine whether two or more independent samples differ in their distributions on a single variable of interest. One common use of this test is to compare two or more groups or conditions on a categorical outcome. A significant test statistic would indicate that the groups differ on the distribution of the variable of interest but does not indicate which of the groups are different or where the groups differ. The null and alternative hypotheses are as follows:

$Hypothesis_0$: The proportions between groups are the same.
$Hypothesis_A$: The proportions between groups are different.

We focus on the practical and important differences between the tests of independence and homogeneity because they are so frequently used in evaluation and applied research studies.

Despite the fact that the formulation of the omnibus test statistic is the same for the test of independence and the test of homogeneity, these two tests differ in their sampling assumptions, null hypotheses, and options following a rejection. The main difference between them is how data are collected and sampled. Specifically, the test of independence collects data on a single sample, and then compares two variables within that sample to determine the relationship between them. The test of homogeneity collects data on two[1] or more distinct groups intentionally, as might be the case in a treatment or intervention study with a comparison group. The two samples are then compared on a single variable of interest to test whether the proportions differ between them. Wickens (1989) presents a thoughtful and succinct description of these tests, as well as their sampling assumptions and hypotheses. In addition to the tests of homogeneity and independence, Wickens presents an additional alternative where both margins are fixed, which he refers to as ''test of unrelated classification.''

When data are collected using only a single sample, only the test of independence is valid and only interpretations of association between variables can be made. When data on two or more samples are collected, the test of homogeneity is appropriate and comparisons of proportions can be made across the multiple groups. When sampling occurs from multiple populations, and thus the homogeneity hypothesis appropriate, it is also reasonable (although less interesting) to ask the independence question.

In the above example regarding hair color and ice cream preference, if the researcher defined the population by hair color and eye color and collected information on 500 brunettes and 500 blondes, these would constitute two independent samples. Comparisons of proportions of blondes and brunettes by their ice cream preferences would be valid. When random assignment is used to assign participants to two or more conditions, these groups are by definition independent and the test of homogeneity may be used to test for differences between the groups.

Table 1. Chi-Square Tests and Attributes

Chi-Square Test Attribute	Test of Independence	Test of Homogeneity	Test of Goodness of Fit
Sampling type	Single dependent sample	Two (or more) independent samples	Sample from population
Interpretation	Association between variables	Difference in proportions	Difference from population
Null hypothesis	No association between variables	No difference in proportion between groups	No difference in distribution between sample and population

Perhaps, these distinctions can be best illustrated by the null hypothesis tested in each of these two tests. The chi-square test of independence null hypothesis states no association between two categorical variables. It can be written as $H_0 : \phi = 0$ or $H_0 : \nu = 0$. This states that the association between two categorical variables, as measured by a Phi (ϕ) correlation for 2×2 contingency tables or with Kramer's V for larger tables, is zero or the variables are independent.

$$\begin{array}{ll} H_0 : \phi = 0 & \\ H_A : \phi \neq 0 & \text{or} \end{array} \quad \begin{array}{l} H_0 : V = 0, \\ H_{\mathrm{A}} : V \neq 0. \end{array}$$

The chi-square test of homogeneity compares the proportions between groups on a variable of interest. The null hypothesis is presented in matrix form:

$$H_0 := \begin{bmatrix} p_{11} = p_{12} = \ldots = p_{1k} \\ p_{21} = p_{22} = \ldots = p_{2k} \\ p_{31} = p_{32} = \ldots = p_{3k} \\ p_{k1} = p_{k2} = \ldots = p_{kk} \end{bmatrix}$$

H_A : The null is false

Rejection of the null hypothesis in the case of three or more groups only allows the researcher to conclude that the proportions between the groups differ, not which groups are different. Table 1 summarizes the distinction between the three types of chi-square tests—specifically, the sampling required for each test, the correct interpretation of each test, and the null hypothesis assumed of each test.

One common misinterpretation of chi-square tests comes from not distinguishing between these three specific tests. Indeed, when most researchers declare that they "utilized a chi-square test," they are typically referring to the chi-square test of independence. This lack of specificity often leads researchers to use interpretations of one test where another was actually conducted. For example, researchers will more often feel compelled to compare the proportions between groups, regardless of how the data were drawn. As is most often the case, the data on two categorical variables are collected from a single sample (e.g., survey data), where the assumptions for chi-square test of homogeneity are not met, and an interpretation comparing proportions between groups is not valid.

Even in those situations where data are drawn from multiple samples and the test of homogeneity is appropriate, researchers seem unaware that procedures exist to specifically follow-up after the rejection of the omnibus test. Consider the following null hypothesis:

$$H_0 : \begin{bmatrix} p_{11} = & p_{12} = & p_{13} \\ p_{21} = & p_{22} = & p_{23} \end{bmatrix}.$$

Table 2. Use of Statistical Tests in Journal Articles

	Total Number of Articles	Articles Using Inferential Statistics	Articles Using Chi- Square Test	Proportion of Articles Using Chi-Square Test (%)
American Journal of Evaluation	65	16	3	18.75
Evaluation Review	61	30	11	36.67
Educational Evaluation and Policy Analysis	52	35	6	17.14
Evaluation and Program Planning	114	26	12	46.15
Total	292	107	32	29.91

A rejection in this case indicates that at least one proportion is different from at least one other proportion.[2] Often, a researcher will conduct a chi-square test, find a significant value, and then look for the cells with the largest disparity in proportions or frequencies to make a substantive interpretation. The proper procedure would involve conducting post hoc comparisons after the omnibus chi-square test to determine where the significant differences actually are. Post hoc procedures for chi-square tests are discussed in a later section.

Chi-square Tests in Recent Evaluation Literature

A brief survey of recent evaluation literature was conducted in order to obtain a general sense of how often chi-square tests are used and how often researchers misinterpret the results.

Surveying the evaluation literature is an approach that has been used by several researchers as a method for better understanding the methods and strategies used in evaluation practice. For example, Greene, Caracelli, and Graham (1989) included published evaluation studies in their sample when reviewing 57 empirical mixed-methods evaluations. Findings from the empirical study were used to refine a mixed-methods conceptual framework that had originally been developed from the theoretical literature and was intended to inform and guide practice. More recently, Miller and Campbell (2006) studied empowerment evaluation in practice by examining 47 case examples published from 1994 through June 2005 to determine the extent to which empowerment evaluation could be distinguished from evaluation approaches emphasizing similar elements, and the extent to which empowerment evaluation led to empowered outcomes for program beneficiaries.

For the current study, four prominent evaluation journals were selected for review: *American Journal of Evaluation*, *Evaluation Review*, *Educational Evaluation and Policy Analysis*, and *Evaluation and Program Planning*. Every article published in these four journals between January 2008 and August 2010 was reviewed. These journals and periods were not intended to be a comprehensive search of the evaluation literature, but mainly to obtain a picture of the prevalence of chi-square tests and the extent to which these tests are incorrectly interpreted. The vast majority of chi-square tests and misinterpretations probably exist in evaluation reports that are never read beyond a small circle of intended users, but we believe that the proliferation of chi-square test misinterpretations is exacerbated by evaluation literature that is read by a larger audience.

After book reviews, section introductions, memoranda, and other editorial content were excluded, there were a total of 292 articles available for review. Two graduate student researchers coded each article on a variety of measures, including whether inferential statistics were used and whether a chi-square test was used. For articles that used a chi-square test, additional codes identified whether the article contained the correct interpretation given the sampling procedure, whether post hoc interpretations were used, and whether post hoc tests were conducted.

Table 2 details the number of articles in each journal as well as how many used inferential quantitative statistics. Overall, just over a third (36.6%; $n = 107$) of the articles used some sort

Table 3. Description of Articles Using Chi-Square Analyses

	Number of Chi-Square Articles	Number of Articles that Used a Valid Chi-Square Test Interpretation		Number of Articles that Used a Post Hoc Interpretation	
	N	*N*	%	*N*	%
American Journal of Evaluation	3	3	100.00	1	33.33
Evaluation Review	11	4	36.36	4	36.36
Educational Evaluation and Policy Analysis	6	2	33.33	2	33.33
Evaluation and Program Planning	12	5	41.67	2	16.67
Total	32	14	43.75	9	28.13

of inferential statistic, ranging from a simple *t* test to more advanced structural equation models. Of the 107 articles that used inferential statistics, 32 articles (29.9%) also used a chi-square test in the Karl Pearson family. *Evaluation and Program Planning* had the most articles employing a chi-square test ($n = 12$) while the *American Journal of Evaluation* had the fewest ($n = 3$).

The 32 articles that used chi-square tests were further reviewed to determine whether the interpretations were justified. Often, researchers were not specific about which chi-square tests were being used (only one of the 32 articles correctly specified the type of chi-square test conducted). To make the determination, then, coders reviewed the Method section in each article to identify which chi-square test would have been appropriate given the sampling design used. The interpretations from the chi-square tests presented in each article were then coded for the types of interpretation used, that is, whether an association claim was made between variables or whether a comparison of proportions was made between groups. This allowed the researchers to determine the type of chi-square test used by the researchers in each article. Any discrepancy between a study's sampling design and the type of chi-square test used was coded as a nonvalid interpretation of the chi-square test. In addition, each of the 32 chi-square articles was coded on whether a post hoc interpretation was used, meaning that the author made comparisons across select rows and columns of the table.

The results from these additional analyses are presented in Table 3. Overall, less than half of the chi-square articles (43.75%; $n = 14$) had interpretations that were justified by the type of chi-square test used. All three articles in the *American Journal of Evaluation* included the correct usage of the chi-square test, whereas only a third (two out of six) of the articles in *Educational Evaluation and Policy Analysis* did so. As shown in Table 3, 9 of the 32 articles that used chi-square (28.1%) included a post hoc interpretation. None of the articles used any post hoc analyses to justify their claims.

Hypothetical Example: Support Components for At-Risk Families

We offer a hypothetical example to illustrate the concepts described above and to guide readers through a proper chi-square post hoc analysis. In this scenario, suppose that researchers are investigating the impact of various family support components for families at risk for child abuse and neglect. Study participants were randomly assigned to receive either parent education/life skills, connections to community resources, or wraparound services made up of the previous components plus case management. Using the county data system, a sample was drawn from each of these three conditions. The dependent variable of interest consisted of 4 outcomes measures 12 months after the families' initial involvement with Child Protective Services (CPS): (a) a CPS rereferral; (b) a substantiated allegation; (c) the child's removal from home; or (d) no further involvement with CPS.

Table 4. Involvement with CPS and Service Conditions

	Parent Education *N*, Col %	Community Resources *N*, Col %	Wraparound *N*, Col %	Total *N*, Col %
Rereferral to CPS	38, 20.43	42, 22.34	49, 13.73	129, 17.65
Substantiated allegation	24, 12.9	18, 9.57	35, 9.8	77, 10.53
Child removed	27, 14.52	8, 4.26	15, 4.2	50, 6.84
No new involvement with CPS	97, 52.15	120, 63.83	258, 72.27	475, 64.98
Total	186	188	357	731

Note. CPS = child protective services.

While randomization is often used to form independent groups, it is not a prerequisite for the appropriate use of the test for homogeneity. What is required is that the groups are identified and sampled intentionally. Table 4 shows the distribution with involvement with CPS across the three conditions. The null hypothesis is as follows:

$$H_0 : \begin{bmatrix} p_{11} = p_{12} = p_{13} \\ p_{21} = p_{22} = p_{23} \\ p_{31} = p_{32} = p_{33} \\ p_{41} = p_{42} = p_{43} \end{bmatrix},$$

$$H_A : \text{The null is false.}$$

The obtained $X_6^2 = 36.77$ is significant at the conventional α level of .05. The justified interpretation following the rejection of the null hypothesis would be to conclude that the proportions are not equal across the three groups.

Often at this point, researchers will conclude that the proportions are not equal and will want to compare specific conditions. For example, they might examine the ''no new involvement'' row and conclude that the wraparound condition (72.3%) is preferable to the parent education (52.2%) or community resources (63.8%) condition. Alternatively, a researcher may be interested in comparing the proportion of children removed across the conditions. It might be tempting to conclude that parent education (14.5%) is significantly different from community resources (4.26%) and wraparound (4.2%). However, this interpretation would be incorrect because there is no statistical justification for these claims based solely on the results of the omnibus test; the omnibus test indicates only that the conditions are significantly different but not which conditions are different.

Because the chi-square test is an omnibus test, post hoc procedures would need to be conducted in order to compare individual conditions. As previously mentioned, the procedure for comparing conditions or groups was developed by Goodman (1963).[3] Similar to the comparison procedures following an analysis of variance (ANOVA), several different approaches—including Scheffé, Holm,[4] and Dunn-Bonferroni—are available for selecting the appropriate critical value. Also similar to the ANOVA, the comparison often takes on the name associated with formulation of the critical value. For purposes of this article, the Scheffé post hoc values are presented because this represents the most conservative approach. For an alternative approach based on Dunn-Bonferonni, see Marasculio and Serlin (1988).

The Goodman procedure is described below. The test statistic for each contrast is as follows:

$$\frac{\hat{\psi}}{\sqrt{SE_{\psi}^2}} = Z.$$

The same equation in an expanded form is as follows:

$$\frac{\hat{\psi}}{\sqrt{SE_{\psi}^2}} = \frac{w_1(p_1) - w_2(p_2)}{\sqrt{w_1^2\left(\frac{p_1q_1}{n_1}\right) + w_2^2\left(\frac{p_2q_2}{n_2}\right)}} = Z,$$

where $\hat{\psi}$ represents the linear combination of weights (W_k) and proportions (θ_k) of the specific contrast:

$$\psi = W_1\theta_1 + W_2\theta_2 + \cdots + W_k\theta_k,$$
where
$$W_1 + W_2 + \cdots + W_k = 0.$$

And the numerator of the test is the square root of the weighted standard error of the contrast:

$$SE_{\psi}^2 = W_1^2SE_{\theta_1}^2 + W_2^2SE_{\theta_2}^2 + \cdots + W_k^2SE_{\theta_k}^2.$$

The standard error of each column is the standard error of an estimated proportion:

$$SE_{\theta}^2 = \frac{p_kq_k}{N_k}.$$

Once the obtained test statistic is found for a comparison of interest, it is compared to a critical value. The Scheffé critical value is found by taking the square root of the critical value in the original omnibus chi-square analysis. In the above example, the chi-square omnibus critical value at the conventional α level of .05 with $(r-1)(c-1) = (4-1)(3-1) = 6$ degrees of freedom is 12.59. The square root of this critical value is $S^* = \sqrt{\chi^2_{v:1-\alpha}} = \sqrt{12.59} = \pm 3.55$ which represents the Scheffé critical value for all contrasts.

Referring back to our previous example, comparing wraparound (72.3%) to parent education (52.2%) on "no new involvement" leads to the following hypothesis:

$$Hypothesis_0 : p_{\text{No new involvement/wraparound}} = p_{\text{No new involvement/parent education}},$$
$$Hypothesis_A : p_{\text{No new involvement/wraparound}} \neq p_{\text{No new involvement/parent education}}.$$

The appropriate test statistic is as follows:

$$\frac{\left(\frac{357}{357}\right)(.7227) - \left(\frac{186}{186}\right)(.5215)}{\sqrt{\left(\frac{357}{357}\right)^2\left(\frac{(.7227)(.2773)}{357}\right) + \left(\frac{186}{186}\right)^2\left(\frac{(.5215)(.4785)}{186}\right)}} = \frac{.2012}{.0436} = 4.61.$$

Since this is a pairwise comparison, the weights $\frac{357}{357}$ and $\frac{186}{186}$ equal 1, and essentially dropout of the equation both in the numerator and in the denominator. Given $4.61 > \pm 3.55$, we reject and conclude that there is a statistically significant difference between these conditions.

Comparisons can be performed within any row. If the researcher wanted to compare wraparound (4.2%) to parent education (14.5%) on whether a child was removed, "child removed," the test statistic is given by

Table 5. Pairwise Contrasts from Hypothetical Example

	ψ	*SE*	TS
Rereferral			
Wraparound versus parent education	−.0670	.0347	−1.931
Wraparound versus community resources	−.0861	.0354	−2.432
Parent education versus community resources	−.0191	.0424	−0.451
Substantiated abuse			
Wraparound versus parent education	−.0310	.0292	−1.062
Wraparound versus community resources	.0023	.0306	0.075
Parent education versus community resources	.0333	.0326	1.020
Child removed			
Wraparound versus parent education	−.1031	.0279	−3.693
Wraparound versus community resources	−.0005	.0182	−0.030
Parent education versus community resources	.1026	.0297	3.451
No new case opened			
Wraparound versus parent education	.2012	.0436	4.612
Wraparound versus community resources	.0844	.0423	1.995
Parent Education versus community resources	−.1168	.0507	−2.304

$$\frac{\left(\frac{357}{357}\right)(.042)-\left(\frac{186}{186}\right)(.1452)}{\sqrt{\left(\left(\frac{357}{357}\right)\right)\left(\frac{(.042)(.958)}{357}\right)+\left(\frac{186}{186}\right)\left(\frac{(.1452)(.8548)}{186}\right)}}=\frac{-.1031}{.0278}=-3.69.$$

Given $-3.69 > \pm 3.55$, we reject and conclude that there is a statistically significant difference between these conditions. A comparison between community resources (4.26%) and parent education (14.5%) produces a test statistic of 3.45 and is not significant due to the differing sample sizes and their impact on the standard error. This is an instance where simply examining the difference between the proportions, without conducting the appropriate post hoc test, might lead to a statistically unsupported conclusion. In both of these, the comparisons the difference between the parent education and the other two conditions were .10. However, in one case, there was a significant difference and in the other there was no difference based on the critical value. A complete listing of all pairwise comparisons is available in the Table 5 at the end of article.

As noted previously, comparisons under this model are not limited to being pairwise. The post hoc procedure can also be used to test complex contrasts. Suppose you want to compare wraparound to the combination of parent education and community resources.

$$\frac{\left(\frac{357}{357}\right)(.1373)-\left[\left(\frac{186}{374}\right)(.2043)+\left(\frac{188}{374}\right)(.2234)\right]}{\sqrt{\left(\frac{357}{357}\right)^2\left(\frac{(.1373)(.8657)}{357}\right)+\left[\left(\frac{186}{374}\right)^2\left(\frac{(.2043)(.7957)}{186}\right)+\left(\frac{188}{374}\right)^2\left(\frac{(.2234)(.7766)}{188}\right)\right]}}$$

$$=\frac{-.0766}{.0273}=-2.81.$$

Unlike with the previous pairwise contrast weights, the combination of parent education and community resources needs to be weighted for their respective contributions. Once this is done, the

test statistic is calculated as it was before. Given $-2.81 < \pm 3.55$, we do not reject and conclude that there is not a statistically significant difference between the wraparound condition and the combination of parent education and community resources.

Discussion

Common misconceptions of the chi-square test were clarified in this article. Specifically, we have distinguished between the members of the Karl Pearson family of chi-square tests and presented post hoc procedures. Evaluators often need to examine the association between categorical variables or to compare groups or conditions on a categorical outcome, which explains their prevalence in evaluation literature and reports. However, effective use of the chi-square test, or any other statistical test for that matter, is dependent on a clear understanding of the assumptions of the test and what is actually being tested (null hypothesis) in the statistical procedure.

A correct interpretation of the chi-square test or of other statistical procedures is often dependent on factors outside of distributional assumptions and characteristics of the data itself—for example, individual observations must be independent from other observations in the contingency table. When this is this case, an interpretation of the chi-square test is based on sampling procedures and how data were collected. Furthermore, since the asymptotic approximation of the chi-square test is less precise at the extreme end of the distribution, expected values of cells need to be greater than five.

The review of the evaluation literature reveals that in about half of the instances where a chi-square test was used, the wrong interpretation was presented. The appropriate interpretation of the results is directly tied to the null hypothesis under test and the interpretation—whether independence or homogeneity—is limited to that hypothesis. More commonly, researchers prefer to interpret the chi-square test of homogeneity by comparing groups across a variable of interest. However, the sampling procedure precludes the researcher from making this claim and has thus misinterpreted the results of the chi-square test.

Researchers also tend to over interpret the results of statistical tests. An omnibus chi-square test informs us that the distribution of observed values deviates from expected values, but does not tell us where the discrepancy is located in the contingency table. Often, researchers will make naïve comparisons between two or more groups without conducting any post hoc tests to determine whether the contrasts were significant.

Many more complex statistical models exist and we have faith that these procedures are still being faithfully and thoughtfully applied. Although the chi-square tests were found to be commonly misinterpreted in recent evaluation literature, the results of these studies are not wrong. Rather, the problem is simply that there is often no statistical justification for some of the claims being made. However, Goodman's procedure is computationally simple and there is little reason it cannot be conducted to justify significant contrasts. Our hope in this article is that researchers and evaluators will be more thoughtful in using common statistical procedures and more carefully consider what their results actually say.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The two-sample test of proportions, which uses the Z distribution, is a special case of the test of homogeneity, employed when you have only two groups.

2. Comparisons in this context are limited to pairwise contrasts. It is perfectly feasible that Groups 2 and 3 combined are from Group 1 and responsible for the significant result.
3. The approach presented here builds logically on the post hoc procedures following multiple group comparisons in analysis of variance (ANOVA) models. Goodman's approach is not the only one available for addressing pairwise comparisons, however. See Seaman and Hill (1996), Gardner (2000), and Delucchi (1993).
4. Information on the use of the Holm procedure, see Holm, 1979.

References

Delucchi, K. L. (1993). On the use and misuse of chi-square. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 295–319). Hillsdale, NJ: Lawrence Erlbaum.

Gardner, R. C. (2000). *Psychological statistics using SPSS for Windows*. Upper Saddle River, NJ: Prentice Hall.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*, 255–274.

Goodman, L. (1963). Simultaneous confidence intervals for contrasts among multinomial populations. *The Annals of Mathematical Statistics*, *35*, 716–725.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

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

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation*, *27*, 296–319. doi:10.1177/109821400602700303

Seaman, M. H., & Hill, C. C. (1996). Pairwise comparisons for proportions: A note on Cox and Key. *Educational and Psychological Measurement*, *56*, 452–459.

Stigler, S. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Wickens, T. D. (1989). *Multiple contingency tables analysis for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.