Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 8, April 2015 ISSN 1531-7714

Your Chi-Square Test is Statistically Significant: Now What?

Donald Sharpe, *University of Regina*

Applied researchers have employed chi-square tests for more than one hundred years. This paper addresses the question of how one should follow a statistically significant chi-square test result in order to determine the source of that result. Four approaches were evaluated: calculating residuals, comparing cells, ransacking, and partitioning. Data from two recent journal articles were used to illustrate these approaches. A call is made for greater consideration of foundational techniques such as the chi-square tests.

Congratulations! After collecting frequency or categorical data, you want to know if more cases fell into one category (i.e., goodness of fit) or if two variables are related based on the distribution of cases (i.e., independence). How do you answer these types of questions? For more than 100 years the choice has been clear --- the chi-square tests.[1] Chi-square tests remain popular. In a survey by Bakker and Wicherts (2011) of six randomly selected psychology journals for 2008, 642 chi-square tests were reported. So you conducted a chi-square test on your frequency data and the result is statistically significant. Is that all there is to it or is there something more that needs to be done?

Omnibus Test

When a chi-square test result is associated with more than one degree of freedom (i.e., larger than a 2 x 2 contingency table for the chi-square test of independence; three or more cells for the chi-square test of goodness of fit), the source of a statistically significant result is unclear. For a chi-square test of goodness of fit with three cells *a*, *b,* and *c*, is statistical significance the product of a difference between cells *a* and *b*? Or cells *a* and *c*? Or cells *b* and c? For a 2(r) x 3(c) chi-square test of independence, is the source of dependence between *r1* and *r2* versus *c1* and *c*2? Or *r1* and *r2* versus *c1* and *c3*? Or *r1* and *r2* versus *c2* and *c3*? While Thompson (1988), Delucchi (1993), and Franke, Ho, and Christie (2012) acknowledged the *omnibus* nature of the chi-square tests, none of these authors made post-hoc testing their focus. Similarly, authors of popular statistics textbooks (e.g., Gravetter & Wallnau, 2013) are largely silent on follow-up tests to chi-square analyses.

While textbook authors are largely silent, perhaps researchers address the omnibus nature of the chi-square tests in their articles? Abstracts of journals published by the *American Psychological Association* for 2012, 2013, and early 2014 were searched via *PsycINFO* for the phrase *chi-square*. Thirteen articles were identified. Taken together, the authors of the thirteen articles conducted 121 chi-square tests (one article had 32 chi-square tests; another 21 chi-square tests). Of these 121 chi-square tests, 34 tests had greater than one degree of freedom. In almost all of those 34 cases, authors did nothing further, ignoring Beasley and Schumacker's (1995) assertion that "no chi-square test should stop with the computation of an omnibus chi-square statistic" (p. 80).

A few of the thirteen authors followed their omnibus test results by *eyeballing the data*. For example, Landis, Barrett, and Galvin (2013) reported a statistically significant chi-square test with eight degrees of freedom based on a 5 x 3 contingency table. Landis

and colleagues were interested in different models of care in a family medicine residency program. The authors compared a standard Collocated Behavioral Health Services (CL) model to a more integrated Primary Care Behavioral Health (PCBH) model and to a Blended Model (BM) that combines elements from the PCBH model with the use of a dedicated manager. One of several socio-demographic variables evaluated in this context was how a patient's care was funded: Medicare, Medicaid, Medicare/Medicaid, Commercial insurance, or Self-Pay. Our Table 1 replicates the data from Landis et al.[2] Of their 15 cells, Landis and colleagues zeroed in on the column percentage of *one* cell as being the source of the statistically significant chi-square: "[P]atients in the BM group were most likely to have commercial insurance (Pearson χ^2 = 18.89, df = 8, p = .015; see Table 1)" (p. 268).

Thompson (1988) regards it to be "logically inconsistent for a researcher to declare that the omnibus null hypothesis must be evaluated statistically, but then to decide that the cell counts in the contingency table will be evaluated by subjective inspection to determine if the null hypothesis was rejected because of counts in a particular cell or in some aggregate of cells" (p. 42). *Subjective inspection* is *eyeballing the data*. MacDonald and Gardner (2000) concur with Thompson. They regard it to be a "serious abuse" to fail to "empirically evaluate individual cell contributions to a statistically significant chi-square result" (p. 737). Yet many applied researchers appear to side with Ludbrook (2011) who argues for eyeballing the data and sees further statistical analysis of chi-square contingency tables to be a "waste of time" (p. 925).

There are at least four approaches available to investigate further a statistically significant omnibus chi-square test result.[3] The first and easiest of the four procedures is *calculating residuals*. A residual analysis identifies those specific cells making the greatest contribution to the chi-square test result. A second procedure, *comparing cells*, evaluates whether specific cells differ from each other. Calculating residuals and comparing cells work for both chi-square tests of goodness of fit and independence. A third procedure, *ransacking*, involves testing the 2 x 2 interactions of greatest interest based on *post-hoc* examination of cell frequencies or *a priori* hypotheses. A fourth procedure, *partitioning*, is the systematic collapsing of the complete r x c contingency table into an orthogonal set of 2 x 2 tables and then testing those 2 x 2 tables for statistical significance.

Table 1. Edited SPSS Output Based on Data from Landis et. al. (2013)

		Column			
Row		CL	PCBH	BM	Marginals
Medicare	Obs	24	53	3	80
	Exp	24.6	50.6	4.8	
	Column %	14.1%	15.1%	9.1%	
	Res	-.6	2.4	-1.8	
	Std. Res	-.1	.3	-.8	
	Adj. Res	-.2	.6	-.9	
Medicaid	Obs	44	57	5	106
	Exp	32.6	67.1	6.3	
	Column %	25.9%	16.3%	15.2%	
	Res	11.4	-10.1	-1.3	
	Std. Res	2.0	-1.2	-.5	
	Adj. Res	**2.7**	**-2.3**	-.6	
Medc/ Meda	Obs	19	18	2	39
	Exp	12.0	24.7	2.3	
	Column %	11.2%	5.1%	6.1%	
	Res	7.0	-6.7	-.3	
	Std. Res	2.0	-1.3	-.2	
	Adj. Res	**2.5**	**-2.3**	-.2	
Com- mercial	Obs	63	165	20	248
	Exp	76.2	157.0	14.8	
	Column %	37.1%	47.1%	60.6%	
	Res	-13.2	8.0	5.2	
	Std. Res	-1.5	.6	1.4	
	Adj. Res	**-2.5**	1.4	1.9	
Self-Pay	Obs	20	57	3	80
	Exp	24.6	50.6	4.8	
	Column %	11.8%	16.3%	9.1%	
	Res	-4.6	6.4	-1.8	
	Std. Res	-.9	.9	-.8	
	Adj. Res	-1.2	1.6	-.9	
Marginals		170	350	33	553

Note. Adjusted residuals in bold are those that exceed +/- 2. CL = collated care, PCBH = primary care behavioral health, BM = blended model.

Calculating Residuals

Delucchi (1993) recommends a researcher identify those cells with the largest *residuals*. A residual is the difference between the observed and expected values for a cell. The larger the residual, the greater the contribution of the cell to the magnitude of the resulting chi-square obtained value. As stated by Agresti (2007), "a cell-by-cell comparison of observed and estimated expected frequencies helps us to better

understand the nature of the evidence" and cells with large residuals "show a greater discrepancy…than we would expect if the variables were truly independent" (p. 38).

Raw residuals are the product of subtracting expected from observed values. Turning again to Table 1, the BM by Commercial cell highlighted by Landis and colleagues (2013) had an observed value of 20 and an expected value of 14.8. Thus, the raw residual for that cell is 5.2. However, cells with the largest expected values also produce the largest raw residuals. To overcome that redundancy, a *standardized* or *Pearson* residual is calculated by dividing the raw residual by the square root of the expected value as an estimate of the raw residual's standard deviation:

$$Std\ Residual = (O - E) / \sqrt{E} \qquad 1.1$$

For the BM by Commercial cell, the standardized residual is equal to 20 minus 14.8 and that sum divided by the square root of 14.8 equals 1.4. The sum of all squared standardized residuals is the chi-square obtained value. There is also what Agresti (2013) calls a standardized residual but SPSS calls an *adjusted standardized residual* of the form:

$$Adj\ Residual = \frac{(O-E)}{\sqrt{E*(1-RowMarginal/n)*(1-ColumnMarginal/n)}} \qquad 1.2$$

The RowMarginal refers to the row marginal for the cell. The ColumnMarginal refers to the column marginal for the cell. The *n* refers to the total number of cases across all cells. The denominator of the adjusted residual equation is the estimated standard error rather than the estimated standard deviation of the residual. For the BM by Commercial cell, the adjusted standardized residual is 20 minus 14.8 divided by the square root of 14.8 times (1 – 248/553) times (1 – 33/553) which equals 1.9.

Table 1 presents raw, standardized, and adjusted residuals derived from the data of Landis et al. (2013). According to Agresti (2007; see Haberman, 1973), "a[n adjusted] standardized residual having absolute value that exceeds about 2 when there are few cells or about 3 when there are many cells indicates lack of fit of H_o in that cell" (p. 38). Five cells were associated with adjusted residuals greater than +/- 2 (no cells produced residuals greater than +/- 3). Four of those five cells were the product of CL or PCBH by Medicaid or Medc/Meda (the combination of Medicare and Medicaid). The two cells associated with CL had positive adjusted residual values, indicating that there were *more* participants in the CL condition for Medicaid and Medc/Meda than would be expected by chance. Conversely, the two cells associated with PCBH had negative adjusted residual values, indicating that there were *fewer* participants in the PCBH condition for Medicaid and Medc/Meda than would be expected by chance. Recall Landis et al. (2013) emphasized the BM by Commercial cell, presumably because that cell had the largest column percentage (60.6%). However, that cell's adjusted residual value is 1.9, which fails to exceed the +/- 2 criteria.

MacDonald and Gardner (2000) suggest a Bonferroni adjustment to the z critical of 1.96 (from which the +/- 2 criteria is derived) if the number of cells in the contingency table is large. In the Landis et al. (2013) example, there are 15 cells in their 3 x 5 contingency table. Thus, alpha should be set at .05/15 or .003 which translates into a critical value of +/- 2.96 (or approximately +/- 3). However, if the magnitude of the residuals merely serves as a guide to what cells might be of interest, then arguably no adjustment is necessary or one could choose a more conservative alpha value than .05 such as .01 (+/- 2.58). SPSS provides raw, standardized, and adjusted residuals; see Field (2013, pp. 743-744).

Comparing Cells

A second approach compares specific cells for a statistically significant difference. Comparing cells is an approach that works for chi-square tests of goodness of fit and independence, and the approach can be conceptualized as *a priori* or *post-hoc* depending on whether or not it is preceded by an omnibus chi-square test.

Marascuilo and Serlin (1988; chapter 28), Delucchi (1993), and Franke et al. (2012) all argue for comparing cells by following the pioneering work of Goodman (1969; 1971). Using the data from Landis et al. (2013; again see Table 1), one might compare for Commercial the observed cell frequencies of 63 for CL versus 20 for BM. To do so, a z test should be calculated:

$$z = (\Psi - 0) / SE_{\Psi} \qquad 1.3$$

with Ψ being the contrast of interest and SE_{Ψ} being the standard error of that contrast. Comparing CL vs. BM for Commercial, the contrast would be of the kind:

$$\Psi_{CL\ vs.\ BM} = p_{CL} - p_{BM} \quad 1.4$$

with p_{CL} being the proportion in the CL cell relative to its column marginal and p_{BM} being the proportion in the BM cell relative to its column marginal or 63/170 – 20/33 = .3706 - .6060 = -.2354. The squared standard error is equal to:

$$SE^2_{\Psi CL\ vs.\ BM} = (1)^2 SE^2_{pCL} + (-1)^2 SE^2_{pBM} \quad 1.5$$

Formula 1.5 converts to:

$$SE^2_{\Psi CL\ vs.\ BM} = \left[1/N_{CL}\left(p_{CL} * q_{CL}\right)\right] + \left[1/N_{BM}\left(p_{BM} * q_{BM}\right)\right] \quad 1.6$$

with p_{CL} again being the proportion of the frequency in a cell associated with CL, q_{CL} being the proportion of the frequency in a cell *not* associated with CL (i.e., 63 cases of the 170 in the CL condition had commercial insurance and thus 107 of the 170 did not), and N_{CL} being the total number of cases in the CL condition. The same follows for BM. Thus, we calculate [1/170 (63/170)(107/170)] + [1/33 (20/33)(13/33)] = .0014 + .0072 = .0086. The square root of .0086 converts SE^2 to SE or .0927. Accordingly, the z obtained value is -.2354/.0927 = -2.54. The z obtained value is *not* tested against a z critical value for alpha = .05 of +/- 1.96; instead it is tested against the square root of the chi-square critical value for the entire contingency table (Marascuilo & Serlin, 1988). Given the square root of the chi-square critical value for 8 degrees of freedom (15.51) from the Landis et al. (2013) example is +/- 3.94 and the z obtained is -2.54, we fail to reject the null hypothesis that the frequencies associated with CL and BM for Commercial differ.

Marascuilo and Serlin (1988) note that determining the chi-square critical value from the entire contingency table is a conservative procedure; "it distributes the risk of Type I error over an infinite set of contrasts, for which only a small number are meaningful and interpretable" (p. 371). Marascuilo and Serlin suggest determining the number of contrasts of interest ahead of time. Given our interest is for Commercial whether BM differs from CL and perhaps whether BM differs from PCBH, the corresponding z critical value is the square root of the chi-square critical value for two degrees of freedom or the square root of 5.99 which is 2.45. Given our z obtained of -2.54 is larger than +/- 2.45, we conclude that for Commercial, the proportion for BM is indeed greater than the proportion for CL.

The most recent versions of SPSS have an option within Crosstabs under Cells to calculate z tests for column proportions for each row in a chi-square contingency table. There is also an option to adjust those z tests for each row using a Bonferroni correction.

Ransacking

A third approach to post-hoc analysis of a contingency table is *ransacking* (Goodman, 1969). One might look for a 2 x 2 table of interest within a larger r x c contingency table and then evaluate that 2 x 2 table for statistical significance. DeViva (2014) compared military veterans for treatment engagement: never seen for therapy, seen but not completing therapy, and completed therapy. One variable that was crossed by treatment engagement was marital status: single/divorced or married. Turning to the data from DeViva reproduced in Table 2, the Never Seen and Completed Therapy by Single/Divorced and Married interaction is of greatest interest according to the adjusted residuals.

Table 2. Edited SPSS Output Based on Data from DeViva (2014)

		Column			
		Never Seen	Seen, Didn't Complete	Com-pleted	Mar-ginals
Single/ Divorced	Obs	57	53	11	121
	Exp	48.9	56.2	15.9	
	Col %	77.0%	62.4%	45.8%	
	Res	8.1	-3.2	-4.9	
	Std. Res	1.2	-0.4	-1.2	
	Adj. Res	**2.6**	-1.0	**-2.3**	
Married	Obs	17	32	13	62
	Exp	25.1	28.8	8.1	
	Col %	23.0%	37.6%	54.2%	
	Res	-8.1	3.2	4.9	
	Std. Res	-1.6	0.6	1.7	
	Adj. Res	**-2.6**	1.0	**2.3**	
Marginals		74	85	24	183

Note. Absolute residuals in bold are those that exceed +/- 2.

Looking at the Never Seen and Completed columns, the *observed odds* for Single/Divorced is 57/11 = 5.18. Conversely, the observed odds for Married is 17/13 = 1.31. Thus, the *odds ratio* is 5.18/1.31 = 3.95. The *log odds ratio* (the natural log of 3.95) or *G* is 1.37. If intervention status by marital status are independent

(unrelated), then the odds ratio should be 1 and the log odds ratio should be 0.

The test of the interaction via the log odds ratio is a z test of the form G (the log odds ratio) divided by the standard error of G. The standard error of G (SE_G) is calculated by:

$$SE_G = \left(h_{11} + h_{22} + h_{12} + h_{21}\right)^{½} \qquad 1.7$$

with h_{11} being the inverse of the frequency in row1, column1, and so on. In our example, SE_G would be equal to $(1/57 + 1/11 + 1/17 + 1/13)^{1/2} = .4941$. Thus, the z obtained value is 1.37/.4941 = 2.77. If we proposed testing this 2 x 2 contingency table *a priori*, then the z critical value of 1.96 would be appropriate. However, if we selected those four cells *post-hoc* from a 3 x 2 contingency table with 2 degrees of freedom, then the critical value for z should not be 1.96 but rather the square root of the chi-square critical value for 2 degrees of freedom (5.99) or 2.45. Given the obtained value of 2.77 exceeds the critical value of 2.45, we would reject the null hypothesis of independence for this 2 x 2 contingency table.

Marscuilo and Serlin (1988), following from Goodman (1969), provide an alternative means of calculating G that is more consistent with how interaction contrasts from a factorial ANOVA are calculated using weights of +1, -1, -1, and +1 (see Jaccard & Guilamo-Ramos, 2002). Marscuilo and Serlin (1988) calculate an interaction contrast for the 2 x 2 contingency table of the form:

$$G = ln\, p_{11} - ln\, p_{12} - ln\, p_{21} + ln\, p_{22} \qquad 1.8$$

with ln p_{11} being the natural log of the frequency in row 1, column 1, and so on. In our example, G = ln 57 – ln 11 – ln 17 + ln 13 = 4.04 – 2.40 – 2.83 + 2.57 = 1.37.

Does this seem like a lot of work? Set aside the *Seen, Didn't Complete* column and run a Likelihood Ratio chi-square test on the resulting 2 x 2 contingency table. The Likelihood Ratio chi-square test is a long-standing alternative to the Pearson chi-square that compares observed frequencies with frequencies predicted by a model based on expectations estimated by maximum likelihood (Cochran, 1952; see Ruxton & Neuhauser, 2010, for a comparison of the Likelihood Ratio and Pearson chi-square tests). Like Pearson's chi-square, the Likelihood Ratio chi-square is available in SPSS and other statistical packages. The 2 x 2 Likelihood Ratio chi-square from DeViva's 2014 data calculated by SPSS is $L\chi^2$ (1) = 7.86. The square root of this Likelihood Ratio chi-square (subject to rounding) is our z obtained value of 2.77.

One issue with ransacking as described above is that it *pretends* the 2 x 2 contingency table is the original data source rather than the 2 x 3 contingency table. In the analogous procedure for factorial ANOVA (again, see Jaccard & Guilamo-Ramos, 2002), one would calculate the denominator for the interaction contrast using a standard error calculated for *all* cells (i.e., the 2 x 3 contingency table). Thus, for any interaction contrast calculated following a factorial ANOVA, the numerator would differ depending on the specific cells implicated, but the denominator would be a constant. Following this logic, an adjustment might be made to Formula 1.7 to include all the cells from the 2 x 3 contingency table. Finally, if ransacking is done multiple times (especially *post-hoc*) for a large contingency table, some adjustment should be made for alpha inflation (e.g., Bonferroni).

Partitioning

The fourth approach is *partitioning*, an approach that involves dividing contingency tables of greater than 2 x 2 into a *set* of smaller 2 x 2 subtables. According to Fisher (1925), there are many ways to partition a table mathematically, with only some partitions being of interest. However, there may be value in systematically creating a set of *orthogonal* partitions which will be uncorrelated or independent from each other. One advantage of orthogonal over non-orthogonal partitions is that the Type I error rate can be known precisely for the orthogonal partitions. There are two disadvantages, however: the number of orthogonal partitions is limited by the degrees of freedom for the original contingency table and many orthogonal partitions may be of little substantive interest (see Thompson, 1990). Nonetheless, Hays (1994) stated in the general case, whenever a researcher conducts more than one comparison from a set of data, "the questions involved in the respective comparisons cannot be given truly separate and unrelated answers unless the comparisons are statistically independent of each other" (pp. 433-434).

Lancaster (1949) provided a method for partitioning large chi-square contingency tables and also the means for determining whether the partitioned subtables are orthogonal. Again turning to DeViva's

(2014) data presented in Table 2, the 2 x 3 contingency table with two degrees of freedom allows for two orthogonal 2 x 2 subtables. These subtables are depicted in Table 3. The upper 2 x 2 subtable in Table 3 is from the left corner of the 2 x 3 contingency table (although the choice of starting corner is arbitrary). A chi-square test of independence is calculated on those four cells. The lower 2 x 2 subtable in Table 3 results from collapsing the cells already tested and comparing those collapsed cells against the remaining two cells. Again, a chi-square test of independence is calculated on these four cells.

Table 3. Collapsing a 2 x 3 Table Based on Data from DeViva (2014)

	Never Seen	Seen Didn't Complete	Completed
Single/Divorced	57	53	11
Married	17	32	13

$\chi^2 = 3.99, p < .046; L\chi^2 = 4.05, p < .044$

	Never Seen	Seen Didn't Complete	Completed
Single/Divorced	57	53	11
Married	17	32	13

$\chi^2 = 5.08, p < .024; L\chi^2 = 4.81, p < .028$

Note. Cells analyzed are represented by boxes with thick lines.

According to Lancaster (1949), the sum of the chi-square obtained values for appropriately collapsed subtables will equal the chi-square obtained value for the contingency table as a whole. However, the Pearson chi-square for DeViva's (2014) 2 x 3 contingency table is $\chi^2 = 8.88$; the sum of the two Pearson chi-squares in Table 3 (3.99 + 5.08) equals 9.07, not 8.88. Shaffer (1973) argues that the resulting sum of the partitioned chi-squares will only *approximate* the overall chi-square obtained value unless one calculates Likelihood Ratio chi-square tests rather than Pearson chi-square tests. The Likelihood Ratio chi-squares of 4.05 and 4.81 sum to the value of the Likelihood Ratio chi-square for the complete Table 2 of $L\chi^2 = 8.86$.

Lancaster's (1949) approach works for larger contingency tables. Turning again to the Landis et al. (2013) data, assume that Medicare, Medicaid and Medc/Meda by CL, PCBH, and BM from Table 1 form a 3 x 3 contingency table. The Pearson chi-square for the 3 x 3 contingency table is $\chi^2 = 5.14$ and the Likelihood Ratio chi-square is $L\chi^2 = 5.19$. Table 4 presents the partitioning of that 3 x 3 contingency table into 2 x 2 subtables, and chi-square tests derived according to Lancaster's partitioning. Again, the

Table 4. Collapsing a 3 x 3 Table Based on Data from Landis et al. (2013)

	CL	PCBH	BM	
Medicare	24_a	53_b	3_c	r1
Medicaid	44_d	57_e	5_f	r2
Medc/Meda	19_g	18_h	2_i	r3
	c1	c2	c3	

$\chi^2 = 2.84, p < .09; L\chi^2 = 2.87, p < .09$

	CL	PCBH	BM
Medicare	24	53	3
Medicaid	44	57	5
Medc/Meda	19	18	2

$\chi^2 = .10, p < .75; L\chi^2 = .105, p < .75$

	CL	PCBH	BM
Medicare	24	53	3
Medicaid	44	57	5
Medc/Meda	19	18	2

$\chi^2 = 2.20, p < .14; L\chi^2 = 2.165, p < .14$

	CL	PCBH	BM
Medicare	24	53	3
Medicaid	44	57	5
Medc/Meda	19	18	2

$\chi^2 = .05, p < .82; L\chi^2 = .05, p < .82$

Note. Cells analyzed are represented by boxes with thick lines. Subscripts are used to identify specific cells or row/column marginals. CL = collated care, PCBH = primary care behavioral health, BM = blended model.

Likelihood Ratio chi-squares for the subtables sum to the Likelihood Ratio chi-square for the complete table; however, the Pearson chi-squares do not.

Agresti (2013; see also Iversen, 1979) summarizes the rules for partitioning based on Lancaster (1949) and also Goodman (1969). According to Agresti (2013), the first rule is "The df for the subtables must sum to the df for the full table" (p. 84). In our example (see Table 4), the degrees of freedom for the 3 x 3 full table is four --- and we have four subtables after partitioning. The second rule is "Each cell count in the full table must be a cell count in one and only one subtable" (p. 84). Turning again to Table 4, our first subtable addresses cells a, b, d, and e; our second subtable addresses cells c and f; our third subtable addresses cells g and h; our fourth subtable addresses cell i. The third rule is "Each marginal total of the full table must be a marginal total for one and only one subtable" (p. 84). Our first subtable addresses no marginal totals; our second subtotal addresses row marginal totals r1 and r2; our third subtable addresses column marginal totals c1 and c2; our fourth subtable addresses column marginal total c3 and row marginal total r3. Finally, Agresti cautions that "for a certain partitioning, when the subtable df values sum properly but the G^2 [Likelihood Ratio] values do not, the components are not independent" (p. 84). Our Likelihood Ratio chi-squares sum properly.

Conclusion

In their discussion of chi-square tests, Lewis and Burke (1949) identified a common error: *Questionable or Incorrect Categorizing*. Lewis and Burke wrote "In any investigation where the χ^2 test is to be applied, the categories must be established in a logically defensible and reliable manner before the data are collected, if possible" (p. 463). That was sound advice in 1949 and it remains sound today. If you can avoid chi-square contingency tables with greater than one degree of freedom, you should do so. For example, a researcher might collapse or discard low frequency cells after collecting the data but prior to conducting a chi-square test.

If one cannot avoid chi-square contingency tables with greater than one degree of freedom, this paper presented four approaches for addressing the issue of omnibus chi-square testing. Typically post-hoc procedures for chi-square are predicated on a statistically significant omnibus chi-square test. Criticism of Null Hypothesis Significance Testing (NHST) has been widespread and vigorous, with increasing emphasis on the reporting of effect size statistics in addition to (or in place of) *p* values. Like many statistics, the chi-square statistic is a measure of effect size confounded by sample size (Haddock, Rindskopf & Shadish, 1998). Frequently authors report the phi coefficient as the measure of effect size following a chi-square test. However, the phi coefficient (as well as Cramer's *V*, identical to the phi-coefficient for a 2 x 2 contingency table) is strongly affected by differences in the row and column marginals and therefore underestimates the magnitude of the effect (see Breaugh, 2003; Haddock et al., 1998). For example, if the columns of a 2 x 2 contingency table represent gender and 90% of the participants are female, the resulting phi coefficient is attenuated such that it mathematically cannot approach its maximum value of one regardless of the strength of the relationship between the column variable (i.e., gender) and a row variable. Haddock et al. (1998) recommend reporting the odds ratio over a phi coefficient as a measure of effect size for 2 x 2 contingency tables. As noted by Kline (2013), the odds ratio "may be the least intuitive of the comparative risk effect sizes, but it probably has the best overall statistical properties" (p. 169). For example, an odds ratio of one rather than zero as with the phi coefficient is indicative of no relationship between two variables. Odds ratios of 1.49, 3.45, and 9 are equivalent to Cohen's (1992) .10, .30, and .50 for the phi coefficient (Oliver & Bell, 2013), but Ferguson (2009) suggests odds ratios of 2, 3, and 4 better correspond with small, medium, and large effects. See Kline (2013) for other related measures of effect size for categorical outcomes.

Chi-square tests are by far the most popular of the non-parametric or distribution free tests and the default choice when applied psychological researchers analyze categorical data. Chi-square tests along with correlations, t-tests, and ANOVA, are foundational techniques, covered in introductory statistics textbooks and introductory statistics classes. Nevertheless, these foundational techniques and especially the chi-square tests are rarely discussed in journals devoted to advanced statistical methods. Iverson (1979) spoke of partitioning of chi-square contingency tables as a *forgotten technique* more than 35 years ago, attributing this forgetting to a lack of awareness by applied researchers of that approach. Instead, applied researchers and

methodologists alike are distracted by newer, *sexier* statistics. While the chi-square tests will never be considered sexy, these tests remain important and useful methods for applied researchers seeking to evaluate categorical data.

Footnotes

[1] The focus here is on Pearson's (1900) chi-square tests --- not the other uses of the chi-square *distribution*, for example in logistic regression or structural equation modeling.

[2] A reviewer commented that the Landis et al. (2013) example has cells with expected frequencies less than five. The *no cells with expected frequencies less than five rule* can be traced to Fisher (1925). Cochran (1954) regarded that rule to be "too conservative" (p. 418), resulting in unacceptable power loss, and recommended instead a set of working rules such that no cells should have an expected frequency less than one and no more than 20% of cells should be between one and five. Delucchi (1993) regards Cochran's rule to be "a fair balance between practicality and precision" (p. 301); Ruxton and Neuhauser (2010) concur but note the situation is complex and "it is not easy to come up with a rule-of-thumb that captures this complexity, is not overly restrictive or liberal and is easy to apply" (p. 1507). The Landis et al. (2013) example has exactly 20% of cells (three of fifteen) with expected frequencies less than five so Cochran's (1954) rule is not violated.

[3] One additional approach to the omnibus test problem recommended by Shaffer (1973), Delucchi (1983), and Streiner and Lin (1998) is to replace chi-square testing with log-linear analysis. Log-linear analysis resembles analysis of variance. It works for r x c contingency tables as well as for multidimensional contingency tables. More than thirty years ago, Delucchi (1983) wrote: "It is not difficult to argue that log-linear models will eventually supersede the use of Pearson's chi-square in the future because of their similarity to [the familiar] analysis of variance (ANOVA) procedures and their extension to higher order tables" (p. 169). While log-linear analysis is available in statistical packages such as SPSS and is discussed in popular sources such as Field (2013), Delucchi's (1983) prediction has failed to come to pass in so far as chi-square tests remain the default choice for analyzing categorical data.

References

Agresti, A. (2007). An introduction to categorical data analysis. Hoboken, NJ: Wiley.

Agresti, A. (2013). Categorical data analysis (3rd ed.). Hoboken NJ: Wiley.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. Behaviour Research Methods, 43, 666-678. doi.org/10.1037/a0032008

Beasley, T. M., & Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables; Post hoc and planned comparison procedures. Journal of Experimental Education, 64, 79-93. doi.org/10.1080/00220973.1995.9943797

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. Journal of Management, 29, 79-97. doi.org/10.1177/014920630302900106

Cochran, W. (1952). The χ^2 test of goodness of fit. Annals of Mathematical Statistics, 23, 315-345. doi.org/10.1214/aoms/1177729380

Cochran, W. (1954). Some methods for strengthening the common χ^2 tests. Biometrics, 10, 417-451. dx.doi.org/10.2307/3001616

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159. doi.org/10.1037/0033-2909.112.1.155

Delucchi, K. L. (1983). The use and misuse of chi-square – Lewis and Burke revisited. Psychological Bulletin, 94, 166-176. doi.org/10.1037/0033-2909.94.1.166

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. 294-319). Hillsdale, NJ: Lawrence Erlbaum.

DeViva, J. C. (2014). Treatment utilization among OEF/OIF veterans referred for psychotherapy for PTSD. Psychological Services, 11, 179-184. doi.org/10.1037/a0035077

Ferguson, C. (2009). An effect size primer: A guide for clinicians and researchers. Professional Psychology: Research and Practice, 40, 532-538. doi.org/10.1037/a0015808

Field, A. (2013). Discovering statistics using SPSS (4th ed.). London, UK: Sage.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver & Boyd.

Franke, T. M., Ho, T., & Christie, C. A. (2012). The chi-square test: Often used and more often misinterpreted. American Journal of Evaluation, 33, 448-458. doi.org/10.1177/1098214011426594

Goodman, L. A. (1969). How to ransack mobility tables and other kinds of cross-classification tables. American Journal of Sociology, 75, 1-40. doi.org/10.1086/224743

Goodman, L. A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. Journal of the American Statistical Association, 66, 339-344. doi.org/10.1080/01621459.1971.10482265

Gravetter, F. J., & Wallnau, L. B. (2013). Statistics for the behavioral science (9th ed.). Belmont, CA: Wadsworth.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. Biometrics, 29, 205-220. doi.org/10.2307/2529686

Haddock, C. K., Rindskopf, D. & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. Psychological Methods, 3, 339-353. doi.org/10.1037/1082-989X.3.3.339

Hays, W. L. (1994). Statistics (5th ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Iversen, G. R. (1979). Decomposing chi-square: A forgotten technique. Sociological Methods and Research, 8, 143-157. doi.org/10.1177/004912417900800202

Jaccard, J., & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical and adolescent psychology: Issues and recommendations. Journal of Clinical Child and Adolescent Psychology, 31, 130-146. doi.org/10.1207/153744202753441747

Kline, R. (2013). Beyond significance testing: Statistics reform in the behavioral sciences (2nd ed.). Washington, DC: American Psychological Association.

Lancaster, H. O. (1949). The derivation and partition of χ^2 in certain discrete distributions. Biometrika, 36, 117-129. doi.org/10.2307/2332535

Landis, S. E., Barrett, M, & Galvin, S. L. (2013). Effects of different models of integrated collaborative care in a family medicine residency program. Families, Systems and Health, 31, 264-273. doi.org/10.1037/a0033410

Lewis, D., & Burke, C. J. (1949). The use and misuse of the chi-square test. Psychological Bulletin, 46, 433-489. doi.org/10.1037/h0059088

Ludbrook, J. (2011). Is there still a place for Pearson's chi-squared test and Fisher's exact test in surgical research? Australia and New Zealand Journal of Surgery, 81, 923-926. doi.org/10.1111/j.1445-2197.2011.05906.x

MacDonald, P. L., & Gardner, R. C. (2000). Type 1 error rate comparisons of post hoc procedures for I J chi-square tables. Educational and Psychological Measurement, 60, 735-754. doi.org/10.1177/00131640021970871

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

Oliver, J., & Bell, M. L. (2013). Effect sizes for 2 x 2 contingency tables. PLOS One, 8, e5877. doi.org/10.1371/journal.pone.0058777

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 reasonably supposed to have risen from random sampling. Philosophical Magazine Series 5, 50, 157-175. doi.org/10.1080/14786440009463897

Ruxton, G. D. & Neuhauser, M. (2010). Good practice in testing for an association in contingency tables. Behavior Ecology and Sociobiology, 64, 1505-1513. doi.org/10.1007/s00265-010-1014-0

Shaffer, J. P. (1973). Testing specific hypotheses in contingency tables: Chi-square partitioning and other methods. Psychological Reports, 33, 343-348. doi.org/10.2466/pr0.1973.33.2.343

Siegel, S., & Castellan, N. J. (1988). Non-parametric statistics for the behavioral sciences (2nd ed.). McGraw Hill.

Streiner, D. L., & Lin, E. (1998). Life after chi-squared: An introduction to log-linear analysis. Canadian Journal of Psychiatry, 43, 837-842.

Thompson, B. (1988). Misuse of chi-square contingency-table test statistics. Educational and Psychological Research, 8, 39-49.

Thompson, B. (1990). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: The neo-classical perspective. Retrieved from ERIC database. (ED318753).

Acknowledgement:

The author thanks Nick Carleton, Alyssa Counsell, Rob Cribbie, Cathy Faye, and Sarah Sangster for their helpful comments and suggestions on earlier versions of this article.

Citation:

Sharpe, Donald (2015). Your Chi-Square Test is Statistically Significant: Now What? *Practical Assessment, Research & Evaluation*, 20(8). Available online: http://pareonline.net/getvn.asp?v=20&n=8

Author:

Donald Sharpe
Department of Psychology
University of Regina
Regina, SK
Canada S4S 0A2.

E-mail: sharped [at] uregina.ca