TYPE I ERROR RATE COMPARISONS OF POST HOC PROCEDURES FOR *I* × *J* CHI-SQUARE TABLES

PAUL L. MACDONALD AND ROBERT C. GARDNER
University of Western Ontario

The authors used Monte Carlo methods to assess the per-contrast and experimentwise Type I error rates of two post hoc tests of cellwise residuals and four post hoc tests of pairwise contrasts in 3 × 4 chi-square contingency tables. The six post hoc procedures were evaluated under three sample sizes and under the null hypotheses of independence and homogeneity. Results of the study indicate that the cellwise adjusted residual method provided adequate experimentwise Type I error rate control when appropriate adjustments to the alpha level were made, and the Gardner pairwise post hoc procedure provided several advantages over the other pairwise procedures. This was true for both the independence and homogeneity models.

One of the oldest statistical tools still being used by social scientists is Karl Pearson's chi-square test for two-dimensional *I* × *J* contingency tables. It has been suggested that the chi-square test may be the best known and most important nonparametric statistical procedure (Mouly, 1978; Popham & Sirotnik, 1973). In fact, early empirical studies (Edgington, 1974) of actual use of the chi-square procedure in seven American Psychological Association (APA) journals found that the chi-square test was employed in 10% to 19% of the published articles reviewed between the years 1948 and 1972. Reviews since then have found the chi-square procedure used in 10% and 17% of the research articles in two educational journals (Goodwin & Goodwin, 1985) and 21% of the articles in a personality assessment journal (Schinka,

This research was conducted while the first author was supported by a University of Western Ontario Scholarship and while the second author was funded by a Social Science and Humanities Research Council of Canada Grant 410-99-0147. We thank Mike Ashton for helpful comments on previous versions of this article. Correspondence should be addressed to Paul MacDonald, Department of Psychology, University of Western Ontario, London, Ontario, Canada N6A 5C2; e-mail: macdonal@julian.uwo.ca.

Educational and Psychological Measurement, Vol. 60 No. 5, October 2000 735-754

LaLone, & Broeckel, 1997). This relative popularity appears not to have waned. We recently reviewed four journals (*Journal of Applied Social Psychology*, *Journal of Experimental Social Psychology*, *Journal of Consulting and Clinical Psychology*, and *Journal of Educational Psychology*) and found that between 10% and 39% of the articles published between the years 1995 and 1997 still made use of chi-square analyses of frequency data.

In practice, the omnibus chi-square test can be conceptualized in either of two ways, both of which evaluate the distribution of cell observations within an $I \quad J$ contingency table. In the first perspective (which is referred to as the homogeneity model), the J columns of the contingency table are formed by sampling from J separate and distinct populations for which observations are made on a single variable with I levels (e.g., students from different countries classified by college major). As a result, the column totals are fixed (not necessarily equal), and the omnibus chi-square test on the $I \quad J$ contingency table can be considered a test of the homogeneity of the proportions for a given level across the populations. In the second perspective (which is referred to as the independence model), the contingency table is formed by using two variables (i.e., I and J) to cross classify members of one population (e.g., college students cross classified by major and political affiliation). In this case, the omnibus chi-square test can be considered a test of independence (or association) between two cross-classified variables within the contingency table (Marascuilo & Serlin, 1988). For both the homogeneity model and the independence model, the omnibus chi-square test statistic for an $I \quad J$ table takes the following form:

$$\chi^2 \quad \sum_{i\ 1}^{I} \sum_{j\ 1}^{J} \frac{f_{ij} \quad F_{ij}^{\ 2}}{F_{ij}},$$

where f_{ij} and F_{ij} are the observed and expected frequency values in a cell. The expected frequency for each cell of the $I \quad J$ contingency table is calculated as the product of the respective row and column frequencies ($f_{i.}$ and $f_{.j}$, respectively) and divided by N (the total number of observations). Under the null hypothesis, this test statistic is distributed approximately as the chi-square theoretical probability distribution with $\nu = (I-1)(J-1)$ degrees of freedom.

The null hypothesis for the omnibus test of homogeneity of proportions can be expressed as

$$H_o \quad \begin{matrix} 11 & 12 & \cdots & 1J \\ 21 & 22 & \cdots & 2J \\ \cdots & \cdots & \cdots & \cdots \\ I1 & I2 & \cdots & IJ \end{matrix},$$

where $_{ij}$ is defined as the proportion of the observations located in the ith row and jth column of the chi-square table to the total number of observations in

the *j*th column. The value of $_{ij}$ can be estimated by the sample data from the contingency table with p_{ij}, which involves the ratio of f_{ij}/N_j, where f_{ij} is the frequency of observations in the *i*th row and *j*th columns of the table and N_j is the sum of the observations in the *j*th column. If the null hypothesis for the omnibus test of homogeneity is rejected at the specified alpha level of statistical significance, it refers to the entire table but provides no information as to *why* it is false.

The null hypothesis for the omnibus test of independence states that the cell frequencies obey the marginal restrictions. If this null hypothesis is rejected, it again provides little information as to *why* it is false.

The focus of the present article is to assess the per-contrast and experimentwise Type I error rates associated with methods that have been proposed to further investigate a statistically significant omnibus chi-square statistic. The per-contrast error rate is defined in terms of the total number of contrasts made. Thus,

$$\text{Per-Contrast Error} = \frac{\text{Number of falsely significant contrasts}}{\text{Total number of contrasts}}.$$

The experimentwise error rate is defined in terms to the number of experiments in which at least one false significant contrast occurs. Thus,

$$\text{Experimentwise Error} = \frac{\text{Number of experiments in which at least one falsely significant contrast occurs}}{\text{Total number of experiments}}.$$

Our intent is to provide a review and demonstration of two principal chi-square post hoc approaches that can be found in statistical textbooks. The first approach involves tests of statistical significance for individual cells of the $I \quad J$ contingency table, and the second approach involves pairwise comparison tests. Our tests are applied to both types of omnibus tests, those assessing homogeneity of proportions and those evaluating association. Although the omnibus chi-square test has the same form for both, the underlying sampling model is different. The question is, does this difference affect the validity of subsequent post hoc procedures?

Individual Cells

Thompson (1988) cites several common misuses of the omnibus chi-square test statistic, including one serious abuse, namely, the failure to empirically evaluate individual cell contributions to a statistically significant chi-square result. He further added that researchers often evaluate cell contributions subjectively (e.g., a visual observation that a particular cell's obtained value is higher or lower than its expected value) but neglect to explore the contributions statistically. A review of the literature on post hoc

comparisons for the omnibus chi-square test has revealed two procedures directed at individual cells of the $I \quad J$ contingency table.

Standardized Residual Method

The first procedure is based on work by Haberman (1973, 1978) and has been described elsewhere (see Beasley & Schumacker, 1995; Hinkle, Wiersma, & Jurs, 1988; Marascuilo & McSweeney, 1977; Reynolds, 1984; Siegel & Castellan, 1988). Frequency data within a contingency table can be used to calculate the standardized residual for each cell as

$$e_{ij} \quad \frac{f_{ij} \quad F_{ij}}{\sqrt{F_{ij}}},$$

which Beasley and Schumacker (1995) claim approximates a unit normal distribution. The discrepancy between the observed and expected frequencies can be used to determine which cells within the contingency table generate residual scores that are larger in magnitude than might be expected by chance. Standardized cell residuals that exceed a greater-than-two rule of thumb (suggested originally by Haberman, 1973) are considered to contribute to the rejection of the null hypothesis for the omnibus chi-square test to a statistically significant degree.

Beasley and Schumacker (1995) caution that the unit normal table should only be used to calculate test statistics after an appropriate adjustment to maintain the experimentwise Type I error rate. They recommended the Sidak (1967) method to adjust the alpha level with

$$\alpha_{adj} = 1 - (1 - \alpha)^{1/c},$$

where c equals the number of cells within the omnibus chi-square table. The adjusted alpha level can then be used to calculate the appropriate critical test statistic to determine which cells contributed to the omnibus chi-square statistic at a statistically significant level.

Adjusted Residual Method

The second post hoc procedure developed to test individual cells within a contingency table is also an extension of the earlier work by Haberman (1973, 1978). The test statistic for the second method approximates a standard normal distribution with the form

$$z_{ij} \quad \frac{f_{ij} \quad F_{ij}}{\sqrt{F_{ij}(1 \quad f_{i.}/N)(1 \quad f_{j}/N)}},$$

where f_{ij} and F_{ij} are the observed and expected cell frequencies, and the respective row and column marginals are depicted by $f_{i.}$ and $f_{.j}$ (Agresti, 1996). After the adjusted residual statistics are calculated for each cell, the greater-than-two rule of thumb is again applied to determine which cells contribute to the rejection of the omnibus chi-square test statistic at a statistically significant level.

Pairwise Comparisons

An alternative approach to investigating the cellwise residuals of a statistically significant omnibus chi-square test is to conduct multiple pairwise comparisons for tests of proportions. For example, consider the case of an I J contingency table. In the chi-square homogeneity model (i.e., in which the columns represent samples from J distinct populations), the J column observations can be contrasted as $\frac{I}{2}$ pairwise comparisons of the proportions of observations in each row relative to the column total. As a result, there would be $I \quad \frac{I}{2}$ possible post hoc pairwise column contrasts within an $I \quad J$ contingency table. Alternatively, in the case of the chi-square independence model, a researcher may be interested in not only column-wise contrasts of proportions of the column totals but also of row-wise contrasts of the proportions of row totals. As a result, for each level of the column variable, the I levels of the second variable can be contrasted as $\frac{I}{2}$ pairwise row contrasts.

To illustrate this distinction, consider a 3 4 contingency table formed with samples from four distinct populations (e.g., college students from four different countries) that were measured on three levels of a variable of interest (e.g., college major). In total, there would be six pairwise tests contrasting the four countries in terms of the proportion of students in each country taking that major. As there are three levels of college major, this would create a total of 18 pairwise post hoc contrasts for the complete contingency table under the homogeneity model. Whereas consider a 3 4 contingency table created by cross classifying two variables (e.g., college students on major and political affiliation). There would be a total of 18 possible column-wise tests contrasting the four political affiliations at each level of major and a total of 12 row-wise tests contrasting the three majors at each level of political affiliation. This would create a total of 30 pairwise post hoc contrasts for the complete contingency table under the independence model.

A review of the literature on pairwise post hoc contrasts following significant omnibus chi-square tests reveals three distinct approaches to testing pairwise differences of proportions. Each approach is distinguishable by the particular test statistic it uses and the method for controlling the experimentwise Type I error rates. Of primary interest to this research are the

particular characteristics of the test statistics and their ability to maintain experiment Type I error rates at nominal levels.

The first approach is demonstrated in Delucchi (1983, 1993), who attributes the procedure to work found in Gart (1962), Gold (1963), and Goodman (1964b). This approach involves calculating a *Z* statistic based on the formula

$$Z = \frac{p_{ij} - p_{ij'}}{\sqrt{\frac{p_{ij}(1 - p_{ij})}{n_j} - \frac{p_{ij'}(1 - p_{ij'})}{n_{j'}}}}$$

where p_{ij} is defined as the proportion of the observations located in the *i*th row and *j*th column of the chi-square table to the total number of observations in the *j*th column. Furthermore, $p_{ij'}$ is defined as the proportion of the observations located in the *i*th row and *j* th column of the chi-square table to the total number of observations in the *j* th column. This is a homogeneity procedure that contrasts the proportions of individuals in the two cells of interest for a specific level (row or column) of a particular variable. To control Type I error rates at nominal levels, Delucchi (1983) recommends using $\sqrt{\chi^2 v, 1 - \alpha}$ as the critical value, where *v* is defined as $(J - 1)$ degrees of freedom, where *J* equals the number of populations sampled in tests of homogeneity. In the case of the 3×4 contingency table, the critical value for this approach would be $\sqrt{\chi^2 3, .95}$, which equals 2.796 for all column-wise contrasts. If this test was applied to the independence model, the row-wise contrasts would also be of interest with the critical value reflecting $(I - 1)$ degrees of freedom. For the 3×4 contingency table, the critical value for all pairwise contrasts within the rows of the $I \times J$ contingency table would be $\sqrt{\chi^2 2, .95}$, which equals 2.448.

The second approach to performing pairwise post hoc contrasts is demonstrated in Marascuilo and Serlin (1988), who base their procedure on the work of Goodman (1964a). Nearly identical to the first approach, the only difference between the Marascuilo and Serlin approach and the Delucchi (1983) approach is the critical value for the statistic. Whereas Delucchi uses a critical value from a chi-square distribution with $(J - 1)$ [or $(I - 1)$] degrees of freedom, the second approach uses a chi-square critical value based on $(I - 1)(J - 1)$ degrees of freedom. In the case of the 3×4 contingency table, the resulting critical value would be $\sqrt{\chi^2 6, .95}$, which is equal to 3.549 for all column-wise and row-wise contrasts.

The third approach to the post hoc contrasts can be found in Seaman and Hill (1996), who also credit their procedure to Goodman (1964a). Similar to the omnibus chi-square test statistic, this method also involves calculating a

statistic of the form $\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{ij} - F_{ij})^2}{F_{ij}}$, where $F_{ij} = (f_{\cdot j} \quad f_{i\cdot})/N$, the expected value in a cell. An important distinction of this approach is that the chi-square statistic is calculated on a reduced 2 2 contingency table created by selecting only the two columns of the omnibus chi-square contingency table of interest. In the event that $I > 2$, the rows of the 2 2 reduced table are formed by selecting the specific row of the original contingency table and collapsing the remaining $(I - 1)$ rows. The critical value for this procedure is calculated as $\chi^2(v, 1 - \alpha)$, in which v is equal to $(I - 1)(J - 1)$, the degrees of freedom for the omnibus chi-square test of homogeneity. In the case of the 3 4 contingency table, the critical value would be $\chi^2(6, .95)$, which is equal to 12.592 for all contrasts.

A fourth approach, which has not been considered previously as a post hoc procedure, involves a Bonferroni adjustment to the alpha level of the analysis. This approach was suggested by Gardner (in press) and has demonstrated some success in preliminary simulations (Lam, 1996). This method calculates a post hoc chi-square statistic of a collapsed 2 2 contingency table, as identified in the Seaman and Hill (1996) approach. The Gardner procedure can be distinguished from the third approach by its Bonferroni type critical value, in which the probability of a Type I error is spread over all possible pairwise contrasts. Recall that in the case of a 3 4 omnibus contingency table, it was calculated that there were 18 possible column-wise post hoc comparisons of proportions and 12 possible row-wise post hoc comparisons. As such, the Type I error rate is spread over all 18 contrasts in the case of the chi-square homogeneity model to create a two-tailed Bonferroni critical value for $\chi^2(1, 1 - .05/18)$ equal to 8.948. In the case of the chi-square independence model, the error rate is spread over all 30 possible contrasts to create a two-tailed Bonferroni critical value of $\chi^2(1, 1 - .05/30)$ equal to 9.885.

Overview of the Present Study

In the present study, Monte Carlo methods were employed to determine the extent to which different post hoc procedures were able to maintain the experimentwise Type I error rates at nominal levels for chi-square contingency tables based on the homogeneity and independence models. Monte Carlo investigators have not yet evaluated the empirical error rates of these procedures, and such findings are important for researchers who want to perform post hoc tests following a statistically significant omnibus chi-square test statistic. The intent of this research is to estimate the per-contrast and experimentwise Type I error rates of the six distinct post hoc procedures available to researchers, which include (a) two cellwise contrasts and (b) four pairwise contrasts.

Method

For the independence model, the random contingency tables were created by first generating a 3 4 chi-square contingency table whose cells were labeled 1 through 12 and in which each cell had a starting observed frequency value of zero. To assign *N* observations (e.g., 100) to the chi-square table, the computer program generated *N* numbers varying from 1 to 12. For each number, the cell whose label corresponded to the randomly generated value had a count of one added to its observed frequency score. This step was repeated until the sum of the observations for the omnibus chi-square contingency table reached the desired sample size.

For the homogeneity model, the random contingency tables were created by first generating a 3 4 contingency table whose cells within each column were labeled 1 through *I* and had a starting observed frequency value of zero. To assign the *N*/*J* observations to the first column, the computer program randomly selected a number from 1 through *I*. The cell in the first column whose label corresponded to the randomly generated value had a count of one added to its observed frequency. This step was repeated until all *N*/*J* observations were assigned to the first column of the contingency table. This process was repeated sequentially for the remaining columns of the contingency table until all *N* observations were randomly assigned to the *J* columns.

For the independence and homogeneity models, the omnibus chi-square was computed on each contingency table, as well as the two cellwise post hoc procedures for each of the 12 cells in the 3 4 table and the four pairwise post hoc procedures for the 30 possible column and row pairwise contrasts. The number of test statistics that exceeded the corresponding critical value for that procedure was summed for each table. The number of statistically significant tests for each procedure was then summed across all 10,000 random chi-square contingency tables to yield the total number of rejections. This value was divided by the total number of statistics calculated (e.g., 120,000 for the cellwise procedures) to compute the per-contrast Type I error rate. In addition, a count was made of the number of tables with at least one statistically significant test. This value was then divided by the total number of tables (i.e., 10,000) to yield the experimentwise Type I error rate. This process was performed for sample sizes of 100, 300, and 500 for each procedure.

Results

Overall Chi-Square

The results for the omnibus chi-square test are presented in Table 1 and Table 2. Table 1 presents the mean, variance, skewness, and kurtosis of the empirical distribution for sample sizes of 100, 300, and 500 for both the inde-

Table 1
Empirical Distribution Characteristics of the Omnibus Chi-Square Test Statistic

	Statistics			
Model/Sample Size	Mean	Variance	Skewness	Kurtosis
Independence model				
100	6.072	11.761	1.158	5.288
300	6.034	12.146	1.173	5.165
500	6.051	12.397	1.211	5.217
Homogeneity model				
100	6.091	11.681	1.074	4.554
300	5.971	11.736	1.192	5.455
500	6.009	11.655	1.131	4.953

Note. Entries are calculated on the obtained test statistic (χ^2) based on 10,000 randomly generated chi-square contingency tables.

Table 2
*Empirical Type I Error Rates for the Omnibus 3 ∞4 Chi-Square Test for the Nominal .05 Significance Level (*df = *6)*

Model/Sample Size	Number of Significant Chi-Square Tables	Type I Error Rate
Independence model		
100	486	.049
300	519	.052
500	537	.054
Homogeneity model		
100	526	.053
300	463	.046
500	464	.046

Note. Entries are based on a count of the obtained test statistics (χ^2) that exceeded the nominal critical value at the .05 significance level for 10,000 randomly generated chi-squares for each simulated condition.

pendence model and the homogeneity model. For the 3 4 omnibus chi-square test, the expected values are 6 for the mean, 12 for the variance, 1.15 for the skewness, and 5 for the kurtosis (Evans, Hastings, & Peacock, 1993). As can be seen, the obtained values are very close to the expected values for both the independence and homogeneity models.

Table 2 illustrates the empirical Type I error rates for the omnibus chi-square test for the independence and homogeneity models at the nominal .05 level. In the case in which the total sample size was 100, we found that 486 of the 10,000 randomly generated contingency tables in the independence model obtained a statistically significant chi-square test statistic. As a result, the empirical Type I error rate was found to be .049. For all sample sizes in the independence model, the empirical value for the Type I error rate closely

approaches the .05 nominal level, which was expected given that the null hypothesis was true. The stability of the chi-square Type I error rate was also demonstrated in the homogeneity model for all sample sizes in this study.

Cellwise Methods

The results of the standardized and adjusted residual methods are illustrated in Table 3 and Table 4 for the independence and homogeneity models. Table 3 presents the mean, variance, skewness, and kurtosis of the empirical distribution for sample sizes of 100, 300, and 500. For a standardized normally distributed statistic, the expected values are 0, 1, 0, and 3, respectively (Evans et al., 1993). As can be seen, the variance for the standardized residual method was appreciably less than 1.0 for all sample sizes, indicating that this statistic did not have a standard unit normal distribution for either the independence or homogeneity models. The obtained values for the adjusted residual method were much closer to the value of 1.0 expected for the variance. All other statistics approximate their expected values for both models.

Table 4 presents the number of rejections and the per-contrast and experimentwise Type I error rates for the two tests of cell frequencies (i.e., the standardized residual and the adjusted residual methods) for both the independence model and the homogeneity model. The table presents a count of the number of cell residuals that were statistically significant (i.e., cellwise rejections) and the number of tables that had at least one statistically significant cell residual (i.e., tables with at least one rejection). The former was used to calculate the per-contrast error rate by dividing this value by 12,000 (i.e., the number of cells 10,000). Thus, for the standardized residual approach within the independence model, the per-contrast error rate for the sample size of 100 was .005 (648/120,000). The number of tables with at least one rejection was used to compute the experimentwise error rate. Where the sample size was equal to 100, we found that 581 of the 10,000 randomly generated contingency tables obtained at least one residual post hoc contrast whose obtained Z value exceeded the critical value of 1.96. This corresponded to an experimentwise Type I error rate of .058 for the standardized residual post hoc method. When a Bonferroni adjustment was made to reflect the number of cellwise contrasts, the new critical value ($Z_{crit} = 2.86$) resulted in the standardized residual method becoming overly conservative. The experimentwise error rate with the adjusted critical value was less than .001 for all sample sizes and models.

The second cellwise approach, the adjusted residual method, yielded per-contrast Type I error rates that were close to the nominal value but experimentwise Type I error rates that were considerably higher than the nominal alpha level for the unadjusted critical value ($Z_{crit} = 1.96$). Note in Table 4 that for all sample sizes, the experimentwise error rate exceeded .36

Table 3
Empirical Distribution Characteristics of the Cellwise Residual Test Statistic

	Standardized Residual Method				Adjusted Residual Method			
Model/Sample Size	Mean	Variance	Skewness	Kurtosis	Mean	Variance	Skewness	Kurtosis
Independence model								
100	.000	.504	.097	2.966	.000	1.007	.090	2.943
300	.000	.506	.041	2.981	.000	1.012	.041	2.977
500	.000	.505	.038	2.987	.000	1.010	.037	2.985
Homogeneity model								
100	.000	.500	.089	2.946	.000	1.000	.085	2.928
300	.000	.502	.052	3.002	.000	1.005	.051	2.996
500	.000	.500	.041	2.997	.000	1.001	.041	2.993

Note. Entries are calculated on the obtained test statistic (Z) based on the 12 cells in the 10,000 randomly generated chi-square contingency tables.

Table 4
The Empirical Type I Error Rates for the Standardized and Adjusted Residual Post Hoc Tests of Proportions

		Standardized Residual Method				Adjusted Residual Method			
				Type I Error Rates				Type I Error Rates	
Model/Sample Size	Z_{crit}	Rejections	Tables With at Least One Rejection	Per Contrast	Experimentwise	Rejections	Tables With at Least One Rejection	Per Contrast	Experimentwise
Independence model									
100	1.96	648	581	.005	.058	6,020	3,726	.050	.373
300		672	570	.006	.057	6,246	3,760	.052	.376
500		671	581	.006	.058	6,147	3,798	.051	.380
100	2.86	2	2	.000	.000	513	450	.004	.045
300		2	2	.000	.000	498	429	.004	.043
500		8	8	.000	.001	513	450	.004	.045
Homogeneity model									
100	1.96	595	529	.005	.053	6,012	3,762	.050	.376
300		699	605	.006	.061	6,084	3,764	.051	.376
500		668	574	.006	.057	6,043	3,686	.070	.369
100	2.86	4	4	.000	.000	441	398	.004	.040
300		4	4	.000	.000	527	463	.004	.046
500		3	3	.000	.000	505	446	.004	.045

Note. Entries are based on a count of the obtained test statistics (Z) that exceeded the nominal critical values associated with .05 significance level, based on 10,000 randomly generated chi-square contingency tables.

for both the chi-square model of homogeneity and independence. In contrast, the Bonferroni adjustments to the critical value (Z_{crit} = 2.86) generated very conservative per contrast error rates but experimentwise error rates that were maintained at the nominal level. For both chi-square models and for all sample sizes in the study, the Bonferroni experimentwise error rates for the adjusted residual method were just slightly below the nominal .05 alpha level. These error rates ranged from a low of .040 to a high of .046.

The results of the two cellwise post hoc procedures provide important information for researchers. In the event that researchers are interested in analyzing further a statistically significant omnibus chi-square test, the experimentwise error rate can be maintained at nominal levels by both the standardized and adjusted residual methods if the correct adjustment to the alpha level is performed. In the case of the standardized residual method, the Bonferroni adjustment to the error rate should not be performed. Whereas for the adjusted residual method, the Bonferroni adjustment should be performed to maintain the experimentwise Type I error rate at nominal levels. Once the appropriate alpha adjustment procedures are performed (or not) for the residual post hoc methods, both procedures provide similar control of the experimentwise error rates. The standardized residual method had slightly inflated experimentwise error rates, whereas the adjusted residual method had slightly conservative experimentwise error rates.

Pairwise Method

The results of the pairwise post hoc procedures are presented in Tables 5 through 7. Table 5 presents the mean, variance, skewness, and kurtosis of the empirical distributions for sample sizes of 100, 300, and 500 for the independence and homogeneity chi-square models. For the *Z* statistic method, the expected values are 0, 1, 0, and 3, respectively, whereas the expected values for the chi-square statistic method are 1, 2, 2.83, and 15 (Evans et al., 1993). As can be seen, the obtained values for the *Z* statistic and the chi-square statistic were very close to the expected values.

The empirical Type I error rates for the four pairwise post hoc procedures for the chi-square homogeneity model are presented in Table 6. Inspection of the probabilities reveal that for all procedures, the per-contrast error rates were conservative, whereas the experimentwise error rates varied from one procedure to another. The Delucchi (1983) procedure provided experimentwise error rates that were inflated, whereas the procedures proposed by Marascuilo and Serlin (1988) and Seaman and Hill (1996) provided experimentwise error rates that were substantially below the .05 nominal level. The procedure proposed by Gardner (in press) yielded experimentwise error rates that were closer to the nominal value. In the case in which the sample size was 100, the experimentwise error rates were .116 for the Delucchi procedure, .016 for the Marascuilo and Serlin procedure, .004 for the Seaman and Hill procedure, and .040 for the Gardner procedure.

(text continues on p. 751)

Table 5
Empirical Distribution Characteristics of the Pairwise Post Hoc Test Statistics

	Z Statistic Method				χ^2 Statistic Method			
Model/Sample Size	Mean	Variance	Skewness	Kurtosis	Mean	Variance	Skewness	Kurtosis
Independence model								
100	.000	1.085	.003	3.203	1.021	2.014	2.738	14.370
300	.000	1.037	.001	3.053	1.009	2.020	2.801	14.746
500	.000	1.023	.010	3.028	1.010	2.048	2.873	15.638
Homogeneity model								
100	.000	1.076	.004	3.163	1.032	2.059	2.685	13.421
300	.000	1.029	.002	3.065	1.002	1.983	2.774	14.431
500	.000	1.015	.002	3.044	1.008	2.005	2.767	14.130

Note. Entries are calculated on the obtained test statistic (*Z*) based on the 12 cells in the 10,000 randomly generated chi-square contingency tables.

Table 6
Empirical Type I Error Rates for the Pairwise Post Hoc Procedures Residuals Tests in the Homogeneity Model

Method/Sample Size	Critical Value	Rejections	Tables With at Least One Rejection	Type I Error Rates	
				Per Contrast	Experimentwise
Delucchi (1983) procedure					
100	$Z = 2.796$	1,594	1,163	.009	.116
300		1,140	832	.006	.083
500		1,060	798	.006	.080
Marascuilo and Serlin (1988) procedure					
100	$Z = 2.796$	171	157	.001	.016
300		110	96	.001	.010
500		93	90	.001	.009
Seaman and Hill (1996) procedure					
100	$\chi^2 = 12.592$	39	36	.000	.004
300		63	58	.000	.006
500		68	62	.000	.006
Gardner (in press) procedure					
100	$\chi^2 = 8.948$	489	396	.003	.040
300		461	367	.003	.037
500		502	402	.003	.040

Note. Entries are based on a count of the obtained test statistics (Z or χ^2) that exceeded the nominal critical values associated with .05 significance level, based on 10,000 randomly generated chi-square contingency tables.

Table 7
Empirical Type I Error Rates for the Pairwise Post Hoc Procedures in the Independence Model

Method/Sample Size	Critical Value	Rejections	Tables With at Least One Rejection	Type I Error Rates	
				Per Contrast	Experimentwise
Delucchi (1983) procedure					
100	$Z = 2.796$ and 2.448	4,094	3,010	.137	.301
300		3,166	2,389	.106	.239
500		3,012	2,301	.100	.230
Marascuilo and Serlin (1988) procedure					
100	$Z = 3.549$	372	337	.012	.034
300		157	141	.005	.014
500		149	137	.005	.014
Seaman and Hill (1996) procedure					
100	$\chi^2 = 12.592$	98	87	.003	.009
300		102	95	.003	.010
500		127	116	.004	.012
Gardner (in press) procedure					
100	$\chi^2 = 9.885$	445	376	.015	.038
300		521	454	.017	.045
500		554	483	.019	.048

Note. Entries are based on a count of the obtained test statistics (Z or χ^2) that exceeded the nominal critical values associated with .05 significance level, based on 10,000 randomly generated chi-square contingency tables. In the Delucchi (1983) procedure, the first critical value ($Z = 2.796$) is used for all column-wise contrasts, and the second critical value ($Z = 2.448$) is used for all row-wise contrasts. The column-wise and row-wise rejections are summed to obtain empirical error rates.

In Table 7, the per-contrast and experimentwise Type I error rates for the four pairwise post hoc procedures are provided for the chi-square independence model. It will be noted that the per-contrast error rates were inflated for the Delucchi (1983) procedure but conservative for all other procedures. Moreover, the Delucchi procedure resulted in an overly inflated experimentwise error rate (e.g., .301 for $N = 100$), whereas the Marascuilo and Serlin (1988) and the Gardner (in press) procedures provided experimentwise error rates that were closer to the nominal value (.034 and .038, respectively), whereas the Seaman and Hill (1996) procedure was much more conservative (.009). With increased sample size, the error rate for the Gardner method remained close to the nominal level, whereas the error rate for the Marascuilo and Serlin procedure became more conservative.

Discussion

A statistically significant omnibus chi-square test leads a researcher to reject the null hypothesis. However, in the event that the contingency table being tested is larger than 2 2, the researcher is unable to interpret the results any further. Post hoc tests can be used to facilitate further interpretation, and as demonstrated above, these can adopt one of two strategies: a cellwise post hoc analysis or a pairwise post hoc analysis contrasting cell proportions based on row or column totals.

Cellwise Analysis

If the researcher decides to investigate a statistically significant omnibus test further using a cellwise analysis, both the standardized residual and the adjusted residual methods provided adequate experimentwise Type I error rate control if and only if appropriate critical values are selected. For the adjusted residuals method, the distributional characteristics of the statistic were normal, and as a result, researchers should adopt critical values that *are* adjusted for the number of contrasts being considered. Failing to adjust the critical value appropriately results in experimentwise Type I error rates that greatly exceed nominal levels. For the standardized residuals method, the distribution of the statistic was *not* unit normal as the obtained variance was appreciably less than 1.0, even though the other distributional statistics were as expected for standard normal distributions. As a result, researchers should be cautious when considering this approach. The standardized residual is not distributed as a standard normal variate, and as such, it is not reasonable to use the standardized normal distribution to obtain critical values.

Pairwise Analysis

If the researcher is interested in pursuing a statistically significant omnibus chi-square test with a pairwise post hoc analysis, the selection of the "best" procedure is somewhat simpler. Based on the simulation methods, we can conclude that the Marascuilo and Serlin (1988) procedure and the Seaman and Hill (1996) procedure provide experimentwise Type I error rate control that is overly conservative, thereby making them poor choices for analyses involving both the homogeneity and independence models. On the other hand, the Delucchi (1983) procedure provides error rates that are overly liberal, thereby making it a poor choice as well.

The remaining pairwise post hoc procedure has several advantages over the other procedures. First, the Gardner (in press) procedure maintains the experimentwise Type I error rates near nominal levels for both the independence and homogeneity chi-square models. The results of this simulation, based on a 3 4 contingency table, replicate the findings on previous computer simulations for a 3 3 contingency table (Lam, 1996). These combined results suggest that the Bonferroni adjustment to the alpha level will maintain adequate error rate control for a variety of contingency table sizes, although this has not yet been confirmed.

A second advantage of the Gardner (in press) procedure is the computational simplicity of the statistic. Rather than manually computing the *Z* statistic for this procedure (as one would have to do for the Marascuilo and Serlin [1988] and the Delucchi [1993] procedures), a researcher can simply perform a chi-square test on the selected columns and collapsed rows (or vice versa) of the data. The obtained chi-square statistic, with one degree of freedom, can then be compared to a critical value with an appropriately adjusted alpha level set by $\alpha/(I \quad \frac{J}{2})$ for the homogeneity model and $\alpha/(I \quad \frac{J}{2} + J \quad \frac{I}{2})$ for the independence model. If tables for the chi-square critical values do not include this adjusted alpha level, then the square root of the obtained chi-square statistic can be compared to an alpha level adjusted critical *Z* statistic.

Limitations of the Study

A limitation of any Monte Carlo investigation is that strictly speaking, the results pertain only to the simulation investigated. Thus, the present results could be said to generalize only to a 3 4 table in which (a) the null hypothesis is true, (b) either an independence or homogeneity model is under consideration, and (c) one of the six procedures is performed. The similarity of the results obtained in this study to those obtained by Lam (1996) suggest, however, that comparable results could be expected for other tables of different dimensions under the conditions indicated. These results apply, however,

only to Type I error rates and are silent with respect to Type II error rates. We chose to investigate Type I error rates because we believe investigators should be aware of the potential Type I errors associated with post hoc tests that are often used to further investigate $I \quad J$ tables.

We chose not to investigate Type II error rates primarily because of the logistical issues involved. Even given a 3 4 table, there are a multitude of ways in which the null hypothesis could be false. One cell could be disproportionate, or more than one cell could be disproportionate, and the degree of disproportion could be the same or different. Also, one 2 2 table computed from the larger table could describe an association, or more than one such table could, and so forth. Power statistics would differ for the various violations that could be investigated, making generalizations extremely difficult, thus severely limiting the utility of the results. The results obtained in the present study do, however, pertain directly to an important consideration in data analysis, namely, that of Type I error control, and it seems very likely that they can be generalized to different sized tables.

Summary

In summary, a review of the literature reveals that several post hoc procedures are available to researchers interested in analyzing further a statistically significant omnibus chi-square test. The results of this investigation demonstrated, however, that these tests differ in how well they maintain the experimentwise Type I error rate at the nominal level (e.g., .05) for both the independence and homogeneity models. If the researcher is interested in a cellwise approach to the post hoc analysis, the adjusted residual method is normally distributed and provides adequate control of the experimentwise Type I error rates (given an adjusted critical value). Researchers interested in making pairwise comparisons between the sampled populations should consider the Gardner (in press) pairwise post hoc procedure. This procedure consistently maintained the experimentwise error rate at near-nominal levels and had the added advantage of being able to perform the analysis with a chi-square test rather than requiring the researcher to make the computations manually. Taken together, either of these two methods can provide researchers valuable information when confronted with a statistically significant omnibus chi-square test statistic.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley.

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*(1), 79-93.

Delucchi, K. L. (1983). The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin*, *94*(1), 166-176.

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: Statistical issues*. Hillsdale, NJ: Lawrence Erlbaum.

Edgington, E. S. (1974). A new tabulation of statistical procedures used in APA journals. *American Psychologist*, *29*, 25-26.

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions* (2nd ed.). New York: John Wiley.

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

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 auxillary to tests in a Markov chain. *Annals of Mathematical Statistics*, *32*, 535-548.

Goodman, L. A. (1964a). Simultaneous confidence intervals for contrasts among multinomial populations. *Annals of Mathematical Statistics*, *35*, 716-725.

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

Goodwin, L. D., & Goodwin, W. L. (1985). Statistical techiniques in *AERJ* articles, 1979-1983: The preparation of graduate students to read the educational research literature. *Educational Researcher*, *14*, 5-11.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, *29*, 205-220.

Haberman, S. J. (1978). *Analysis of qualitative data* (Vol. 1). London: Academic Press.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1988). *Applied statistics for the behavioral sciences* (2nd ed.). Boston: Houghton Mifflin.

Lam, T. C. (1996). *On the interpretation of an $R \times C$ chi-square: A Monte Carlo investigation*. Unpublished senior undergraduate honors thesis, University of Western Ontario, London, Ontario, Canada.

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.

Mouly, G. J. (1978). *Educational research: The art and science of investigation*. Boston: Allyn & Bacon.

Popham, W. J., & Sirotnik, K. A. (1973). *Educational statistics: Use and interpretation* (2nd ed.). New York: Harper & Row.

Reynolds, H. T. (1984). *Analysis of nominal data*. Beverly Hills, CA: Sage.

Schinka, J. A., LaLone, L., & Broeckel, J. A. (1997). Statistical methods in personality assessment research. *Journal of Personality Assessment*, *68*(3), 487-496.

Seaman, M. H., & Hill, C. C. (1996). Pairwise comparisons for proportions: A note on Cox and Key. *Educational and Psychological Measurement*, *56*(3), 452-459.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626-633.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Thompson, B. (1988). Misuse of chi-square contingency-table test statistics. *Educational and Psychological Research*, *8*(1), 39-49.