AN ADJUSTMENT TO THE CHI-SQUARE TEST IN CASES OF SPARSE CONTINGENCY TABLES

By

MICAH WARD

Bachelor of Science

Oklahoma State University

Stillwater, Oklahoma

1994

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 1996

AN ADJUSTMENT TO THE CHI-SQUARE TEST IN CASES

OF SPARSE CONTINGENCY TABLES

Thesis Approved:

P. Larry Claypool
Thesis Advisor

Mark E Payton

Melinda H. McCann

Thomas C. Collins
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to thank Dr. P. Larry Claypool, my major advisor, for his advice and encouragement on this thesis and throughout the past few years. I thank Dr. Mark Payton and Dr. Mindy McCann for their suggestions and help for my project. I also thank Brenda Masters for giving me the opportunity to teach the joys of statistics.

My thanks would not be complete unless I let my fellow graduate students know how much I appreciated their support and help over the past two years. Thank you Laura Coombs, Carrie Duvall, Marla Eason, Anthony Miller, and Valerie Skaggs for your friendship and support.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Statement of Problem

Many times in research, scientists present their data in the form of a contingency table and use a chi-square test based on frequencies to analyze their data. In many cases when a statistical software package is used, a warning may be printed which advises the scientist that the analysis may not be valid due to an over-abundance of "small" cells in the contingency table. More specifically, many of these software packages may consider the table to be sparse and give the warning noted above when more than 20% of the cells in the table have an expected frequency of less than 5. At this point, the scientist may either ignore the warning or perform an alternative statistical analysis; however, unless the scientist has knowledge of advanced statistical methods, the possible alternatives may be either unknown or beyond the level of his/her competence and understanding.

Although other test procedures exist, such as Fisher's Exact Test and the test procedures based on loglinear models, these are beyond the scope of statistical knowledge for many researchers. When such researchers are faced with the dilemma of an analysis that "may not be valid" due to a sparse table, their next option is to look for a simple modification of the chi-square test or to find an easy-to-apply alternative analysis.

Purpose and Objectives

This study will address the experimental situation in which a single random sample is taken and each observation is categorized into one of several nominal categories for each of two variables. Hence, the chi-square test for independence would be an appropriate data analysis procedure. The purpose of this paper is three-fold. First, it is desired to find a simple modification of the chi-square test for independence which may be applied when the data set produces a sparse contingency table. Secondly, determine whether the literature contains a suitably simple alternative test procedure that may be applied in the sparse table situation. Finally, any procedures which result from the first two objectives will be compared to determine which test procedure is better.

CHAPTER II

REVIEW OF THE LITERATURE

OVERVIEW

Numerous articles exist where suggestions of how to analyze a sparse contingency table are given. Some are Haberman (1988), Simonoff (1986), Read (1984), Fienberg (1979), and Haberman (1977). Haberman's 1988 article is very mathematical in nature and deals with the bias of the chi-square test that occurs when too many of the cells are small. Simonoff discusses nonparametric techniques for estimating the variance for a statistic that is not necessarily asymptotically χ^2 due to many sparse cells. Read examines the small-sample properties of the power divergence family of goodness-of-fit statistics to show that the power of the G^2 and χ^2 can be improved by choosing other statistics from the family. This is particularly important if the table is sparse. Fienberg compares the chi-square test to its likelihood ratio test when used in large, sparse tables. Finally, in 1977 Haberman writes about using loglinear models to analyze contingency tables which have small cell expected frequencies. These ideas are not considered in this paper because they use loglinear models or other mathematical methods, which may be beyond the scope of understanding for the researcher whose background may include only one or two elementary statistics courses.

ADJUSTED CHI-SQUARE STATISTIC

The adjustment discussed in this paper was first used in Mellina (1984). She used this adjustment only when more than 20% of the cells in the contingency table had expected frequencies less than 5 and the usual chi-square statistic rejected the null hypothesis of independence. This adjustment resulted from discussions between herself and her advisor, Dr. P. Larry Claypool, a Professor of Statistics at Oklahoma State University. The basic concept of the adjustment is to ask, "What if all cells in a table made the same average contribution to the calculated χ^2 statistic as the 'large cells' from the table?" Specifically, calculate the chi-square contribution for each cell in the table. Find the cells which have expected frequencies of less than five and ignore their corresponding chi-square contributions. Next average the remaining chi-square contributions and insert this average "large cell" contribution as the cell

contribution for each cell which has an expected frequency of less than five (the one's whose chi-square contribution was previously ignored). Finally, sum all the chi-square contributions and use this sum as the test statistic to test for independence. Alternatively, the same adjusted value of the test statistic would be obtained by simply multiplying the average "large cell" contribution by the total number of cells in the table (r*c).

ZELTERMAN'S D^2

Zelterman (1987) proposed a statistic specifically for the sparse table situation which has two appealing properties. First, it is easy to apply and secondly, it is asymptotically normally distributed. He named it D^2 and the equation is as follows:

$$D^2 = \sum [(n_{ij} - \hat{\lambda}_{ij})^2 - n_{ij}] / \hat{\lambda}_{ij}$$

where n_{ij} is the cell count in the i^{th} row and the j^{th} column and $\hat{\lambda}_{ij}$ is the estimated cell expected frequency for the i^{th} row and the j^{th} column. Also, the $\hat{\lambda}_{ij}$'s are found by multiplying the i^{th} row total by the j^{th} column total and dividing by the total number of observations in the table. That is:

$$\hat{\lambda}_{ij} = (\sum_j n_{ij})(\sum_i n_{ij}) / (\sum_{ij} n_{ij}).$$

The D^2 formula looks just like the usual chi-square statistic formula, except that in the numerator the cell count is subtracted before the division by the denominator. This statistic is approximately normally distributed and has decent power. The mean and variance of the D^2 are given in Mielke and Berry (1988).

DEGREES OF FREEDOM

For purposes explained later it was postulated that the degrees of freedom for the adjusted chi-square statistic might have to be modified. Therefore, literature that explained how to alter the degrees of freedom when a parameter (or, in this case, a cell probability) is to be estimated was desired. The customary procedure has been to decrease the degrees of freedom by 1 for each parameter to be estimated, but the question arises as to whether this decrease should be exactly 1. Perhaps an interval around 1.0 would be more appropriate for different situations. A review of the literature did not show any detailed proof of how the value of 1 was obtained. Fisher (1922, 1924) simply refers to this as an accepted fact.

Cramer (1946) gives some discussion on the degrees of freedom. but still does not outline a formal proof as to why the integer 1 is used. It was anticipated that these results would be useful later in the methodology of the research.

CHAPTER III

METHODOLOGY

This is a simulation study to compare the relative merits of the usual chi-square statistic, the adjusted chi-square statistic, and Zelterman's D^2 statistics. SAS (1990) was used to perform all the simulations to compare the three statistics (see Appendix C for an example program). The data were simulated for tables having seven different dimensions and two different percentages of sparse cells. Each table size had a 25% small cell case and a 50% small cell case for a total of 14 unique tables. The number of times each statistic rejected the null hypothesis of independence for each of these 14 tables was compared. The seven different sizes of tables are 4x3, 4x5, 4x8, 5x8, 4x9, 3x10, and 5x10. For each table size probability structures were assigned corresponding to 25% sparse cells and to 50% sparse cells (see Appendix A for probability structures). In order to ensure these sparseness percentages, enough data was generated to guarantee an average of 10 observations per cell. For instance, since the 4x3 table has 12 cells, 120 observations were generated. Each of the 14 tables was generated 1000 times. Each table had a probability structure that insured independence. The row and column marginal proportions were determined (or assigned) so that i) both the row and column marginal probabilities add to 1.0; ii) the row and column marginal probabilities are listed in a random order; iii) the individual cell probabilities are found from the product of the corresponding row and column marginal probabilities; and iv) either 25% or 50% of the cells had expected frequencies less than 5 when n=10*r*c. For example, referring to the first 4x3 table in Appendix A, the arbitrarily assigned row probabilities are 0.30, 0.10, 0.40, and 0.20. The corresponding assigned column probabilities are 0.25, 0.60, and 0.15. The assignment of probabilities in this case results in 3 cells (or 25%) having an expected frequency less than five, since n=120. Corresponding to each observation a random number, x, was generated from a uniform (0,1) distribution using a seed based on the internal clock. This value would increase the tally by one for a specific cell of the table according to the following algorithm. If $x \leq p_{11}$ (the probability of the (1,1) cell which is 0.075 in this example), then add 1 to the tally of the (1,1) cell. If $p_{11} < x \leq p_{11} + p_{12}$ then add 1 to the tally of the (1,2) cell. Note: $p_{12} = 0.18$ in this example. If $p_{11} + p_{12} < x \leq p_{11} + p_{12} + p_{13}$ then add

1 to the tally of the (1,3) cell. Continue this process across each row until finally, if $1 - p_{rc} < x \leq 1$, then add 1 to the tally of the (r,c) cell; here if $0.97 < x \leq 1.0$, add 1 to the tally of the (4,3) cell.

For each of the 1000 tables generated within each size by sparseness combination, the value for the usual chi-square statistic, the adjusted chi-square statistic, and Zelterman's D^2 were calculated and the observed significance levels were determined. Note that while specific probabilities were assigned for simulation purposes, each statistic was calculated using only the information generated for the table; that is, the expected frequencies are always estimated. Both the usual chi-square statistic and the adjusted chi-square statistic are compared to the χ^2 distribution with (r-1)*(c-1) degrees of freedom. Since the D^2 statistic is approximately normally distributed, it will be standardized using the mean and variance found in Mielke and Berry (1988) and then compared to the standard normal distribution. Next, the results from the 1000 tables were summarized in terms of the number of times each of the three statistics rejected the null hypothesis of independence. Also, the number of times any two of these statistics rejected the null hypothesis for the same table was recorded. Finally, since the adjusted chi-square statistic ignored some of the cells and the corresponding observations, the average proportion of observations used and the average proportion of cells used were calculated for the 1000 tables generated for each size by sparseness combination. Since only the adjusted chi-square statistic ignored some of the information, these proportions refer only to the adjusted chi-square statistic.

COMPUTER SIMULATIONS

DEGREES OF FREEDOM PATTERN

The next step was to find a pattern, if any, between degrees of freedom and table size, proportion of observations used, proportion of cells used, or any other information that could be gathered. The degrees of freedom for the adjusted chi-square statistic were altered from those of the usual chi-square ((r-1)*(c-1)) in four ways: i) the degrees of freedom for the usual chi-square statistic were reduced by 1 for each cell that had its chi-square contribution estimated; ii) the degrees of freedom for the usual chi-square statistic were altered in a trial-and-error fashion until the number of rejections was approximately 50 (5% of 1000); iii) the degrees of freedom for the usual chi-square statistic were multiplied by the proportion of

cells used; and iv) the degrees of freedom for the usual chi-square statistic were multiplied by the proportion of observations used. No special formula was applied to alter the degrees of freedom in ii). The complete simulation of 1000 tables was simply repeated with different values for the degrees of freedom until the adjusted chi-square statistic gave approximately 50 (5%) rejections. Then, after the tables were analyzed it was hoped that a pattern for the degrees of freedom of the adjusted chi-square statistic would be found. Hopefully, the pattern would follow from information gathered from one or more of the modifications mentioned previously.

DIFFERENT ARRANGEMENTS OF PROBABILITIES

Another idea was that a degrees of freedom pattern and the number of rejections for each statistic might depend on the arrangement of the marginal (row and column) cell probabilities. The original set of tables had marginal probabilities that were used in a haphazard order. That is, the marginal probabilities were set in a random order. Therefore, the original set of tables will be referred to as the haphazard tables, since there was no attempt to order the marginal probabilities in any way. After using this structure the marginal probabilities would be arranged in a decreasing order; that is, decreasing across the top and decreasing down the side (see Appendix B for probability structures). For example, the 4x3 table with 25% small cells would now have the probabilities of 0.6, 0.25, and 0.15 across the top, and 0.4, 0.3, 0.2, and 0.1 down the side in that listed order. Now, of course, the cell probabilities would not change for each table, but would be rearranged and, hence, the tally algorithm would have different values inserted into it. The decreasing probability tables were simulated using the same algorithm as utilized for the haphazard tables with the new "accumulated" probabilities bounding each cell probability inserted in the algorithm. It was anticipated that the degrees of freedom for the adjusted chi-square statistic would change very little, if at all, using the decreasing probabilities structure.

Another table structure that was simulated had uniform probabilities. In other words, a table would have the same probability in each cell; that is, each row was assigned marginal probabilities of 1/r and each column was assigned marginal probabilities of 1/c. For example the 4x3 table, having 12 cells, would have a probability of 1/12 in each cell due to independence. The purposes were to see the effect on degrees of freedom and the effect on the number of rejections by ignoring 25% and then 50% of the cells of

an independent table when calculating the adjusted chi-square statistic. The criteria for calculating the value of the adjusted chi-square statistic for these tables had to be different than that of the two previous arrangements of probabilities, because none of the cells would have an expected frequency less than 5. Now, for each table size 75% of the cells were used and then 50% of the cells were used to calculate the value of the adjusted chi-square statistic. In other words, 25% of the cells were selected to be ignored and then 50% of the cells were selected to be ignored, regardless of the content of the cells. Since the data were generated randomly, deleting a row or two would be equivalent to choosing cells at random and ignoring them. Ignoring the selected cells was done to emulate the 25% small cell case and the 50% small cell case, respectively, for the 7 table sizes. Again, the simulations were repeated and the degrees of freedom for the adjusted chi-square statistic were altered as in procedure ii) above, until the adjusted chi-square statistic gave approximately 5% rejections for the 1000 tables.

Finally, a set of tables was simulated where each table in the set began with the same seed to generate all the observations. The purpose here was to see the effect on the degrees of freedom for the adjusted chi-square statistic and the effect on whether a table gives a rejection or not when the same cell probabilities were moved to different cells. Since the same seed was used, the data would be the same for each table. The seed value was 1000000 for each table. The 4x5 tables with haphazard probabilities, descending probabilities, and uniform probabilities were used. Also, both the 25% small cell case and the 50% small cell case were used for each probability structure, for a total of 6 tables . The marginal and cell probabilities used are the same ones that are found on the 4x5 tables in Appendix A and Appendix B. Very little difference would be anticipated between the haphazard and decreasing probability structures within either of the sparseness levels.

CHAPTER IV

FINDINGS

The findings discussed here are the results of the computer simulations described in Chapter III. Since 1000 tables were generated for each table size by sparseness level by probability structure combination, it was desired to have approximately 50 rejections for each statistic to give a 0.05 significance level test.

HAPHAZARD PROBABILITIES

The first set of simulations were run with the degrees of freedom for the adjusted chi-square statistic kept at (r-1)*(c-1) to see if there would be any need for modification. Table 1 below shows the number of rejections for each statistic for this first set of simulations. The notation "C" denotes the number of rejections for the usual chi-square statistic; "K" denotes the number of rejections for the adjusted chi-square statistic; and "Z" denotes the number of rejections for the Zelterman's D^2. Also, the notation "C vs. K" denotes the number of times both the usual and adjusted chi-square statistics rejected the same table; "C vs. Z" denotes the number of times both the usual chi-square statistic and Zelterman's D^2 rejected the same table; and "K vs. Z" denotes the number of times both the adjusted chi-square statistic and Zelterman's D^2 rejected the same table. The previous nomenclature applies to Table 1 and all other tables, which follow. The number of rejections for the most part are too low for the smaller dimensioned tables and either just about right or too large for the larger tables. For example, in the 4x3 table the adjusted chi-square statistic (K in Table 1) had rejection rates of 30 and then 14 out of 1000 for the 25% small cell case and the 50% small cell case, respectively. These were considered to be too low. However, the 5x10 table for the same statistic had rejection rates of 51 and 66 out of 1000 for the 25% small cell case and the 50% small cell case respectively. These results were considered to be either about right or a little too large. In any event, the fact that some of the rejections for the three statistics are less than 50 suggest that the degrees of freedom for the adjusted chi-square statistic should be modified from (r-1)*(c-1).

Table 1. Number of rejections for three statistics for a given dimension and sparseness level using the haphazard probabilities and the usual degrees of freedom for the statistic K.

Dimension	Small Cells	DF	C[a]	K[b]	Z[c]	C[d] vs. K	C[e] vs. Z	K[f] vs. Z
4x3	25%	6	38	30	21	18	21	12
4x3	50%	6	46	14	20	6	20	3
4x5	25%	12	31	36	26	19	22	20
4x5	50%	12	44	43	36	14	34	13
4x8	25%	21	36	84	24	27	19	22
4x8	50%	21	42	42	28	10	24	9
5x8	25%	28	46	56	42	31	42	27
5x8	50%	28	44	55	42	25	42	24
4x9	25%	24	58	49	43	26	40	23
4x9	50%	24	52	17	34	8	32	5
3x10	25%	18	54	54	28	31	28	19
3x10	50%	18	61	66	28	27	26	17
5x10	25%	36	48	51	40	32	40	27
5x10	50%	36	49	66	41	22	40	20

[a] C denotes the usual chi-square statistic
[b] K denotes the adjusted chi-square statistic
[c] Z denotes Zelterman's D^2
[d] C vs. K denotes the number of times the usual and adjusted chi-square statistics rejected the same table
[e] C vs. Z denotes the number of times the usual chi-square statistic and Zelterman's rejected the same table
[f] K vs. Z denotes the number of times the adjusted chi-square statistic and Zelterman's rejected the same table

Three of the procedures for modifying the degrees of freedom for the adjusted chi-square statistic were found to be not very useful. First, reducing degrees of freedom by 1 for each cell chi-square contribution that was estimated resulted in too many rejections. Second, multiplying the usual degrees of freedom ((r-1)*(c-1)) by the proportion of cells used also resulted in too many rejections. Thirdly, multiplying the usual degrees of freedom ((r-1)*(c-1)) by the proportion of observation used resulted in too few rejections. Hence, these particular modifications were abandoned early in the simulation study.

Therefore, the degrees of freedom modification where the usual degrees of freedom were just altered in a trial-and-error manner until the number of rejections was approximately 50 was used in the simulation of all the tables. This involved guessing at a value for the degrees of freedom and then generating the entire set of 1000 tables to see if the adjusted chi-square statistic gave approximately 50

rejections. If the number of rejections was not close to 50, another guess was made and the simulation was repeated. Otherwise, the simulation was repeated three more times at the specified degrees of freedom to ensure that it would give around 50 rejections each time. The degrees of freedom found using this process will henceforth be called the modified degrees of freedom for the adjusted chi-square statistic. Such modifications apply only to the adjusted chi-square statistic.

The modified degrees of freedom found using the haphazard tables changed very little from the usual degrees of freedom ((r-1)*(c-1)). Table 2 below shows the values for the modified degrees of freedom for the adjusted chi-square statistic. In addition this table shows the number of rejections for each statistic and comparisons discussed above for Table 1; however, these values are results from the simulations which used the modified degrees of freedom. For example, the 4x8 table needed degrees of freedom of 20.5 and 18 for the adjusted chi-square statistic for the 25% small cell case and the 50% small cell case, respectively. On the other hand, the degrees of freedom for the larger tables needed no modification, because for the larger tables the number of rejections for the adjusted chi-square statistic were already close to or more than fifty. For instance, the 5x10 table used 36 and 36 degrees of freedom for the 25% small cell case and 50% small cell case, respectively. These are the same as the usual chi-square statistic degrees of freedom ((r-1)*(c-1)). Therefore, some of the results listed in Table 2 are exactly the same (represent the same simulations) as the corresponding results from Table 1. These results are identified by a '*'. When discrepancies between modified degrees of freedom and the usual degrees of freedom occurred, the larger of these discrepancies were associated with tables with 50% small cells. So it seemed that the degrees of freedom for the adjusted chi-square statistic was not much different than those for the usual chi-square statistic, at least for smaller tables. Also, notice that Table 2 contains values for the average proportion of observations used (MPROPO) and the average proportion of cells used (MPROPC). They are listed because later it was found that they may affect the modified degrees of freedom in a nonlinear fashion.

Table 2. Number of rejections for three statistics for a given dimension and sparseness level using the haphazard probabilities and the modified degrees of freedom for the statistic K.

									M[g]	M[h]
									P	P
									R	R
						C[d]	C[e]	K[f]	O	O
	Small	MOD				vs.	vs.	vs.	P	P
Dimension	Cells	DF	C[a]	K[b]	Z[c]	K	Z	Z	O	C
4x3	25%	5.25	41	56	29	22	29	17	0.91	0.68
4x3	50%	4	64	46	43	18	41	17	0.9	0.49
4x5	25%	12	34	52	33	26	28	28	0.99	0.75
4x5	50%	11	45	53	33	17	33	13	0.93	0.5
4x8	25%	20.5	48	47	33	27	32	25	0.99	0.75
4x8	50%	18	41	53	28	17	22	10	0.94	0.51
5x8*	25%	28	46	56	42	31	42	27	0.92	0.76
5x8*	50%	28	44	55	42	25	42	24	0.86	0.49
4x9*	25%	24	58	49	43	26	40	23	0.98	0.73
4x9	50%	21	67	53	47	16	42	14	0.94	0.5
3x10*	25%	18	54	54	28	31	28	19	0.91	0.71
3x10*	50%	18	61	66	28	27	26	17	0.91	0.46
5x10*	25%	36	48	51	40	32	40	27	0.93	0.75
5x10*	50%	36	49	66	41	22	40	20	0.89	0.5

[a] C denotes the usual chi-square statistic
[b] K denotes the adjusted chi-square statistic
[c] Z denotes Zelterman's D^2
[d] C vs. K denotes the number of times the usual and adjusted chi-square statistics rejected the same table
[e] C vs. Z denotes the number of times the usual chi-square statistic and Zelterman's rejected the same table
[f] K vs. Z denotes the number of times the adjusted chi-square statistic and Zelterman's rejected the same table
[g] MPROPO denotes the average proportion of observations used in the 1000 tables
[h] MPROPC denotes the average proportion of cells used in the 1000 tables
* denotes that the results in table 2 are duplicated from table 1 because the degrees of freedom required no adjustment

As well as finding a degrees of freedom pattern, it was desired to compare the number of rejections for the three statistics. In general, the adjusted chi-square statistic rejected the null hypothesis more than the usual chi-square statistic and more than Zelterman's D^2. Zelterman's D^2 always rejected the null hypothesis fewer times than the usual chi-square statistic did. Table 2 shows the number of rejections for each statistic. For example, the 4x3 table with 25% small cells had 41 rejections for the usual chi-square statistic, 56 rejections for the adjusted chi-square statistic, and 29 rejections for Zelterman's D^2. The cause of individual rejections for any of the three statistics was usually either one or

two small cells giving large contributions to the statistic or several small cells giving moderate contributions to the statistic. These large and moderate cell contributions tended to inflate the overall statistic, thus rejecting the null hypothesis at a 0.05 significance level.

Recall, the program also checked how many times any two statistics rejected the null hypothesis for the same table. The adjusted chi-square statistic rejected the same table about half the time that the usual chi-square statistic rejected. The adjusted chi-square statistic also rejected the same table about half the time than Zelterman's statistic did. Finally, Zelterman's statistic rejected virtually every time that the usual chi-square statistic did. Again, Table 2 lists these rejection comparisons. Look at the 4x3-25% table as an example. The usual and adjusted chi-square statistics rejected the same table 22 times, which is approximately half of the 41 times that the usual chi-square statistic rejected and a little less than half of the 56 times that the adjusted chi-square statistic rejected. The adjusted chi-square statistic and Zelterman's D^2 rejected the same table 17 times, which is a little more than half of the 29 times that the Zelterman's D^2 rejected. Finally, the usual chi-square statistic and Zelterman's D^2 rejected the same table 29 times, which implies each of the 29 times that Zelterman's D^2 rejected, the usual chi-square statistic had also rejected.

Usually, when two statistics rejected the same table the cause of the rejections were one or two small cells that were giving large contributions to the various statistics. The problem cells would be the same cells for both statistics. Looking at the 4x3-25% table as an example, if cell (4,1) gave a large contribution to the adjusted chi-square statistic it almost always gave a large contribution to the Zelterman statistic. The same thing happens when the usual chi-square and Zelterman statistics reject the same table. The same cells are giving large contributions to the various statistics. When the usual and adjusted chi-square statistic rejected the same table it was for a different reason. In the calculation of the adjusted chi-square statistic the large contributions from the small cells should have been ignored. However, some large cells, that is, cells with expected frequencies more than 5, were giving large contributions to both the usual and adjusted chi-square statistics. For these cells the expected frequencies would be large, but the actual count would be small, thus a large contribution would get added into both statistics, inflating them.

The cases when some of the statistics disagreed deserves some mention. Sometimes the usual chi-square statistic would reject a table, but the adjusted chi-square would not. The reason is that the adjusted chi-square statistic would ignore the large contributions given by the small cells. Sometimes the adjusted chi-square statistic would reject a table, but the usual chi-square would not. One reason is that the degrees of freedom for the adjusted chi-square statistic might be reduced from that of the usual chi-square statistic. Therefore, even if the adjusted chi-square statistic is smaller than the usual chi-square statistic the smaller modified degrees of freedom would cause the rejection. However, there were times when the modified degrees of freedom were the same as the usual degrees of freedom. Whenever this was the case and the adjusted chi-square statistic rejected a table that the usual chi-square did not, the cause was a high average cell contribution for the large cells. This high average would cause the adjusted chi-square statistic to be larger than the usual chi-square statistic. Thus the table would be rejected for the larger adjusted chi-square statistic, but not the smaller usual one. It should be noted here that in Mellina (1984) the adjusted chi-square statistic was used only when the usual chi-square rejected and the software gave the sparse table warning. Mellina expected that the usual chi-square statistic would reject too frequently(more than 5%) and that the adjusted chi-square statistic would reject less frequently than the usual chi-square did. However, that would not help in the cases where the adjusted chi-square statistic rejected and the usual chi-square statistic did not. The results of this study would make that point moot, because the usual chi-square statistic did not reject too frequently, in general.

Another comparison involves the cases where the usual chi-square and Zelterman statistics do not reject in the same table. The reason for the disagreement is based on the fact that even though the same cell will give a large contribution to both statistics, the large contributions to Zelterman's statistic are not as big as the corresponding contribution to the usual chi-square statistic. Also, since Zelterman's statistic was standardized, there are positive and negative contributions to the statistic. When the usual chi-square statistic rejected and the Zelterman's statistic did not it was found that the usual chi-square statistic was significant at the 0.05 level, but the Zelterman statistic was significant at a slightly higher level, say 0.10. The cell contributions for the Zelterman statistic are just not enough to give a rejection for a 0.05 level test. The last situation is when Zelterman's D^2 rejects, but the usual chi-square statistic does

not. This was very rare. When this did happen the large contributions for Zelterman's D^2 were mostly positive, which resulted in a small observed significance level. In these few instances the usual chi-square statistic was large enough to reject at a 0.10 level test, but not the desired 0.05 level test.

DESCENDING PROBABILITIES

Just like the modified degrees of freedom found from using the haphazard probabilities, the modified degrees of freedom found from using the descending probabilities changed very little from the usual degrees of freedom. This was not surprising since the same probabilities that were used in this set of tables were used in the haphazard tables. Table 3 shows the modified degrees of freedom for the adjusted chi-square statistic when using the descending probabilities. For example, the modified degrees of freedom for the 4x3 25% small cell case is 5.5 in Table 3 and 5.25 in Table 2. If any table structure had a modified degrees of freedom that changed much from the usual degrees of freedom they were the small tables with 50% small cells. Again, the adjusted degrees of freedom for the larger tables are the same as the usual degrees of freedom.

As anticipated, the number of rejections for the three statistics for these tables were similar to the number of rejections found using the haphazard probabilities. Table 3 shows the number of rejections for the three statistics. The adjusted chi-square statistic usually rejected more often than the usual chi-square statistic; the adjusted chi-square statistic always rejected more often than the Zelterman statistic; and the Zelterman statistic always rejected less often than the usual chi-square statistic. For example, the 4x3 table with 25% small cells had 52 rejections for the usual chi-square statistic, 55 rejections for the adjusted chi-square statistic, and 35 rejections for the Zelterman statistic. Again, the proportion of times that any two statistics rejected the same table was similar to that of the haphazard probabilities. Look at the 4x3-25% table as an example. The usual and adjusted chi-square statistics rejected the same table 24 times, which is roughly half of the 52 times that the usual chi-square statistic rejected. The adjusted chi-square statistic and Zelterman's D^2 rejected the same table 18 times which is about half of the 35 times that Zelterman's D^2 rejected. The usual chi-square and Zelterman statistics rejected the same table 35 times which again implies that Zelterman's D^2 rejected only when the usual chi-square statistic did. So it appeared that the arrangement of the marginal probabilities had very little effect beyond the expected

Table 3. Number of rejections for three statistics for a given dimension and sparseness level using the decreasing probabilities and the modified degrees of freedom for the statistic K.

									M^g	M^h
									P	P
									R	R
						C^d	C^e	K^f	O	O
	Small	MOD				vs.	vs.	vs.	P	P
Dimension	Cells	DF	C^a	K^b	Z^c	K	Z	Z	O	C
4x3	25%	5.5	52	55	35	24	35	18	0.9	0.69
4x3	50%	3.75	39	53	21	11	20	9	0.9	0.49
4x5	25%	11.5	31	54	34	26	29	30	0.99	0.75
4x5	50%	10.5	43	51	33	17	32	15	0.93	0.5
4x8	25%	20.5	50	50	38	29	34	26	0.99	0.75
4x8	50%	18	58	46	32	17	30	11	0.94	0.51
5x8	25%	28	92	113	85	66	85	63	0.92	0.76
5x8	50%	28	45	49	40	17	39	17	0.86	0.49
4x9	25%	22.5	41	48	29	18	24	14	0.98	0.73
4x9	50%	21	53	48	39	14	37	11	0.94	0.5
3x10	25%	18	54	48	29	27	29	18	0.91	0.7
3x10	50%	18	41	49	21	11	19	12	0.91	0.46
5x10	25%	36	53	63	46	38	46	36	0.93	0.75
5x10	50%	36	40	60	42	23	40	25	0.89	0.5

[a] C denotes the usual chi-square statistic
[b] K denotes the adjusted chi-square statistic
[c] Z denotes Zelterman's D^2
[d] C vs. K denotes the number of times the usual and adjusted chi-square statistics rejected the same table
[e] C vs. Z denotes the number of times the usual chi-square statistic and Zelterman's rejected the same table
[f] K vs. Z denotes the number of times the adjusted chi-square statistic and Zelterman's rejected the same table
[g] MPROPO denotes the average proportion of observations used in the 1000 tables
[h] MPROPC denotes the average proportion of cells used in the 1000 tables

variability on either the modified degrees of freedom used with the adjusted chi-square statistic or the number of rejections that each statistic had. Therefore, the precaution of using this probability structure would probably not be necessary useless in further studies of the adjusted chi-square statistic.

UNIFORM PROBABILITIES

Initially, the modified degrees of freedom for the adjusted chi-square statistic using the uniform probability structure seemed to follow a pattern, but that proved to be a disappointment. For the smaller tables the modified degrees of freedom for the adjusted chi-square statistic were approximately the proportion of cells used multiplied by the degrees of freedom for the usual chi-square statistic. Table 4

Table 4. Number of rejections for three statistics for a given dimension and sparseness level using the uniform probabilities and the modified degrees of freedom for the statistic K.

									M[g]	M[h]
									P	P
									R	R
						C[d]	C[e]	K[f]	O	O
	Cells	MOD				vs.	vs.	vs.	P	P
Dimension	Used	DF	C[a]	K[b]	Z[c]	K	Z	Z	O	C
4x3	75%	4.25	47	47	30	32	30	23	0.75	0.75
4x3	50%	3	50	55	37	27	37	22	0.5	0.5
4x5	75%	8.75	45	49	34	28	34	22	0.75	0.75
4x5	50%	5.75	49	52	41	26	41	22	0.5	0.5
4x8	75%	15.4	41	48	32	26	32	22	0.75	0.74
4x8	50%	10	44	46	35	25	35	21	0.5	0.5
5x8	75%	28	42	69	39	32	39	31	0.75	0.75
5x8	50%	28	41	95	38	22	38	21	0.5	0.5
4x9	75%	24	39	54	29	27	29	22	0.75	0.75
4x9	50%	24	48	91	36	33	36	25	0.5	0.5
3x10	75%	18	42	62	14	34	14	12	0.77	0.77
3x10	50%	18	53	113	28	33	28	20	0.5	0.5
5x10	75%	36	53	71	49	42	49	40	0.76	0.76
5x10	50%	36	55	105	48	34	48	30	0.5	0.5

[a] C denotes the usual chi-square statistic
[b] K denotes the adjusted chi-square statistic
[c] Z denotes Zelterman's D^2
[d] C vs. K denotes the number of times the usual and adjusted chi-square statistics rejected the same table
[e] C vs. Z denotes the number of times the usual chi-square statistic and Zelterman's rejected the same table
[f] K vs. Z denotes the number of times the adjusted chi-square statistic and Zelterman's rejected the same table
[g] MPROPO denotes the average proportion of observations used in the 1000 tables
[h] MPROPC denotes the average proportion of cells used in the 1000 tables

has the modified degrees of freedom for the adjusted chi-square statistic. For example, the 4x3 table with 75% of the cells used had 4.25 for its degrees of freedom. Seventy five percent of the usual 6 degrees of freedom for a 4x3 table is 4.5. This is very close to the 4.25 found running the uniform probability programs. The modified degrees of freedom for the larger tables, however, followed no such pattern. In fact, some of the modified degrees of freedom would need to be larger than the usual degrees of freedom in order to give approximately 50 rejections. Thus, it looked like the degrees of freedom for the adjusted chi-square statistic was a function of table size and proportion of cells used. However, this function appears to be fairly complicated.

The pattern of rejections for the three statistics was the same for the uniform probabilities as it was for the haphazard and descending probabilities. Table 4 lists the number of rejections for the three statistics. The number of rejections for the adjusted chi-square statistic is more than that for the usual chi-square statistic; either chi-square statistic rejected more often than Zelterman's D^2 ; and Zelterman's D^2 rejected almost every time that the usual chi-square statistic did. For example, the 4x3 table with 50% cells used has 50 rejections for the usual chi-square statistic, 55 rejections for the adjusted chi-square statistic, and 30 rejections for Zelterman's D^2. Also the proportion of rejections for the times when any two statistics reject the same table is similar to that of the haphazard and descending probabilities. Look at the 4x3 table with 50% cells used. The usual and adjusted chi-square statistics rejected the same table 27 times, which is about half the 55 rejections for the adjusted chi-square statistic. Since there were no small cells present, this results suggests that the comparison between the usual chi-square and adjusted chi-square statistics should be valid for any contingency table. That is, both statistics could be used for almost any table, whether or not it was sparse. The adjusted chi-square and the Zelterman statistics rejected the same table 22 times which is a little less than half of the 37 rejections for the Zelterman statistic. The usual chi-square and Zelterman's D^2 rejected the same table 37 times which, as seen before, implies that Zelterman's D^2 rejected only when the usual chi-square statistic did. From the number of rejections in Table 4 it would appear that the adjusted chi-square statistic can be used in a table with no sparse cells, but degrees of freedom smaller than those of the usual chi-square statistic are needed for small tables and degrees of freedom larger than those of the usual chi-square statistic are needed for large tables. However, a degrees of freedom pattern is unavailable.

SAME SEED

The table size used in this case was the 4x5 with 25% small cells and 50% small cells. All three probability structures, haphazard, descending, and uniform, were used with both sparseness conditions for a total of 6 tables. The seed utilized here was 1000000. As expected, the modified degrees of freedom for the adjusted chi-square statistic were almost identical to the modified degrees of freedom found using the

Table 5. Number of rejections for three statistics for a 4x5 table with each sparseness level within each probability structure using the modified degrees of freedom for the statistic K and the same seed to start simulation.

									M^g	M^h
									P	P
									R	R
						C^d	C^e	K^f	O	O
Probability	Small	MOD				vs.	vs.	vs.	P	P
Structure	Cells	DF	C^a	K^b	Z^c	K	Z	Z	O	C
Haphazard	25%	12	48	50	43	36	40	35	0.99	0.75
Uniform	25%	8.75	55	53	44	33	44	27	0.75	0.75
Descending	25%	12	41	50	40	31	32	33	0.99	0.75
Haphazard	50%	11	36	49	26	18	26	15	0.93	0.5
Uniform	50%	5.75	55	48	44	26	44	22	0.5	0.5
Descending	50%	11.5	41	49	32	19	30	17	0.93	0.5

[a] C denotes the usual chi-square statistic
[b] K denotes the adjusted chi-square statistic
[c] Z denotes Zelterman's D^2
[d] C vs. K denotes the number of times the usual and adjusted chi-square statistics rejected the same table
[e] C vs. Z denotes the number of times the usual chi-square statistic and Zelterman's rejected the same table
[f] K vs. Z denotes the number of times the adjusted chi-square statistic and Zelterman's rejected the same table
[g] MPROPO denotes the average proportion of observations used in the 1000 tables
[h] MPROPC denotes the average proportion of cells used in the 1000 tables

internal clock as the seed. That is, the modified degrees of freedom found here is very similar to the modified degrees of freedom found for the 4x5 tables in the last three sections. Table 5 shows the modified degrees of freedom for the adjusted chi-square statistic when using the same seed case. For example, the haphazard probability structure with 25% small cells had 12 degrees of freedom, while the degrees of freedom for the 4x5 table with 25% small cells from table 2 was also 12. Thus, it appeared that moving the cell probabilities around a table for the same data set does not really affect the modified degrees of freedom.

The pattern of rejections here is very comparable to the pattern observed in previous sections. Table 5 shows the number of rejections for the three statistics. For example, the haphazard probability structure with 50% small cells had 36 rejections for the usual chi-square statistic, 49 rejections for the adjusted chi-square statistic, and 26 rejections for the Zelterman statistic. The usual and adjusted chi-

square statistic rejected the same tables 18 times, which is a little less than half of the 49 rejections for the adjusted chi-square statistic. The adjusted chi-square and Zelterman statistics rejected the same table 15 times, which is a little more than half of the 26 rejections for the Zelterman statistic. Finally, the usual chi-square statistic and Zelterman's D^2 rejected the same table 26 times, which is the same 26 rejections that the Zelterman statistic had. All the results found with this case followed the patterns that were seen before; which was anticipated. Also the results for the haphazard and descending probability structures look alike except for the random variability due to the algorithm used to assign the "observations" to the individual cells; which was expected. Although it seems logical that the results from these 4x5 tables would apply to the other table sizes in the simulation, it is not known for a fact that the results are similar.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

From the data acquired it appears that the modified degrees of freedom for the adjusted chi-square statistic does not change substantially from the usual degrees of freedom. Of course, they do change a little, but a formula to calculate the modified degrees of freedom could not be found. However, it appears that any pattern would depend on table size and on the proportion of cells that are used to calculate the adjusted degrees of freedom. Also, the usual chi-square statistic still rejected 50 times or less, no matter what the probability structure or table size or sparseness level were in the simulation. This leads one to wonder if maybe the usual chi-square statistic is still valid even when the table has as much as 50% sparse cells. Therefore, as long as the contingency table that is used is one of the same dimensions that were studied in this paper and as long as the sparseness level is either 25% or 50% the usual chi-square statistic should still be valid in this limited range. Also, since Zelterman's rejects almost every time that the usual chi-square statistic rejects it could be used to analyze sparse tables.

Further research would include a mathematical approach to finding a pattern for the modified degrees of freedom for the adjusted chi-square statistic, as well as, finding the distribution of the adjusted chi-square statistic. The reason is that this study assumed that the adjusted chi-square statistic followed a central chi-square statistic distribution and it may not actually have a central chi-square distribution. Also, more table sizes with larger patterns of sparseness should be simulated so the three statistics can be compared for more situations.

SELECTED BIBLIOGRAPHY

Beatty, G. (1983), "Salary Survey of Mathematicians and Statisticians," *Proceedings of the Section on Survey Methods, American Statistical Association*, 743-747.

Birnbaun, Z. W. (1962), Probability and Mathematical Statistics. 252-253. New York, NY: Harper & Brothers.

Cramer, H. (1946), Mathematical Methods of Statistics. 424-434. Princeton, NJ: Princeton University Press.

Dawson, R. B. (1954), "A Simplified Expression for the Variance of the χ^2 Function on a Contingency Table," *Biometrika*, 41, 280.

Fienberg, Stephen E. (1979), "The Use of Chi-Squared Statistics for Categorical Data Problems," *Journal of the Royal Statistical Society*, 41, 54-64.

Fisher, R. A. (1922), "On the Interpretation of χ^2 From Contingency Tables and the Calculation of P," *Journal of the Royal Statistical Society*, 85, 87-94.

Fisher, R. A. (1924), "The conditions Under which χ^2 Measures the Discrepancy Between Observation and Hypothesis," *Jounal of the Royal Statistical Society*, 87, 442-450.

Haberman, Shelby J. (1977), "Log-Linear Models and Frequency Tables with Small Expected Cell Counts," *The Annals of Statistics*, 5, 1148-1169.

Haberman, Shelby J. (1988), "A Warning on the Use of Chi-Square Statistics With Frequency Tables With Small Expected Cell Counts," *American Statistical Asssociation*, 83, 555-560.

Mellina, Catherine Mary (1984), "Families and Work: Employment Policies and Benefits Survey," unpublished Masters report, Oklahoma State University, Department of Statistics.

Mielke, P. W. and Berry, K. J. (1988), "Cumulant Methods for Analyzing Independence of the r-way Contingency Tables and Goodness-of-Fit Frequency Data," *Biometrika*, 75, 790-793.

Read, Timothy R. C. (1984), "Small-Sample Comparisons for the Power Divergence Goodness-of-Fit Statistics," *American Statistical Association*, 79, 929-935.

SAS Language: Reference. Cary, NC: SAS Institute Inc., 1990.

Simonoff, Jeffery S. (1986), "Jackknifing and Bootstrapping Goodness-of-Fit Statistics in Sparse Multinomials," *American Statistical Association*, 81, 1005-1011.

Zelterman, D. (1987), "Goodness-of-Fit Tests for Large Sparse Multinomial Distributions," *American Statistical Association*, 82, 624-629.

APPENDICIES

APPENDIX A

HAPHAZARD TABLES

The 4x3 table with 25% small cells

		V1		
	0.25	0.6	0.15	
0.3	0.075	0.18	0.045	
V2 0.1	0.025	0.06	0.015	
0.4	0.1	0.24	0.06	
0.2	0.05	0.12	0.03	

The 4x3 table with 50% small cells

		V1		
	0.05	0.2	0.75	
0.1	0.005	0.02	0.075	
V2 0.25	0.0125	0.05	0.1875	
0.5	0.025	0.1	0.375	
0.15	0.0075	0.03	0.1125	

The 4x5 table with 25% small cells

			V1			
	0.16	0.21	0.2	0.25	0.18	
0.25	0.04	0.0525	0.05	0.0625	0.045	
V2 0.3	0.048	0.063	0.06	0.075	0.054	
0.44	0.0704	0.0924	0.088	0.11	0.0792	
0.01	0.0016	0.0021	0.002	0.0025	0.0018	

The 4x5 table with 50% small cells

			V1			
	0.16	0.21	0.2	0.25	0.18	
V2 0.4	0.064	0.084	0.08	0.1	0.072	
0.53	0.0848	0.1113	0.106	0.1325	0.0954	
0.05	0.008	0.0105	0.01	0.0125	0.009	
0.02	0.0032	0.0042	0.004	0.005	0.0036	

The 4x8 table with 25% small cells

				V1					
	0.25	0.1	0.15	0.13	0.08	0.09	0.1	0.1	
0.4	0.1	0.04	0.06	0.052	0.032	0.036	0.04	0.04	
V2 0.29	0.0725	0.029	0.0435	0.0377	0.0232	0.0261	0.029	0.029	
0.3	0.075	0.03	0.045	0.039	0.024	0.027	0.03	0.03	
0.01	0.0025	0.001	0.0015	0.0013	0.0008	0.0009	0.001	0.001	

The 4x8 table with 50% small cells

				V1					
	0.25	0.1	0.15	0.13	0.08	0.09	0.1	0.1	
0.4	0.1	0.04	0.06	0.052	0.032	0.036	0.04	0.04	
V2 0.54	0.135	0.054	0.081	0.0702	0.0432	0.0486	0.054	0.054	
0.05	0.0125	0.005	0.0075	0.0065	0.004	0.0045	0.005	0.005	
0.01	0.0025	0.001	0.0015	0.0013	0.0008	0.0009	0.001	0.001	

The 5x8 table with 25% small cells

				V1					
	0.25	0.11	0.15	0.045	0.09	0.046	0.13	0.179	
0.2	0.05	0.022	0.03	0.009	0.018	0.0092	0.026	0.0358	
0.15	0.0375	0.0165	0.0225	0.0068	0.0135	0.0069	0.0195	0.0269	
V2 0.21	0.0525	0.0231	0.0315	0.0095	0.0189	0.0097	0.0273	0.0376	
0.17	0.0425	0.0187	0.0255	0.0077	0.0153	0.0078	0.0221	0.0304	
0.27	0.0675	0.0297	0.0405	0.0122	0.0243	0.0124	0.0351	0.0483	

The 5x8 table with 50% small cells

				V1					
	0.15	0.045	0.084	0.54	0.046	0.02	0.03	0.085	
0.2	0.03	0.009	0.0168	0.108	0.0092	0.004	0.006	0.017	
0.15	0.0225	0.0068	0.0126	0.081	0.0069	0.003	0.0045	0.0128	
V2 0.21	0.0315	0.0095	0.0176	0.1134	0.0097	0.0042	0.0063	0.0179	
0.17	0.0255	0.0077	0.0143	0.0918	0.0078	0.0034	0.0051	0.0145	
0.27	0.0405	0.0122	0.0227	0.1458	0.0124	0.0054	0.0081	0.023	

				The 4x9 table with 25% small cells							
						V1					
		0.07	0.15	0.1	0.08	0.07	0.2	0.11	0.12	0.1	
	0.25	0.0175	0.0375	0.025	0.02	0.0175	0.05	0.0275	0.03	0.025	
V2	0.48	0.0336	0.072	0.048	0.0384	0.0336	0.096	0.0528	0.0576	0.048	
	0.26	0.0182	0.039	0.026	0.0208	0.0182	0.052	0.0286	0.0312	0.026	
	0.01	0.0007	0.0015	0.001	0.0008	0.0007	0.002	0.0011	0.0012	0.001	
				The 4x9 table with 50% small cells							
						V1					
		0.07	0.15	0.1	0.08	0.07	0.11	0.2	0.12	0.1	
	0.44	0.0308	0.066	0.044	0.0352	0.0308	0.0484	0.088	0.0528	0.044	
V2	0.05	0.0035	0.0075	0.005	0.004	0.0035	0.0055	0.01	0.006	0.005	
	0.5	0.035	0.075	0.05	0.04	0.035	0.055	0.1	0.06	0.05	
	0.01	0.0007	0.0015	0.001	0.0008	0.0007	0.0011	0.002	0.0012	0.001	
				The 3x10 table with 25% small cells							
						V1					
		0.12	0.06	0.15	0.13	0.125	0.027	0.03	0.118	0.121	0.119
	0.15	0.018	0.009	0.0225	0.0195	0.0188	0.0041	0.0045	0.0177	0.0182	0.0179
V2	0.55	0.066	0.033	0.0825	0.0715	0.0688	0.0149	0.0165	0.0649	0.0666	0.0655
	0.3	0.036	0.018	0.045	0.039	0.0375	0.0081	0.009	0.0354	0.0363	0.0357
				The 3x10 table with 50% small cells							
						V1					
		0.168	0.01	0.015	0.19	0.177	0.02	0.025	0.19	0.2	0.005
	0.3	0.0504	0.003	0.0045	0.057	0.0531	0.006	0.0075	0.057	0.06	0.0015
V2	0.1	0.0168	0.001	0.0015	0.019	0.0177	0.002	0.0025	0.019	0.02	0.0005
	0.6	0.1008	0.006	0.009	0.114	0.1062	0.012	0.015	0.114	0.12	0.003
				The 5x10 table with 25% small cells							
						V1					
		0.15	0.03	0.05	0.09	0.1	0.02	0.12	0.14	0.13	0.17
	0.2	0.03	0.006	0.01	0.018	0.02	0.004	0.024	0.028	0.026	0.034
	0.15	0.0225	0.0045	0.0075	0.0135	0.015	0.03	0.018	0.021	0.0195	0.0255
V2	0.21	0.0315	0.0063	0.0105	0.0189	0.021	0.0042	0.0252	0.0294	0.0273	0.0357
	0.27	0.0405	0.0081	0.0135	0.0243	0.027	0.0054	0.0324	0.0378	0.0351	0.0459
	0.17	0.0255	0.0051	0.0085	0.0153	0.017	0.0034	0.0204	0.0238	0.0221	0.0289
				The 5x10 table with 50% small cells							
						V1					
		0.22	0.03	0.025	0.199	0.3	0.017	0.09	0.029	0.08	0.01
	0.15	0.033	0.0045	0.0038	0.0299	0.045	0.0026	0.0135	0.0044	0.012	0.0015
	0.2	0.044	0.006	0.005	0.0398	0.06	0.0034	0.018	0.0058	0.016	0.002
V2	0.21	0.0462	0.0063	0.0053	0.0418	0.063	0.0036	0.0189	0.0061	0.0168	0.0021
	0.17	0.0374	0.0051	0.0043	0.0338	0.051	0.0029	0.0153	0.0049	0.0136	0.0017
	0.27	0.0594	0.0081	0.0068	0.0537	0.081	0.0046	0.0243	0.0078	0.0216	0.0027

APPENDIX B

DESCENDING TABLES

The 4x3 table with 25% small cells

		V1		
		0.6	0.25	0.15
	0.4	0.24	0.1	0.06
V2	0.3	0.18	0.075	0.045
	0.2	0.12	0.05	0.03
	0.1	0.06	0.025	0.015

The 4x3 table with 50% small cells

		V1		
		0.75	0.2	0.05
	0.5	0.375	0.1	0.025
V2	0.25	0.1875	0.05	0.0125
	0.15	0.1125	0.03	0.0075
	0.1	0.075	0.02	0.005

The 4x5 table with 25% small cells

		V1				
		0.25	0.21	0.2	0.18	0.16
	0.44	0.11	0.0924	0.088	0.0792	0.0704
V2	0.075	0.075	0.063	0.06	0.054	0.048
	0.0625	0.0625	0.0525	0.05	0.045	0.04
	0.0025	0.0025	0.0021	0.002	0.0018	0.0016

The 4x5 table with 50% small cells

		V1				
		0.25	0.21	0.2	0.18	0.16
V2	0.53	0.1325	0.1113	0.106	0.0954	0.0848
	0.4	0.1	0.084	0.08	0.072	0.064
	0.05	0.0125	0.0105	0.01	0.009	0.008
	0.02	0.005	0.0042	0.004	0.0036	0.0032

The 4x8 table with 25% small cells

		V1						
	0.25	0.15	0.13	0.1	0.1	0.1	0.09	0.08
0.4	0.1	0.06	0.052	0.04	0.04	0.04	0.036	0.032
V2 0.3	0.075	0.045	0.039	0.03	0.03	0.03	0.027	0.024
0.29	0.0725	0.0435	0.0377	0.029	0.029	0.029	0.0261	0.0232
0.01	0.0025	0.0015	0.0013	0.001	0.001	0.001	0.0009	0.0008

The 4x8 table with 50% small cells

		V1						
	0.25	0.15	0.13	0.1	0.1	0.1	0.09	0.08
0.54	0.135	0.081	0.0702	0.054	0.054	0.054	0.0486	0.0432
V2 0.4	0.1	0.06	0.052	0.04	0.04	0.04	0.036	0.032
0.05	0.0125	0.0075	0.0065	0.005	0.005	0.005	0.0045	0.004
0.01	0.0025	0.0015	0.0013	0.001	0.001	0.001	0.0009	0.0008

The 5x8 table with 25% small cells

		V1						
	0.25	0.179	0.15	0.13	0.11	0.09	0.046	0.045
0.27	0.0675	0.04833	0.0405	0.0351	0.0297	0.0243	0.01242	0.01215
0.21	0.0525	0.03759	0.0315	0.0273	0.0231	0.0189	0.00966	0.00945
V2 0.2	0.05	0.0358	0.03	0.026	0.022	0.018	0.0092	0.009
0.17	0.0425	0.03043	0.0255	0.0221	0.0187	0.0153	0.00782	0.00765
0.15	0.0374	0.02685	0.0225	0.0195	0.0165	0.0135	0.0069	0.00675

The 5x8 table with 50% small cells

		V1						
	0.54	0.15	0.085	0.084	0.046	0.045	0.03	0.02
0.27	0.1458	0.0405	0.02295	0.02268	0.01242	0.01215	0.0081	0.0054
0.21	0.1134	0.0315	0.01785	0.01764	0.00966	0.00945	0.0063	0.0042
V2 0.2	0.108	0.03	0.017	0.0168	0.0092	0.009	0.006	0.004
0.17	0.0918	0.0255	0.01445	0.01428	0.00782	0.00765	0.0051	0.0034
0.15	0.081	0.0225	0.01275	0.0126	0.0069	0.00675	0.0045	0.003

				The 4x9 table with 25% small cells							
						V1					
		0.2	0.15	0.12	0.11	0.1	0.1	0.08	0.07	0.07	
	0.48	0.096	0.072	0.0576	0.0528	0.048	0.048	0.0384	0.0336	0.0336	
V2	0.26	0.052	0.039	0.0312	0.0286	0.026	0.026	0.0208	0.0182	0.0182	
	0.25	0.05	0.0375	0.03	0.0275	0.025	0.025	0.02	0.0175	0.0175	
	0.01	0.002	0.0015	0.0012	0.0011	0.001	0.001	0.0008	0.0007	0.0007	
				The 4x9 table with 50% small cells							
						V1					
		0.2	0.15	0.12	0.11	0.1	0.1	0.08	0.07	0.07	
	0.5	0.1	0.075	0.06	0.055	0.05	0.05	0.04	0.035	0.035	
V2	0.44	0.088	0.066	0.0528	0.0484	0.044	0.044	0.0352	0.0308	0.0308	
	0.05	0.01	0.0075	0.006	0.0055	0.005	0.005	0.004	0.0035	0.0035	
	0.01	0.002	0.0015	0.0012	0.0011	0.001	0.001	0.0008	0.0007	0.0007	
				The 3x10 table with 25% small cells							
						V1					
		0.15	0.13	0.125	0.121	0.12	0.119	0.118	0.06	0.03	0.027
	0.55	0.0825	0.0715	0.06875	0.06655	0.066	0.06545	0.0649	0.033	0.0164	0.01485
V2	0.3	0.045	0.039	0.0375	0.0363	0.036	0.0357	0.0354	0.018	0.009	0.0081
	0.15	0.0225	0.0195	0.01875	0.01815	0.018	0.01785	0.0177	0.009	0.0045	0.00405
				The 3x10 table with 50% small cells							
						V1					
		0.2	0.19	0.19	0.177	0.168	0.025	0.02	0.015	0.01	0.005
	0.6	0.12	0.114	0.144	0.1062	0.1008	0.015	0.012	0.009	0.006	0.003
V2	0.3	0.06	0.057	0.057	0.0531	0.0504	0.0075	0.006	0.0045	0.003	0.0015
	0.1	0.02	0.019	0.019	0.0177	0.0168	0.0025	0.002	0.0015	0.001	0.0005
				The 5x10 table with 25% small cells							
						V1					
		0.17	0.15	0.14	0.13	0.12	0.1	0.09	0.05	0.03	0.02
	0.27	0.0459	0.0405	0.0378	0.0351	0.0324	0.027	0.0243	0.0135	0.0081	0.0054
	0.21	0.0357	0.0315	0.0294	0.0273	0.0252	0.021	0.0189	0.0105	0.0063	0.0042
V2	0.2	0.034	0.03	0.028	0.026	0.024	0.02	0.018	0.01	0.006	0.004
	0.17	0.0289	0.0255	0.0238	0.0221	0.0204	0.017	0.0153	0.0085	0.0051	0.0034
	0.15	0.0255	0.0225	0.021	0.0195	0.018	0.015	0.0135	0.0075	0.0045	0.003
				The 5x10 table with 50% small cells							
						V1					
		0.3	0.22	0.199	0.09	0.08	0.03	0.029	0.025	0.017	0.01
	0.27	0.081	0.0594	0.05373	0.0243	0.0216	0.0081	0.00783	0.00675	0.00459	0.0027
	0.21	0.063	0.0462	0.04179	0.0189	0.0168	0.0063	0.00609	0.00525	0.00357	0.0021
V2	0.2	0.06	0.044	0.0398	0.018	0.016	0.006	0.0058	0.005	0.0034	0.002
	0.17	0.051	0.0374	0.03383	0.0153	0.0136	0.0051	0.00493	0.00425	0.00289	0.0017
	0.15	0.045	0.033	0.02985	0.0135	0.012	0.0045	0.00435	0.00375	0.00255	0.0015

APPENDIX C

SAMPLE PROGRAM

Sample Program for the 4x3, 25% Small,
Haphazzard Probability Case

```
DM'OUTPUT;CLEAR;LOG;CLEAR;';
OPTIONS PS=60 LS=80 NODATE;
DATA TEST1;
DO J=1 TO 1000;
DO I=1 TO 120;
X=RANUNI(0);
 IF X <= .075 THEN DO;
  V1=1; V2=1; CNT=1;
 END;
 IF .075 < X <= .255 THEN DO;
  V1=2; V2=1; CNT=1;
 END;
 IF .255 < X <= .30 THEN DO;
  V1=3; V2=1;  CNT=1;
 END;
 IF .30 < X <= .325 THEN DO;
  V1=1; V2=2;  CNT=1;
 END;
 IF .325 < X <= .385 THEN DO;
  V1=2; V2=2;  CNT=1;
 END;
 IF .385 < X <= .40  THEN DO;
  V1=3; V2=2;  CNT=1;
 END;
 IF .40  < X <= .50  THEN DO;
  V1=1; V2=3;  CNT=1;
 END;
 IF .50  < X <= .74 THEN DO;
  V1=2; V2=3;   CNT=1;
 END;
 IF .74 < X <= .80 THEN DO;
  V1=3; V2=3;   CNT=1;
 END;
 IF .80 < X <= .85  THEN DO;
  V1=1; V2=4;   CNT=1;
 END;
 IF .85  < X <= .97  THEN DO;
  V1=2; V2=4;   CNT=1;
 END;
 IF .97  < X <= 1.0 THEN DO
  V1=3; V2=4;   CNT=1;
 END;
OUTPUT;
END;
END;
```

```
DATA TEST2;
 DO J=1 TO 1000;
  DO V2=1 TO 4;
   DO V1=1 TO 3;
    CNT=0;
    OUTPUT;
   END;
  END;
 END;

DATA TEST; SET TEST1 TEST2;
PROC SORT;
BY J V2 V1;
PROC MEANS NOPRINT;
BY J V2;
VAR CNT;
OUTPUT OUT=ROWT SUM=RT;



DATA TWO;
MERGE TEST ROWT;
BY J V2;
PROC SORT DATA=TWO;
BY J V1;
PROC MEANS DATA=TWO NOPRINT;
BY J V1;
VAR CNT;
OUTPUT OUT=COLT SUM=CT;

PROC MEANS DATA=TWO NOPRINT;
BY J V1 V2;
VAR CNT;
OUTPUT OUT=CNTT SUM=COUNT;

PROC SORT DATA=CNTT;
BY J V2;

DATA THREE;
MERGE CNTT ROWT;
BY J V2;

PROC SORT DATA=THREE;
BY J V1;

DATA FOUR;
MERGE THREE COLT;
BY J V1;
EXPF = RT*CT/120;
IF RT = 0 THEN RT = 0.000001;
IF CT = 0 THEN CT = 0.000001;
INVR = 1/RT;
INVC = 1/CT;
IF EXPF = 0 THEN EXPF = 0.000001;
```

```
ZCELL = (((COUNT - EXPF)**2) - COUNT)/EXPF;
CELLCHI2 = ((COUNT-EXPF)**2)/EXPF;
IF EXPF >= 5 THEN CELLKP=CELLCHI2;
ELSE CELLKP=.;
IF EXPF >= 5 THEN COUNT2=COUNT;
ELSE COUNT2=0;
IF EXPF >= 5 THEN CELL=1;
ELSE CELL=0;

PROC MEANS DATA=FOUR NOPRINT;
BY J;
VAR CELLKP CELLCHI2 ZCELL INVR INVC COUNT2 CELL;
OUTPUT OUT=FIVE MEAN=MK M2 MZ MR MC MC2 MCELL SUM=SM S2 SZ SR SC SC2 SCELL;
RUN;


DATA SIX; SET FIVE;
ADJCHI2=12*MK;
N = 3; M = 4; TOT = 120;
MU = (N-1)*(TOT-N)/(TOT-1);
NU = (M-1)*(TOT-M)/(TOT-1);
SIGMA = (TOT*SR - N*N)/(TOT-2);
TAU = (TOT*SC - M*M)/(TOT-2);
VAR = ABS(2*TOT/(TOT-3)*(NU-SIGMA)*(MU-TAU) + 4*SIGMA*TAU/(TOT-1));
MEAN = TOT/(TOT-1)*(N-1)*(M-1) - N*M;
STANZ = (SZ - MEAN)/SQRT(VAR);
PROPO = SC2/TOT;
PROPC = SCELL/(N*M);
OSL2=1-PROBCHI(S2,6);
OSLK=1-PROBCHI(ADJCHI2,5.25);
OSLZ=1-PROBNORM(SZ);


 IF OSL2 < 0.05 THEN R2=1;
ELSE R2=0;
 IF OSLK < 0.05 THEN RK=1;
ELSE RK=0;
 IF OSLZ < 0.05 THEN RZ=1;
ELSE RZ=0;
 IF R2=1 AND RK=1 THEN R2K=1;
ELSE R2K=0;
 IF R2=1 AND RZ=1 THEN R2Z=1;
ELSE R2Z=0;
 IF RK=1 AND RZ=1 THEN RKZ=1;
ELSE RKZ=0;
 IF R2=1 AND RK=0 THEN R2NK=1;
ELSE R2NK=0;
 IF R2=1 AND RZ=0 THEN R2NZ=1;
ELSE R2NZ=0;
 IF RK=1 AND RZ=0 THEN RKNZ=1;
ELSE RKNZ=0;
```

```
PROC MEANS DATA=SIX NOPRINT;
VAR R2 RK RZ R2K R2Z RKZ R2NK R2NZ RKNZ PROPO PROPC;
OUTPUT OUT=SEVEN MEAN=MR2 MRK MRZ MR2K MR2Z MRKZ MR2NK MR2NZ MRKNZ
  MPROPO MPROPC
SUM=REJECT2 REJECTK REJECTZ RJCT2K RJCT2Z RJCTKZ
        RJCTR2NK RJCTR2NZ RJCTRKNZ SPROPO SPROPC;
PROC PRINT;
VAR REJECT2 REJECTK REJECTZ RJCT2K RJCT2Z RJCTKZ RJCTR2NK RJCTR2NZ
  RJCTRKNZ MPROPO MPROPC;
RUN;
```

VITA

Micah Ward

Candidate for the Degree of

Master of Science

Thesis: AN ADJUSTMENT TO THE CHI-SQUARE TEST IN CASES OF SPARSE CONTINGENCY TABLES

Major Field: Statistics

Biographical:

Education: Graduated from Del City High School, Del City, Oklahoma in May 1990; received Bachelor of Science in Statistics from Oklahoma State University, Stillwater, Oklahoma in May 1994. Completed the requirements for the Master of Science degree with a major in Statistics at Oklahoma State University in December, 1996.

Experience: Employed by Oklahoma State University, Department of Statistics as an undergraduate teaching assistant in 1994 and as a graduate teaching assistant, 1994 to present.

Professional Memberships: American Statistical Association.