A COMPARISON OF THE CHI-SQUARE TEST FOR 1 df AND THE FISHER EXACT TEST

Anders Sweetland

RAND Corporation
Santa Monica, California

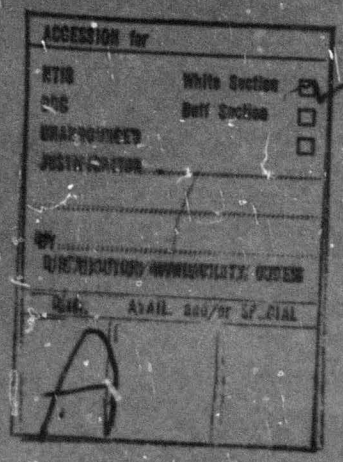
June 1972

A COMPARISON OF THE CHI-SQUARE TEST FOR 1 df AND THE FISHER EXACT TEST

Anders Sweetland

June 1972

P-4850



SUMMARY

The chi-square test was compared with the Fisher exact test using Ns ranging from 3 to 69. Contrary to expectations, there was closer agreement between the two tests with the smaller Ns. The chi-square test gave very good approximations of the true probabilities--especially when the two groups being compared were nearly equal in sample size--but only if it was used as a one-tailed test. If the groups being compared had very unequal sample sizes (ratios of 7:1 and greater), the chi-square test sometimes gave questionable results. A table of one-tailed probabilities is provided.

ACKNOWLEDGMENTS

I am indebted to Phan Phi Long, who wrote the computer program used in the testing, and to his brother, Phan Phi Hung, who calculated the table of one-tailed probabilities.

CONTENTS

I. INTRODUCTION

Writers of statistical texts are in almost unanimous agreement that small theoretical frequencies should be avoided when computing chi-square for 1 df. The most frequent recommendation is that no cell should have a theoretical frequency of less than 5.0; and most writers generally add: "and preferably not less than 10.0." (An excellent study by Roscoe and Byars [1] includes a summary of the research on this topic.)

The stand against small frequencies was made so forcibly by Lewis and Burke [2] that some people quit using the chi-square test.

A minority report was filed by Edwards [3] in his reply to Lewis and Burke. Edwards included six examples that violated the small-frequency fiat. He computed both the Fisher test and the chi-square test corrected for continuity. In all six comparisons, the differences were so small as to have no practical significance: five of the six differed in the third decimal place. Edwards sensibly concluded that the preferred strategy was first to use the chi-square test and then compute the Fisher only if chi-square probabilities fell in the critical region.

II. METHODS

The author has long been intrigued by the close agreement between the Fisher and the chi-square test shown by Edwards. Recent access to a mini-computer made possible a more complete exploration. Accordingly, a program was written that

(1) Computed the Fisher exact test.
(2) Computed the chi-square test corrected for continuity.
(3) Converted the chi-square results to exact probabilities.
(4) Divided the chi-square probabilities by 2. (Fisher probabilities are one-tailed, while chi-square probabilities are two-tailed.)

To simplify discussion, the following arrangement is always assumed:

	Responses	
	No	Yes
Experimentals	A	B
Controls	C	D

N

Note that the rows refer to entities (or independent variables) and the columns to attributes (or dependent variables). A, B, C, and D identify the four cells of the tetrad. N is the total frequency count. All probabilities, except when specifically noted, are one-tailed.

The frame of reference is that of a person in social science research; the major concern is whether or not to reject the null hypothesis. Consequently, we were interested only in the critical region defined as $P \leq .10$ (one-tailed).

It was easiest to first create all of the tetrads for a fixed set of marginal totals and then, beginning with the extreme cases, to work

toward the middle until both probabilities exceeded .10. An example is shown for the marginal totals 8-4, 8-4:

	A	B	C	D	E
	4 4	5 3	6 2	7 1	8 0
	4 0	3 1	2 2	1 3	0 4
Fisher	.141		.406	.067	.002
Chi-square	.140		.414	.065	.002

Tetrad A is tested first. Since the p are greater than .10, testing this end of the set of tetrads is discontinued. Tetrad E is tested next, then D, then C. Tetrad C exceeds the cutoff of .10, so testing is stopped. This tetrad set provides two comparisons: D and E.

III. FINDINGS

The first study tested all tetrad sets N = 3 through N = 20. The results are summarized in Table 1 and in Table 2 (on page 5). (In these and the following tables, chi-square probabilities are always subtracted from Fisher probabilities; hence a positive difference indicates a larger Fisher probability.)

Table 1

DIFFERENCES BETWEEN CHI-SQUARE AND FISHER PROBABILITIES

	Sample Size			
Difference	3-8	13	20	3-20
.014			1	1
.013			1	1
.012				
.011			1	4
.010			2	10
.009		1		5
.008			1	7
.007		2	1	11
.006	1		1	8
.005	1	1	3	13
.004	4		1	9
.003		1	2	14
.002		1	7	21
.001		2	5	38
.000		4	38	157
-.001	4	3	26	118
-.002	1	4	11	56
-.003	2	2	4	37
-.004		1	3	22
-.005		1	3	14
-.006			3	8
-.007			1	1

The entries in the difference column are midpoints. Ninety-nine percent of the differences fall within the range of -.007 to +.010; 90 percent within the range ±.005; and 70 percent within ±.002. The special case of the tetrad

N-1	0
0	1

is shown in Table 3 (on page 7).

The major surprise in this phase was that the smaller Ns yielded smaller differences, which is exactly the opposite of what might be expected. Plots similar to Table 1 were made for each N. The results were the same: increasing N increased the range of the differences.

The differences were then plotted for each of the row totals within N. The results for N = 20 are shown in Table 2. The peculiar shape of this distribution was characteristic of all N, except for N less than 14. The smaller Ns had too few entries to delineate the pattern clearly. A cross-check was made using N = 25. The same pattern as before emerged.

Table 2

DISTRIBUTIONS OF DIFFERENCES SEPARATED BY ROW TOTALS; N = 20

	Row Totals[a]								
Differences	10-10	11-9	12-8	13-7	14-6	15-5	16-4	17-3	18-2
.014									1
.013								1	
.012									
.011									
.010						1		1	
.009									
.008							1		
.007					1				
.006						1			
.005						1	1		1
.004					1				
.003					1	1			
.002				4	2		1		
.001			2		1	1		1	
.000	10	8	7	7	3	2	1		
-.001	14	9	3						
-.002	4	5	2						
-.003	4								
-.004		2	1						
-.005		1	1	1					
-.006			1	1	1				
-.007				1					

[a]Row totals 19-1 not shown. See Table 3.

As noted, a peculiar distribution that can be likened in appearance to a spoon with a bent handle was characteristic of all but the smaller Ns. The following characteristics appeared in all distributions including the smallest (N = 3-20, 25):

(1) When the row totals were equal or nearly equal, the differences, when they occured, were small and negative (smaller than -.005).
(2) When row totals were approximately 2:1, differences were about equally divided between plus and minus.
(3) When row totals were 3:1 and greater, the differences were always positive--i.e., the chi-square probabilities were always smaller than the Fisher probabilities.

Although the findings shown in Tables 1 and 2 suggest that the chi-square test could be used with a reasonable degree of impunity, we decided that it might be prudent to look more closely at the larger differences appearing in the "bowl of the spoon." These are the differences that occur when row totals differ by a factor of 3:1 or greater. In looking at these differences, we discovered some consistent peculiarities. Certain "patterns" of the frequency counts in cells B, C, and D (and particularly in cells C and D) were contributing all the larger differences. For example, with the row total fixed at C + D = 4, there are five possible permutations of C and D. But only the combination C = 2, D = 2 yielded a difference greater than .010. This provoked another search that included all of the differences of .010 or larger shown in Table 2. The results are shown in Table 3.

The reader should not try to infer too much from the patterns shown in Table 3. They tell only part of the story, and little significance should be given to the fact that there is no more security in the larger Ns than in the smaller. It is doubtful that much social science research will use, say, a control group of 3 contrasted with an experimental group of 47.

The frustrating finding is that the relationship between the size of the difference and the size of N is nonmonotonic: As N increases,

Table 3

THE EFFECT OF PATTERNS ON THE DIFFERENCES BETWEEN THE TWO METHODS

	Patterns					
N	A 0 2 2	A 1 1 2	A 0 1 2	A 0 3 2	A 0 1 1	A 0 0 1
50	.005	.007	.002	.007	.031	.020
40	.007	.010	.004	.010	.032	.024
30	.010	.013	.006	.012	.028	.029
25	.011	.014	.008	.012	.023	.032
20	.011	.013	.010	.010	.014	.033
15	.009	.008	.010	.006	(a)	.031
10	(a)	(a)	.006	(a)	(a)	.020

[a]Not computed because the probability was greater than .10. Note that all the differences are positive.

the differences rise to some peak and then fall. There is no predicting when the peak will occur nor how great it will be.

The findings shown in Table 3 indicated that it would be necessary to explore the tails of the distributions using larger Ns and unequal row totals.

The first phase looked at row total ratios of 2:1 for N = 30, 35, and 60. The agreement under these conditions was excellent. The largest difference was between a Fisher probability of .009 and a chi-square probability of .004. All other differences were .003 or less, with differences of .000 and .001 in the great majority.

In the second phase, the row totals were fixed at 3:1 with Ns of 32, 44, and 60. The agreement between the two methods was still very good. The largest difference was between a Fisher probability of .042 and a chi-square probability of .054. The remaining differences were all less than .010. Again, differences of .000 and .001 predominated but not so strongly as in the 2:1 case.

The third phase consisted of a search in the tails for additional patterns contributing large differences. A score were discovered all

Table 4

ADDITIONAL PATTERNS SHOWING LARGE DIFFERENCES

N	A 0 2 1	A 1 1 1	A 0 3 1	A 0 4 1
69	.031	.028	.030	.024
65	.032	.027	.029	.022
50	.029	.017	.024	.011
35	.019	(a)	(a)	(a)
20	(a)	(a)	(a)	(a)

[a]Probability greater than .10.
Note that a peak was discovered only in the first pattern.

occurring with ratios of 7:1 or greater. Those peaking at .020 and greater are shown in Table 4.

Before concluding this part of the study, it is desirable to introduce one additional finding that gives insight into the behavior of the chi-square probabilities when the row totals are varied. This has to do with the shape of their distributions. The finding is best shown by contrasting two examples. The first shows a segment of adjacent tetrads with row totals set at a ratio of 2:1:

	15 5 5 5	16 4 4 6	17 3 3 7	18 2 2 8	19 1 1 9
Fisher	.1688	.0387	.0048	.0003	.0000
Chi-square	.1689	.0375	.0046	.0003	.0000
Difference	-.0002	.0012	.0002	.0000	.0000

Notice how the chi-square probabilities fluctuate between being larger and smaller than the Fisher probabilities. This behavior is markedly different when row total ratios are extreme:

	53 5 / 5 2	54 4 / 4 3	55 3 / 3 4	56 2 / 2 5
Fisher	.1612	.0229	.0016	.0001
Chi-square	.1678	.0121	.0002	.0000
Difference	-.0066	.0183	.0014	.0001

When the row totals are very unequal beginning with the extreme cases, the chi-square probabilities start much lower, rise more rapidly, and eventually reach a crossover point beyond which the difference becomes negative. When this relationship is expressed as a ratio, the agreement in the extremes is very poor but constantly improves as one approaches the middle of the distributions. In this example, the right-hand tetrad has a Fisher-to-chi-square ratio of approximately 500:1. This improves to 8:1, 2:1, and finally 1:1. Unfortunately the differences, which are the critical element, do not behave so nicely.

The problem areas can almost always be identified by two characteristics: (1) the chi-square probability is about .01 to .02 and (2) cells B, C, and D have nearly equal frequencies. The differences are never greater than .033 and are always positive. They occur only when row totals are very unequal.

IV. PRACTICAL CONSIDERATIONS

There is one circumstance in which the chi-square approximation is seriously in error. This occurs when the tetrad has a probability greater than .50:

	A	B	C	D	E
	10 4	11 3	12 2	13 1	14 0
	11 0	10 1	9 2	8 3	7 4
p tetrad	.079	.317	.396	.183	.026
Fisher	.079	.396		.209	.026
Chi-square	.083	.388	.388	.208	.028

The Fisher probability for the left-hand tail of tetrad C is .792 = .396 + .317 + .079. For the right-hand tail, it is .605 = .396 + .183 + .026. Neither of these agrees with the chi-square approximation of .388.

Most sets of tetrads contain one case in which the probability exceeds .50. There is a simple test for determining when this occurs: compute the diagonal cross-products and correct for continuity. If the result is negative, the probability is greater than .50, and the chi-square approximation will be incorrect. In the previous example,

$$\begin{aligned} Q &= |BC - AD| - N/2 \\ &= |(12(2) - (2)(9)| - 25/2 \\ &= -6.5 \ . \end{aligned}$$

Since the major concern is whether to reject the null hypothesis, there is really no need to compute chi-square when this test yields a negative quantity. If you must know the probability of this tetrad, it can be approximated using the chi-square test. This may cause a mild psychological wrench because one is used to thinking of the test as being two-tailed. It should be remembered, however, that the chi-square probabilities have been divided by two to make them comparable to the Fisher test.

In the example, the chi-square probability of tetrad B (.388) represents the probability of B and all the more extreme cases to the left. Similarly, the probability of tetrad D (.208) represents the same case for the right-hand tail. We can determine the "tail-less" probability of C by subtracting these two from 1.0:

$$1.0 - .388 - .208 = .404\ (.396).$$

The probabilities of the left- and right-hand tails are then determined:

$$\text{Left tail} \quad .404 + .388 = .792\ (.792)$$
$$\text{Right tail} \quad .404 + .208 = .612\ (.605)$$

The approximations compare favorably with the true probabilities in the parentheses. Several similar approximations follow, with C = chi-square probability and F = Fisher probability.

Tetrad	Chi-square	Left	Right
7 6	.417	C .723	.583
6 6		F .764	.582
8 6	.467	C .533	.728
7 4		F .533	.770
13 2	.353	C .800	.647
9 1		F .802	.654
12 2	.388	C .612	.792
9 2		F .604	.791

By now the sharp-eyed reader will have noticed that the chi-square probabilities in the second column are equivalent to 1.0 minus the smaller tail: .417 = 1.0 - .583, etc. It is no coincidence that the two identical chi-square probabilities (tetrads B and C) occured when the correction for continuity resulted in a negative quantity. This always happens and led us to question the current usage of the chi-square test, which in turn, led to the following.

As currently used, the chi-square test is two-tailed. The tabled values, in effect, are obtained by computing for one tail and doubling this quantity. The assumption is that the sets of tetrads are symmetrical. But symmetry is the exception. It occurs only when row totals are equal--an infrequent occurrence. For example, there are 286 different tetrads for N = 20. Of these, only 26 have mirror images--i.e., there are only 52 tetrads that yield a correct two-tailed calculation. For N = 21, there are 384 tetrads. None of these have mirror images. As N increases, the proportion of mirror images decreases. A conservative statement would be that fewer than 10 percent of the tetrads have mirror images for Ns that are typically used in social science research.

To illustrate the magnitude of the error caused by the assumption of symmetry, consider the case of N = 49. The calculations are for the row totals of 25 and 24. These are very close to being equal, so the tails should show close agreement. The one-tailed Fisher probabilities are shown for the 6th through the 10th tetrads:

Left tail	.00013	.00115	.00707	.03085	.09847
Right tail	.00004	.00042	.00305	.01563	.05790

Even in this case where, presumably, we should get very good agreement between the tails, the discrepancy is large.

Two more examples are given for comparison. The first consists of row totals 25 and 23--with a difference of 2. The 7th through the 10th tetrads are shown:

Left tail	.00410	.02028	.07203	.18971
Right tail	.00064	.00445	.02158	.07550

The third example is for row totals 26 and 23, the same N as the first example. In effect, one case has been taken out of the control group and put into the experimental group. The 7th through the 10th positions are shown:

Left tail	.00641	.02873	.09366	.22894
Right tail	.00044	.00320	.01633	.06010

Even this modest change has a serious effect on the symmetry of the probabilities.

We have compared the tails of several hundred sets of tetrads. The overwhelming majority show less agreement than the examples that were deliberately selected to show the "best cases." In brief, the two-tailed test yields very poor approximations of the true probability unless the row totals are equal. In contrast, the one-tailed test gives excellent results. The possible exception occurs when the row totals are very unequal (ratios greater than 5:1), but in this case, the two-tailed test is useless because the two tails have no resemblance to each other.

We no longer follow the conventional use of chi-square. We use only the one-tailed test. It is much more accurate. Of greater importance, it is more consonant with the research we do--e.g., determining differences between drug abusers and nonabusers, communists and noncommunists, neurotics and nonneurotics. In these cases, we always have some idea about the direction of the responses. To us, this direction is more important than absolute alpha level. Toward this objective the following one-tailed chi-square probabilities are given in Table 5 on the following page. They can also be used to infer the probability of the tetrad with the negative correction for continuity.

Even with computing facilities available, we do not compute the Fisher probabilities because of the extensive labor involved. Consider the following real-life data. Fifty-one defectors are compared with 84 prisoners of war:

	No	Yes
Defectors	19	32
Prisoners	62	22

Computing the Fisher requires inputting 20 tetrads; computing the chi-square requires only 1. And this is simply one item out of 78 used to develop the scale. The complete analysis using the Fisher would have required inputting more than 1500 tetrads.

Table 5

CHI-SQUARE PROBABILITIES: ONE-TAILED, df = 1

p		p		p	
.49	.0006	.29	.307	.16	.989
	.0015		.323		1.031
.48	.003	.28	.340	.15	1.075
	.004		.358		1.120
.47	.006	.27	.377	.14	1.168
	.008		.395		1.217
.46	.010	.26	.414	.13	1.269
	.013		.435		1.325
.45	.016	.25	.455	.12	1.389
	.019		.476		1.441
.44	.023	.24	.500	.11	1.505
	.027		.522		1.572
.43	.031	.23	.546	.10	1.643
	.036		.571		1.718
.42	.041	.22	.597	.09	1.798
	.046		.623		1.883
.41	.052	.21	.651	.08	1.975
	.058		.679		2.074
.40	.064	.20	.710	.07	2.178
	.071		.739		2.297
.39	.078	.19	.772	.06	2.418
	.086		.804		2.555
.38	.094	.18	.838	.05	2.706
	.102		.874		2.875
.37	.110	.17	.911	.04	3.065
	.120		.949		3.282
.36	.129			.03	3.538
	.139				3.843
.35	.149			.02	4.217
	.160				4.707
.34	.171			.01	5.412
	.182				
.33	.194			.005	6.630
	.206				
.32	.219			.001	9.550
	.233				
.31	.246				
	.261				
.30	.275				
	.291				

Unless row totals are equal, the one-tailed probabilities should not be doubled to make a two-tailed test. If a two-tailed test is required and the row totals are unequal, the only satisfactory solution is to determine the more extreme probability of the opposite tail. In the first illustration in this section (see page 10), if one wished to make a two-tailed test for tetrad A (.083), he would add the more extreme case of tetrad E (.028). The two-tailed probability then is .111 = .083 + .028, which is considerably different from (2) (.083) = .166 gotten by the doubling convention.

Three additional pairs of tails are given so the reader may make his own comparisons. In each case, the row totals differ by only 1--i.e., the agreement between tails is the best one can expect when row totals differ:

N = 19	Left	.00011	.00449	.05126	.24221
	Right	.00001	.00099	.01852	.12764
N = 35	Left	.00048	.00498	.03029	.11709
	Right	.00013	.00173	.01342	.06405
N = 65	Left	.00430	.01699	.05284	.13164
	Right	.00200	.00892	.03111	.08627

As has been shown, the largest errors occur when the row totals are unequal. The largest error that one can make is .033, and because this error is always positive--occuring only when row totals differ by ratios of 5:1 and greater--it can be used to establish the bounds of the alpha level if desired.

A much more sensible approach would be to divide the larger group into several smaller groups, making the row totals nearly equal. This would yield several tests, and the chi-square approximations would be very accurate. As an example, if the experimental group contained 20 cases and the control group 62, we would split the latter into three smaller groups and make three one-tailed tests.

REFERENCES

1. Roscoe, J. T., and J. A. Byars, "Sample Size Restraints Commonly Imposed on the Use of the Chi-square Statistic," *Journal of the American Statistical Association*, Vol. 66, 1971, pp. 755-759.

2. Lewis, D., and C. J. Burke, "The Use and Mis-use of the Chi-square Test," *Psychological Bulletin*, Vol. 46, 1949, pp. 433-489.

3. Edwards, A. L., "On 'The Use and Mis-use of the Chi-Square Test,'" *Psychological Bulletin*, Vol. 47, 1950, pp. 341-346.