Rapid Approximations for some Chi Square and Derived Correlational Statistics Used in the Social and Biological Sciences.[1]

HANS W. WENDT, Valparaiso University[2]

Abstract

Short-cut expressions are given which approximate basic forms of Chi Square first and second order interactions, the fourfold point coefficient of correlation, and the contingency coefficient. This strictly empirical approach assumes that cell frequencies in tabled data distributions are roughly symmetrical. Under these conditions, differences among marginal totals of 2 by k tables are small and little error is introduced by substituting, in effect, an average value for pairs of unequal diagonal cell entries. The estimates are compared with conventional solutions for 20 examples, with two thirds of these falling within approximately $\pm$ 10% of the value of the statistic in question as ordinarily computed. The short cuts approach the more precise solutions as the cell entries involved become more fully symmetrical. Criteria are offered for predicting the direction of the error resulting from the simplified substitutions.

Introduction

Some time ago a short cut method was pointed out for dealing with special cases of the Chi Square situation (6), intended to complement a number of standard discussions of nonparametric data treatments (1, 2, 3, 5, 7). Extensions of the original procedure have since been used in experimental work and in teaching students to run rapid over-all checks on conventional calculations. We have found them helpful in brief treatments of the significance and correlation topics in courses for non-specialists who needed some basic statistical criteria that could be used in field settings and would permit quick decisions without incurring excessive error.

Summaries of the modified procedures are given below. Analytically or empirically, criteria might be derived by means of which some equivalent of confidence intervals would be available for these estimates. However, this complication could easily defeat their basic purpose. Instead, we are presenting a number of practical examples (not systematically selected except for the constraints discussed below) where standard computational procedures and short cuts have been used side by side, to show the type or error that can be expected. The investigator will decide for himself in which situations the rapid estimates should merely complement the more rigorous solutions (as in preliminary screening of data from field work and gross checks on more elaborate work); or where circumstances suggest that they might be used independently.

---

[1] The author is indebted to Waldemar C. Gunther for suggesting possible applications of this methodology, originally based on behavioral studies, to biological and zoological research.

[2] Now at Macalester College.

The 2 by 2 table

Experimental and survey data are often cast in the form of a fourfold table of classes of events or subjects, with observed frequencies labelled

A	B
C	D

The formula customarily used for determining interaction is abbreviated from the definitional expression for Chi Square, or (without correction for continuity),

$$\chi^2 = \frac{N\,(BC-AD)^2}{(A+B)\,(C+D)\,(A+C)\,(B+D)}$$

While simpler than another (basic) expression, computation for any N over approximately 50 magnitude becomes laborious. Furthermore, decimal errors are often made by students who are not thoroughly practised with contingency tables.

If possible, the 2 by 2 table is set up such that all four marginal totals are equal. This design assures the most powerful test of the null hypothesis and minimizes the problem of small expected frequencies. Where data from continuous distributions are to be collapsed into two categories on each variable this means dichotomizing around the respective medians. Given the special situation of successful median splits, however, the conventional formula reduces to

$$\chi^2 = \frac{(|A-C| + |B-D|)^2}{N}$$

Moreover, this expression is still satisfactorily close to the long solution where only the general criterion of "ordinal symmetry" is satisfied, that is, where observed cell frequencies are arranged as

$$\begin{array}{ccc} A & > & B \\ \vee & & \wedge \\ C & < & D \end{array}$$

Given such a basic pattern (or its mirror image),

$$\chi^2 \approx \frac{(|A-C| + |B-D|)^2}{N}$$

Approximation for Chi Square in 2 by 2 table[3]

In effect, we simply square the sum of the left and right column (absolute) differences and divide by total cases. The procedure can be equally well phrased, of course, in terms of row differences as well as differences between diagonal sums. We have found the formulation above the most practical and least confusing to the student, and it is consistent with the approximation proposed for special 2 by 2k tables.

[3] It was pointed out to the author by G. Lienert that the analytical basis for the validity of the above assumption has been worked out earlier by M. H. Quenouille (*Rapid statistical calculations*, London, Griffin, 1959), who arrived at an equivalent formula for this special 2 by 2 case.

If ordinal symmetry in the table deviates greatly from complete symmetry in metric terms, that is, where any two diagonal cells are greatly different although the basic condition stated earlier is satisfied, χ^2 will be systematically over- or under-estimated depending on whether the larger or the smaller diagonal is involved (see later).

The examples in Table 1, some taken from the texts quoted, compare the rapid approximation with the standard method and show the consequences for accepting or rejecting the null hypothesis. The last two examples illustrate the type of error which results from gross imbalances in a diagonal.

TABLE 1.

Basic 2 by 2 table	χ^2, standard procedure	P	χ^2, estimate	P	Error, % of true value
16 8 11 25	7.59	<.01	8.1	<.01	+7
17 12 13 24	3.62	<.10	3.9	<.05	+8
37 66 52 49	5.02	<.05	5.0	<.05	0
16 28 30 25	3.24	<.10	2.9	<.10	—11
75 113 98 89	5.93	<.02	5.9	<.02	—1
24 17 20 22	0.99	>.30	1.0	>.30	+1
34 37 94 35	12.40	<.001	19.2	<.001	+55[a]
19 12 1 18	15.40	<.001	11.5	<.001	—25[a]

[a]Note asymmetry in diagonal cells.

The 2 by 2k table

In the standard treatment, individual terms of the type $\frac{(O-E)^2}{E}$ are summed after expected (E) frequencies in each cell are computed from the marginal sub-totals of the observed (O) frequencies. There is a special case where (a) row totals are equal, and (b) where observed cell frequencies assume only two different values. Here the procedure can be much simplified. For example, if a contingency table

A	B	C	D
E	F	G	H

consists of the following hypothetical cell frequencies

15	25	15	25
25	15	25	15

an analogy exists to the 2 by 2 case, and

$$\chi^2 = \frac{(|A—E| + |B—F| + |C—G| + |D—H)^2}{N}$$

$$= \frac{40^2}{160} = 10.0 \ (df=3, P < .02).$$

This result is identical with the conventional long solution. If the stated equalities are satisfied only approximately, the identity no longer holds in the general case. However, it can often be written, without excessive error,

$$\chi^2 \approx \frac{(|A—E| + |B—F| + |C—G| + |D—H|)^2}{N}$$ *Approximation for Chi Square in 2 by 4 table*

The solution extends to 2 x 2k tables. (With constraints stated, the number of columns is necessarily even.) There we add all observed column differences, square the total, and divide by *N*.

Table 2 compares some solutions for cases of moderate deviations from the desired pattern.

TABLE 2.

Basic 2 by 4 table				χ^2, standard procedure	P	χ^2, estimate	P	Error, % of true value
7	15	18	6	21.76	<.001	21.2	<.001	—3
17	8	6	23					
13	8	12	10	8.69	<.05	8.7	<.05	0
6	14	7	20					

The 2 by 2 by 2 and Higher Order Interactions

One standard treatment, after M. S. Bartlett, for the second order (2 by 2 by 2) interaction consists in solving a cubic equation (4) and is rarely given in textbooks. The reasons may be the forbidding amount of labor required where no computer is available and a relatively low efficiency of this statistic. Consequently, other measures have come to be preferred. As the Ns in research often do not exceed the magnitude of 100 it can be argued that there is mainly one situation where a gross significance estimate for an interaction may still be desired. Consider

the case where two 2 by 2 tables show suggestive and opposite trends in the observed cell entries, as in the example,

9	40
38	5

X

44	3
7	51

Here the condition is satisfied that

$$\begin{array}{ccc} A < B & A' > B' & \text{where } A'+B'+C'+D'=N' \\ \wedge \quad \vee & \vee \quad \wedge & A+B+C+D=N, \\ C > D & C' < D' & \text{and } N \sim N' \end{array}$$

Given this type of situation (and only where opposite trends are evident from inspection) a workable estimate of the interaction, with df=1, is possible as

$$\chi^2 \approx \frac{(|A-C| + |B-D| + |A'-C'| + |B'-D'|)^2}{N + N'}$$

Approximation for second order interaction Chi Square

The amount of time saved over the precise long method tends to be substantial.

The examples in Table 3 compare solutions by way of the standard method and the approximation, respectively. It will be noted that the conditions stated above are violated to some extent in both cases. However, the basic opposition of diagonal trends is still visible.

TABLE 3.

Basic 2 by 2 by 2 table				χ^2, standard procedure	P	χ^2, estimate	P	Error, % of true value
63	70	43	19	5.51	<.02	4.6	<.05	—17
70	55	35	31					
61	77	61	31	11.89	<.001	13.9	<.001	+17
101	82	42	51					

(Speculatively it would seem that the approximation given for the 2 by 2 by 2 case could be generalized to interactions of third or higher order. Precise computation of this type of statistic by way of solving the appropriate equations is, to our knowledge, not usually attempted. Occasionally, however, a significance estimate might be sought by an investigator if a high degree of symmetry is obvious among the trends of all sub-tables. For example, in the following hypothetical example an

interaction may be suspected among such variables as sex, birth order, affiliation need, and resulting anxiety:

9	12
15	8

15	10
11	20

X

17	10
12	12

7	15
16	15

Extending the earlier reasoning by analogy, the resulting 2 by 2 by 2 by 2 interaction might be approximated from the respective eight column differences in the four tables as being of the order of

$$\chi^2 \sim \frac{41^2}{205} \sim 8.2$$

We can offer no empirical check on this estimate precisely because no "standard solution" seems to be readily available.)

Correcting for Continuity

The Yates correction is usually required for both 2 by 2 and 2 by 2 by 2 tables where small expected frequencies would inflate the χ^2 estimate. It can be applied for the approximations discussed above. For the regular 2 by 2 table, a constant of 2 is subtracted from the sum of the column differences before squaring the total. For the 2 by 2 by 2 table, a constant of 4 should first be subtracted in the numerator.

The Fourfold Point Coefficient of Correlation

The definitional formula for r_p suggests that computation can be simplified wherever the cell entries are symmetrical in nature, that is, result from median splits along both variables. The standard formula,

$$r_p = \frac{BC - AD}{\sqrt{(A+B)\ (C+D)\ (A+C)\ (B+D)}}$$

in the case of equality among marginal sub-totals reduces to

$$r_p = \frac{|A - C| + |B - D|}{N}$$

The expression disregards the sign of the coefficient which is easily determined from the actual cell distribution. If symmetry is not com-

pletely achieved by reason of tied scores in the original data, or because a larger number of cells cannot be conveniently combined,

$$r_p \sim \frac{|A-C| + |B-D|}{N}$$ *Approximation for fourfold point coefficient of correlation*

In other words, we add the absolute values of the two column differences and divide the sum by N. This version of a correlation coefficient is readily computed without as much as a slide rule wherever N is manageable (e.g., an even number and under 50). The ease would seem to make it unique among a number of other shortcuts since it is faster than the cosine pi approximation of the tetrachoric, or the Chown-Moran and Mosteller approximations to the product-moment measure (cf. 5). The method may be particularly helpful for preliminary item analyses and gives good precision with test data where near-median splits can be achieved. It also serves as a general check where r_p or possibly r are computed by means of the standard procedures.

Table 4 illustrates the adequacy of the simplified formula in some empirical cases.

TABLE 4.

Basic fourfold table		r_p, standard procedure	r_p, estimate	Error, % of true value
16	64	0.319	0.28	—12
51	49			
20	70	0.470	0.49	+4
41	18			
23	67	0.280	0.30	+7
37	33			
2	10	0.544	0.53	—3
13	5			
7	64	0.624	0.61	—2
44	19			
22	95	0.790	0.77	—3
73	0			

The Contingency Coefficient for Four Cells

The standard formula is usually written as

$$C = \sqrt{\frac{\chi^2}{N+\chi^2}}$$

and, in the special case of the fourfold table,

$$C = \sqrt{\frac{\dfrac{N(BC-AD)^2}{(A+B)\ (C+D)\ (A+C)\ B+D)}}{N+\dfrac{N(BC-AD)^2}{(A+B)\ (C+D)\ (A+C)\ (B+D)}}}$$

Where all observed cell frequencies are reasonably symmetrical the original expression reduces to

$$C \sim \frac{|A-C| + |B-D|}{\sqrt{(|A-C| + |B-D|)^2 + N^2}}$$ *Approximation for four-cell contingency coefficient*

The two examples of Table 5 illustrate the adequacy of the estimate.

TABLE 5.

Basic fourfold table		*C*, standard procedure	*C*, estimate	Error, % of true value
3	23	0.612	0.61	0
17	2			
11	32	0.258	0.30	+16
14	13			

Direction and Size of Error

In the 20 illustrations presented above approximately two thirds fall within —12 and +8% of the statistic in question, and one fourth within —1 and +1%. This evaluation is based on a small sample of five somewhat heterogeneous situations. It is also open to the criticism that the cell frequencies were selected so as to satisfy, primarily, the constraints stated at the outset. In place of the gross summary above, however, the short cuts can also be examined in terms of the direction and amount of error to be expected a priori.

It can be shown, and is demonstrated by inspection of the sixteen examples in the 2 by 2 category, that the approximation tends to overestimate the exact value of the statistic whenever the deviation from true symmetry is (proportionately) greater within the larger of the two pairs of diagonal cells. The exact value is underestimated whenever discrepancies are greater between the two smaller cells. At the same time, asymmetry in the smaller pair of diagonals generally carries less weight in the overall statistic than a comparable degree of asymmetry involving the larger pair. Furthermore, it may be argued that the composition of diagonals in 2 by 2 tables ("symmetrical" in the sense used above) is distributed as the result of sampling from fixed diagonal sums. Intuitively, then, larger degrees of asymmetry (in terms of per cent discrepancy) would normally occur within the lesser cell diagonal because of the

small numerical values, which readily form extreme proportions as they approach zero. This would not apply equally to the cell frequencies observed within the other, larger diagonal. If this analysis is correst we would expect positively skewed error distributions in large scale applications of the short cut methods. That is, (a) underestimates of χ^2, r_p and C will be more common, and their amount will be small—the approximations tend to be conservative; (b) overestimates will be less common but will be more extreme when they occur.

Literature Cited

1. LIENERT, G. A. 1962. Verteilungsfreie Methoden in der Biostatistik. Weinheim, Germany: Beltz.
2. MCNEMAR, Q. 1955, 1962. Psychological statistics. New York, Wiley.
3. SIEGEL, S. 1956. Nonparametric statistics for the behavioral sciences. New York, McGraw-Hill.
4. SNEDECOR, G. W. 1946. Statistical methods, 4th edition. Ames, Iowa, Iowa State College Press.
5. TATE, M. W., and R. C. CLELLAND. 1957. Nonparametric and shortcut statistics in the social, biological, and medical sciences. Danville, Illinois, Interstate Printers.
6. WENDT, H. W. 1958. Naeherungsloesungen fuer Chi-Quadrat-Werte: Wechselwirkungen erster und zweiter Ordnung in Vierfeld-Tafeln. Psychologie und Praxis 2:39-44.
7. WILCOXON, F. 1949. Some rapid approximate statistical procedures. New York, American Cyanamid Co.