5-1-2008

Tests for Independence in Two-Way Contingency Tables with Small Samples

Stephen Sharp

University of Edinburgh, stephen.sharp@ed.ac.uk

Tests for Independence in Two-Way Contingency Tables with Small Samples

Stephen Sharp
University of Edinburgh

When testing the null hypothesis of independence in a two-way contingency table, the likelihood ratio test statistic is approximately distributed as Chi-squared d for large sample sizes (*N*) but may not be for small samples. This paper presents expressions which match the mean of the statistic to Chi-squared d as far as N^{-1} and N^{-2}, derives a method of estimating the expressions from observed data and evaluates them using Monte Carlo simulations. It is concluded that using appropriate dividing factors, rejection rates after matching are more accurate than for either the unadjusted likelihood ratio statistic or the Pearson approximation which is the main alternative statistic. Minimum cell frequencies necessary for high test accuracy are smaller than those commonly given in textbooks.

Key words: Contingency tables, likelihood ratios, small samples, dividing factors.

Introduction

A common requirement in social science research is to test the null hypothesis of independence between the two axes of a contingency table. It is well known that this can be tested either by using Pearson's Chi-squared approximation based on squares of differences between observed and expected values, or by the likelihood ratio test statistic (*LR*) originally proposed by Neyman and Pearson (1928). Many widely used statistical packages like SPSS give both statistics. Both tend asymptotically to a Chi-squared d distribution as sample size increases. Tabachnick and Fidell (2004, p. 251) pointed out that from a theoretical point of view, *LR* is preferable because it is available for testing overall fit, screening, and testing for differences among hierarchical models. However *LR* has the relative disadvantage that it converges to Chi-squared d more slowly than

Stephen Sharp is a Visiting Scholar at the Centre for Educational Leadership in the Moray House School of Education. His research interests include educational assessment, evaluation and monitoring; teacher careers and professional development; and statistical applications in educational research. E-mail him at stephen.sharp@ed.ac.uk.

the Pearson test and that for small samples it has values which are stochastically larger than its asymptote (i.e. it errs in the 'wrong' direction). This is because *LR* is based on the function $\sum p \log p$ where the summation goes over a complete set of probabilities, and as this is a downward concave function, replacing probabilities by their estimates leads to bias in estimating the sum. A similar effect has been noted where the same function is used to estimate entropy in physical systems (Victor 2000).

Starting in the 1950s, statisticians have tried to find ways of adjusting the *LR* test to speed up its rate of convergence to Chi-squared and hence overcome its main limitation. The classic papers of Bartlett (1954) and Lawley (1957) developed a general method which applies to all continuous likelihood functions. The Bartlett-Lawley adjustment takes the form of a number, which, if used as a divisor for the *LR* statistic, matches all its moments to those of Chi-squared as far as terms in N^{-1}, where *N* is the sample size, thus accelerating the convergence. However Frydenberg and Jensen (1989) cast doubt on whether Lawley's method is effective at all when applied to discrete data (of which contingency tables are an example). They point out that the Lawley method assumes that the *LR* statistic can be written as a function of a continuous random variable, which is not the case with contingency tables with discrete

cell frequencies. The method however may still be valid to the extent that discrete data approximates continuous data, which will increasingly be the case as the number of cells in the table increases. Frydenberg and Jensen presented evidence that the method is seriously in error for one-dimensional frequency tables with three and four categories, where the data is at its most discrete, as it were. But other writers have argued that this view is overly pessimistic. Pierce and Peters (1992) explored one-parameter functions of exponential families, finding that excellent approximations can be obtained from simple adjustments to the signed square root of the likelihood ratio statistic with one degree of freedom. They did not however consider alternative hypotheses of a more generalized nature, as the present paper does.

Williams (1976, 1978) derived first-order adjustments for the *LR* statistic for one-way, two-way and five-way contingency tables, though he did this not by using the Bartlett-Lawley method directly but by expanding $n \log n$ (where N is an observed cell frequency) as a Taylor series around its mean. Also, he did not offer any empirical evaluation of the expressions he derived. Subsequently Smith *et al* (1981) also used a Taylor expansion around the mean and, by taking more terms in the series, derived a second-order expression which matches the first moment of *LR* to its asymptotic value as far as terms in N^{-2} and all others as far as terms in N^{-1}. Smith *et al* considered only one-way tables with the null hypothesis of equal probabilities but produced evidence that their more accurate adjustment did indeed model upper cut-offs more accurately than that of Williams. Bayo Lawal (1984) showed that the Pearson test performs as well as the Williams-adjusted *LR* for one-way tables with 3 and 4 cells, though he did not consider the adjustment of Smith et al.

The aim of the present paper is to extend Williams' expression for two-way tables from first order (as far as N^{-1}) to second order (as far as N^{-2}) levels of accuracy for the mean of the test statistic. Put another way, it is to extend Smith et al's second-order expression from one-way to two-way tables. The paper also presents Monte Carlo simulations to evaluate various adjustments to the *LR* statistic against each other and against the Pearson test and considers various practical issues concerning the implementation of the adjustments.

Methodology

First-order adjustments for two-way tables

To evaluate the Taylor series which results from expanding $n \log n$, it is necessary to assume how N is distributed. Williams took the Poisson distribution while Smith *et al* used the multinomial. The latter is used throughout this paper for consistency. The distributions are closely linked and lead to the same answer for the first-order divisor, which Williams showed to be given by the expression

$$1 + \frac{\left(\sum_i r_i^{-1} - 1\right)\left(\sum_j c_j^{-1} - 1\right)}{6N(r-1)(c-1)} \tag{1}$$

where the table has r rows and c columns with marginal probabilities r_i (i = 1, 2, ...r) and c_j (j = 1, 2, ...c). Williams pointed out that the effect of the adjustment will be minimized where all the r_i equal $1/r$ and all the c_j equal $1/c$. In this case, $\sum r_i^{-1} = r^2$ and $\sum c_j^{-1} = c^2$ so the above expression can be written simply as

$$1 + \frac{(r+1)(c+1)}{6N} \tag{2}$$

This is undoubtedly safe but perhaps the adjustment may be made more accurate by estimating the sums of the reciprocals of the probabilities from the data. Neither Williams nor Smith *et al* considered the practicalities of doing this, the former because he attempted no empirical validation of the expression and the latter because they considered only the null hypothesis of uniformity where the parameters are known and do not have to be estimated.

The naive estimate of r_i^{-1} (an analogous argument applies to the column probabilities) is simply the reciprocal of its maximum likelihood estimate ie R_i / N, where R_i is the *i*th row total.

However N/R_i is not an unbiased estimate of r_i^{-1}. In fact the expected value of N/R_i is undefined if R_i follows a Poisson or multinomial distribution as there is a finite probability that R_i equals zero. In practice when using contingency tables, rows and columns with no observations at all are deleted from the analysis, the degrees of freedom being reduced accordingly. For present purposes however the problem is to estimate r_i^{-1} where R_i might be zero. To do this, we consider the expected value of $(R_i+1)^{-1}$. From the density function of the binomial distribution, this is given by

$$E\frac{1}{R_i+1} = \sum_{R_i=0}^{N} \frac{N!}{(R_i+1)!(N-R_i)!} r_i^{R_i}(1-r_i)^{N-R_i}$$

We make the binomial series complete again by multiplying by $r_i(N+1)$ and adding in a term for $R_i=-1$. Rearranging yields

$$E\left(\frac{N+1}{R_i+1}\right) = \frac{1}{r_i}\left[1-(1-r_i)^{N+1}\right]$$

(3)

The left hand side is an underestimate of r_i^{-1}, the error being $(1-r_i)^{N+1}$. However this is less than 5% if the expected value of R_i is around three and less than 1% if it is around five, values which should be exceeded comfortably by sample sizes used in practice in research. An analogous argument leads to $(N+1)/(C_j+1)$ as an estimate of c_j^{-1}.

Second-order adjustments for two-way tables

This is achieved in the same way as the first-order adjustments except that more terms are taken from the Taylor series. The expression derived by Smith *et al* for the second-order divisor for one-way tables was

$$1+\frac{\sum p_i^{-1}-1}{6N(k-1)}+\frac{\sum p_i^{-2}-\sum p_i^{-1}}{6N^2(k-1)} \qquad (4)$$

where there are k categories with probabilities p_i. The same method can be applied to the null hypothesis of independence in a two-way table rather than that of specified *p*-values in a one-way table. The resulting algebra is laborious but straightforward. It leads to the rather ungainly expression

$$1+\frac{\left(1-\frac{1}{N}\right)\left(\sum_i r_i^{-1}-1\right)\left(\sum_j c_j^{-1}-1\right)}{6N(r-1)(c-1)}+$$
$$\frac{\left(\sum_i r_i^{-2}-1\right)\left(\sum_j c_j^{-2}-1\right)}{6N^2(r-1)(c-1)} \qquad (5)$$

which is clearly a combination of (1) and (4). In the second-order case, the use of the Poisson or multinomial assumption makes a difference. The above version is the multinomial one. The term $1-1/N$ in the numerator of the middle part of (5) disappears in the Poisson version, but in practice the difference between the two will be negligible if N has a value which is reasonable for research purposes.

Again there is a 'safe' version of this based on the assumption that all the r_i equal $1/r$ and all the c_j equal $1/c$. The result is

$$1+\frac{(1-1/N)(r+1)(c+1)}{6N}+$$
$$\frac{(r^2+r+1)(c^2+c+1)}{6N^2} \qquad (6)$$

Again, the $1-1/N$ term is absent if the Poisson distribution is assumed. Following an argument analogous to that used above, we estimate r_i^{-2} by considering the expected value of the reciprocal of $(R_i+1)(R_i+2)$:

$$E\frac{1}{(R_i+1)(R_i+2)} = \sum_{R_i=0}^{N} \frac{N!}{(R_i+2)!(N-R_i)!} r_i^{R_i}(1-r_i)^{N-R_i}$$

This time we complete the binomial series by multiplying by $r_i^2(N+1)(N+2)$

and adding in terms for $R_i = -1$ and $R_i = -2$. Rearranging yields

$$E \frac{(N+1)(N+2)}{(R_i+1)(R_i+2)} =$$

$$\frac{1}{r_i^2}\left\{1-(1-r_i)^{N+1}\left[1+r_i(N+1)\right]\right\} \qquad (7)$$

The extent of the underestimation of r_i^{-2} is greater than for r_i^{-1} but it is still less than 5% if the expected value of R_i is at least five and less than 1% if it is at least seven, and these are also values which should be exceeded comfortably by sample sizes used in practice in research. An analogous argument leads to $(N+1)(N+2)/\left[(C_j+1)(C_j+2)\right]$ as an estimate of c_j^{-2}. However it is not certain that a second-order expression based on estimated parameters will be successful in improving the accuracy of the method. Victor (2000) found that this approach did not always lead to greater accuracy when trying to derive improved estimates for entropy in physical systems. The figures reported in the next section throw light on the accuracy of the various adjustments.

Results

Monte Carlo methods were used to assess the accuracy of six different tests which are summarized in Table 1. The choice of which sort of simulated data to use is inevitably to some extent arbitrary. The choice used here is based on the advice offered by most statistical text books (e. g., Tabachnick & Fidell 2004, p. 223) that the Pearson test should not be used unless all expected cell frequencies in the table are greater than one and not more than one-fifth of them are less than five. Tabachnick and Fidell do not give the source of this advice and there seems to be no corresponding advice for the *LR* test.

All the contingency tables used in the simulations had five columns with the number of rows being two, three, four and five. In each row the expected value of the first cell was one while all other cells had the expected value *M* where *M*

had the values two, three, four and five. Thus in all cases, one-fifth of cells have an expected value of one and all other cells have an expected value of *M*. The case where *M* equals five is the criterion case for the advice given in statistics texts. The aim is to investigate whether any of the adjusted tests perform well with values of *M* less than five.

For each of the 16 (four numbers of rows by four values of *M*) versions of the table, 10,000 sets of data were simulated where the null hypothesis was true. The results are contained in table 2. For ease of interpretation, some of the entries in this table are in bold face. If the actual and nominal rejection rates for a test are the same, the percentage of rejections at a level of significance *p* (e. g., 0.05) has an expected value of $100p$ and variance $p(1-p)$. Entries in the tables where the observed percentage is within two standard deviations of the nominal percentage are in bold type. This is a stringent criterion as the only deviation which it allows from the nominal rejection levels is that expected on the basis of sampling error.

The main comparison is between tests 2 to 6 (i. e., the various *LR* tests). Test 1, the Pearson approximation, acts as a benchmark. It is immediately clear from table 2 that test 2 (the unadjusted *LR* test) is seriously in error, the rejection rate being well above the nominal rate for all levels of significance, especially the less stringent ones. The smallest dividing factor is test 3, the *LR* test with first order adjustment based on equal marginal parameters. This is a marked improvement on test 2 but still has a tendency to over-reject slightly at the 10% and 5% levels. Larger adjustments are provided by tests 4 and 5 and these have higher concentrations of accurate rejection rates. There is little to choose between them except at the very smallest sample size where *M*=2 and each row has an expected frequency of just nine. Here, test 4 over rejects slightly at the 10% level but is more accurate than test 5 at the 5% and 1% levels (the levels most often used in social science research). Test 6, the second-order correction with estimated parameters, has the largest adjustment of all but appears to be a step too far, as it were. Its performance is similar to test 1 (the Pearson approximation), i. e., safe but

conservative, especially at the more stringent significance levels, but not as accurate as tests 4 or 5.

It might seem counterintuitive that test 6, which has the strongest theoretical rationale, is not the most accurate but in fact this is not so surprising. As Smith et al. (1981) pointed out, the use of a scaling factor to match moments is by its very nature a fairly crude device whose effects at a fine level of detail may not always match closely with theoretical expectations. Also, the adjustments are designed to match the first moment of the *LR* statistic with its asymptotic mean, and the mean values (not reported here) observed in the simulations do indeed show that test 6 usually produces the mean closest to the number of degrees of freedom. However the criterion used here (and the one in which test users are interested) is the accuracy with which each test models not the mean but the upper cut-off scores and this depends on characteristics of the distribution other than the mean.

Conclusion

On the basis of the arguments and data reported above, these conclusions are offered:

 (i) the unadjusted *LR* test should not be used for small samples;
 (ii) the Pearson approximation is safe but conservative;
(iii) the view of Frydenberg and Jensen is overly pessimistic in the context of two-way contingency tables where the use of rescaling factors can result in improved test accuracy;
(iv) if a second-order adjustment assuming equal marginal probabilities is used based on (6) above, then at the 10% and 5% levels of significance, accurate rejection rates are achieved where all expected values are at least one and not more than one-fifth are less than three; and
 (v) if a first-order adjustment is used with marginal probabilities estimated from the data using the method based on (1) and (3) above, accurate rejection rates are achieved where all expected values are at least one and not more than one-fifth are less than two (for the 5% level of significance) or three (for the 1% level).

References

Bartlett, M. S. (1954). A note on the multiplying factors in various Chi-squared approximations. *Proceedings of the Cambridge Philosophical Society, 47,* 86-95.

Bayo Lawal, H. (1984). Comparisons of the X^2, Y^2, Freeman-Tukey and Williams' improved G^2 test statistics in small samples of one-way multinomials. *Biometrika, 71,* 415-418.

Frydenberg, M. & Jensen, J. L. (1989). Is the 'improved likelihood ratio statistic' really improved in the discrete case? *Biometrika, 76,* 655-661.

Lawley, D. N. (1957). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika, 43,* 295-303.

Neyman, J. & Pearson, E. S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20,* 175-240 and 263-294.

Pierce, D. A. & Peters, D. (1992) Practical use of higher order asymptotics for multiparameter exponential families. *Journal of the Royal Statistical Society, B54,* 701-737.

Smith, P. J., Rae, D. S., Manderscheid, R. W. & Silbergeld, S. (1981). Approximating the moments and distribution of the likelihood ratio statistic for multinomial goodness of fit. *Journal of the American Statistical Association, 76,* 737-740.

Tabachnick, B. G. & Fidell, L. S. (2004). *Using multivariate statistics.* (4th edition) New York: Allyn and Bacon.

Victor, J. D. (2000). Asymptotic bias in information estimates and the exponential (Bell) polynomials. *Neural Computation, 12,* 2797-2804.

Williams, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika, 63,* 33-37.

Williams, D. A. (1978). Amendments and corrections: improved likelihood ratio tests for complete contingency tables. *Biometrika, 65,* 77.

Table 1: Adjustments to *LR* tests and their divisors

Test Description	Divisor
1 Pearson approximation.	-
2 Unadjusted *LR* test.	-
3 *LR* test with first order adjustment and equal marginal parameters.	expression (2)
4 *LR* test with first order adjustment and estimated marginal parameters.	expressions (1) and (3)
5 *LR* test with second order adjustment and equal marginal parameters.	expression (6)
6 *LR* test with second order adjustment and estimated marginal parameters.	expressions (5) and (7)

Table 2: Rejection rates for six tests and various tables sizes and cell frequencies. Bold entries are within sampling variation of the nominal rate.

	rows	M	10%	5%	1%	0.1%		rows	M	10%	5%	1%	0.1%
Test 1	2	2	8.58	3.07	0.12	0.00	Test 2	2	2	18.99	10.42	1.71	**0.06**
Test 1	3	2	8.56	3.73	0.52	0.03	Test 2	3	2	19.78	10.51	1.87	**0.13**
Test 1	4	2	8.98	4.34	0.67	0.02	Test 2	4	2	20.97	11.65	2.30	**0.14**
Test 1	5	2	9.29	4.15	0.59	0.03	Test 2	5	2	22.88	12.57	2.56	0.20
Test 1	2	3	9.06	3.74	0.43	0.00	Test 2	2	3	16.31	8.99	2.06	0.21
Test 1	3	3	**9.60**	4.07	0.55	0.02	Test 2	3	3	17.81	9.93	1.88	**0.12**
Test 1	4	3	**9.44**	4.47	**0.85**	**0.08**	Test 2	4	3	19.20	10.58	2.48	0.23
Test 1	5	3	9.25	4.49	**0.80**	**0.08**	Test 2	5	3	19.00	10.27	2.48	0.29
Test 1	2	4	8.59	3.96	0.52	0.03	Test 2	2	4	14.10	7.52	1.87	0.24
Test 1	3	4	9.11	4.19	0.76	**0.05**	Test 2	3	4	15.42	8.05	2.04	0.24
Test 1	4	4	9.12	4.30	**0.80**	0.04	Test 2	4	4	15.90	8.69	1.90	0.21
Test 1	5	4	9.04	4.12	**0.83**	**0.08**	Test 2	5	4	16.66	8.73	2.01	0.27
Test 1	2	5	**9.47**	4.24	0.74	0.02	Test 2	2	5	13.64	7.50	1.83	0.23
Test 1	3	5	**9.89**	4.48	0.61	**0.04**	Test 2	3	5	14.57	7.76	1.66	0.18
Test 1	4	5	**9.80**	**4.78**	0.95	**0.08**	Test 2	4	5	14.84	8.36	1.88	0.24
Test 1	5	5	9.04	4.03	**1.00**	**0.08**	Test 2	5	5	14.81	7.70	1.98	0.24

	rows	M	10%	5%	1%	0.1%		rows	M	10%	5%	1%	0.1%
Test 3	2	2	11.55	**5.04**	0.49	0.00	Test 4	2	2	11.96	**5.23**	0.53	0.00
Test 3	3	2	11.01	**4.80**	0.60	0.00	Test 4	3	2	10.98	**4.78**	0.62	0.00
Test 3	4	2	11.37	**5.08**	0.61	0.01	Test 4	4	2	11.02	**4.84**	0.60	0.01
Test 3	5	2	11.51	**5.10**	0.54	0.01	Test 4	5	2	11.07	**4.97**	0.55	0.00
Test 3	2	3	11.94	5.97	**1.05**	**0.10**	Test 4	2	3	11.82	5.81	**0.99**	**0.11**
Test 3	3	3	12.35	6.09	**0.86**	0.01	Test 4	3	3	11.62	5.68	0.78	0.01
Test 3	4	3	12.59	6.20	**1.03**	**0.06**	Test 4	4	3	11.65	5.72	**0.89**	**0.06**
Test 3	5	3	11.83	5.95	1.24	**0.08**	Test 4	5	3	10.72	**5.44**	**1.04**	0.07
Test 3	2	4	10.99	5.49	**1.11**	**0.11**	Test 4	2	4	**10.47**	**5.30**	**1.03**	**0.09**
Test 3	3	4	11.19	5.80	**1.11**	**0.11**	Test 4	3	4	**10.36**	**5.23**	**1.01**	**0.10**
Test 3	4	4	11.46	**5.17**	**1.12**	0.03	Test 4	4	4	**10.31**	4.69	**0.96**	0.02
Test 3	5	4	11.42	**5.37**	**1.06**	**0.11**	Test 4	5	4	**10.10**	4.81	**0.92**	**0.09**
Test 3	2	5	11.30	5.66	1.25	**0.11**	Test 4	2	5	10.87	**5.39**	**1.18**	**0.09**
Test 3	3	5	11.13	5.51	**1.01**	**0.08**	Test 4	3	5	**10.16**	4.82	0.79	**0.04**
Test 3	4	5	11.65	5.95	1.23	**0.12**	Test 4	4	5	**10.43**	**5.30**	**1.03**	**0.10**
Test 3	5	5	10.84	**5.38**	1.26	**0.12**	Test 4	5	5	9.38	**4.54**	**1.01**	**0.09**

	rows	M	10%	5%	1%	0.1%		rows	M	10%	5%	1%	0.1%
Test 5	2	2	8.46	3.14	0.21	0.00	Test 6	2	2	**9.51**	3.79	0.32	0.00
Test 5	3	2	7.44	2.82	0.25	0.00	Test 6	3	2	7.56	2.98	0.28	0.00
Test 5	4	2	7.48	2.99	0.26	0.01	Test 6	4	2	7.00	2.77	0.26	0.00
Test 5	5	2	7.32	2.83	0.24	0.00	Test 6	5	2	6.73	2.54	0.21	0.00
Test 5	2	3	**10.27**	**4.88**	0.78	**0.05**	Test 6	2	3	**10.05**	**4.74**	0.70	**0.09**
Test 5	3	3	**10.55**	**4.88**	0.68	0.01	Test 6	3	3	9.32	4.19	0.53	0.01
Test 5	4	3	**10.52**	**4.88**	0.72	**0.05**	Test 6	4	3	8.71	4.03	0.58	0.03
Test 5	5	3	**9.70**	**4.74**	0.81	**0.06**	Test 6	5	3	8.03	3.70	0.63	0.03
Test 5	2	4	**10.01**	**5.06**	0.90	**0.08**	Test 6	2	4	9.17	**4.60**	0.76	**0.07**
Test 5	3	4	**10.16**	**5.14**	0.95	**0.10**	Test 6	3	4	8.70	4.27	0.79	**0.08**
Test 5	4	4	**10.30**	4.68	0.93	0.01	Test 6	4	4	8.13	3.60	0.60	0.00
Test 5	5	4	**10.11**	4.75	0.93	**0.09**	Test 6	5	4	7.64	3.52	0.66	**0.05**
Test 5	2	5	10.89	**5.40**	**1.16**	**0.09**	Test 6	2	5	**9.80**	**4.80**	**1.00**	**0.06**
Test 5	3	5	**10.50**	**5.13**	0.88	**0.05**	Test 6	3	5	8.78	3.98	0.55	0.02
Test 5	4	5	10.90	5.59	**1.10**	**0.11**	Test 6	4	5	8.66	4.16	0.77	**0.06**
Test 5	5	5	**10.02**	**4.92**	**1.12**	**0.10**	Test 6	5	5	7.35	3.54	0.74	0.03