PSYCHOMETRIKA—VOL. 44, NO. 1.
MARCH, 1979

EXACT CONDITIONAL TESTS FOR CROSS-CLASSIFICATIONS: APPROXIMATION OF ATTAINED SIGNIFICANCE LEVELS

ALAN AGRESTI, DENNIS WACKERLY, AND JAMES M. BOYETT

UNIVERSITY OF FLORIDA

A procedure is proposed for approximating attained significance levels of exact conditional tests. The procedure utilizes a sampling from the null distribution of tables having the same marginal frequencies as the observed table. Application of the approximation through a computer subroutine yields precise approximations for practically any table dimensions and sample size.

Key words: contingency tables, independence, chi-square, Kruskal-Wallis, computer algorithm.

1. Introduction

Several recent articles have outlined methodologies for exact conditional analyses of data which are summarized in an $r \times c$ matrix of counts. Depending upon the purposes of the experimenter, several types of null and alternative hypotheses are appropriate. Some of the procedures [Agresti & Wackerly, 1977; Freeman & Halton, 1951] test for independence between two variables, while others [Klotz & Teng, 1977] are designed to compare several treatments with respect to observed responses on an ordinal categorical variable. Extentions of Fisher's [1971] now famous experiments involving the tea-tasting lady have also been considered [Wackerly, McClave & Rao, 1978].

When testing the null hypothesis of independence of two variables in a cross-classification table, various types of alternative hypotheses may be appropriate. For example, if the two variables are measured on a strictly nominal scale, we might be interested in the broad alternative of "statistical dependence". A natural test statistic in that case is the standard chi-square statistic or a nominal measure of association such as Goodman and Kruskal's lambda or tau. Alternately, if both variables are ordinal categorical, we might wish to detect whether there is a monotonic relationship between the variables. In that case we might use the alternative hypothesis that the proportion of concordant pairs of observations is unequal to the proportion of discordant pairs, and employ Kendall's tau-*b* as the test statistic. In the case of comparing several treatments on an ordinal categorical variable, the null hypothesis of independence corresponds to homogeneity of the treatments. We might then wish to use a Kruskal-Wallis type statistic for detecting response shifts among those treatments.

In practice, cross-classification tables often occur in which the overall sample size or the cell frequencies are too small to employ asymptotically derived sampling distributions for these test statistics. In such cases, as an alternative procedure we can conduct an exact test of independence, conditional on the observed marginal frequencies. Probably the best known test of this nature is Fisher's exact test of independence for 2×2 tables. However,

The authors gratefully acknowledge helpful discussions with Professor John G. Saw which lead to the procedure for sampling the tables. Gratitude is also extended to the referees for their helpful comments and to the Northeast Regional Data Center at the University of Florida for computer support.

Requests for reprints should be sent to James M. Boyett, Department of Statistics, Nuclear Sciences Center, University of Florida, Gainesville, Florida, 32611.

0033-3123/79/0300-0075$00.75/0

the same principle applies for any table size $r \times c$ and for whatever statistic is used to detect the condition listed in the alternative hypothesis. The approach is to

1. calculate the appropriate test statistic for every $r \times c$ array of non-negative integers having the same marginal frequencies as the observed table,
2. calculate the null probability of each table conditional on the marginal frequencies,
3. define the attained significance level to be the sum of the null probabilities of those tables which are at least as favorable to the alternative hypothesis (as measured by the test statistic) as the observed table.

Let n_{ij} denote the frequency observed in the cell falling in the i^{th} row and j^{th} column ($1 \leq i \leq r$, $1 \leq j \leq c$) of the cross-classification table. The main difficulty in carrying out exact conditional tests is the sheer number of calculations involved. One must generate all $r \times c$ arrays of nonnegative integers in the set

$$S = \left\{ n'_{ij} : \sum_i n'_{ij} = n_{\cdot j},\ \sum_j n'_{ij} = n_{i\cdot} \quad \text{for all } i, j \right\} \tag{1.1}$$

having the same marginal frequencies as the observed table. The conditional probability under the null hypothesis as well as the value of the test statistic must be computed for each table in S. The number of tables in S, which we denote by $|S|$, increases very rapidly as a function of the sample size. Especially for relatively large table dimensions, $|S|$ is too large for the practical implementation of the exact tests even when asymptotic approaches would be crude. To illustrate, for a 4×4 table, the maximum number of tables in S when the sample size is $n = 10$ is 626; when $n = 20$, it is 40,176; but when $n = 30$ an approximate maximum is 574,249. (The maxima for $n = 10$ and $n = 20$ are from Table 5 of the paper by Agresti and Wackerly [1977]. The figure for $n = 30$ is based on an approximation for $|S|$ given in Good's [1976] paper. The value is 672,156 if Gail and Mantel's [1977] approximation is used.) For that table dimension, a computer such as the IBM 370/165 can handle an exact test with approximately 100,000 tables in a minute of CPU time. Thus, it would be infeasible to perform an exact conditional test on data such as in Table 1, for which Klotz and Teng [1977] give $|S|$ to be 12,798,781. Some guidelines on the sample sizes that could be managed for various table sizes were presented in Table 5 of Agresti and Wackerly [1977].

In the next section, we present a method which can be utilized when

a. there is doubt about whether an asymptotic approximation for the distribution of the test statistic is valid, and
b. there is doubt about whether an exact conditional test can be economically implemented.

Instead of analyzing all the tables in the set S, we randomly generate sufficiently many of them so that the attained significance level of the test can be estimated as accurately as is practically necessary. A similar approach has been utilized in permutation tests to analyze data for which all possible permutations cannot practically be considered (see Forsythe & Frey, 1970, and Boyett & Shuster, 1977). Sampling procedures have also been applied recently in attempts to provide probabilistic proofs to propositions that would take too long to prove or disprove by deductive argument, even on a computer. One such application involves showing "beyond a reasonable doubt" whether a given large number is a prime [Kolata, 1976].

2. Approximating Attained Significance Levels in Exact Tests

All of the exact conditional test procedures discussed above focus on the set S of all

TABLE 1

$|S| = 12{,}798{,}781$ in Exact Conditional Kruskal-Wallis Test.[a]

Group	Response				
	Very Low	Low	High	Very High	Total
1	1	5	3	3	12
2	1	6	6	4	17
3	5	7	1	1	14
4	1	9	2	1	13
Total	8	27	12	9	56

[a] Klotz and Teng [1977] reported $|S|$ for these marginal distributions.

tables with the same row marginal counts $\{n_{i\cdot}, 1 \le i \le r\}$ and the same column marginal counts $\{n_{\cdot j}, 1 \le j \le c\}$ as the observed table. In all instances, conditional upon these marginal entries, the null probability of observing a table in S with entries $\{n'_{ij}\}$ is shown [Lehmann, 1975, p. 384] to be of the generalized hypergeometric form,

$$P(\{n'_{ij}\} \mid \{n_{i\cdot}, n_{\cdot j}\}) = \frac{\prod_{i=1}^{r} n_{i\cdot}! \prod_{j=1}^{c} n_{\cdot j}!}{n! \prod_{i} \prod_{j} n'_{ij}!} . \tag{2.1}$$

Instead of calculating the exact conditional significance level α by considering every table in S, we propose estimating the significance level by utilizing information from a random sample oftables in S. Since the tables in S occur with different relative frequencies [see (2.1)], it is necessary to impose a sampling procedure which produces tables according to these probabilities.

Refer to Table 2 for definiteness. There is a total of n observations, $n_{\cdot 1}$ in the first column, $n_{\cdot 2}$ in the second column, and $n_{\cdot 3}$ in the third column. Corresponding to this identification of the observations by column we set aside a total of n objects, $n_{\cdot 1}$ of which are denoted A, $n_{\cdot 2}$ of which are B's and $n_{\cdot 3}$ of which are C's. We then make a random permutation of these objects and partition the set into groups of the first $n_{1\cdot}$, the next $n_{2\cdot}$, and the remaining $n_{3\cdot}$. This partition can be achieved in $n!/(n_{1\cdot}!n_{2\cdot}!n_{3\cdot}!)$ distinct and equally likely ways. Now for each of the $n_{1\cdot}$ objects alloted to Group 1, we count the number of A's, B's and C's. Let n_{11} denote the number of A's, n_{12} the number of B's and n_{13} the number of C's. We repeat the above for the $n_{2\cdot}$ in Group 2 and $n_{3\cdot}$ in Group 3. The result will be a matrix with the appropriate row and column marginals. The number of ways the internal portion of the matrix can be generated is

$$\binom{n_{\cdot 1}}{n_{11}\ n_{21}\ n_{31}} \times \binom{n_{\cdot 2}}{n_{12}\ n_{22}\ n_{32}} \times \binom{n_{\cdot 3}}{n_{13}\ n_{23}\ n_{33}} .$$

TABLE 2

Table Used to Illustrate Sampling Scheme

				Total
	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Dividing this number by the total number of partitions just given, we see that the probability of any particular matrix is in fact as given in (2.1).

A random sample of M distinct tables from S can be achieved by repeating this procedure M times. After each table has been generated, we calculate the value of the desired statistic and compare it to the value of the statistic for the table actually observed. If the values of the statistic for X of the sample tables provide at least as much evidence in favor of H_a as the value of the statistic for the observed table, then the estimated exact conditional significance level, $\hat{\alpha}$, is simply X/M.

Now, the number of "tail" test statistic values X is a binomially distributed variable with M trials and success probability α. For M large, $\hat{\alpha} = X/M$ is approximately normally distributed with mean α and variance $(1 - \alpha)\alpha/M$. Thus if we desire to estimate α within B units with $(1 - \delta)100\%$ confidence, we require

$$M \doteq \frac{(Z_{\delta/2})^2}{B^2}(\alpha)(1 - \alpha)$$

where Z_δ denotes the $(1 - \delta)^{\text{th}}$ quantile of the standard normal distribution. Since $\alpha(1 - \alpha) \leq \frac{1}{4}$ for all α,

$$M \geq \tfrac{1}{4}\frac{(Z_{\delta/2})^2}{B^2}$$

will be sufficient for our purposes for any α. For example, if we desire to estimate α to within .01 with 99% confidence, we require

$$M \geq \tfrac{1}{4}\frac{(2.576)}{.01} = 16{,}589.44.$$

Thus, we see that $M = 17{,}000$ is more than sufficient to estimate α to within .01 with 99% confidence. Similarly, a sample of just $M = 1700$ tables is adequate to estimate α to within .02 with 90% confidence. These values of M lead to inexpensive analyses for tables where the magnitude $|S|$ is so large that the exact analysis is not feasible.

To illustrate the approximate conditional test, we re-analyzed some tables for which exact conditional test results were reported by Agresti and Wackerly [1977] and Klotz and Teng [1977]. The exact chi-square test on Table 3 was reported by Agresti and Wackerly to yield $\alpha = .004$. (To four decimal places, the level is .0038). Our approximate test yielded $\hat{\alpha} = .0039$ based on sampling 17,000 tables, and $\hat{\alpha} = .0047$ based on sampling $n = 1700$ tables. Alternate test statistics, such as the likelihood ratio statistic or a nominal measure

TABLE 3

Table Used to Illustrate Approximation of α in Exact Conditional Chi-square Test

10	1	6
3	5	0
5	0	1

of association, could lead to different α and $\hat{\alpha}$ values. In the examples we have studied though, these test criteria lead to very similar results. An exception to this is the Freeman-Halton test, in which the tables are ordered (often in an anomalous manner) by their probabilities, rather than by an index of their deviations from the null hypothesis. Klotz and Teng reported $\alpha = .044055$ for the exact Kruskal-Wallis test on Table 4, for which $|S| = 32{,}194$. Using the approximate test with 1700 tables, we obtained $\hat{\alpha} = .042941$. These $\hat{\alpha}$ values are well within the limits of what would be expected due to sampling error. Of course, the most important application of the approximate test is to tables for which the exact test is impractical. We give such an application at the end of the next section.

3. Guideline For Implementation Of The Procedure

In this section we develop some guidelines concerning when the approximation procedure should be used and the ease with which it can be applied in those situations. If the sample size is large enough in a particular table that an asymptotic test is clearly appropriate, then it is probably simplest to use it. For example, in testing for association between two nominal variances, the chi-square test of independence might be used if all the expected frequencies exceed five. On the other hand, if we doubt the appropriateness of the asymptotic test, but the exact conditional test seems feasible, we would use it. For example, if we wish to conclude the test within one minute of computer time (on a computer comparable to the IBM 370/165), we could use the exact conditional test when we are confident that $|S|$ is less than about 100,000 for small tables and less than about 50,000 for larger tables (say with degrees of freedom exceeding ten).

It is not simple to calculate $|S|$ exactly, a priori, in order to gauge whether an exact test can be economically implemented. In the general $r \times c$ case, no closed form expression is available for $|S|$ as a function of the marginal frequencies. An upper bound for $|S|$ for

TABLE 4

Table Used to Illustrate Approximation of α in Exact Conditional Kruskal-Wallis Test[a]

		Response (ordered)			
		A	B	C	D
Group	1	1	4	2	5
	2	1	9	3	1
	3	4	6	3	0

[a]Exact test conducted in Klotz and Teng [1977].

various table dimensions and sample sizes is given by Agresti and Wackerly [1977] in their Table 5. Klotz and Teng [1977] gave a geometric method for evaluating $|S|$. However, this method itself could require considerable computer time. For example, they report that it took 17 seconds of CPU time on the UNIVAC 1110 to determine $|S|$ for Table 1. Gail and Mantel [1977] obtained a recursive relationship which makes it possible to determine $|S|$ through a procedure for which $|S|$ is obtained iteratively for a sequence of submatrices of sizes $1 \times c$, $2 \times c$, $\cdots$, $r \times c$. Their procedure also requires the use of a computer to obtain a solution, except for small tables with very small sample sizes.

For the purpose of choosing between the exact and the approximate conditional tests, it is sufficient to calculate an approximation for $|S|$. Good [1976] conjectured an approximation for $|S|$ of

$$|S| \approx \frac{1.3n^4B}{rc\sum n_{i\cdot}^2 n_{\cdot j}^2},$$

where

$$B = \frac{\prod_i \binom{n_{i\cdot} + c - 1}{n_{i\cdot}} \prod_j \binom{n_{\cdot j} + r - 1}{n_{\cdot j}}}{\binom{n + rc - 1}{n}}.$$

This approximation seems to perform well when the row or column marginal frequencies are equal. The ratio of $|S|$ to the approximation fell between .75 and 1.1 for all tables studied by Good in which the rows margins were equal. Another approximation to $|S|$ was given by Gail and Mantel [1977], based on a Central Limit Theorem argument in which the vectors $(n_{i1}, \cdots, n_{ic-1})$, $i = 1, \cdots, r$, are treated as independent. If the table is arranged so that $r \geq c$, then their approximation is

$$|S| \approx \left[\prod_i \binom{n_{i\cdot} + c - 1}{n_{i\cdot}}\right]\left[\frac{c(c+1)}{2\pi\sum_i n_{i\cdot}(n_{i\cdot} + c)}\right]^{\frac{c-1}{2}} c^{1/2} \exp\left[\frac{c(c+1)\left(\sum_j n_{\cdot j}^2 - \frac{n^2}{c}\right)}{2\sum_i n_{i\cdot}(n_{i\cdot} + c)}\right]$$

Gail & Mantel suggest that this approximation improves as r and the $\{n_{i\cdot}\}$ increase in size. The results of utilizing these approximations on some tables for which $|S|$ is known are presented in Table 5. In this table, we compare the Good and the Gail-Mantel approximations to $|S|$ for Tables 1, 3, and 4 of this paper, and for the largest sample sizes for which the exact test was conducted by Agresti and Wackerly [1977], for various table dimensions with uniform marginal frequencies. In general, both approximations seem to be reasonably adequate with neither establishing a tendency to be more adequate than the other. Our experience has been that the approximations tend to be less accurate when the marginal frequencies are markedly non-uniform.

In practice, we recommend that both of the above approximations for $|S|$ be calculated and used to gauge the order of magnitude of $|S|$—whether $|S|$ is in the thousands or hundreds of thousands, for example. If both approximations give values which make an exact test seem feasible (according to Table 5 of Agresti & Wackerly, 1977), then we suggest using an exact test.

In most applications, use of Klotz and Teng's geometric construction or Gail and Mantel's recursive formulas for calculating the exact size of $|S|$ would be unnecessary, since an indication of the relative magnitude of $|S|$ is sufficient for gauging the feasibility of the exact test. If we doubt the adequacy of the asymptotic approximation, and if the exact conditional test appears to be too time-consuming, then the approximation of the exact conditional significance level should be obtained. We investigated the degree of

TABLE 5

Comparison of Exact Number of Tables Having Given Marginals ($|S|$) to Approximations Suggested by Good [1976] and by Gail and Mantel [1977].

	Table 1	Table 3	Table 4	← Uniform marginals						→		
Table size	4×4	3×3	3×4	2×3	2×4	2×5	2×6	2×7	3×3	3×4	3×5	4×4
Sample size	56	31	39	100	100	100	50	40	70	40	20	20
$\|S\|$	12,798,781	728	32,194	884	11,726	116,601	38,802	46,398	47,450	110,328	16,250	40,176
Good approx.	14,242,526	562	34,046	958	12,502	124,550	41,297	49,244	47,398	110,311	16,138	38,711
Gail-Mantel approx.	6,956,153	490	39,143	941	12,154	120,333	39,839	47,492	52,358	119,783	17,437	46,254

precision with which the attained significance level in the exact chi-square test of independence can be estimated on the Amdahl 470V/6 11 computer. Specifically, we observed the number of seconds of CPU time which is required to obtain an estimate of α using 17,000 random tables, for which $P(|\hat{\alpha} - \alpha| \leq .01) \geq .99$, for several different table dimensions and sample sizes. Special emphasis was given to arrays that cannot be easily handled using the exact test. Table 6 reports these results, for the cases in which all marginal frequencies are equal or within at most one of each other. These times include compiling time for the program as well as the time spent in the subroutine.

In comparing the approximation procedure to the exact conditional test, two differences are suggested by Table 6. First, unlike the exact test, for a fixed sample size the approximation procedure is practically as feasible on a table of large dimensions as on one of small dimensions. This is because the same number of tables (17,000) is generated in either case for the approximation procedure, whereas $|S|$ is dramatically larger for implementing the exact test on the larger table dimensions. Secondly, the CPU time is approximately linearly related to the sample size n in the approximation procedure, for fixed table dimensions. This is not surprising, since the number of operations required to generate each of the 17,000 tables is roughly proportional to the number of elements there are to be allocated to the cells of the table. For the exact test, on the other hand, $|S|$ blows up dramatically as n increases, for fixed table dimensions. In summary, comparing Table 6 to Table 5 in Agresti and Wackerly [1977], we see that the approximation procedure can in practice be used economically for those sample size-table dimensions combinations for which the exact test is impractical and the asymptotic test is questionable. Also, good estimates of α can be obtained in much less time than indicated in Table 6, if necessary. For example, $P(|\hat{\alpha} - \alpha| \leq .02) \geq .90$ if we generate 1700 tables, which takes roughly one-tenth the CPU time indicated in that table.

We mentioned in Section 1 that Table 1 could not be feasibly handled with the exact Kruskal-Wallis test, since $|S| = 12{,}798{,}781$. The evaluation of that table using the approximation procedure required 21.19 seconds of CPU time for generating 17,000 tables, and yielded $\hat{\alpha} = .020$. The corresponding approximation for 1700 tables required only 3.46 seconds, and yielded $\hat{\alpha} = .024$.

TABLE 6

Amdahl 470V/6 11 CPU Time for Approximating α in the Exact Chi-square Test of Independence Using 17,000 Tables

Table Dimensions	Sample Size 20	30	50	100	200
2 × 7	10.58	14.14	21.72	39.39	78.10
3 × 4	10.08	13.07	19.58	36.05	69.51
4 × 4	10.42	13.64	20.28	36.79	70.16
5 × 5	11.70	15.56	22.02	39.59	72.81
6 × 6	13.37	17.19	24.14	41.70	78.50

NOTE: Tables on right of line cannot be evaluated in less than one minute CPU time using the exact conditional test.

Note that the values 17,000 and 1700 for the number of tables to be generated are actually upper bounds for the number of tables required to obtain the desired accuracy. These quantities were calculated by using the worst possible value for α, namely $\alpha = .50$. If α is expected to be closer to zero or one, the actual number of tables required will be substantially less than these upper bounds. Further savings in computer time could be accomplished by doing the sampling sequentially. For example, if it is necessary simply to distinguish whether $\underline{\alpha} \leq .05$ or $\alpha > .05$, the sampling process is likely to be terminated more quickly unless α is close to .05.

Although we have limited our discussion of exact conditional tests to analyses of bivariate cross-classifications, clearly the concept of estimating attained significance levels when asymptotic approximations are poor can be extended to other situations. In particular, exact tests of independence, of no interaction or of no partial association, are of interest in multidimensional cross-classification tables with small sample sizes. Gail & Mantel's approximation for the number of 3-dimensional tables having a given set of marginal frequencies can be used to decide whether the α-levels should be estimated rather than exactly calculated.

A copy of the Fortran subroutine used for estimating attained significance levels in the Kruskal-Wallis test and in the chi-square test of independence is available from the authors.

REFERENCES

Agresti, Alan, & Wackerly, Dennis. Some exact conditional tests of independence for $r \times c$ cross-classification tables. *Psychometrika*, 1977, *42*, 111–125.

Boyett, James M., & Shuster, J. J. Nonparametric one-sided tests in multivariate analysis with medical applications. *Journal of the American Statistical Association*, 1977, *72*, 665–668.

Fisher, R. A., *The design of experiments* (9th ed.) New York: Hafner Press, 1977.

Forsythe, Alan B., & Frey, H. S. Tests of significance from survival data. *Computers and Biomedical Research*, 1970, *3*, 124–132.

Freeman, G. H., & Halton, J. H. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 1951, *38*, 141–149.

Gail, M., & Mantel, N. Counting the number of $r \times c$ contingency tables with fixed margins. *Journal of the American Statistical Association*, 1977, *72*, 859–862.

Good, I. J. On the application of symmetric dirichlet distributions and their mixtures to contingency tables. *The Annals of Statistics*, 1976, *4*, 1159–1189.

Klotz, Jerome, & Teng, James. One-way layout for counts and the exact enumeration of the Kruskal-Wallis H distribution with ties. *Journal of the American Statistical Association*, 1977, *72*, 165–169.

Kolata, Gina Bari. Mathematical proofs: The genesis of reasonable doubt. *Science*, 1976, *192*, June 4, 989–990.

Lehmann, E. *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden-Day Inc., 1975.

Wackerly, D. D., McClave, J. T., & Rao, P. V. Measuring nominal scale agreement between a judge and a known standard. *Psychometrika*, 1978, *43*, 213–223.

Manuscript received 2/24/78
Final version received 7/3/78