A Test for Detecting Outlying Cells in the Multinomial Distribution and Two-Way Contingency Tables

CAMIL FUCHS and RON KENETT*

A test based on the maximum adjusted residual from multinomial models, namely, the M test, is proposed. Sharp bounds on its critical values are provided for the multinomial case, while for the two-way table we present conservative critical values. The new test can be used to test null hypotheses against one- or two-sided alternatives and to detect outliers simultaneously with the rejection of the null hypothesis. Under certain alternatives, the M test is asymptotically more powerful than the chi-squared test.

KEY WORDS: Outlying cells; Adjusted residuals; M test; Multinomial models; Contingency tables.

1. INTRODUCTION

Chi-squared tests for hypotheses concerning multinomial and product multinomial distributions are among the most frequently used in the statistical literature. As Cochran (1952) pointed out, however, ". . . the test is commonly used when we do not have a clear cut alternative in mind" (p. 323). We focus here on specific alternatives characterized by the presence of outlying cells and propose a test that can be used to test null hypotheses against one- or two-sided alternatives and to detect outlying cells simultaneously with the rejection of the null hypothesis.

Several methods have been proposed in recent years for detecting outliers in two-way contingency tables (Fienberg 1969; Haberman 1973; Brown 1974). The last two authors proposed criteria for detecting outlying cells that are based on adjusted residuals. Brown's method (which has been incorporated in the BMDP computer package) proceeds to the identification of significant cells only after rejection of the null hypothesis by the chi-squared test and stops when the chi-square for quasi-independence is nonsignificant.

In this work we propose a new test (the M test) based on the maximum adjusted residual and provide its critical values. In some cases this new test is asymptotically more powerful than the common chi-squared test (particularly in the presence of a single outlier). The use of the M test may therefore lead to detection of outlying cells that would remain unidentified by the methods that use the chi-squared test as a stopping rule for detecting outliers.

A key finding in this study is that the use of the simple Bonferroni approximation in the multinomial case leads to sharp critical values for the M test. We could have incorporated in our title the statement "Bonferroni wins again" as in Bohrer et al. (1979). Similarly, in a review on the developments in multiple comparisons, Miller (1977) remarks that "over the course of the past ten years I have become even more impressed with the tightness of the (Bonferroni) bound" (p. 779).

2. THE M TEST

Let $\mathbf{n}$ be a random vector from a multinomial distribution, $\mathbf{n} = \{n_i: 1 \le i \le k\} \sim \text{mult}(N, \mathbf{p})$, $N = \sum n_i$, $p_i \ge 0$, $\sum p_i = 1$. We test H_0: $\mathbf{p} = \mathbf{p}^{(0)}$ against H_1: $\mathbf{p} \ne \mathbf{p}^{(0)}$, where $\mathbf{p}^{(0)}$ is a prespecified frequency vector. Under the null hypothesis, n_i is asymptotically normally distributed with mean $Np_i^{(0)}$ and variance

$$Np_i^{(0)}(1 - p_i^{(0)}) .$$

The adjusted residuals Z_i are defined as

$$Z_i = \frac{n_i - Np_i^{(0)}}{(Np_i^{(0)}(1 - p_i^{(0)}))^{\frac{1}{2}}}, \quad i = 1, \ldots, k . \quad (2.1)$$

Under the null hypothesis, H_0: $\mathbf{p} = \mathbf{p}^{(0)}$, the random vector $\mathbf{Z}$ converges in distribution to a vector $\mathbf{Z}^{(0)} = (Z_1^{(0)}, \ldots, Z_k^{(0)})$, which has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{R} = (R_{ij})$, where

$$r_{ij} = -\left\{\frac{p_i^{(0)}p_j^{(0)}}{(1 - p_i^{(0)})(1 - p_j^{(0)})}\right\}^{\frac{1}{2}} \quad \text{for} \quad i \ne j$$

$$= 1 \quad \text{for} \quad i = j .$$

The proposed M test for two-sided alternatives, at significance level α, rejects the null hypothesis H_0 if

$$\max_i |Z_i| > M^* ,$$

where $\{\Pr \max_i |Z_i| > M^* | H_0\} = \alpha$. (2.2)

In some cases the researcher might have a priori knowledge that the outliers (if any) have positive devia-

* Camil Fuchs is Assistant Professor in the Statistics Department, University of Wisconsin, Madison, WI 53706. Ron Kenett is a member of the technical staff, Bell Laboratories, 6 Corporate Place, Piscataway, NJ 08854. This article was written while both authors were at the University of Wisconsin. An early version of this article was presented by the first author at the 138th annual meeting of the American Statistical Association, San Diego, California. The authors wish to thank an associate editor whose pertinent remarks and references helped considerably improve this article.

Reprinted from: © Journal of the American Statistical Association
June 1980, Volume 75, Number 370
Theory and Methods Section
Pages 395-398

tions from the expected values under H_0. The test for such one-sided alternatives is to reject H_0 when

$$\max_i Z_i > M_+^* ,$$

$$\text{where} \quad \Pr\{\max_i Z_i > M_+^* | H_0\} = \alpha . \quad (2.3)$$

By symmetry, the critical value for detecting negative outliers is $M_-^* = -M_+^*$, and H_0 is rejected if $\min_i Z_i < M_-^*$. Note that all the techniques mentioned in the previous section for detecting outliers in multinomial models test for two-sided alternatives only.

3. COMPUTATION OF CRITICAL VALUES

In this section we find bounds on the critical value M_+^* which satisfies

$$\Pr(\max_i Z_i^{(0)} \geq M_+^*) = \alpha. \quad (3.1)$$

For large samples, M_+^* can be used as an approximate critical value for the M test. The bounds on M_+^* are derived as a corollary to a basic inequality whose proof requires the following two lemmas.

Lemma 1 (Feller 1970, IV.5):

Let E_i, $i = 1, \ldots, k$ be arbitrary events. Then

$$\sum_i \Pr(E_i) - \sum_{i<j} \Pr(E_iE_j) \leq \Pr(\bigcup_{i=1}^k E_i) \leq \sum_i \Pr(E_i) . \quad (3.2)$$

Let E_i be the event $Z_i^{(0)} \geq M_+^*$, where M_+^* is a positive constant. Then $\bigcup_{i=1}^k E_i = \max_i Z_i^{(0)}$. Since the $Z_i^{(0)}$ are identically distributed we can define

$$\theta = \Pr(E_i) , \quad i = 1, \ldots, k .$$

If we use these definitions, (3.2) yields

$$k\theta - \sum_{i<j} \Pr(E_iE_j) \leq \Pr(\max_i Z_i^{(0)}) \leq k\theta . \quad (3.3)$$

Lemma 2:

Let E_i, $i = 1, \ldots, k$ and θ be defined as shown before. Then

$$\sum_{i<j} \Pr(E_iE_j) \leq \frac{k(k-1)}{2}\theta^2 . \quad (3.4)$$

The proof of the inequality makes use of the negative correlations between $Z_i^{(0)}$ and $Z_j^{(0)}$ for $i \neq j$. (See Šidák 1968 for a more general discussion of the dependence of the multivariate normal probabilities on correlations.)

Basic Inequality:

Combining Lemmas 1 and 2, we get the inequality:

$$k\theta - \frac{k(k-1)}{2}\theta^2 \leq \Pr(\max_i Z_i^{(0)} \geq M_+^*) \leq k\theta . \quad (3.5)$$

Corollary:

The lower and upper bounds on M_+^* are

$$\Phi^{-1}\left\{1 - \frac{k - [k^2 - \alpha k(k-1)]^{\frac{1}{2}}}{k(k-1)}\right\} , \quad \Phi^{-1}\left\{1 - \frac{\alpha}{k}\right\} , \quad (3.6)$$

respectively, where $\Phi(\cdot)$ is the cumulative standard normal distribution.

Proof: Let Z have a univariate standard normal distribution and $\theta = \Pr(Z \geq M_+^*)$. We can rewrite (3.5) as

$$k \Pr(Z \geq M_+^*) - \frac{k(k-1)}{2}[\Pr(Z \geq M_+^*)]^2 \leq \Pr(\max_i Z_i^{(0)} \geq M_+^*) \leq k \Pr(Z \geq M_+^*) . \quad (3.7)$$

Equating each side of (3.7) with a significance level α, we obtain by the inverse normal transformation the upper and the lower bound on M_+^*.

The use of the upper bound alone was suggested by Kozelka (1956) for the case $p_i^{(0)} = 1/k$, $1 \leq i \leq k$. In that work it is shown that the exact values of M_+^* for $k = 3, 4, 5$, $N = 3(1)12$ and $\alpha = .05$ agree with the values obtained by using the upper bound of the asymptotic approximation. (Also see David 1970.)

The bounds for the critical value M_+^* calculated in the preceding paragraphs refer to the one-sided test for positive outliers. The critical value for the one-sided test with negative outliers is $-M_+^*$. For the two-sided test, an upper bound on M^* is obtained by replacing α by $\alpha/2$ in the upper bound on M_+^*. A slight improvement on the upper bound for the two-sided test can be achieved through an inequality presented in Šidák (1968), namely,

$$\Pr(\max_i |Z_i^{(0)}| \geq M^*) \leq \Phi^{-1}\left[\frac{1 + (1-\alpha)^{1/k}}{2}\right] . \quad (3.8)$$

4. EXTENSION TO TWO-WAY TABLES

The M test is easily extended to cover two-way contingency tables with entries n_{ij}, $i = 1, \ldots, I$; $j = 1, \ldots, J$. The adjusted residuals for a $I \times J$ table, Z_{ij}, are defined (Haberman 1974) as

$$Z_{ij} = \frac{n_{ij} - n_{i+}n_{+j}/N}{[n_{i+}n_{+j}(N - n_{i+})(N - n_{+j})/N^3]^{\frac{1}{2}}} , \quad (4.1)$$

$$i = 1, \ldots, I ; \quad j = 1, \ldots, J ,$$

$$N = \sum n_{ij} , \quad n_{i+} = \sum_j n_{ij} , \quad n_{+j} = \sum_i n_{ij} .$$

Define $p_{ij} = E(n_{ij}/N)$. Then the M test for testing the hypothesis of independence, H_0: $p_{ij} = p_{i+}p_{+j}$ versus H_1: $p_{ij} = p_{i+}p_{+j}$, rejects H_0 at a level α if

$$\max_{i,j} |Z_{ij}| > M^* ,$$

$$\text{where} \quad \Pr(\max_{i,j} |Z_{ij}| > M^* | H_0) = \alpha . \quad (4.2)$$

When positive residuals are of interest the test rejects H_0 if

$$\max_{i,j} Z_{ij} > M_+^* ,$$

$$\text{where} \quad \Pr(\max Z_{ij} > M_+^* | H_0) = \alpha . \quad (4.3)$$

Similarly, for negative residuals we reject H_0 if

$$\min_{i,j} Z_{ij} < M_-^* = -M_+^* .$$

Testing one-sided alternatives might reflect particular interest in detecting cells exhibiting positive (negative) association.

An upper bound on M_+^* is $\Phi^{-1}(1 - (\alpha/k))$, the Bonferroni bound, where $k = IJ$. Lemma 2 in Section 3 does not hold here, however, because not all residuals in a two-way table are negatively correlated. To derive lower bounds for the critical values of the M test in this case we therefore have to rely on inequality (3.3) and solve for M_+^*:

$$k \Pr(A_l) - \sum_{l<l'} \Pr(A_l A_{l'}) = \alpha , \quad (4.4)$$

where A_l is the event $Z_{ij} > M_+^*$ following the vectorization of the matrix $\mathbf{Z}$. If we use tables for the bivariate normal distribution (National Bureau of Standards 1959) for each of the correlations among pairs of residuals (see Haberman 1974, Ch. 4.3 for a derivation of the mathematical structure of the correlation matrices), critical values satisfying (4.4) can be found by trial and error or any other interpolation procedure. For the two-sided test, an upper bound on M^* can be obtained by using the methods from Section 3 with $k = IJ$.

5. DISCUSSION

In this study, the M test, which is based on adjusted residuals, is proposed. Its critical values are derived from inequalities relevant to simultaneous test procedures (e.g., see Dunn 1961 and Goodman 1964).

An evaluation of the performance of the M test for a variety of cases by means of a numerical comparison of its power with the power of chi-squared test can be found in Fuchs (1978). For both the multinomial and the two-way contingency table cases the probability vectors $\mathbf{p}$ were assumed to satisfy

$$\mathbf{p} - \mathbf{p}^0 = \mathbf{d}/\sqrt{N} , \quad \mathbf{d} = (d_1, \ldots, d_k) , \quad (5.1)$$

where the d_i's are constants. ((5.1) can be relaxed to $\mathbf{p} \to \mathbf{p}^0 = \mathbf{d}(N)/\sqrt{N}$ with $\mathbf{d}(N) \to \mathbf{d}$ as $N \to \infty$). This structure of alternatives has been considered in investigations of the limiting power of the chi-squared test (e.g., see Cochran 1952, Chapman and Meng 1966, and Haberman 1974).

Among all the cases, the maximum difference between the bounds on the power of the M test, as calculated by using the upper Bonferroni bound and the lower bound in (3.6), was less than .004. Therefore, in the case of a single marginal constraint the Bonferroni approximation turns out to be very accurate. In the multinomial case the computed lower bounds of the asymptotic power of the proposed M test exceed the power of the chi-squared test in the presence of a single outlier for moderate- and large-number groups. In the case of several outliers, the M test has higher power provided that the number of outliers is limited (to less than 5 to 10 percent of the number of cells) and the deviations are fairly unequal. When a single outlier is present in a two-way contingency table the probability of its correct identification by the M test exceeds the power of the chi-squared test for almost all the cases of interest.

Therefore, it seems appropriate to suggest that the adjusted residuals be computed and the M test be used as a supplement to the chi-squared test whether the null hypothesis is retained. The possibility of testing H_0 against one-sided alternatives and the ability to identify one or more outliers simultaneously with the rejection of the null hypothesis are notable advantages of the new test. The upper bounds on the critical values of the M test are simple to compute and result in a conservative test. By computing the lower bounds the accuracy of these upper bounds can be evaluated.

The calculation of the bounds on the critical values of the M test rely only on the asymptotic bivariate normal distribution of pairs of residuals (and on their marginal univariate normal distributions). Equation (4.4) can be solved for a variety of models for which the M test can also be applied in order to detect outliers and to test null hypotheses.

[*Received October 1978. Revised July 1979.*]

REFERENCES

Bohrer, R., Chow, W., Faith, R., Joshi, V., and Wu, C.F. (1979), "Multiple Decision Rules for Factorial Simple Effects: Bonferroni Wins Again!," unpublished manuscript.

Brown, M.B. (1974), "Identification of the Sources of Significance in Two-Way Contingency Tables," *Applied Statistics*, 23, 405–413.

Chapman, D.G., and Meng, R.C. (1966), "The Power of Chi-Square Tests for Contingency Tables," *Journal of the American Statistical Association*, 61, 965–975.

Cochran, W.G. (1952), "The χ^2 Test of Goodness of Fit," *Annals of Mathematical Statistics*, 23, 315–345.

David H.A. (1970), *Order Statistics*, New York: John Wiley & Sons.

Dunn, O.J. (1961), "Multiple Comparisons Among Means," *Journal of the American Statistical Association*, 52, 52–64.

Feller, W. (1970), *An Introduction to Probability Theory and Its Application* (Vol. 1), New York: John Wiley & Sons.

Fienberg, S.E (1969), "Preliminary Graphical Analysis and Quasi-Independence for Two-Way Contingency Tables," *Applied Statistics*, 18, 153–168.

Fuchs, C. (1978), "Multinomial Distribution: A Test for the Null Hypothesis Based on Adjusted Residuals," *Proceedings of the American Statistical Association*, Social Statistics Section, 409–414.

Goodman, L.A. (1964), "Simultaneous Confidence Limits for Cross-Product Ratios in Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B, 26, 86–102.

Haberman, S.J. (1973), "The Analysis of Residuals in Cross-Classified Tables," *Biometrics*, 29, 205–220.

——— (1974), *The Analysis of Frequency Data*, Chicago: University of Chicago Press.

Kozelka, R.M. (1956), "Approximate Upper Percentage Points for Extreme Values in Multinomial Sampling," *Annals of Mathematical Statistics*, 27, 507–512.

Miller, R.G. (1977), "Developments in Multiple Comparisons 1966–1976," *Journal of the American Statistical Association*, 72, 779–788.

National Bureau of Standards (1959), *Tables of the Bivariate Normal Distribution and Related Functions*, Applied Mathematics Series 50, Washington, D.C.: U.S. Government Printing Office.

Šidák, Z. (1968), "On Multivariate Normal Probabilities on Rectangles: Their Dependence on Correlations," *Annals of Mathematical Statistics*, 32, 1425–1434.