Testing for Association between Categorical Variables with Multiple-Response Data

Dan Nettleton, University of Nebraska-Lincoln
917 Oldfather Hall, Lincoln, Nebraska 68588-0323
dnettlet@mathstat.unl.edu

Key Words: Categorical Data, Chi-Square Test, Homogeneity, Multiple-Response, Rao-Scott Correction.

Abstract:

Some survey questions provide respondents with a list of possible answers and instructions to "mark all that apply." This paper presents methods for determining when the responses to a mark-all-that-apply question are associated with the responses to a standard question whose answers fall in one of several mutually exclusive categories. Such data originate from many sources including large-scale social, political, and economic surveys; work place studies; market research; and health care analyses among others. Recent approaches to the problem are briefly reviewed, a related alternative procedure proposed, and extensions that allow for the proper analysis of multiple-response data collected through complex sampling are suggested.

1 Introduction

Loughin and Scherer (1998) consider testing for association with multiple-response data. They study a survey of 262 Kansas livestock farmers who were asked to specify their education level (1. high school or less, 2. vocational school, 3. two-year college, 4. four-year college, or 5. other) and primary sources of veterinary information (1. professional consultant, 2. veterinarian, 3. state or local extension service, 4. magazines, and 5. feed company representatives). The 262 farmers – who were instructed to mark one education level and as many information sources as applicable – provided 453 responses to the veterinary information question. The survey researchers were interested in determining if the proportion of farmers using each information source is constant across varying levels of education.

In general, suppose X is a categorical random variable with I levels and $\mathbf{Y} \equiv (Y_1, \ldots, Y_J)'$ is a random vector of binary responses, i.e., $Y_j \in \{0, 1\}$ for all $j = 1, \ldots, J$. Suppose we have n independent observations,

$$\{(X_k, \mathbf{Y}_k')' \equiv (X_k, Y_{k1}, \ldots, Y_{kJ})'\}_{k=1,\ldots,n},$$

from the joint distribution of X and $\mathbf{Y}$. The vector $\mathbf{Y}$ can take any of 2^J values, in general, or any of $2^J - 1$ values when $(0, 0, \ldots, 0)'$ is not a valid response. For $i = 1, \ldots, I$ and $j = 1, \ldots, J$; let m_{ij} denote the number of observations for which $X = i$ and $Y_j = 1$. In the survey of livestock farmers, $I = 5$, $J = 5$, X_k takes value i if the kth farmer reports education level i, and Y_{kj} would be coded as 1 if the j^{th} information source is selected by farmer k and 0 otherwise. The count m_{ij} is simply the number of farmers who report education level i and indicate information source j among their sources of veterinary information. The complete livestock farmer survey data can be found in Loughin and Scherer (1998).

Loughin and Scherer consider testing

$$H_0 = \cap_{j=1}^{J} H_{0j} \text{ against } H_1 : \cup_{j=1}^{J} H_{1j} \qquad (1)$$

where, for $j = 1, \ldots, J$;

$$H_{0j} : P(X = i,\ Y_j = 1) = P(X = i)P(Y_j = 1)$$

for all $i = 1, \ldots, I$ and

$$H_{1j} : P(X = i,\ Y_j = 1) \neq P(X = i)P(Y_j = 1)$$

for some i. The null hypothesis H_0 is equivalent to "X is independent of Y_j for all $j = 1, \ldots, J$." Agresti and Liu (1999) refer to this null hypothesis as the hypothesis of multiple marginal independence which is conveniently abbreviated MMI. The alternative hypothesis H_1 is equivalent to "X and Y_j are associated for at least one $j \in \{1, \ldots, J\}$."

Recent approaches to testing MMI are sketched in Section 2. As indicated by Agresti and Liu (1999), a problem common to many of these proposals is a lack of invariance to the coding of 0 and 1 values in the components of $\mathbf{Y}$. For example, it is possible for a procedure to suggest a departure from MMI when positive responses are coded as 1 but agreement with MMI when positive responses are coded as 0. This

Partial funding provided by the Gallup Research Center

problem is examined in detail in Section 3, and a simple invariant test statistic given by the sum of the J Pearson chi-square tests of independence between X and each Y_j is proposed. The asymptotic null distribution of the proposed statistic is derived in Section 4, and a simple procedure for carrying out an approximate test is developed. Extensions of these results that include the analysis of multiple-response data collected through complex sampling are mentioned in Section 5.

2 Previous Work

Loughin and Scherer (1998) develop a bootstrap procedure for testing MMI based on the modified chi-square statistic:

$$X_M^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(m_{ij} - F_{ij})^2}{F_{ij}} \text{ where } F_{ij} = n_i m_{+j}/n$$

and n_i denotes the number of observations for which $X = i$ and $m_{+j} \equiv \sum_{i=1}^{I} m_{ij}$. A with-replacement sample of size n is drawn from the observed data, $X_1, \ldots, X_n$. Likewise, a with-replacement sample of size n is drawn from the observed vectors, $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. The two samples are randomly paired and a bootstrap replication of X_M^2 is computed with the resampled data. This process is repeated B times and the significance of the observed statistic is determined by comparison to the distribution of bootstrap values. A p-value, for example, is estimated by the proportion of bootstrap-replicated test statistics that exceed the observed statistic. Loughin and Scherer apply the modified chi-square testing procedure to the data from the survey of livestock farmers. The modified procedure provides significant evidence of association with an estimated p-value = 0.047 based on $B = 5000$).

Umesh (1995) considers the analysis of multiple-response data on gender and automobile preference using a "pseudo chi-square statistic". This pseudo chi-square statistic - which is equal to Loughin and Scherer's X_M^2 - is computed with observed data and compared to a chi-square distribution with $(I-1)(J-1)$ degrees of freedom. Umesh (1995) cautions that the chi-square approximation is not valid since the multiple responses of a single individual are not independent of each other. Loughin and Scherer (1998) show that X_M^2 is distributed as a weighted sum of independent chi-square random variables with weights that depend on the unknown joint distribution of X and $\mathbf{Y}$ under independence. Loughin and Scherer (1998) also provide an example that shows that Umesh's procedure can be very conservative.

Agresti and Liu (1999) discuss of variety of procedures for testing MMI including fitting generalized loglinear models of the type considered by Lang and Agresti (1994), a generalized estimating equation approach based on the work of Liang and Zeger (1996), the weighted least squares method of Koch, et al. (1977), a likelihood ratio or Pearson chi-square test, and a Bonferroni approach. These approaches use approximations of the test statistic's null distribution based on asymptotic results. Agresti and Liu (1999) note that Loughin and Scherer's bootstrap procedure is appealing because no asymptotic approximation is necessary. They point out, however, that resampling is done under the null hypothesis of independence between X and $\mathbf{Y}$ which is narrower than the MMI null hypothesis.

Decady and Thomas (1999) consider testing

$$H_0 = \cap_{j=1}^{J} H_{0j} \text{ against } H_1 : \cup_{j=1}^{J} H_{1j}$$

where, for $j = 1, \ldots, J$;

$$H_{0j} : P(Y_j = 1 \mid X = i) = P(Y_j = 1)$$

for all $i = 1, \ldots, I$ and

$$H_{1j} : P(Y_j = 1 \mid X = i) \neq P(Y_j = 1).$$

This testing problem is similar to a test of MMI but different in that the analysis focuses on the conditional distribution of Y_j given X, rather than the joint distribution of Y_j and X. MMH (for multiple marginal homogeneity) will be used to refer to this null hypothesis.

Decady and Thomas (1999) attempt to find a method of testing MMH that is computationally simpler than the bootstrap approach of Loughin and Scherer and the generalized-loglinear-model methods described by Agresti and Liu. They emphasize the need for methods of analysis that can be applied easily by practitioners to data that has been summarized from its raw form,

$$\{(X_k, \mathbf{Y}_k')' \equiv (X_k, Y_{k1}, \ldots, Y_{kJ})'\}_{k=1,\ldots,n},$$

to count data, n_i and m_{ij} ($i = 1, \ldots, I$; $j = 1, \ldots J$). Methods for analysis of count data are necessary, Decady and Thomas argue, because the sparseness of observed $\mathbf{Y}$ data in the 2^J-dimensional set of possible responses can lead to failure of the asymptotic approximations used in the direct analysis of the raw data. In addition, multiple-response data is typically published in summarized form and raw data is often publicly unavailable.

Decady and Thomas (1999) advocate a test of MMH based on a first-order Rao-Scott correction to

X^2_M (see Scott and Rao, 1981; Rao and Scott, 1981, 1984, 1987). The corrected test statistic is given by

$$X^2_M/(1-\frac{m_{++}}{nJ}), \text{ where } m_{++} \equiv \sum_{i=1}^{I}\sum_{j=1}^{J} m_{ij}.$$

Significance is assessed by comparing the corrected statistic to a chi-square distribution with $(I-1)J$ degrees of freedom. This chi-square distribution often serves as a reasonable approximation to the distribution of the corrected test statistic under MMH, and the procedure should prove useful when computation must be kept to a minimum, raw data is unavailable, or sparseness issues in the raw data call into question the appropriateness of asymptotic approximations used in more complex procedures. The main drawback to this approach is a lack of invariance to the coding of the **Y** data discussed in the next section.

3 Invariance Issues

Agresti and Liu (1999) argue that a good procedure for testing MMI should be invariant to whether Y_{kj} is coded as 1 if the jth item is selected by the kth individual and 0 otherwise as opposed to 0 if the jth item is selected by the kth individual and 1 otherwise. Such invariance is clearly a desirable property since Y_j is unassociated with X for all $j=1,\ldots,J$ if and only if $1-Y_j$ is unassociated with $1-Y_j$ for all $j=1,\ldots,J$. It is easy to show that none of the procedures mentioned in Section 2 based on the X^2_M statistic possess the invariance property.

The statistic X^2_M is closely related to the sum of the J Pearson chi-square statistics for testing the independence of X with Y_j $(j=1,\ldots,J)$. For $i=1,\ldots,I$; $j=1,\ldots,J$; and $\ell=0,1$; let $C_{i\ell j}$ denote the i,ℓ^{th} cell count in the $I\times 2$ contingency table corresponding to the cross classification of X and Y_j. Let $E_{i\ell j}$ denote the expected count (assuming independence of X and Y_j) for the i,ℓ^{th} cell in the j^{th} table, i.e.,

$$E_{i\ell j} = \frac{C_{i+j}C_{+\ell j}}{C_{++j}}$$

where $C_{i+j} \equiv C_{i0j}+C_{i1j}$, $C_{+\ell j} \equiv \sum_{i=1}^{I} C_{i\ell j}$, and $C_{++j} \equiv \sum_{i=1}^{I}(C_{i0j}+C_{i1j})$. Note that $C_{i1j}=m_{ij}$ and $E_{i1j}=F_{ij}$. Thus,

$$\begin{aligned} X^2_M &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(C_{i1j}-E_{i1j})^2}{E_{i1j}} \\ &= \sum_{j=1}^{J}\frac{C_{+0j}}{C_{++j}}\sum_{i=1}^{I}\frac{C_{++j}(C_{i1j}-E_{i1j})^2}{C_{+0j}E_{i1j}} \\ &= \sum_{j=1}^{J}\frac{C_{+0j}}{C_{++j}}\sum_{i=1}^{I}\sum_{\ell=1}^{2}\frac{(C_{i\ell j}-E_{i\ell j})^2}{E_{i\ell j}} \\ &= \sum_{j=1}^{J}\hat{\eta}_j X^2_j, \qquad (2) \end{aligned}$$

where $\hat{\eta}_j$ is the proportion of observations for which $Y_j=0$ and X^2_j is the Pearson chi-square statistic for testing the independence of X and Y_j.

Let X^2_{M1} denote the modified chi-square statistic computed with the original data, and let X^2_{M0} denote the statistic computed with $1-Y_{kj}$ in place of Y_{kj} for all $k=1,\ldots,n$; $j=1,\ldots,J$. Since X^2_j is invariant to recoding of the **Y** data, expression (2) implies that

$$X^2_{M0} = \sum_{j=1}^{J}(1-\hat{\eta}_j)X^2_j$$

and

$$X^2_{M0}+X^2_{M1} = \sum_{j=1}^{J}X^2_j. \qquad (3)$$

In view of these relationships, it seems logical to base a test of MMH on $X^2_{\text{SUM}} \equiv \sum_{j=1}^{J} X^2_j$. The statistic is invariant to the coding of the **Y** data and provides an intuitive measure of the degree to which the data depart from MMH. Agresti and Liu (1999) compute

$$X^2_{\text{SUM}} = 5.96+7.89+4.62+1.42+10.95 = 30.84$$

for the veterinary data. Since each X^2_j term in the sum has an asymptotic chi-square distribution with degrees of freedom $(I-1)=4$ under MMH, they compare the observed value of X^2_{SUM} to the chi-square distribution with $(I-1)J=20$ degrees of freedom to obtain 0.06 as an approximate p-value. Of course $X^2_1,\ldots,X^2_J$ are dependent, so X^2_{SUM} is not asymptotically chi-square with $(I-1)J$ degrees of freedom under the null. Realizing this, Agresti and Liu label the approach naive but note that the observed test statistic for the veterinary data (30.84) is quite similar to legitimate 20-degree-of-freedom chi-square statistics obtained through more computationally complex maximum likelihood methods. The work of the next section explains why comparing X^2_{SUM} to a chi-square distribution with $(I-1)J$ is a reasonable procedure supported by asymptotic theory.

4 The Asymptotic Null Distribution of X^2_{SUM}

Following the notation of Decady and Thomas (1999), let $\tau_j = P(Y_j=1)$ and $\tau_{ij} = P(Y_j = 1 \mid$

$X = i$). Let $\hat{\tau}_j = m_{+j}/n$ and $\hat{\tau}_{ij} = m_{ij}/n_i$ denote the corresponding sample estimates. Use η in place of τ to represent the $Y_j = 0$ quantities. For example, $\hat{\eta}_j = (n_i - m_{ij})/n_i$, the proportion of observations for which $Y_j = 0$, as in the previous section. The definitions of X^2_{M0} and X^2_{M1} along with expression (3) imply

$$\begin{aligned} X^2_{\text{SUM}} &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_i(\hat{\eta}_{ij} - \hat{\eta}_j)^2}{\hat{\eta}_j} \\ &\quad + \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_i(\hat{\tau}_{ij} - \hat{\tau}_j)^2}{\hat{\tau}_j} \\ &= \hat{\gamma}'\hat{\boldsymbol{\Gamma}}^{-1}\hat{\gamma}, \end{aligned}$$

where $\hat{\gamma}$ is a $2IJ$-dimensional vector with elements

$$\{\sqrt{n_i}(\hat{\eta}_{ij} - \hat{\eta}_j)\}_{i=1,\ldots,I;j=1,\ldots,J}$$

and

$$\{\sqrt{n_i}(\hat{\tau}_{ij} - \hat{\tau}_j)\}_{i=1,\ldots,I;j=1,\ldots,J},$$

and $\hat{\boldsymbol{\Gamma}}$ is a $2IJ \times 2IJ$ diagonal matrix with diagonal elements

$$\hat{\eta}_1, \ldots, \hat{\eta}_J, \ldots, \hat{\eta}_1, \ldots, \hat{\eta}_J, \hat{\tau}_1, \ldots, \hat{\tau}_J, \ldots, \hat{\tau}_1, \ldots, \hat{\tau}_J.$$

Note that under MMH, $\hat{\gamma}$ converges in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and rank $(I-1)J$ variance-covariance matrix $\boldsymbol{\Sigma}$ that has diagonal elements

$$\{(1 - \pi_i)\eta_j(1 - \eta_j)\}_{i=1,\ldots,I;j=1,\ldots,J}$$

and

$$\{(1 - \pi_i)\tau_j(1 - \tau_j)\}_{i=1,\ldots,I;j=1,\ldots,J}$$

where π_i is the limit of n_i/n that is assumed to exist for all i. Also note that $\hat{\boldsymbol{\Gamma}}$ converges almost surely to a diagonal matrix $\boldsymbol{\Gamma}$ with nonzero elements

$$\eta_1, \ldots, \eta_J, \ldots, \eta_1, \ldots, \eta_J, \tau_1, \ldots, \tau_J, \ldots, \tau_1, \ldots, \tau_J$$

so that, under MMH, X^2_{SUM} converges in distribution to the distribution of $\mathbf{U}'\boldsymbol{\Gamma}^{-1}\mathbf{U}$, where $\mathbf{U}$ is multivariate normal with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The following lemma leads to a useful characterization of the distribution of $\mathbf{U}'\boldsymbol{\Gamma}^{-1}\mathbf{U}$. (For proof see page 36 of Mathai and Provost, 1992, for example.)

Lemma. *Suppose* $\mathbf{U} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, *where* $\boldsymbol{\Sigma} = \mathbf{BB}'$ *for* $\mathbf{B}$, *a* $p \times r$ *matrix of rank* r. *Let* $\boldsymbol{\Gamma}$ *be a* $p \times p$ *symmetric matrix of rank* p. *Then* $\mathbf{U}'\boldsymbol{\Gamma}^{-1}\mathbf{U}$ *is distributed as* $\sum_{t=1}^{r} \lambda_t W_t$, *where* λ_t *are the nonzero eigenvalues of* $\mathbf{B}'\boldsymbol{\Gamma}^{-1}\mathbf{B}$ *and* $W_1, \ldots, W_r$ *are independent and identically distributed single-degree-of-freedom chi-square random variables.*

If we let $\mathbf{B}$ denote a $2IJ \times (I-1)J$ matrix such that $\mathbf{BB}' = \boldsymbol{\Sigma}$, an application of the lemma shows that, under MMH, X^2_{SUM} converges in distribution to $\sum_{t=1}^{(I-1)J} \lambda_t W_t$ where $\lambda_1, \ldots, \lambda_{(I-1)J}$ are the nonzero eigenvalues of $\mathbf{B}'\boldsymbol{\Gamma}^{-1}\mathbf{B}$ and $W_1, \ldots, W_{(I-1)J}$ are independent and identically distributed single-degree-of-freedom chi-square random variables. This asymptotic null distribution could be approximated through simulation. Estimates of $\tau_1, \ldots, \tau_J$ and an estimate of $\boldsymbol{\Sigma}$ could be used to obtain estimates of the eigenvalues, say $\hat{\lambda}_1, \ldots, \hat{\lambda}_{(I-1)J}$. Observations from the conditional distribution of $\sum_{t=1}^{(I-1)J} \hat{\lambda}_t W_t$ given $\hat{\lambda}_1, \ldots, \hat{\lambda}_{(I-1)J}$ could be used to approximate the asymptotic null distribution and assess the significance of X^2_{SUM}. Such simulation approaches are seldom popular with practitioners, and a simpler and faster approach is desirable.

An alternate strategy is to consider a first-order Rao-Scott correction to X^2_{SUM} of the type used by Decady and Thomas (1999) for the X^2_M statistic. Note that the average of $\lambda_1, \ldots, \lambda_{(I-1)J}$ is given by

$$\begin{aligned} \frac{\text{trace}(\mathbf{B}'\boldsymbol{\Gamma}^{-1}\mathbf{B})}{(I-1)J} &= \frac{\text{trace}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma})}{(I-1)J} \\ &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(1-\pi_i)(1-\eta_j+1-\tau_j)}{(I-1)J} \\ &= \frac{\sum_{i=1}^{I}(1-\pi_i)J}{(I-1)J} = 1. \end{aligned}$$

Thus, the Rao-Scott correction factor in this case is 1. It follows that X^2_{SUM} has an asymptotic null distribution that is approximately chi-square with $(I-1)J$ degrees of freedom whenever there is not large variation among the eigenvalues $\lambda_1, \ldots, \lambda_{(I-1)J}$. This suggests that the naive procedure that compares X^2_{SUM} to a chi-square distribution with $(I-1)J$ degrees of freedom is actually a reasonable method of obtaining an approximate test of MMH that requires only summarized count data and simple computation.

5 Discussion

A simulation study conducted by C. R. Bilder (manuscript in preparation) indicates that the X^2_{SUM} test exhibits type I error rates close to nominal in a wide variety of situations. There are cases where the test is too liberal, and further study is needed to characterize situations in which the test is unreliable. A second-order Rao-Scott correction should be considered as a possible means of improving the test's performance. An investigation of the power of the test for various alternatives is also needed.

There are clearly limitations to the X^2_{SUM} test, but it possesses a many positive features. The X^2_{SUM} test is a simple function of chi-square statistics that are familiar to practitioners and requires little computational sophistication. The test can be conducted with count data that is more likely to be available in secondary analysis than the raw data required by many other procedures. The distribution of X^2_{SUM} may be approximated by its asymptotic distribution for smaller sample sizes than are required by more complex procedures that make more complete use of the data. X^2_{SUM} should behave like a weighted sum of independent chi-square random variables when there is not sparseness in any of the tables cross-classifying X and Y_j ($j = 1, \ldots, J$). When the hypothesis of MMH is rejected, the individual chi-square statistics that are summed to obtain X^2_{SUM} can provide specific information about which components of $\mathbf{Y}$ are associated with $\mathbf{X}$.

The X^2_{SUM} test can be extended to the analysis of data collected through a complex sampling design. The first-order Rao-Scott correction factor is no longer 1 for complex sampling designs, but consistent estimates of the variances of the elements of $\hat{\gamma}$ under MMH are the only variance estimates required. For all $i = 1, \ldots, I; j = 1, \ldots, J$; let $\hat{u}_{ij}$ and $\hat{v}_{ij}$ denote consistent estimates (under MMH) of the variance of $\sqrt{n_i}(\hat{\eta}_{ij} - \hat{\eta}_j)$ and $\sqrt{n_i}(\hat{\tau}_{ij} - \hat{\tau}_j)$, respectively. The first-order Rao-Scott corrected X^2_{SUM} statistic is given by

$$\frac{X^2_{\text{SUM}}}{\hat{\lambda}.}, \text{ where } \hat{\lambda}. \equiv \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{u}_{ij}/\hat{\eta}_j + \hat{v}_{ij}/\hat{\tau}_j)}{(I-1)J}.$$

The X^2_{SUM} test can also be extended to situations in which Y_j is a categorical variable with k_j categories ($j = 1, \ldots, J$). As an example, one might wish to test whether the distribution of responses is the same for males and females for all the questions of a test that contains several multiple choice questions, each with a potentially different number of possible answers. Extensions that allow for the analysis of this type of data collected through complex sampling designs are also possible.

6 References

Agresti, A. and Liu, I.-M. (1999). Modeling a categorical variable allowing arbitrarily many category choices. To appear in *Biometrics*.

Decady, Y. J. and Thomas, D. R. (1999). Testing hypotheses on multiple response tables: a Rao-Scott adjusted chi-square approach. To appear in *Proceedings of the Administrative Sciences Association of Canada: Marketing Division*, ed. by A. Lavack. N.B. : ASAC, St. John.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33**, 133-158.

Lang, J. B. and Agresti, A. (1994). Simultaneous modelling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89**, 625-632.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Loughin, T. M. and Scherer, P. N. (1998). Testing for association in contingency tables with multiple column responses. *Biometrics* **54**, 630-638.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic Forms in Random Variables*. Macel Dekker, New York.

Rao, J. N. K. and Scott, A. J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *J. Amer. Statist. Assoc.* **76**, 221-230.

Rao, J. N. K. and Scott, A. J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. *Ann. of Statist.* **12**, 46-60.

Rao, J. N. K. and Scott, A. J. (1987). On Simple Adjustments to Chi-Square Tests with Sample Survey Data. *Ann. Statist.* **15**, 385-397.

Scott, A. J. and Rao, J, N. K. (1981). Chi-Squared Tests for Contingency Tables with Proportions Estimated from Survey Data. pp. 247-266 in *Current Topics in Survey Sampling*, ed. by D. Krewski, R. Platek, and J. N. K. Rao. Academic Press, New York.

Umesh, U. N. (1995). Predicting nominal relationships with multiple response. *Journal of Forecasting.* **14**, 585-596.