ANALYSIS OF SPARSE REPEATED CATEGORICAL MEASUREMENT DATA

Alan Agresti, University of Florida
Stuart Lipsitz, Harvard University
Joseph B. Lang, University of Florida

ABSTRACT

The feasibility of maximum likelihood (ML) analyses of marginal distributions of repeated categorical measurement data diminishes as the numbers of response occasions and response categories increases. This article describes alternative approaches that are much more feasible. To estimate model parameters, we recommend a "pseudo ML" approach that treats repeated responses as independent and uses a jackknife to estimate the covariance matrix of those estimates. Tests of hypotheses about response distributions (e.g., marginal homogeneity) use Wald statistics or adapted score statistics from the independent-samples case. We illustrate these analyses with a seven-dimensional table having 78,125 cells. Simulation results show no substantive loss of efficiency from using pseudo ML estimates.

1. Introduction

Many studies involve observing a response variable for each subject at several occasions -- for instance, at several time points or under several conditions. Such "repeated measurement" data are common in health-related applications. For example, a clinician might evaluate patients at weekly intervals regarding whether a new drug treatment is successful. When the response is categorical (e.g., success vs. failure), data can be displayed in a contingency table having the same categories for each dimension. When we observe each of n subjects at T occasions on an I-category response, a T-dimensional contingency table having I^T cells cross-classifies the T responses for those subjects.

The "occasions" for a repeated response need not refer to different times. For instance, a biomedical response might be measured at T locations on a subject's body, or by T raters. To illustrate, consider Table 1, based on data presented by Landis and Koch (1977). This table presents classifications on a 5-level ordinal scale regarding carcinoma in situ of the uterine cervix, for seven pathologists evaluating n = 118 slides. Here I = 5 and T = 7, and there are 5^7 = 78,125 possible joint ratings patterns for the seven raters. Table 1 shows that 77 distinct patterns occurred for these 118 observations. The data can be organized in a contingency table having 5^7 cells, where each cell represents a possible rating pattern. For Table 1, only 77 cells have positive counts.

A key feature of repeated measurement data is within-subject dependence of observations. In Table 1, for instance, since each rater evaluates the same subjects (slides), the seven sample distributions of ratings must be treated as dependent rather than independent samples. These sample distributions are the first-order marginal distributions of the 5^7 contingency table.

At occasion g, let $\phi_h(g)$ denote the probability that a subject makes response h. The probabilities $\{\phi_h(g), b = 1,...,I\}$ form the gth first-order marginal distribution of the response. There is marginal homogeneity, which we denote by MH, if

$$\phi_h(1) = \phi_h(2) = ... = \phi_h(T), \text{ for } h = 1,...,I. \qquad (1.1)$$

The hypothesis of marginal homogeneity states that the T first-order marginal distributions of the response are identical. This article discusses ways of comparing marginal distributions for large, sparse contingency tables. For instance, for Table 1, we will see how to test whether the seven pathologists have identical response distributions.

Madansky (1963) gave a likelihood-ratio test of MH. It assumes a multinomial likelihood for the I^T cells, and it compares the likelihood maximized subject to constraint (1.1) to the likelihood maximized in the unrestricted case. Lipsitz (1988) showed how to conduct this test using standard software such as SAS. Another likelihood-based approach tests MH in the context of the quasi-symmetry model. It tests the hypothesis that the quasi-symmetry model holds with MH (i.e., that there is symmetry) against the alternative that quasi symmetry holds without MH, by comparing the maximized likelihoods for the symmetry and quasi-symmetry models. See Darroch (1981) and Agresti (1990, Sec. 11.2) for details on these and other methods for testing MH.

Even in this age of computers, these likelihood-ratio tests are infeasible when I and T are moderately large, because of the huge number of cells and the extreme sparseness of the table. In Table 1, for instance, many sums of cell counts that are sufficient statistics for the symmetry and quasi-symmetry models equal zero, and regular ML estimates do not exist for these models. Madansky's ML test must maximize a multinomial likelihood defined over the 78,125 cells, subject to constraints for the first-order marginal distributions. Methodology for doing this has been available for some time (Aitchison and Silvey 1958), but published examples of such analyses (e.g., Haber 1985) have dealt only with small tables.

This article describes simple strategies for comparing marginal distributions of large, sparse contingency tables. We test MH in the context of a model for the marginal distributions, such that MH is a special case of the model. This leads naturally to post-test description and inference regarding the nature of the marginal heterogeneity. Also, models can be generalized to incorporate explanatory variables, so that effects of those variables can also be analysed or so one can make adjusted comparisons of marginal distributions. For instance, one might want to analyze whether changes across occasions in marginal distributions differ according to gender, age, or treatment. It is often sensible to use a directed alternative to MH corresponding to a parsimonious model, so that the test statistic has fewer degrees of freedom, and hence potentially greater power. This is particularly true when the response is ordinal. Koch et al. (1977) and Agresti (1989) described ways of modeling marginal distributions, and we consider some models for Table 1 in Sections 5 and 6.

The simplicity of our approach results from using ML to estimate model parameters under the naive assumption that the repeated responses are independent. For Table 1, in treating the 7 marginal distributions for the 118 observations with 5-category response as independent, we apply standard ML methods to 7×118 = 826 observations in cells of a 7×5 table. Problems of sparseness and complex computations then disappear. We use the jackknife technique to obtain an appropriate estimated covariance matrix of the estimates. To test MH, we conduct a Wald test using these estimates, or

simply modify the covariance structure in score statistics for comparison of independent multinomial distributions.

Sections 2-4 present the strategies for comparing marginal distributions. We illustrate their use for ordinal classifications in Section 5, and apply them in Section 6 to compare the marginal distributions of Table 1. Section 7 gives results of a simulation study that suggests the naive estimates are surprisingly efficient. Section 8 briefly describes use of the methods for nominal classifications, and Section 9 describes complications resulting from missing data.

2.Pseudo ML Estimation Assuming Independent Multinomials

For large, sparse tables, one can easily fit models for the T first-order marginal distributions by treating sample counts from different margins as statistically independent. Liang and Zeger (1986) used this naive approach for univariate longitudinal data problems. Consider the T×I table consisting of the T sample marginal distributions. A single observation in the original I^T table is replaced by T observations in this T×I table. One obtains parameter estimates by using ML to fit the model to this table, treating the rows as having independent multinomial distributions. The resulting estimates are not truly ML, since those distributions are not truly independent and the function maximized is not the true likelihood. But, the consistency of the sample estimators of the marginal probabilities implies that these "pseudo ML" estimators are consistent, assuming that the model holds. The estimated covariance matrix obtained by treating the margins as independent is not consistent for the true covariance matrix of the estimators, however.

For tables that are too large for ordinary ML methods, we recommend estimating model parameters using the pseudo ML estimates and estimating the covariance matrix of those estimators using the jackknife method. This involves re-fitting the model repeatedly, each time deleting one observation (which corresponds to T observations in the TxI table). Results of Lipsitz et al. (1990a) suggest that for each re-fit of the model, it is preferable to use a one-step jackknife. This uses only the first step of the iterative process for fitting the model, with the pseudo ML estimates as the initial estimates.

White (1982) gave the true asymptotic covariance matrix for ML estimators in models with misspecified likelihood. The one-step jackknife estimator is asymptotically equivalent to an estimator White proposed of that matrix, but is simpler to compute for many models. We outline the reasons for the asymptotic equivalence in the Appendix.

For a model having parameter vector $\underset{\sim}{\beta}$, denote the pseudo ML estimator by $\hat{\underset{\sim}{\beta}}$ and denote the estimator when the jth observation is deleted by $\hat{\underset{\sim}{\beta}}_{-j}$. One form of the jackknife estimator of the covariance matrix of $\hat{\underset{\sim}{\beta}}$ is

$$\sum_j \left(\hat{\underset{\sim}{\beta}}_{-j} - \hat{\underset{\sim}{\beta}}\right)\left(\hat{\underset{\sim}{\beta}}_{-j} - \hat{\underset{\sim}{\beta}}\right)'.$$

We programmed calculation of the pseudo ML estimator and its jackknife estimated covariance matrix for models discussed in this article using IML in SAS (see Table 5).

3. Tests of Marginal Homogeneity

After obtaining model parameter estimates and an estimated covariance matrix, one can apply standard methods of inference. For instance, one can test MH using a Wald test for uniformity of certain parameters across the T occasions. The form of the Wald statistic is $\underset{\sim}{d}'\,[\mathrm{Cov}(\underset{\sim}{d})]^{-1}\underset{\sim}{d}$, where $\underset{\sim}{d}$ is a vector of differences of estimates across occasions.

Alternatively, one can formulate simple test statistics for MH by adapting score statistics for this hypothesis. For the model chosen to reflect possible departures from MH, one obtains the Fisher efficient score vector based on the pseudo likelihood that treats the T marginal distributions as independent. One then estimates the covariance matrix of the efficient score vector using the dependence structure across occasions implied by a multinomial assumption for the I^T table. The test statistic is a quadratic form comparing the efficient score vector to its null expected value, weighted by the inverse estimated covariance matrix. The pseudo score test approach is applicable when the overall sample size is large enough that the score vector is approximately normally distributed, so that the quadratic form has an asymptotic chi-squared distribution.

4. Weighted Least Squares Model-Fitting

Another approach uses weighted least squares (WLS) methodology (Koch et al. 1977). This is also more amenable than standard ML for fitting models to margins of large, sparse contingency tables. We now give some attention to WLS, because it can be more readily implemented with SAS (using CATMOD) than other methods.

In modeling first-order marginal functions, WLS methods require only the second-order marginal tables to estimate the asymptotic covariance structure of those response functions. The second-order marginal counts must be sufficiently large that the sample response functions are approximately normally distributed and their estimated covariance matrix is non-singular. In practice, this usually requires the first-order marginal counts to nearly all exceed about 5. Koch et al. (1977), Landis et al. (1988), and Agresti (1989) gave examples of the use of WLS for analyzing repeated categorical data.

When the model holds, WLS is asymptotically equivalent to ML for the full I^T table. However, pseudo ML methods have the advantage of being applicable in cases when the data are too sparse to support WLS. In particular, unlike WLS, pseudo ML methods apply when there are continuous explanatory variables. Also, the example in Section 6 shows that WLS can be unreliable and highly sensitive to slight changes in the data when some marginal counts are small.

5. Marginal Comparisons of Ordinal Classifications

To illustrate methods for comparing marginal distributions, we discuss a general class of models,

$$\mathrm{Link}_j(g) = \alpha_j - \mu_g,\ j=1,\ldots,I-1,\ g=1,\ldots,T, \qquad (5.1)$$

for ordinal response variables. Two important special cases are (1) the cumulative logit model, whereby

$$\mathrm{Link}_j(g) = \mathrm{logit}[\gamma_j(g)], \qquad (5.2)$$

with $\gamma_j(g) = \phi_1(g) + \ldots + \phi_j(g)$, and (2) the adjacent-categories logit model, whereby

$$\mathrm{Link}_j(g) = \log[\phi_j(g)/\phi_{j+1}(g)]. \qquad (5.3)$$

Models for these logits are easy to fit using CATMOD (with RESPONSE CLOGITS or RESPONSE ALOGITS options), and the cumulative logit is also an option in procedure LOGISTIC.

Model (5.1) implies that the margins are location shifts on some scale, and uses I-1 parameters to describe

marginal heterogeneity. For this model, MH corresponds to $\mu_1 = ... = \mu_T$. Using results from our pseudo ML approach, with jackknife estimated covariance matrix for estimates of $\{\mu_i\}$, we can test MH using a Wald test. The chi-squared asymptotic distribution has df = T-1, rather than df = (T-1)(I-1) as in the most general (unstructured) tests.

For many versions of model (5.1), pseudo score statistics are simple alternatives for testing MH. Let n_{tj} denote the number of subjects who make response j at occasion t. Assuming no missing data, let $n = n_{t+} = \Sigma_j n_{tj}$. For a set of monotone response scores $\{v_j\}$ for the ordinal response scale, let

$$M_t = \Sigma_j v_j n_{tj}/n, \; t = 1,...,T$$

and

$$M = \Sigma_j v_j n_{+j}/nT$$

Then M_t is the "mean" response at occasion t for the response scores $\{v_j\}$, and $M = \Sigma_t M_t/T$ is the mean response for the sum of the single-factor response distributions. For model (5.3) with rows in the TxI table treated as independent multinomials, the efficient score vector for testing MH has components $n(M_t - M)$ with $\{v_j = j\}$. To test MH, we use a statistic that is a quadratic form describing variation among these means, or equivalently variation from 0 among $\{d_t = [(M_t - M) - (M_T - M)] = M_t - M_T, \; t = 1,...,T-1\}$. Such a quadratic form must utilize the true dependence structure in estimating covariances among $\{d_t\}$.

Let $p_h(t) = n_{th}/n$, and let $p_{hi}(tu)$ denote the proportion of subjects making response h at occasion t and response i at occasion u. Let $\underline{d} = (d_1,...,d_{T-1})'$, and for all (j,k) with $j < T$, $k < T$, let $\underline{S}$ denote the matrix having elements

$$s_{jk} = \Sigma_h \Sigma_i \, v_h v_i \, [p_{hi}(jk) - p_{hi}(jT) - p_{hi}(kT) + p_h(T)\delta_{hi}]$$

$$-(M_j - M_T)(M_k - M_T)$$

Then s_{jk}/n is the unrestricted ML estimate of $Cov(d_j, d_k)$, and $n\underline{d}'\underline{S}^{-1}\underline{d}$ is a pseudo score statistic for testing MH using model (5.3), based on df = T-1. When we let the $\{v_j\}$ be ridit scores for margin $\{n_{+j}, \; j = 1,...,I\}$, this is a pseudo score statistic for the cumulative logit model (5.2).

The pseudo score statistic just described is also the WLS goodness-of-fit statistic for testing the null mean response model

$$E(M_t) = \alpha, \; t = 1,...,T$$

assuming multinomial sampling for the I^T table. For T=2, Bhapkar (1970) proposed tests of this form. One can use CATMOD to compute this statistic, as well as to do standard WLS fits of models (5.2) and (5.3) for this multinomial sampling model, at least when the data are not too sparse.

6. Marginal Comparison of Carcinoma Ratings

Table 1 is highly sparse, with 118 observations in 78,125 cells. The first-order marginal counts are much less sparse, varying between 1 and 69, with 23 of the 35 counts exceeding 10. We first tested MH with Wald statistics for model (5.1), using the jackknife to estimate the covariance matrix of pseudo ML estimators. The Wald statistic equals 113.6 for the cumulative logit case and 57.5 for the adjacent-categories logit case. Both statistics are based on df = 6, and give very strong evidence against MH.

Pseudo score statistics for model (5.1) also give strong evidence of marginal heterogeneity. For instance, the version with $\{v_j = j\}$ gives a chi-squared statistic of 161.4, also based on df = 6.

Table 2 gives pseudo ML and jackknifed pseudo ML model parameter estimates for the cumulative logit model, as well as estimated standard errors. The standard errors we report for the pseudo ML estimates are the ones that treat the samples as independent, and are incorrect. The ones for the jackknife recognize the dependence, and hold for the pseudo estimates as well. We included the incorrect ones to show how one can drastically overestimate variability in describing within-subject effects by naively treating the samples as independent.

In addition, we used WLS to fit the cumulative logit model to margins of the table, directly incorporating estimates of dependence from a multinomial structure for the full 5^7 table (We added a count of .001 to cell (4,4,4,4,4,4,5) to obtain a nonsingular covariance matrix). The reliability of these estimates is questionable, since two of the marginal counts equal 1 and two equal 2. The WLS Wald statistic for testing $\mu_1 = ... = \mu_7$ equals 85.8, based on df = 6. The WLS fit has residual chi-squared equal to 98.4, based on df = 18. The model does not fit well, but it detects enough of the departure from MH to also give a very small P-value.

Table 2 also contains the WLS model parameter estimates, as well as pseudo WLS estimates based on treating the samples as independent. The pseudo WLS estimates are similar to the pseudo ML estimates. The standard WLS estimates differ from the others, and have smallest estimated standard errors. A sensitivity analysis revealed that these WLS estimates are unreliable, because of the very small marginal counts. For instance, all raters made rating 4 more often than rating 5 except for rater F, who made rating 4 only once and rating 5 four times. If we change the observation (4,3,3,3,3,5,3) to (4,3,3,3,3,4,3), thus increasing rater F's marginal count for rating 4 from 1 to 2, the WLS Wald test statistic for MH drops from 85.8 to 32.7, and the third column of estimates in Table 2 changes dramatically from (0.37, 0.30, 0.02, -0.38, 0.34, -0.67, 0.02) to (0.28, -0.04, -0.02, -0.27, 0.12, -0.05, -0.02). By contrast the pseudo WLS estimates hardly change at all, the largest change being the first, which changes from 0.520 to 0.528.

For only 8 of the 118 slides did any raters use rating 5, and all ten of the marginal counts that are less than 10 refer to rating 4 or 5. Thus, combining these rating categories should improve the reliability of methods that are highly susceptible to sparseness. When we combined categories 4 and 5, all marginal counts equal at least 5. The WLS model estimates then changed to (0.43, 0.44, -0.19, -0.37, 0.41, -0.84, -0.12), which are much closer to the pseudo WLS and ML estimates in Table 2. In this case, the WLS estimated standard errors also increased to levels close to those reported for the jackknife. Combining columns had trivial results on the other approaches. For instance, the pseudo WLS estimates all changed by less than 0.02 (reflecting the property of invariance to scale collapsings that McCullagh (1980) gave as an important quality of this model), and their naive estimated standard errors all changed by less than 0.001.

Table 3 shows the use of CATMOD for some of these analyses. The first use of it fits the cumulative logit model and conducts the WLS test of MH for it, using the multinomial dependence structure; the second use gives the pseudo score test of MH that is the goodness-of-fit test of the null mean response model, using response scores {1,2,3,4,5}. The design matrix for the cumulative logit model corresponds

to ten parameters, the cutpoint parameters $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ and the marginal effect parameters $\{\mu_1,...,\mu_6\}$, with $\mu_7 = -(\mu_1 + ... + \mu_6)$ determined by the constraint $\Sigma \mu_j = 0$.

7. Efficiency of Pseudo ML and WLS Estimates

It is important to consider whether the pseudo estimates are much less efficient than the ordinary estimates. When responses are strongly correlated across occasions, one would expect that a pseudo ML estimator might have larger mean squared error (MSE) than an ordinary ML estimator, since the pseudo estimator ignores the dependence. However, we performed a small-scale simulation study that gave promising results for the relative efficiency of pseudo estimators. For marginal comparisons using the adjacent-categories logit model, there was no reduction in precision using pseudo estimators.

Because of the extremely time-consuming nature of the ordinary ML estimation process, we limited our investigation to T = 2 occasions and I = 3 categories. We generated independent samples of size n from multinomial distributions defined over the 3x3 table. The marginal probabilities satisfied model (5.3) with $\{\pi_{i+} = 1/3\}$ and with $\{\pi_{+j}\}$ determined by the model, for a fixed value of $\mu_d = \mu_2 - \mu_1$. The cell probabilities in the table were those of an underlying bivariate normal distribution having the given marginal probabilities. Eight combinations of n, correlation ρ, and μ_d were chosen: n = 20 and 50, ρ = 0.2 and 0.8, and μ_d = 0.0 (marginal homogeneity) and 0.4.

The algorithm for calculating constrained ML estimators used techniques developed by Aitchison and Silvey (1958) and Haber (1985). For generated tables in which at least one estimate did not exist, we added 0.00001 to each cell count, which always resulted in existence. Table 4 reports the square root of the MSE estimates for four estimators (ML, WLS, pseudo ML, pseudo WLS), based on 1000 simulations at each setting of (n, ρ, μ_d). With probability .95, for the n = 20 cases, the root MSE estimates are good to within about 0.020 when ρ = .2 and 0.015 when ρ = .8; for the n = 50 cases, they are good to within about 0.012 and 0.009, respectively.

Table 4 shows that, to the degree of accuracy obtained in this simulation study, the four estimators performed equally well. Surprisingly, the pseudo estimates performed adequately even when ρ was large. The WLS estimates performed as well as the ML estimates, though the marginal counts were not small enough to cause the sorts of problems WLS estimates can have with sparse data.

8. Marginal Comparisons of Nominal Classifications

The pseudo ML fitting procedure for models for nominal classifications proceeds in a similar way. For instance, suppose we want to fit a multinomial logit model that has additive occasion and treatment effects as explanatory variables. The pseudo ML estimates, which treat the occasions as independent, are identical to the regular ML estimates for the loglinear no three-factor interaction model fitted to the treatment × occasion × response table.

Suppose we want to construct a pseudo score statistic to test MH for a nominal classification. When there are no covariates, we consider the saturated loglinear model for the TxI table $\{n_{ij}\}$. The components of the efficient score vector for testing MH (i.e., independence for the TxI table) are $\{U_{ij} = n_{ij} - n_{+j}/T, i = 1,...,T-1, j = 1,...,I-1\}$. Note that $\Sigma_i U_{ij} = \Sigma_j U_{ij} = 0$. Let $d_{ij} = U_{ij} - U_{Tj} = n_{ij} - n_{Tj}$, i = 1,...,T-1, j = 1,...,I-1. Then, MH is equivalent to $E(d_{ij}) = 0$ all i and j, and a pseudo score statistic is given by a quadratic form in the vector of (T-1)(I-1) $\{d_{ij}\}$ and their estimated covariances.

One can conduct a WLS test of MH based on the unrestricted ML estimators of marginal probabilities (i.e., the sample marginal proportions) and the unrestricted ML estimator of the covariance matrix of differences of those estimators. See Bhapkar (1973) and Darroch (1981). But this is precisely the same as the pseudo score test just described. That is, the pseudo score test is the WLS goodness-of-fit test of the model of MH for the I^T contingency table, having df = (T-1)(I-1). It can be implemented with CATMOD. The third use of CATMOD in Table 4 implements this test for Table 1, yielding a chi-squared statistic of 303.4 based on df = 24.

9. Missing Data Issues

Although a goal of longitudinal studies is normally to collect data on every subject in the sample at each time of follow-up, it often happens that some subjects are not observed at all occasions. In this case, ML estimates and ML score tests are consistent when data are "missing at random" (Rubin 1976), meaning that the missing data process depends on the observed responses. All other estimators and test statistics discussed in this article require the data to be missing completely at random (Rubin, 1976), which is a stronger assumption, meaning that the missing data process cannot depend on the observed responses.

To be consistent under the appropriate missing data conditions, however, ML also requires the correct specification of the I^T joint multinomial distribution, whereas "pseudo ML" requires only the correct specification of the T marginal distributions. Thus, ML is consistent under weaker missing data conditions and pseudo ML is consistent under weaker conditions about the joint distribution of the responses over time.

Assuming the appropriate missing data conditions hold, the estimates, standard errors, and test statistics discussed change minimally with missing data. The ML estimates can be obtained using either the EM algorithm (Dempster, et. al. 1977) or the Newton-Raphson algorithm (Hocking and Oxspring, 1971) and the asymptotic variance is consistently estimated by the inverse of the observed information. When the data are missing at random, the expected information can only be obtained if we are also willing to specify the missing data process. Fortunately, one need not specify the missing data process to estimate the variance of the ML estimate when the data are missing at random, since the observed information will converge in probability to its expectation over this missing data process and the I^T multinomial distribution.

When using pseudo ML estimates and score statistics with missing data, the rows of the TxI contingency table are still treated as independent, but a row sum will not be identically n, and instead will satisfy $n_{t+} \leq n$. Then, when performing the jackknife to estimate the variance of the pseudo ML estimate, we delete each subject as before (i.e., for subject i, we delete T_i responses, where $T_i \leq T$). In the pseudo score tests, a modification that gives consistent results when data are missing completely at random is

$$M_t = (\Sigma_j v_j n_{tj})/n_{t+},$$

and, when calculating s_{jk},

$$p_h(t) = n_{th}/n_{t+} \text{ and } p_{hi}(tu) = n_{tuhi}/n_{tu++},$$

where n_{tuhi} is the number of subject who have repouse h at occasion t and response i at occasion u.

In modifying WLS with missing data, two-step methods have been proposed by Koch et. al. (1972), Woolson

and Clarke (1984), Landis et al. (1988), and Lipsitz et al. (1990b). The first steps of the approaches are different methods of estimating the probabilities in the TxI table as well as the covariance matrix of these estimates. The second steps of the methods are identical: perform weighted least squares on the estimates from step one to estimate the parameters under the appropriate model. In particular, Koch et. al. (1972) further stratified individuals by their pattern of non-response, and then used weighted least squares to estimate the TxI probabilities. Woolson and Clarke (1984) estimated the marginal probability of response h at occasion t by the proportion of individuals with response h among those who respond at that occasion. To estimate the variance, they proposed an $(I+1)^T$ multinomial distribution, adding one response category at each time point that corresponds to missing. Lipsitz et. al. (1990b) estimated the TxI probabilities using the EM algorithm with the underlying I^T joint multinomial distribution. The Lipsitz et. al. method is consistent when the data are missing at random, whereas the other two require data to be missing completely at random.

10. Discussion

There are yet other ways of comparing marginal distributions and estimating covariance matrices that we have not discussed in this article. For instance, an alternative to the jackknife for estimating the covariance matrix is to adapt an empirical estimator described by Liang and Zeger (1986). Results in Lipsitz et al. (1990a) suggest this is also asymptotically equivalent to using the jackknife. Stram et al. (1988) presented an alternative strategy of estimating a separate set of parameters at each occasion, and then empirically estimating the joint covariance matrix of estimated parameters from different occasions. They then used standard methods such as Wald tests to compare parameters across occasions. This is a special case of the Liang and Zeger approach using the naive "independence" estimates, if one fits a model in which the sets of parameters for different occasions are completely separate. Another approach is to estimate parameters using some assumed structure for the covariance matrix of the sample marginal responses. When we use the covariance structure induced by assuming a multinomial distribution over the full I^T table, this simplifies to the ordinary WLS approach.

In future research, it would be useful to compare various ways of obtaining pseudo estimates. We believe that the simple estimates based on treating occasions as independent will be adequate for most purposes. It is also important in future work to compare ways of estimating the covariance matrix of pseudo estimates.

Finally, we note that when the main focus is purely testing of MH, another strategy uses generalizations of the Cochran-Mantel-Haenszel test. For details, see Agresti (1990, Sections 7.4 and 8.4), Darroch (1981), White et al. (1982), and Landis et al. (1988).

APPENDIX: Asymptotic equivalence of jackknife estimator and White's estimator of covariance matrix .

For a model having parameter vector $\underset{\sim}{\beta}$, the pseudo ML estimate $\hat{\underset{\sim}{\beta}}$ is obtained by setting

$$\underset{\sim}{u}(\hat{\underset{\sim}{\beta}}) = \Sigma\, \underset{\sim}{u}_i(\hat{\underset{\sim}{\beta}}) = \underset{\sim}{0} \qquad (A.1)$$

and solving for $\hat{\underset{\sim}{\beta}}$, where $\underset{\sim}{u}_i(\hat{\underset{\sim}{\beta}})$ denotes the contribution to the score vector (the derivative of the log likelihood with respect to $\hat{\underset{\sim}{\beta}}$) from subject i. Given $\hat{\underset{\sim}{\beta}}$ and deleting the jth subject, the first step of the Newton-Raphson algorithm produces

$$\hat{\underset{\sim}{\beta}}_{-j} = \hat{\underset{\sim}{\beta}} + \left[\sum_{i\neq j} I_i(\hat{\beta})\right]^{-1}\left[\sum_{i\neq j} \underset{\sim}{u}_i(\hat{\underset{\sim}{\beta}})\right] \qquad (A.2)$$

where $\Sigma\, I_i(\underset{\sim}{\beta}) = -\Sigma\, \partial \underset{\sim}{u}_i(\underset{\sim}{\beta})/\partial\underset{\sim}{\beta}$ is the information matrix. From (A.1),

$$\sum_{i\neq j} \underset{\sim}{u}_i(\hat{\underset{\sim}{\beta}}) = -\underset{\sim}{u}_j(\hat{\underset{\sim}{\beta}})$$

so that (A.2) becomes

$$\hat{\underset{\sim}{\beta}}_{-j} - \hat{\underset{\sim}{\beta}} = -\left[\sum_{i\neq j} I_i(\hat{\underset{\sim}{\beta}})\right]^{-1}\left[\underset{\sim}{u}_j(\hat{\underset{\sim}{\beta}})\right]$$

One form of the jackknife estimator of the covariance matrix of $\hat{\underset{\sim}{\beta}}$ is

$$\sum_j (\hat{\underset{\sim}{\beta}}_{-j} - \hat{\underset{\sim}{\beta}})(\hat{\underset{\sim}{\beta}}_{-j} - \hat{\underset{\sim}{\beta}})' \qquad (A.3)$$

$$= \sum_j \left\{\left[\sum_{i\neq j} I_i(\hat{\underset{\sim}{\beta}})\right]^{-1}\left[\underset{\sim}{u}_j(\hat{\underset{\sim}{\beta}})\ \underset{\sim}{u}_j(\hat{\underset{\sim}{\beta}})'\right]\left[\sum_{i\neq j} I_i(\hat{\underset{\sim}{\beta}})\right]^{-1}\right\}$$

Under regularity conditions needed for $\hat{\underset{\sim}{\beta}}$ to be consistent, this is asymptotically equivalent to

$$\left[\sum_j I_i(\hat{\underset{\sim}{\beta}})\right]^{-1}\left[\sum_j \underset{\sim}{u}_j(\hat{\underset{\sim}{\beta}})\underset{\sim}{u}_j(\hat{\underset{\sim}{\beta}})'\right]\left[\sum_i I_i(\hat{\underset{\sim}{\beta}})\right]^{-1} \qquad (A.4)$$

which is the estimator proposed by White. The asymptotic equivalence refers to the sample size times each of these estimators converging to the true asymptotic covariance matrix of $\sqrt{n}\,(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})$. The usual estimator of the covariance matrix, the inverse information $[\Sigma\, I_i(\hat{\underset{\sim}{\beta}})]^{-1}$, is asymptotically equivalent to (A.3) and (A.4) under the additional assumption that

$$E\left[I_i(\underset{\sim}{\beta})\right] = E[\underset{\sim}{u}_i(\hat{\underset{\sim}{\beta}})\underset{\sim}{u}_i(\underset{\sim}{\beta})'],$$

where the expectation is taken with respect to the true distribution (i.e., not the naive "independence" distribution) of the data.

Table 1. Cross-Classification of Seven Pathologists on Five Categories

Pathologist Ratings							Count	Pathologist Ratings							Count	Pathologist Ratings							Count
A	B	C	D	E	F	G		A	B	C	D	E	F	G		A	B	C	D	E	F	G	
1	1	1	1	1	1	1	10	2	3	2	2	2	1	2	1	3	3	3	4	3	2	3	1
1	1	1	1	2	1	1	8	2	3	2	2	3	1	3	1	3	3	3	4	3	2	4	1
1	1	2	1	1	1	1	2	2	3	2	2	3	2	2	1	4	2	3	2	3	2	3	1
1	1	2	1	2	1	1	2	2	3	2	2	3	2	3	2	4	3	1	1	2	1	2	1
1	2	1	1	1	1	1	1	2	3	2	2	4	1	2	1	4	3	1	3	3	2	3	1
1	2	2	1	2	1	2	1	2	3	2	2	4	1	3	1	4	3	3	2	3	2	3	1
1	3	2	1	2	1	1	1	3	2	2	2	2	1	1	1	4	3	3	3	3	2	3	2
1	3	2	2	2	1	2	1	3	2	2	2	2	1	2	1	4	3	3	3	3	3	3	3
2	1	1	1	2	1	1	1	3	3	2	1	3	2	2	1	4	3	3	3	3	5	3	1
2	1	1	2	1	1	1	1	3	3	2	2	2	1	2	1	4	3	3	4	3	3	3	2
2	1	2	1	1	1	1	1	3	3	2	2	2	2	3	1	4	3	3	4	4	3	3	1
2	1	2	1	2	1	1	1	3	3	2	2	3	1	3	4	4	3	4	2	3	3	3	1
2	1	2	2	2	1	2	1	3	3	2	2	3	2	2	1	4	3	4	2	4	1	3	1
2	2	1	1	2	1	1	1	3	3	2	2	3	2	3	2	4	4	3	2	4	1	3	2
2	2	1	1	2	1	2	1	3	3	2	2	3	3	3	1	4	4	3	3	4	3	3	1
2	2	1	2	2	1	2	1	3	3	2	2	4	2	3	1	4	4	3	4	4	3	4	1
2	2	2	1	1	1	2	1	3	3	2	3	2	2	3	1	4	4	4	2	4	3	3	1
2	2	2	1	2	2	2	1	3	3	2	3	3	1	3	1	4	4	4	2	5	1	3	1
2	2	2	2	3	1	2	2	3	3	2	3	3	3	3	2	4	4	4	3	3	3	3	1
2	3	1	1	2	1	1	1	3	3	3	2	3	1	3	2	5	3	3	2	3	2	3	1
2	3	1	1	2	1	2	1	3	3	3	2	3	2	3	3	5	3	3	3	4	1	3	1
2	3	1	1	3	1	1	1	3	3	3	2	3	3	3	1	5	3	4	2	3	4	3	1
2	3	1	2	3	1	3	1	3	3	3	2	4	2	3	2	5	5	1	4	5	5	4	1
2	3	2	1	3	2	2	1	3	3	3	2	4	3	3	1	5	5	5	4	5	5	5	1
2	3	2	2	2	1	3	1	3	3	3	3	3	2	3	5	5	5	5	5	5	5	5	1
2	3	2	2	2	2	2	1	3	3	3	3	3	3	3	4								

Table 2. Parameter Estimates for Cumulative Logit Model

Rater	Pseudo ML	Pseudo WLS	Dependent WLS	Jackknife
A	0.58 (.165)	0.52 (.155)	0.37 (.066)	0.58 (.088)
B	0.51 (.158)	0.49 (.165)	0.30 (.066)	0.51 (.087)
C	-0.19 (.152)	-0.19 (.156)	0.02 (.040)	-0.19 (.087)
D	-0.51 (.153)	-0.47 (.157)	-0.38 (.073)	-0.51 (.083)
E	0.62 (.156)	0.55 (.158)	0.34 (.075)	0.62 (.088)
F	-1.13 (.164)	-1.06 (.164)	-0.67 (.098)	-1.13 (.128)
G	0.12 (.156)	0.16 (.161)	0.02 (.039)	0.12 (.058)

Table 3. SAS Code for WLS Fitting of Cumulative Logit Model and Adjusted Score Test of MH for Table 1.

```
input a b c d e f g count @@;
cards;
1 1 1 1 1 1 1 10
1 1 1 1 2 1 1  8
...
...
5 5 5 5 5 5 5  1
4 4 4 4 4 4 5 .001
;
proc catmod; weight count;
response clogits;
model a*b*c*d*e*f*g =
(1 0 0 0 -1  0  0  0  0  0,
 0 1 0 0 -1  0  0  0  0  0,
 0 0 1 0 -1  0  0  0  0  0,
 0 0 0 1 -1  0  0  0  0  0,
 1 0 0 0  0 -1  0  0  0  0,
 0 1 0 0  0 -1  0  0  0  0,
 0 0 1 0  0 -1  0  0  0  0,
 0 0 0 1  0 -1  0  0  0  0,
 1 0 0 0  0  0 -1  0  0  0,
 0 1 0 0  0  0 -1  0  0  0,
 0 0 1 0  0  0 -1  0  0  0,
 0 0 0 1  0  0 -1  0  0  0,
 1 0 0 0  0  0  0 -1  0  0,
 0 1 0 0  0  0  0 -1  0  0,
 0 0 1 0  0  0  0 -1  0  0,
 0 0 0 1  0  0  0 -1  0  0,
 1 0 0 0  0  0  0  0 -1  0,
 0 1 0 0  0  0  0  0 -1  0,
 0 0 1 0  0  0  0  0 -1  0,
 0 0 0 1  0  0  0  0 -1  0,
 1 0 0 0  0  0  0  0  0 -1,
 0 1 0 0  0  0  0  0  0 -1,
 0 0 1 0  0  0  0  0  0 -1,
 0 0 0 1  0  0  0  0  0 -1,
 1 0 0 0  1  1  1  1  1  1,
 0 1 0 0  1  1  1  1  1  1,
 0 0 1 0  1  1  1  1  1  1,
 0 0 0 1  1  1  1  1  1  1)
(1 2 3 4 = 'cutpoints',
 5 6 7 8 9 10 = 'homo');

proc catmod; weight count;
response means;
model a*b*c*d*e*f*g = (1,1,1,1,1,1,1);

proc catmod; weight count;
response marginals;
model a*b*c*d*e*f*g = _response_;
repeated raters 7;
```

Table 4. Estimated Root Mean Squared Error for Logit Comparison of Margins Based on Underlying Normal Distribution

	n = 20				n = 50			
	$\mu_d = 0$		$\mu_d = .4$		$\mu_d = 0$		$\mu_d = .4$	
Estimator	ρ=.2	ρ=.8	ρ=.2	ρ=.8	ρ=.2	ρ=.8	ρ=.2	ρ=.8
ML	.384	.251	.410	.275	.241	.143	.246	.159
WLS	.372	.233	.400	.272	.239	.140	.243	.157
P-ML	.381	.233	.413	.277	.238	.140	.242	.157
P-WLS	.371	.229	.395	.272	.237	.139	.240	.155

Note: P-ML denotes pseudo ML, P-WLS denotes pseudo WLS

Table 5. Using SAS to Obtain Jackknife Estimated Covariance Matrix

```
DATA PATHOL;
INPUT A B C D E F G COUNT;
cards;
1 1 1 1 1 1 1 10
1 1 1 1 2 1 1  8
...
...
5 5 5 4 5 5 5  1
5 5 5 5 5 5 5  1
;
run;
/* sample size of data */
proc means noprint;
  var count;
  output out=total(drop=_type_ _freq_)
  sum = totals;
run;
%macro cumlog;
DATA PATHOL2(DROP=A B C D E F G );
 set pathol;
Y=A; TIME=1; OUTPUT;
Y=B; TIME=2; OUTPUT;
Y=C; TIME=3; OUTPUT;
Y=D; TIME=4; OUTPUT;
Y=E; TIME=5; OUTPUT;
Y=F; TIME=6; OUTPUT;
Y=G; TIME=7; OUTPUT;
RUN;
PROC SORT; BY TIME Y;
/* summary data (path2) with marg. counts */
/* separate record for each time y count */
proc means noprint; by time y;
     var count;
     output out=path2(drop=_type_ _freq_ )
     sum = counts;
run;
data two;
 do time=1 to 7; do y=1 to 5;
      output;
  end; end;
run;
PROC SORT data=two; BY TIME Y;
 run;
/* data 'three' like 'path2' except 'path2' */
/* will not have sepsrate record when marg. */
/* count is 0, and 'three' will  */
data three;
  merge two path2; by time y;
  if counts=  then counts=0;
  if y=5 then delete;
run;
/* IML to calc. pseudo-MLEs for cum. logit */
PROC IML WORKSIZE=300;
reset noprint nolog;
   USE THREE;
   READ ALL INTO X;
   USE TOTAL;
   READ ALL INTO N;
TIME=X[,1];
Y=X[,2];
COUNTS=X[,3];
xa=designf(time)||design(y);
nbeta=ncol(xa);
a=$1 0 0 0,
   1 1 0 0,
   1 1 1 0,
   1 1 1 1;
ainv=inv(a);
i7=i(7);
ainv=i7@ainv;
J4=J(4,4,1);
  beta = j(6,1,0)//$.05,.075,.1,.125;
   CRIT=$1;
  DO IT=1 TO 12 WHILE(CRIT > .0000001);
   U= J(NBETA,$1,$0);
   DVD= J(NBETA,NBETA,$0);
      PHAT0= EXP(XA*BETA)/(1+ EXP(XA*BETA) );
      V0=diag( DIAG(PHAT0)-PHAT0*PHAT0`);
    PHAT=AINV*PHAT0;
    V= DIAG(PHAT)-PHAT*PHAT`;
    V=(I7@J4)#V;
    DT=XA`*V0*AINV`;
   U=DT*INV(V)*(COUNTS-N#PHAT);
   DVD=N#DT*INV(V)*DT`;
   DELTA= SOLVE(DVD,U);
   BETA=BETA+DELTA;
   CRIT= MAX( ABS(DELTA));
   END;
%mend;
%cumlog;
     vb=inv(dvd); *var. matrix under indep.;
     sebeta=sqrt(vecdiag(vb)); *vector of
       estimated standard errors of beta;
    z=beta/sebeta;          *z-statistics;
    zsq=z#z;
    p=1-probchi(zsq,1);  *two-sided p-value;
variable= $
"time1" "time2" "time3" "time4" "time5" "time6"
"resp1" "resp2" "resp3" "resp4" ;
variable = variable`;
print, $ 'cumulative logit';
print, $ 'STANDARD ERRORS UNDER INDEPENDENCE' ;
     print, variable beta sebeta z p;
contrast= (i(6))||(j(6,4,0));
gsq=(contrast*beta)`*inv(contrast*vb*contrast`)
   *(contrast*beta);
df=nrow(contrast);
p=1-probchi(gsq,df);
print, $ 'WALD STAT FOR MH UNDER INDEPENDENCE';
print, gsq df p;
out=beta`;
coln = variable`;
create qml from out [colname=coln];
append from out;
close qml;
quit;
data jac.obs;
input
 time1   time2   time3   time4   time5   time6
 resp1   resp2   resp3   resp4;
cards;
;
run;
/* use macro ord to calc. est. of pseudo-MLE */
/* dropping each non-empty cell of TxI table */
/* Used to compute jacknife estimates  */
%macro ord(start,stop);
%do i=&start %to &stop;
DATA PATHOLM;

 set pathol;
IF _N_ = &I THEN COUNT=COUNT-1;
 run;
proc means noprint;
     var count;
     output out=total(drop=_type_ _freq_ )
     sum = totals;
run;
%cumlog;
out=beta`;
coln= $
"time1" "time2" "time3" "time4" "time5"
"time6" "resp1" "resp2" "resp3" "resp4" ;
create bet from out [colname=coln];
append from out;
close bet;
quit;
 proc append base=jac.obs data=bet;
 run;
%end;
%mend ;
/* now obtain summary stat's from macro ord */
/* Used to compute jacknife estimates   */
%ord(1,77);
run;
data jac.obs
 (keep=time1 time2 time3 time4 time5 time6
  resp1   resp2   resp3   resp4 count);
 merge pathol jac.obs;
 run;
proc corr nocorr csscp out=var(type=csscp)
     noprint;
 freq count;
 run;
data var(drop=_type_ _name_ );
     set var;
     if (_type_ ne 'CSSCP') then delete;
     RUN;
proc means noprint data=pathol;
     var count;
     output out=total(drop=_type_ _freq_ )
     sum = totals;
run;
```

Table 5 (continued)

```
proc iml;
 reset nolog noprint;
   USE TOTAL;
   READ ALL INTO N;
   USE QML;
   READ ALL INTO BETA;
 beta=beta';
     use var;
     read all into jackvar;
vb=((N-10/7)/N)#jackvar;
   sebeta=sqrt(vecdiag(vb)); *vector of
      estimated standard errors of beta;
     z=beta/sebeta;    *z-statistics;
     zsq=z#z;
     p=1-probchi(zsq,1); *two-sided p-value;
variable= $
"time1" "time2" "time3" "time4" "time5"
"time6" "resp1" "resp2" "resp3" "resp4" ;
variable = variable';
print, $ 'cumulative logit';
print, $ 'STANDARD ERRORS FROM JACKKNIFE' ;
     print, variable beta sebeta z p;
contrast= (i(6))||(j(6,4,0));

gsq=(contrast*beta)'
  *inv(contrast*vb*contrast')*(contrast*beta);
df=nrow(contrast);
p=1-probchi(gsq,df);
print, $ 'WALD STAT FOR MH FROM JACKKNIFE' ;
print, gsq df p;
quit;
```


ACKNOWLEDGMENTS

The work of Agresti and Lang was partially supported by NIH grant GM 43824.

REFERENCES

Agresti, A. A survey of models for repeated ordered categorical response data. Statistics in Medicine 8 (1989) 1209-1224.

Agresti, A. Categorical Data Analysis (1990) New York: Wiley.

Aitchison, J., and S. D. Silvey. Maximum likelihood estimation of parameters subject to restraints. Ann. Math. Statist. 29 (1958) 813-828.

Bhapkar, V. P. Categorical data analogs of some multivariate tests. Pp. 85-110 in Essays in Probability and Statistics, ed. by R. C. Bose et al. (1970) Chapel Hill, NC: Univ. of North Carolina Press.

Bhapkar, V. P. On the comparison of proportions in matched samples. Sankhyā A35 (1973) 341-356.

Darroch, J. N. The Mantel-Haenszel test and tests of marginal symmetry; fixed-effects and mixed models for a categorical response. Internat. Statist. Rev. 49 (1981) 285-307.

Dempster, A. P., N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B 39 (1977) 1-38.

Haber, M. Log-Linear models for correlated marginal totals of a frequency table. Commun. Statist., Theory and Methods, 14 (1985) 2845-2856.

Hocking, R. R. and H. H. Oxspring. Maximum likelihood estimation with incomplete multinomial data. J. Amer. Statist. Assoc. 63 (1971) 65-70.

Koch, G. G., P. B. Imrey, and D. W. Reinfurt. Linear model analysis of categorical data with incomplete response vectors. Biometrics 28 (1972) 663-692.

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

Landis, J. R. and G. G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33 (1977) 363-374.

Landis, J. R., M. E. Miller, C. S. Davis, and G. G. Koch. Some general methods for the analysis of categorical data in longitudinal studies. Statist. Medic. 7 (1988) 109-137.
Liang, K. Y., and S. L. Zeger. Longitudinal data analysis using generalized linear models. Biometrika 73 (1986) 13-22.

Lipsitz, S. Methods for analyzing repeated categorical outcomes. Ph.D. Dissertation, Dept. of Biostatistics (1988), Harvard Univ.

Lipsitz, S. R., N. M. Laird, and D. P. Harrington. Using the jackknife to estimate the variance of regression estimators from repeated measures studies. Commun. Statist., Theory and Methods 19 (1990a) 821-845.

Lipsitz, S. R., N. M. Laird, and D. P. Harrington. Weighted least squares analysis of repeated categorical measurements with outcomes subject to non-response. Tech. Report, Dept. of Biostatistics (1990b), Harvard Univ.

Madansky, A. Tests of homogeneity for correlated samples. J. Amer. Statist. Assoc. 58 (1963) 97-119.

McCullagh, P. Regression models for ordinal data (with discussion). J. Roy. Statist. Soc. B 42 (1980) 109-142.

Rubin, D. B. Inference and missing data. Biometrika 63 (1976) 581-592.

White, A. A., J. R. Landis, and M. M. Cooper. A note on the equivalence of several marginal homogeneity test criteria for categorical data. Internat. Statist. Rev. 50 (1982) 27-34.

White, H. Maximum likelihood estimation under misspecified models. Econometrica 50 (1982) 1-26.

Woolson, R. F., and W. R. Clarke. Analysis of categorical incomplete longitudinal data. J. Roy. Statist. Soc. A 147 (1984) 87-99.

Alan Agresti
Department of Statistics
University of Florida
Gainesville, FL 32611