Computational Statistics & Data Analysis 14 (1992) 55–73
North-Holland

Comparing marginal distributions of large, sparse contingency tables

Alan Agresti
University of Florida, Gainesville, FL 32611, USA

Stuart Lipsitz
Harvard School of Public Health, Boston, MA 02115, USA

Joseph B. Lang
University of Florida, Gainesville, FL 32611, USA

Received November 1990
Revised March 1991

Abstract: The feasibility of maximum likelihood (ML) analyses of models for marginal distributions of contingency tables diminishes as the numbers of margins and response categories increases. This article describes alternative approaches that are much more feasible. We recommend a "pseudo ML" approach that obtains model parameter estimates by treating repeated responses as independent and uses a jackknife to estimate the covariance matrix of those estimates. We test marginal homogeneity using a Wald statistic, or by adapting the efficient score statistic from the independent-samples case. We illustrate these approaches with a seven-dimensional table having 78125 cells, and we give simulation results that show no substantive loss of efficiency from using pseudo ML estimates.

Keywords: Loglinear models; Longitudinal data; Maximum likelihood; Ordinal data; Pseudo maximum likelihood; Score statistic; Weighted least squares.

1. Introduction

At each of T occasions, suppose we observe responses for n subjects on a categorical variable having I levels. A T-dimensional contingency table having I^T cells then cross-classifies the n observations. The hypothesis of *marginal homogeneity*, which we denote by MH, states that the T first-order marginal distributions of the responses are identical. At occasion g, let $\phi_h(g)$ denote the probability that a subject makes response h. There is MH if

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

Correspondence to: Alan Agresti, Department of Statistics, University of Florida, Griffin-Floyd Hall, Gainesville, FL 32611-2049, USA.

This article discusses ways of comparing marginal distributions for large, sparse contingency tables. These methods test MH in the context of a model for the marginal distributions, such that MH is a special case of the model. There are several advantages of making the test model-based. First, this leads naturally to post-test description and inference regarding the nature of the marginal heterogeneity. Second, models can be generalized to incorporate explanatory variables, so that effects of those variables can also be analyzed or so one can make adjusted comparisons of marginal distributions. Finally, it is often sensible to use a directed alternative to MH corresponding to some 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.

There is considerable literature on likelihood methods for testing MH. Madansky (1963) gave a likelihood-ratio test. It assumes a multinomial likelihood for the I^T cells for the table, and it compares the likelihood maximized subject to constraint (1.1) to the likelihood maximized in the unrestricted case. That is, Madansky's test is simply a goodness-of-fit test for the model of MH. The unrestricted alternative hypothesis corresponds to the saturated model for the marginal distributions. Firth and Treat (1988) and Lipsitz (1988) showed how to conduct this test using standard software such as GLIM and SAS. An alternative 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 , Section 11.2) for details on these and other methods for testing MH.

Even in this modern computer age, such likelihood-ratio tests are difficult to implement or even infeasible when I and T are moderately large, because of the huge number of cells and the extreme sparseness of the table. For instance, 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. The ordered response categories are: (1) Negative; (2) atypical squamous hyperplasia; (3) carcinoma in situ; (4) squamous carcinoma with early stromal invasion; (5) invasive carcinoma. The resulting contingency table has $5^7 = 78\,125$ cells. The table satisfies MH if the seven raters have identical response distributions, but it is unclear how one can test MH. 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 and directed methods based on models for marginal distributions must maximize a multinomial likelihood having 78 124 parameters, subject to certain constraints for the first-order marginal distributions. Methodology for maximizing likelihoods subject to constraints 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.

The main purpose of this article is to describe simple strategies for comparing

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	2	2	1	1	1	1	3	3	2	2	2	1	1	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								

marginal distributions of large, sparse contingency tables. We propose a method that uses ML to estimate model parameters under the naive assumption that the repeated responses are independent, but then uses a jackknife to obtain an appropriate estimated covariance matrix of the estimates. 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 $7 \times 118 = 826$ observations in cells of a 7×5 table. Problems of sparseness and awkward computations then disappear. To test MH, we conduct a Wald test using these estimates. We also show how one can test MH by modifying the covariance structure in score statistics (Rao, 1973) 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 with nominal classifications.

Though this article focuses on simple comparisons of marginal distributions, in practice it is usually important to describe the dependence of those distributions on a set of explanatory variables. Also, as we discuss in Section 9, missing data often pose a problem in longitudinal studies. We study the simpler case of no covariates and no missing data primarily to focus attention on basic strategies and to make some simple efficiency comparisons. However, when there are covariates, the feasibility of standard ML approaches becomes even more problematic.

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. This naive approach was used for univariate longitudinal data problems by Liang and Zeger (1986). Consider the $T \times I$ table consisting of the T sample marginal distributions. A single observation in the original I^T table is replaced by T observations in this $T \times I$ 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 chosen to represent the variation in the marginal distributions 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. Calculating the estimated covariance matrix involves re-fitting the model repeatedly, each time deleting one observation, and using the left-hand side of formula (2.3) in the following discussion. If K cells in the full I^T table have non-zero counts, then one need only do K re-fits of the model, weighting the results to obtain the jackknife estimated covariance matrix. Simulation results of Lipsitz et al. (1990a) for modeling repeated binary responses suggest that for each re-fit of the model, it is preferable to use a one-step jackknife rather than a fully-iterated jackknife. This procedure, outlined below, uses only the first step of the Newton–Raphson iterative process for fitting the model, with the pseudo ML estimates as the initial estimates. Fox et al. (1980) and Simonoff and Tsai (1986) discussed advantages of such "linearized" jackknife procedures for use in obtaining standard errors of estimates in general non-linear modeling. The one-step jackknife also saves considerable time in what can be a computationally intensive estimation process.

White (1982) and Gourieroux et al. (1984) gave the true asymptotic covariance matrix for the ML estimator of a model with misspecified likelihood. The one-step jackknife estimator is asymptotically equivalent to an estimator White proposed of that matrix. We next outline the reasons for this asymptotic equivalence.

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

$$\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \sum \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}) = \mathbf{0}, \tag{2.1}$$

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

$$\hat{\boldsymbol{\beta}}_{-j} = \hat{\boldsymbol{\beta}} + \left[\sum_{i \neq j} \boldsymbol{I}_i(\hat{\boldsymbol{\beta}})\right]^{-1} \left[\sum_{i \neq j} \boldsymbol{u}_i(\hat{\boldsymbol{\beta}})\right], \tag{2.2}$$

where $\sum \boldsymbol{I}_i(\boldsymbol{\beta}) = -\sum \partial \boldsymbol{u}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is the information matrix. From (2.1),

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

so that (2.2) becomes

$$\left(\hat{\boldsymbol{\beta}}_{-j} - \hat{\boldsymbol{\beta}}\right) = -\left[\sum_{i \neq j} \boldsymbol{I}_i(\hat{\boldsymbol{\beta}})\right]^{-1} \left[\boldsymbol{u}_j(\hat{\boldsymbol{\beta}})\right].$$

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

$$\sum_j \left(\hat{\boldsymbol{\beta}}_{-j} - \hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_{-j} - \hat{\boldsymbol{\beta}}\right)'$$

$$= \sum_j \left\{\left[\sum_{i \neq j} \boldsymbol{I}_i(\hat{\boldsymbol{\beta}})\right]^{-1} \left[\boldsymbol{u}_j(\hat{\boldsymbol{\beta}}) \boldsymbol{u}_j(\hat{\boldsymbol{\beta}})'\right] \left[\sum_{i \neq j} \boldsymbol{I}_i(\hat{\boldsymbol{\beta}})\right]^{-1}\right\}. \tag{2.3}$$

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

$$\left[\sum_i I_i(\hat{\boldsymbol{\beta}})\right]^{-1}\left[\sum_j \boldsymbol{u}_j(\hat{\boldsymbol{\beta}})\boldsymbol{u}_j(\hat{\boldsymbol{\beta}})'\right]\left[\sum_i I_i(\hat{\boldsymbol{\beta}})\right]^{-1}, \tag{2.4}$$

which is the estimator proposed by White. The asymptotic equivalence refers to each of these estimators converging (when multiplied by n) to the true asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$. The usual estimator of the covariance matrix, the inverse information $[\Sigma I_i(\hat{\boldsymbol{\beta}})]^{-1}$, is aymptotically equivalent to (2.3) and (2.4) under the additional assumption that

$$\mathrm{E}[I_i(\boldsymbol{\beta})] = \mathrm{E}[\boldsymbol{u}_i(\boldsymbol{\beta})\boldsymbol{u}_i(\boldsymbol{\beta})'],$$

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

The left-hand side of (2.3) is often simpler to compute than (2.4). This is particularly true for model (5.2) discussed in Section 5, for which the naive likelihood is not a simple function of model parameters.

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 chi-squared test for uniformity of certain parameters across the T occasions. For each parameter one contributes $T-1$ elements to a vector $\boldsymbol{d}$ of differences between $T-1$ of the occasions and a baseline occasion. The Wald statistic then has form $\boldsymbol{d}'\boldsymbol{V}^{-1}\boldsymbol{d}$, where $\boldsymbol{V}$ is an estimated covariance matrix of $\boldsymbol{d}$.

Alternatively, and computationally more simply, one can formulate test statistics for MH by adapting efficient score statistics for this hypothesis. The efficient score statistic is a quadratic form based on the score vector of partial derivatives of the log likelihood with respect to the parameters of interest, evaluated at the null hypothesis estimates. Cox and Hinkley (1974, Chapter 9) and Rao (1973, Section 6e) described this type of test and showed asymptotic equivalences among it and the Wald and likelihood-ratio tests. Our approach is to calculate the score vector using the pseudo log likelihood (i.e., treating the T marginal distributions as independent), but to calculate its covariance matrix using the true underlying dependence structure.

Specifically, for the model chosen to reflect possible departures from MH, we obtain the score vector by calculating partial derivatives of the pseudo log likelihood with respect to the model parameters describing the marginal differences, evaluated at the estimates obtained under MH. Since it depends only on the first derivative of the log likelihood, the score vector provides a consistent indication of whether MH holds (under the model assumption), regardless of the true dependence structure. We estimate the covariance matrix of the score

vector using the dependence structure across occasions implied by a multinomial assumption for the complete I^T table. The test statistic is a quadratic form comparing the score vector to its null expected value, weighted by the inverse estimated covariance matrix. Thus, the test statistic has the simplicity of a score statistic for independent samples, yet it takes into account the actual dependence.

There are two forms for this pseudo score statistic. One form uses the estimated covariance matrix for the score vector under the assumption of MH, whereas the other form uses the non-null estimated covariance matrix. Under MH, the two estimates converge to the same matrix as n increases, though one expects better approximations to null asymptotic sampling distributions using the null-based estimate. 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 null distribution. This can happen even when data are so highly sparse that regular or pseudo ML estimates of model parameters do not exist.

4. Weighted least squares model-fitting

One can use weighted least squares (WLS) methodology (Koch et al., 1977) to formulate another approach that 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 one can readily implement it with standard computer software.

In modeling first-order marginal functions such as logits, WLS methods require only the second-order marginal tables to estimate the asymptotic covariance structure of those response functions. For many models, one can obtain WLS fitting of the models for marginal distributions using procedure CATMOD in SAS. 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 5–10. 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. When the sample size is large and WLS is feasible but ML is not, WLS may be preferred over pseudo ML methods because of computational simplicity and guaranteed optimal asymptotic efficiency. However, the pseudo ML methods have the advantage of being applicable 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 is unreliable and highly sensitive to slight changes in the data when some marginal counts are small. In such cases, WLS estimates may be much poorer than pseudo ML estimates.

5. Marginal comparisons of ordinal classifications

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

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

for ordinal response variables. The link is some function of the first-order marginal probabilities. Two important special cases are: (1) the *cumulative logit* model, whereby

$$\text{Link}_j(g) = \text{logit}\big[\gamma_j(g)\big], \tag{5.2}$$

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

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

We use the latter model partly because of its equivalence to a simple loglinear model, the *row effects* model. In model (5.1), $\{\alpha_j\}$ are nuisance parameters, and $\{\mu_g\}$ describe variation in the marginal distributions. For this model, MH corresponds to $\mu_1 = \cdots = \mu_T$. For identifiability, $\{\mu_g\}$ satisfy a constraint such as $\Sigma\mu_g = 0$. For the cumulative logit link, it follows from Anderson and Philips (1981) that if there are underlying continuous logistic distributions for the marginal distributions, differing only in location, then this model holds and differences among $\{\mu_g\}$ describe the location shifts.

Our pseudo ML approach fits the model to the $T \times I$ table of sample marginal distributions, treating the samples as independent multinomials. We estimate a covariance matrix for the estimates of $\{\mu_i\}$ using (2.3) for the one-step jackknife. Using those results, we can test MH using a Wald test, letting the vector of differences be the pseudo ML estimates of $(\mu_1 - \mu_T, \ldots, \mu_{T-1} - \mu_T)$. The chi-squared asymptotic distribution for the Wald statistic has $df = T - 1$, rather than $df = (T-1)(I-1)$ as in the most general (unstructured) tests. Testing MH using this model gives a test directed towards variation in location among the occasions.

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

$$M_t = \sum_j v_j n_{tj}/n, \qquad t = 1,\ldots,T, \qquad \text{and} \qquad M = \sum_j v_j n_{\cdot j}/nT.$$

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. Our pseudo score statistics are quadratic forms describing variation among these means, or equivalently variation from 0 among $\{d_t = [(M_t - M) - (M_T - M)] = M_t - M_T, \quad t = 1,\ldots,T-1\}$. The quadratic forms use esti-

mated covariances among $\{d_t\}$ generated by assuming multinomial sampling over the full I^T table.

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 $\boldsymbol{d} = (d_1, \ldots, d_{T-1})'$ and for all (j, k) with $j < T$, $k < T$, let $\boldsymbol{S}$ denote the matrix having elements

$$s_{jk} = \sum_h \sum_i v_h v_i [p_{hi}(jk) - p_{hi}(jT) - p_{hi}(kT) + p_h(T)\delta_{hi}] - (M_j - M_T)(M_k - M_T),$$

where $\delta_{hi} = 1$ when $h = i$, and $\delta_{hi} = 0$ when $h \neq i$. Letting Y_{it} denote the response score for subject i at occasion t, we can also express

$$s_{jk} = \sum_e [(Y_{ej} - M_j) - (Y_{eT} - M_T)][(Y_{ek} - M_k) - (Y_{eT} - M_T)]/n.$$

Then s_{jk}/n is the unrestricted ML estimate of $\mathrm{Cov}(d_j, d_k)$, and the quadratic form $Q = n\boldsymbol{d}'\boldsymbol{S}^{-1}\boldsymbol{d}$ is an asymptotic chi-squared statistic for testing MH, based on $df = T - 1$.

For model (5.3) with rows in the $T \times I$ table treated as independent multinomials, the efficient score vector for testing MH has components $n(M_t - M)$, with $\{v_j = j\}$. Hence the statistic Q is a pseudo score statistic for testing MH using that model. When we let $\{v_j = [n_{\cdot 1} + \cdots + n_{\cdot j-1} + (1/2)n_{\cdot j}]/n\}$, the average cumulative proportion or 'ridit' scores for margin $\{n_{\cdot j},\ j = 1, \ldots, I\}$, Q is a pseudo score statistic for the cumulative logit model (5.2).

The pseudo score statistic Q is equivalent to the WLS goodness-of-fit statistic for testing the null *mean response* model

$$\mathrm{E}(M_t) = \alpha, \qquad t = 1, \ldots, T,$$

using a multinomial sampling assumption for the I^T table. One can compute this statistic using procedure CATMOD in SAS. For $T = 2$, Bhapkar (1970), Fleiss and Everitt (1971), Koch and Reinfurt (1971), and Meeks and D'Agostino (1983) have proposed tests of this form. It is also possible to use CATMOD to do standard WLS fits of models (5.2) and (5.3), incorporating the full multinomial dependence structure, at least when the first-order marginal totals are not too small.

The alternative form of the pseudo score statistic uses

$$s_{jk} = \sum_i (Y_{ij} - Y_{iT})(Y_{ik} - Y_{iT})/n,$$

which yields an estimated covariance matrix under the null hypothesis of equal marginal means. When MH holds, one would expect the score statistic using this estimate to have actual distribution more closely approximating the asymptotic chi-squared distribution, for small to moderate n. The two forms of the estimated covariance can give quite different test statistic values, and a good topic for future study is to compare operating characteristics of the two approaches.

Table 2
Estimates of $\{\mu_g\}$ for cumulative logit model

Pathologist	Pseudo ML	Pseudo WLS	Dependent WLS	Jackknife
A	0.58 (0.165)	0.52 (0.155)	0.37 (0.066)	0.58 (0.088)
B	0.51 (0.158)	0.49 (0.165)	0.30 (0.066)	0.51 (0.087)
C	−0.19 (0.152)	−0.19 (0.156)	0.02 (0.040)	−0.19 (0.087)
D	−0.51 (0.153)	−0.47 (0.157)	−0.38 (0.073)	−0.51 (0.083)
E	0.62 (0.156)	0.55 (0.158)	0.34 (0.075)	0.62 (0.088)
F	−1.13 (0.164)	−1.06 (0.164)	−0.67 (0.098)	−1.13 (0.128)
G	0.12 (0.156)	0.16 (0.161)	0.02 (0.039)	0.12 (0.058)

In the meantime, for purposes of testing MH, we recommend using the statistic based on the null estimate.

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.

Table 2 gives pseudo ML and jackknifed pseudo ML estimates of $\{\mu_g\}$ for the cumulative logit model, as well as estimated standard errors. The pseudo ML estimates are the ones obtained by fitting the cumulative logit version of model (5.1) to the 7×5 table of sample marginal distributions, treating the seven rows as independent multinomial samples. The jackknife estimates are the average of the 118 estimates obtained by leaving out, one at a time, each of the 118 observations. 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, based on (2.3), 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 0.001 tc the count for cell (4,4,4,4,4,4,5) to obtain a nonsingular covariance matrix). The reliability of such estimates is questionable, since two marginal counts equal 1 and two equal 2. The WLS Wald statistic for testing $\mu_1 = \cdots = \mu_7$ equals 85.7, based on $df = 6$. The WLS fit has residual chi-squared equal to 98.5, based on $df = 18$. The model does not

appear to 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 using the WLS method to fit the model to the 7×5 table that treats the margins as independent. The pseudo WLS estimates have similar values as the pseudo ML estimates. The WLS estimates that incorporate the dependence are somewhat different from the others, and have smallest estimated standard errors. However, a sensitivity analysis revealed that these estimates are unreliable, because of the presence of 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 85.7 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 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 combine categories 4 and 5, all marginal counts equal at least 5. The WLS Wald test statistic changed from 85.8 to 94.6, and the WLS model estimates changed to (0.43, 0.44, −0.19, −0.37, 0.41, −0.84, −0.12). These 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 has 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.

Similar results occurred for the adjacent-categories logit model. Table 3 reports estimates of $\{\mu_g\}$ and estimated standard errors for that model, using the same four approaches. Estimates are smaller than those for the cumulative

Table 3
Estimates of $\{\mu_g\}$ for adjacent-categories logit model

Pathologist	Pseudo ML	Pseudo WLS	Dependent WLS	Jackknife
A	0.32 (0.084)	0.28 (0.077)	0.20 (0.038)	0.32 (0.054)
B	0.24 (0.084)	0.21 (0.086)	0.11 (0.031)	0.24 (0.049)
C	−0.09 (0.086)	−0.13 (0.090)	0.02 (0.026)	−0.09 (0.050)
D	−0.27 (0.089)	−0.27 (0.092)	−0.20 (0.044)	−0.27 (0.053)
E	0.34 (0.084)	0.29 (0.089)	0.13 (0.045)	0.34 (0.055)
F	−0.59 (0.098)	−0.46 (0.088)	−0.26 (0.052)	−0.59 (0.096)
G	0.05 (0.084)	0.07 (0.089)	0.01 (0.024)	0.05 (0.031)

logit model (about half as large), since differences of estimates refer to *local* log odds ratios rather than log odds ratios utilizing the entire response scale. Again, the standard errors reported for the pseudo estimates are incorrect, being too large. The jackknife estimated standard errors are more reliable, and apply both to the jackknife estimates and to the pseudo ML and WLS estimates.

Pseudo score statistics for model (5.1), like the Wald statistics using the jackknife estimates and their standard errors, give strong evidence of marginal heterogeneity. For instance, the version with $\{v_j = j\}$ (the score statistic for the adjacent-categories logit model) gives a chi-squared statistic of 68.2 ($df = 6$) using the null estimated covariance.

7. Efficiency of pseudo ML and WLS estimates

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

Because of the extremely time-consuming nature of the ordinary ML estimation process, we limited most of our investigation to a single population with $T = 2$ occasions and $I = 3$ possible responses. We generated independent samples of size n from multinomial distributions defined over the 3×3 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 $\mu_d = \mu_1 - \mu_2$. The cell probabilities in the table were based on an underlying bivariate normal distribution having the given marginal probabilities. Specifically, let (X, Y) have a bivariate normal distribution with means 0, variances 1, and correlation ρ. Define cutpoints $\{\alpha_i, \gamma_i, i = 0,1,2,3\}$ by

$$P(\alpha_{i-1} \le X \le \alpha_i) = \pi_{i.}, \qquad P(\gamma_{j-1} \le Y \le \gamma_j) = \pi_{.j},$$

where $\alpha_0 = \gamma_0 = -\infty$ and $\alpha_3 = \gamma_3 = \infty$. The multinomial probabilities are $\{\pi_{ij} = P(\alpha_{i-1} \le X \le \alpha_i, \gamma_{j-1} \le Y \le \gamma_j)\}$. Eight combinations of n, ρ, and μ_d were chosen: $n = 20$ and 50, $\rho = 0.2$ and 0.8, and $\mu_d = 0.0$ (marginal homogeneity) and 0.4.

The program for the simulations used S-PLUS software (1990) and was executed on a DECstation 3100. We utilized the 'runif' and 'cut' functions in S-PLUS to randomly generate the multinomial counts. The algorithm for calculating constrained ML estimators followed the techniques developed by Aitchison and Silvey (1958) and Haber (1985). We obtained an estimated standard error for the ML estimator using the delta method. For generated

Table 4
Estimated root mean squared error (MSE) and standard deviation (SD) values for logit comparison of margins based on discretized bivariate normal distribution

Case	Estimator	$n = 20$			$n = 50$		
		Root MSE	Empirical std dev	Average model SD	Root MSE	Empirical std dev	Average model SD
$\mu_d = 0$	ML	0.251	0.251	0.213	0.143	0.143	0.136
$\rho = 0.8$	WLS	0.233	0.233	0.223	0.140	0.140	0.137
	P-ML[1]	0.233	0.233	0.402	0.140	0.140	0.248
	P-WLS[2]	0.229	0.229	0.406	0.139	0.139	0.248
$\mu_d = 0$	ML	0.384	0.384	0.362	0.241	0.241	0.227
$\rho = 0.2$	WLS	0.372	0.372	0.367	0.239	0.239	0.227
	P-ML	0.381	0.381	0.404	0.238	0.238	0.249
	P-WLS	0.371	0.371	0.410	0.237	0.237	0.250
$\mu_d = 0.4$	ML	0.275	0.272	0.242	0.159	0.159	0.152
$\rho = 0.8$	WLS	0.272	0.272	0.249	0.157	0.157	0.153
	P-ML	0.277	0.275	0.414	0.157	0.156	0.255
	P-WLS	0.272	0.271	0.418	0.155	0.155	0.256
$\mu_d = 0.4$	ML	0.410	0.409	0.378	0.246	0.246	0.234
$\rho = 0.2$	WLS	0.400	0.400	0.387	0.243	0.243	0.235
	P-ML	0.413	0.411	0.419	0.242	0.242	0.256
	P-WLS	0.395	0.394	0.426	0.240	0.240	0.258

[1] P-ML denotes pseudo ML
[2] P-WLS denotes pseudo WLS

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 the four estimators (ML, WLS, pseudo ML, pseudo WLS), based on 1000 simulations at each of the eight settings of (n, ρ, μ_d). With probability 0.95, for the $n = 20$ cases, the root MSE estimates are good to within about 0.020 when $\rho = 0.2$ and 0.015 when $\rho = 0.8$; for the $n = 50$ cases, they are good to within about 0.012 and 0.009.

The root MSE estimates in Table 4 show that, to the degree of accuracy obtained in this simulation study, (1) the four estimators performed equally well, and (2) we can conclude that the true MSE values are not substantively different. Somewhat surprisingly, even when X and Y were strongly correlated, the pseudo estimates performed adequately. Also, the WLS estimates performed as well as the ML estimates, though the sizes of the marginal counts were not small enough to cause the sorts of problems WLS estimates can have with sparse data.

Table 4 also reports the empirical standard deviations of the estimates, as well as the average of the standard errors predicted by the model fits. As expected, the pseudo standard errors behave poorly when there is substantial correlation; treating the margins as independent results in overestimating the degree of variability. If we regard the empirical standard deviation as being close to the true population standard error, we see that the ordinary model-based

estimates are quite good, though there is some evidence that they tend to be too small when n is small.

We have not attempted a comparison of standard deviation estimators, mainly because of the large amount of time it would take to simulate the jackknife. We take comfort from other studies of methods for categorical repeated measures by Bloch and Kraemer (1989) and Lipsitz et al. (1990a) that showed that the jackknife performed very well in estimating standard deviations. We did note empirically that the ML and WLS estimators of standard deviations that allow for correlation are about 50–75% more variable than the estimators of standard deviations that treat the classifications as independent. A useful topic for future research is to check whether standard deviation estimators such as the jackknife or robust estimators proposed by Liang and Zeger (1986) are more stable than the ordinary estimators.

Since 20–50 observations in a 3×3 table do not generate nearly the degree of sparseness encountered in many repeated categorical measurement studies, this study does not shed much light on effects of severe sparseness. It is impractical to simulate ordinary ML for much larger tables, but we did make a comparison of (WLS, pseudo WLS, pseudo ML) for some 10×10 tables with only 20 observations. Similar results held, in that the pseudo estimates were adequate. For instance, when there is marginal homogeneity, the estimated root MSE values for 1000 simulations were (0.103, 0.096, 0.108) when $\rho = 0.2$ and (0.064, 0.067, 0.059) when $\rho = 0.8$.

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 $T \times I$ table $\{n_{ij}\}$. The components of the efficient score vector for testing MH (i.e., independence for the $T \times I$ table) are $\{U_{ij} = n_{ij} - n_{.j}/T,\ i = 1, \ldots, T-1,\ j = 1, \ldots, I-1\}$. Note that $\Sigma_i U_{ij} = \Sigma_j U_{ij} = 0$. Let $d_{ij} = U_{ij} - U_{Tj} = n_{ij} - n_{Tj},\ i = 1, \ldots, T-1,\ j = 1, \ldots, I-1$. Then, MH is equivalent to $E(d_{ij}) = 0$ for all i and j, and a pseudo score statistic is given by a quadratic form in the vector of the $(T-1)(I-1)\{d_{ij}\}$ 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 in SAS.

9. Missing data issues

Although a goal of repeated measures and 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 (obtained, for example, using the EM algorithm: Dempster et al., 1977) and ML score tests are consistent under weak missing data conditions (missing at random; Rubin, 1976). When the data are missing at random, 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. When the data are missing completely at random, 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 complete 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. Further, as we have discussed, pseudo ML can be used with much sparser data.

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 or the Newton–Raphson algorithm (Hocking and Oxspring, 1971) and the asymptotic variance is consistently estimated by the inverse of the observed information (second derivative of the log-likelihood). 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 $T \times I$ contingency table are still treated as independent, but a row sum will not be identically n, and instead will satisfy $n_{t\cdot} \leq n$. 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 $M_t = (\sum_j v_j n_{tj})/n_{t\cdot}$, and, when calculating s_{jk}, $p_h(t) = n_{th}/n_{t\cdot}$ and $p_{hi}(tu) = n_{tuhi}/n_{tu\cdot\cdot}$, where n_{tuhi} is the number of subjects who have response 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 three approaches are different methods of estimating the probabilities in the $T \times I$ table as well as the covariance matrix of these estimates. The second step of the methods is the same; 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 $T \times I$ 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 $T \times I$ 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

In closing, we mention that 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 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. Such approaches also have the advantage over WLS of being valid for smaller sample sizes and sparser data.

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 principle, it is possible to use such an approach in which we model the covariance matrix in terms of some smaller set of parameters. For instance, Liang and Zeger used this approach for univariate responses with structures such as autoregressive or one-step dependence. For categorical responses having more than two categories, simple structures are not so readily apparent.

In future research, it is of interest to compare various ways of obtaining pseudo estimates. We believe that the sample pseudo estimates based on treating occasions as independent will be adequate for most purposes. It is also important in future work to compare the various ways of estimating the covariance matrix of pseudo estimates. This is especially crucial, since our simulation results suggest that it is more difficult to precisely estimate standard errors when there truly is dependence across occasions.

Finally, for completeness, we note there are other strategies one can employ when the main focus is testing of MH, rather than estimating model parameters. One approach uses generalizations of the Cochran–Mantel–Haenszel (C–M–H) test. For details, see Agresti (1990, Sections 7.4 and 8.4), Darroch (1981), White et al. (1982), and Landis et al. (1988). The generalized C–M–H statistic is applied to a $T \times I \times n$ table, in which the T responses for subject k form the kth of n strata or blocks. The stratum for subject k consists of a $T \times I$ table, in which there is a single observation in each row. The $T \times I$ table fitted in Section 2 is the two-way marginal table of this one, collapsed over subjects. Some versions of this statistic correspond, for certain models, to a pseudo score test statistic.

One can use a generalized C–M–H statistic presented by Landis et al. (1988) to test variation among marginal means of an ordinal classification, with $df = T - 1$. For Table 1 it also gives strong evidence of marginal heterogeneity, equaling 202.0 based on $df = 6$ for the case of equally-spaced responses scores. When one uses this generalized statistic with rank scores computed separately in each block, this gives the corrected-for-ties version of Friedman's test (White et al. 1982). For Table 1 this statistic equals 214.2.

Darroch (1981) noted that the generalized C–M–H statistic for nominal classifications, having $df = (T-1)(I-1)$, has the advantage that the inverse of the $(T-1)(I-1) \times (T-1)(I-1)$ covariance matrix of $\{d_{ij}\}$ from Section 8 is determined by the inverse of an $(I-1) \times (I-1)$ matrix. The generalized C–M–H statistics apply to sparse data, and are available using procedure FREQ in SAS. However, they have the disadvantages that they make certain exchangeability assumptions that rarely are suitable when the occasions are times (Landis et al. 1988), and they are directed strictly towards testing MH, rather than estimating the degree of marginal heterogeneity for some specified model.

Sample programs for performing the analyses reported in this paper are available upon request from the authors. The jackknife program for estimating the covariance matrix of the pseudo estimates uses procedure IML in SAS.

Acknowledgments

The work of Agresti and Lang was partially supported by NIH grant GM 43824, and the work of Lipsitz was partially supported by NIH grant GM 29745. The authors thank the referees for helpful comments.

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) (Wiley, New York).

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

Anderson, J.A., and P.R. Philips. Regression, discrimination, and measurement models for ordered categorical variables. *Appl. Statist.* **30** (1981), 22–31.

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

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

Bloch, D.A., and H.C. Kraemer. 2×2 kappa coefficients: Measures of agreement or association. *Biometrics* **45** (1989), 269–287.

Cox, D.R., and D.V. Hinkley. *Theoretical Statistics* (1974) (Chapman and Hall, London).

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.

Firth, D., and B.R. Treat. Square contingency tables and GLIM. *GLIM Newsletter*, **16** (1988), 16–20.

Fleiss, J.L., and B.S. Everitt. Comparing the marginal totals of square contingency tables. *British J. Math. Statist. Psychol.*, **24** (1971), 117–123.

Fox, T., D. Hinkley, and K. Larntz. Jackknifing in non-linear regression. *Technometrics* **22** (1980), 29–33.

Gourieroux, C., A. Monfort, and A. Trognon. Pseudo maximum likelihood methods: theory. *Econometrica* **52** (1984), 681–700.

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., and D.W. Reinfurt. The analysis of categorical data from mixed models. *Biometrics* **27** (1971), 157–173.

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., E.R. Heyman, and G.G. Koch. Average partial association in three-way contingency tables: A review and discussion of alternative tests. *Internat. Statist. Rev.* **46** (1978), 237–254.

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.

Meeks, S.L., and R.B. D'Agostino. A model for comparisons with ordered categorical data. *Commun. Statist.* **A12** (1983), 895–906.

Rao, C.R. *Linear Statistical Inference and its Applications*, 2nd ed. (1973) (Wiley, New York).

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

Simonoff, J.S., and C.L. Tsai. Jackknife-based estimators and confidence regions in non-linear regression. *Technometrics* **28** (1986), 103–112.

S-PLUS User's Manual (1990) Statistical Sciences, Inc. Seattle, WA.

Stram, D.O., L.J. Wei, and J.H. Ware. Analysis of repeated categorical outcomes with possibly missing observations and time-dependent covariates. *J. Amer. Statistic. Assoc.* **83** (1988), 631–637.

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.