ELSEVIER Computational Statistics & Data Analysis 24 (1997) 89–104

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables

Donguk Kim [a], Alan Agresti [b,*]

[a] *Department of Computer Science and Statistics, Hanshin University, Osan, KyungKi-Do, 447-791, South Korea*

[b] *Department of Statistics, University of Florida, Gainesville, FL 32611-8545, USA*

Received May 1995; revised February 1996

Abstract

We discuss a variety of test statistics for the hypothesis of conditional independence in three-way contingency tables. Statistics that are designed to detect association between nominal variables, between ordinal variables, and between nominal and ordinal variables are presented. Tests previously presented by Birch (1965) and Landis et al. (1978) are efficient score statistics for alternative hypotheses corresponding to loglinear models that assume homogeneous associations across levels of the control variable. Additional score statistics are presented for loglinear model alternatives that permit interaction, in the form of heterogeneous associations. Ordinary asymptotic chi-squared inference is well established for both types of alternatives. For small samples or sparse data, however, software for exact conditional inference is currently unavailable for most cases having multiple levels of response variables. Monte-Carlo simulation of exact distributions is computationally simple and quick, even for large sparse tables. It results in precise estimates of P-values for exact conditional tests, and hence 'nearly exact' tests, for cases lacking software or cases that are likely to be computationally infeasible in the indefinite future. The tests with homogeneous association alternatives also can be applied to nearly exact testing of marginal homogeneity for multivariate responses having the same categorical scale for each component.

Keywords: Cochran–Mantel–Haenszel test; Linear-by-linear association; Loglinear model; Monte-Carlo; Ordinal variables; Score statistic

* Corresponding author. E-mail: aa@stat.ufl.edu.

PII S0167-9473(96)00038-2

1. Introduction

Table 1 shows some preliminary results from a double-blind, parallel-group clinical study that is being conducted by Merck Research Laboratories at a large number of centers. The study is designed to compare an active drug with placebo for the treatment of patients suffering from a particular chronic respiratory disease. Patients were randomly assigned to the treatments. At the end of the study, investigators were asked to describe their perception of the patient's change in condition, using the ordinal scale (better, unchanged, worse). Because of time considerations in recruiting subjects, the study used a large number of centers (33). To conserve space, Table 1 shows data from only 10 of them. The data in Table 1 are highly sparse, the three-way table having many strata but few observations per stratum.

For three-way $I \times J \times K$ contingency tables, let X, Y, and Z denote the row, column, and layer classifications. This article discusses several tests of the hypothesis of conditional independence of X and Y, given Z. For Table 1, for instance, we investigate whether treatment has an impact on response, adjusting for center. The tests are also applicable to testing X–Y conditional independence in tables of arbitrary numbers of dimensions, by identifying Z as a composite variable having levels given by all combinations of the other variables. We are particularly concerned with cases,

Table 1
Evaluations of response to active drug and placebo

Stratum	Drug	Response		
		Better	Unchanged	Worse
1	Placebo	0	2	1
	Active	1	1	0
2	Placebo	0	1	0
	Active	1	1	0
3	Placebo	1	1	0
	Active	0	1	0
4	Placebo	1	0	0
	Active	1	1	0
5	Placebo	0	0	1
	Active	0	1	0
6	Placebo	2	0	0
	Active	1	0	1
7	Placebo	0	1	0
	Active	1	0	0
8	Placebo	0	2	0
	Active	1	1	0
9	Placebo	0	1	0
	Active	1	0	0
10	Placebo	0	1	1
	Active	1	0	0

such as Table 1, in which sparseness of data implies that ordinary large-sample chi-squared tests may be invalid.

Let $\{n_{ijk}\}$ denote the cell counts, having total sample size n and expected frequencies $\{m_{ijk}\}$. The counts may follow any of the standard sampling models, such as multinomial or independent Poisson over the entire table, or independent multinomial within each level of Z or each combination of levels of X and Z. For test statistics, the tests that we present use efficient score statistics for loglinear and corresponding logit models representing various alternative hypotheses. Large-sample versions of many of these tests are already 'well-known' and available in standard software.

In practice, the hypothesis of conditional independence is most commonly tested against the alternative represented by the loglinear model of form

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{1}$$

This model, which lacks the three-factor interaction term, assumes that the X–Y association is homogeneous across the levels of Z. Conditional independence of X and Y is the special case in which all $\lambda_{ij}^{XY} = 0$. Another approach, occasionally used when the X–Y association may be highly heterogeneous, tests against the general alternative corresponding to the saturated loglinear model. These tests treat both X and Y as nominal variables.

We consider tests in which the alternatives relate to model (1) or the saturated model, or models that are special cases of these two and permit trends reflecting ordinality of X and/or Y. Section 2 discusses statistics for alternatives that are special cases of model (1). The statistics are designed to detect various types of homogeneous association. Section 3 presents statistics for alternatives that permit three-factor interaction.

For a fixed number of cells, asymptotic chi-squared theory is well developed for the statistics we present. Because multi-way contingency tables are often sparse, we focus on the use of these statistics with the exact conditional distribution. For each test considered, one could use a likelihood-ratio, Wald, or efficient score statistic as the test statistic. We focus on score statistics, because inferential analyses using the exact distribution are then computationally much simpler. The *conditional score* statistic uses the distribution based on conditioning on row and column totals in each stratum, which are sufficient statistics for the unknown nuisance parameters.

Computational algorithms for exact tests have limited availability when I and J exceed two. Section 4 uses a simple Monte-Carlo algorithm to provide nearly exact results for all tests presented in this article. This approach provides highly precise approximations for P-values based on the exact conditional approach. Section 5 presents asymptotic, nearly exact, and exact results for various tests applied to Table 1.

Finally, Section 6 discusses comparisons of marginal distributions of a multivariate categorical response having ordered or unordered categories. One can use the homogeneous association statistics as the basis of nearly exact tests of marginal homogeneity.

2. Tests of conditional independence, assuming no interaction

This section reviews statistics for testing conditional independence, assuming homogeneous association. Nearly all tests of conditional independence in the current literature refer to this case. We see that three statistics presented by Birch (1965) are score statistics for loglinear models. One test treats both X and Y as nominal, one test treats X as nominal and Y as ordinal, and one test treats both as ordinal. In practice, of course, it is unrealistic to expect homogeneous association to hold perfectly, but test statistics derived under that assumption have enhanced power for cases in which the degree of heterogeneity is insubstantial.

As a point of reference, we begin with the case in which X and Y are nominal. Let $\boldsymbol{n}_k$ denote the counts for cells in the first $I-1$ rows and $J-1$ columns for stratum k of Z. Conditional on the row and column totals in that stratum, let $\boldsymbol{m}_k$ denote the null expected value of $\boldsymbol{n}_k$. Then $\boldsymbol{d} = \sum_k (\boldsymbol{n}_k - \boldsymbol{m}_k)$ represents the $(I-1)(J-1) \times 1$ vector having elements

$$d_{ij} = \sum_k \left[n_{ijk} - \left(\frac{n_{i+k} n_{+jk}}{n_{++k}} \right) \right], \quad i = 1,\ldots,I-1, \; j = 1,\ldots,J-1. \tag{2}$$

Let $\boldsymbol{V}_k$ denote the null covariance matrix of $\boldsymbol{n}_k$, where

$$\mathrm{Cov}(n_{ijk}, n_{i'j'k}) = \frac{n_{i+k}(\delta_{ii'} n_{++k} - n_{i'+k}) n_{+jk} (\delta_{jj'} n_{++k} - n_{+j'k})}{n_{++k}^2 (n_{++k} - 1)}, \tag{3}$$

with

$$\delta_{ab} = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\boldsymbol{V} = \sum_k \boldsymbol{V}_k$ is the null covariance matrix of $\boldsymbol{d}$. The conditional score statistic for testing conditional independence against the homogenous association alternative (1) equals

$$T_{\mathrm{N}} = \boldsymbol{d}' \boldsymbol{V}^{-1} \boldsymbol{d}, \tag{4}$$

where the N subscript refers to the treatment of the variables as nominal. The test statistic has a large-sample null chi-squared distribution with $df = (I-1)(J-1)$. For $K = 1$ stratum, T_{N} reduces to the multiple $(n-1)/n$ of the Pearson chi-squared statistic for testing independence. For $I = J = 2$, it is the Cochran–Mantel–Haenszel statistic.

When X and Y are ordinal, it usually makes sense to test conditional independence against a narrower alternative, in order to increase the power of the test. One can do this by describing the X–Y partial association using fewer parameters, in order to concentrate the effect on a small df value. Reasonable test statistics are based on alternative models that are special cases of (1) and utilize the ordinality.

In many applications, one expects the partial association to exhibit, in some sense, a monotone trend. A simple model that implies such a trend is the one having

a linear-by-linear form for the X–Y association that is homogeneous across levels of Z,

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{5}$$

This model replaces the general association term λ_{ij}^{XY} in (1) by a linear-by-linear term $\beta u_i v_j$, where $\{u_i\}$ and $\{v_j\}$ are monotone scores for levels of X and Y. The model implies stochastic orderings of the rows and of the columns with respect to their conditional distribution on the other variable, within each level of Z. The model of conditional independence of X and Y is its special case in which $\beta = 0$. For further discussion of this type of model and other ordinal models discussed in this article, see Goodman (1979), Clogg (1982), and Agresti and Kezouh (1983).

The sufficient statistic for β in this model is $\sum_k[\sum_i \sum_j u_i v_j n_{ijk}]$. Given the marginal totals and under conditional independence of X and Y,

$$E\left(\sum_i \sum_j u_i v_j n_{ijk}\right) = \frac{(\sum_i u_i n_{i+k})(\sum_j v_j n_{+jk})}{n_{++k}},$$

$$Var\left(\sum_i \sum_j u_i v_j n_{ijk}\right) = \frac{1}{n_{++k}-1}\left[\sum_i u_i^2 n_{i+k} - \frac{(\sum_i u_i n_{i+k})^2}{n_{++k}}\right] \times \left[\sum_j v_j^2 n_{+jk} - \frac{(\sum_j v_j n_{+jk})^2}{n_{++k}}\right].$$

To summarize the ordinal information from the K strata, Mantel (1963) proposed the statistic

$$T_O = \frac{\{\sum_k[\sum_i \sum_j u_i v_j n_{ijk} - E(\sum_i \sum_j u_i v_j n_{ijk})]\}^2}{\sum_k Var(\sum_i \sum_j u_i v_j n_{ijk})}. \tag{6}$$

The O subscript refers to the treatment of the variables as ordinal. This is the conditional score statistic for testing conditional independence for model (5). It is sensitive to 'correlation' alternatives to conditional independence and has an asymptotic chi-squared distribution with $df = 1$.

Suppose next that the row variable X is nominal and the column variable Y is ordinal. A useful loglinear model replaces the ordered row scores in model (5) by unordered parameters $\{\mu_i\}$,

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \mu_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{7}$$

Within each level of Z, this model implies a stochastic ordering of the rows with respect to their conditional distributions on Y. The ordering is the same as that of $\{\mu_i\}$, and is identical at each layer of Z. The fully ordinal model (5) is the special case in which these parameters change linearly across the rows. One can also use model (7) when X is ordinal but one does not expect a linear trend. An analogous model with ordered row scores and unordered column parameters applies when X is ordinal and Y is nominal, or when they are both ordinal and one expects the levels of Y to be stochastically ordered on X but without a linear trend of column parameters.

For model (7), the sufficient statistics for $\{\mu_i\}$ are $\sum_j v_j n_{ij+}$, $i = 1,\ldots,I$. Assuming that the model holds, conditional independence is equivalent to $\mu_1 = \mu_2 = \cdots = \mu_I$. Let ℓ denote the $(I-1)\times 1$ vector having elements

$$\ell_i = \sum_k n_{i+k}(\bar{y}_{ik} - \bar{y}_k), \quad i = 1,\ldots,I-1,$$

where $\bar{y}_{ik} = \sum_j n_{ijk}v_j/n_{i+k}$, and $\bar{y}_k = \sum_i \sum_j n_{ijk}v_j/n_{++k}$. Note that $\bar{y}_{ik}$ is the row mean on Y at levels i of X and k of Z, and $\bar{y}_k$ is the kth stratum mean for Y, treating Y as a response with scores $\{v_j\}$. Conditional on the row and column marginal totals for each stratum, let A denote the null covariance matrix of ℓ, which has elements

$$\mathrm{Cov}(\ell_i, \ell_{i'}) = \sum_k \left[\frac{n_{i+k}(\delta_{ii'}n_{++k} - n_{i'+k})}{n_{++k}(n_{++k}-1)} \sum_j n_{+jk}(v_j - \bar{y}_k)^2 \right]. \tag{8}$$

The conditional score statistic for testing conditional independence against the nominal–ordinal alternative (7) is

$$T_{\mathrm{NO}} = \ell' A^{-1} \ell. \tag{9}$$

This statistic is sensitive to differences in row mean scores among the I conditional distributions of Y that are similar at each level of Z. The asymptotic null distribution is chi-squared with $df = I - 1$.

The three statistics just discussed are not new. Presented by Birch (1965), they are special cases of a general statistic proposed by Landis et al. (1978). Their connections with loglinear models do not seem to have been considered. The asymptotic chi-squared tests using these three statistics are available in SAS (CMH option in PROC FREQ). The ordinal statistics can be applied with scores chosen by the user or with ridit (midrank) scores. One can regard the rank-score version of statistic (6) as a stratified Spearman correlation test statistic, and the rank-score version of statistic (9) as a stratified Kruskal–Wallis test statistic.

For Table 1, for the general association alternative, $T_{\mathrm{N}} = 1.90$ $(df = 2)$ has a P-value of 0.39. When the table has only two rows, the statistic (6) for the correlation alternative is identical to the statistic (9) that compares row means. With equally spaced column scores, $T_{\mathrm{O}} = T_{\mathrm{NO}} = 1.81$ $(df = 1)$, for a P-value of 0.18. These P-values are based on the asymptotic chi-squared distributions. For the one-sided alternative of a better response for drug than placebo, one can treat the signed square root of the correlation statistic T_{O} as a normal variate. The P-value then equals 0.09.

3. Tests of conditional independence, permitting interaction

The tests just discussed are directed toward homogeneous association alternatives. In some applications, one might expect the association between X and Y to vary considerably across levels of Z. The test statistic should then relate to a model that permits three-factor interaction. Such statistics combine information from the various strata based on summarizing the association separately in each stratum. Because they focus on broader alternatives, these statistics run the risk of potential lack of

power. To help protect against this, one should use a statistic that recognizes any ordinal classifications and takes into account the likely form of the association in each stratum.

We present conditional score statistics for three alternatives corresponding to models permitting heterogeneous association. They are unified by having the common structure $T^* = \sum T(k)$, where $T(k)$ is one of the three forms of statistic from the previous section applied to stratum k alone.

When X and Y are nominal, the ordinary test has as its alternative the saturated loglinear model. The score statistic is similar to the Pearson statistic for testing conditional independence against that alternative. Letting X_k^2 denote the Pearson statistic for testing independence within the kth level of Z, the conditional score statistic is

$$T_N^* = \sum_k [(n_{++k} - 1)/n_{++k}]X_k^2.$$

The unconditional score statistic, $\sum_k X_k^2$, is the Pearson statistic normally used for the large-sample test. The asymptotic distribution for either statistic is chi-squared with $df = K(I-1)(J-1)$.

A disadvantage of this test is that it can have much less power than the nominal-by-nominal test of the previous section when the degree of three-factor interaction is not severe, particularly when K is large. One can narrow the alternative hypothesis for such a test somewhat by using a statistic that corresponds to a model that provides some pattern for the interaction or for the association in each partial table.

When X and Y are ordinal, suppose that one expects a monotone association between X and Y that changes strength across levels of Z. A relevant loglinear model is then the heterogeneous linear-by-linear association model,

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_k u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{10}$$

Fitting this model is equivalent to fitting a linear-by-linear association model separately at each level of Z. Model (5) is the special case in which $\beta_1 = \cdots = \beta_K$, and conditional independence is the further special case in which that common value is equal to 0.

For this model the sufficient statistic for β_k is $\sum_i \sum_j u_i v_j n_{ijk}$. Let $\boldsymbol{r}$ denote the $K \times 1$ vector having elements

$$r_k = \sum_i \sum_j u_i v_j \left(n_{ijk} - \frac{n_{i+k} n_{+jk}}{n_{++k}} \right).$$

The conditional score statistic for testing $H_0 : \beta_1 = \cdots = \beta_K = 0$ is a quadratic form based on $\boldsymbol{r}$ that simplifies to

$$T_O^* = \sum_k T_O(k),$$

where $T_O(k)$ denotes (6) applied to stratum k alone. Its asymptotic distribution is chi-squared with $df = K$. This statistic is sensitive to correlation alternatives that vary in strength among the strata.

When X is nominal and Y is ordinal, a relevant loglinear model to allow heterogeneity across the strata is

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \mu_{ik} v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{11}$$

The $\{v_j\}$ are fixed monotone scores, and separate row effect parameters $\{\mu_{ik}\}$ pertain to each stratum. In stratum k, the conditional distributions on Y are stochastically ordered according to the values of these parameters. The relevant sufficient statistics are sample means for Y at all the row and stratum combinations. Model (7) is the special case in which $\mu_{i1} = \cdots = \mu_{iK}$ for all i, in which case the comparison of the rows has the same form for each stratum. Model (11) reflects alternatives whereby means on Y vary across levels of X, but in a different way at different levels of Z. Conditional independence is the further special case in which $\mu_{1k} = \cdots = \mu_{Ik}$ for all k.

Let $\boldsymbol{q}$ be the $K(I-1) \times 1$ vector having elements

$$q_{ik} = \sum_j v_j \left(n_{ijk} - \frac{n_{i+k} n_{+jk}}{n_{++k}} \right), \quad i = 1, \ldots, I-1, \; k = 1, \ldots, K.$$

The conditional score statistic for testing $H_0 : \mu_{1k} = \cdots = \mu_{Ik}$ for all k is a quadratic form in $\boldsymbol{q}$ that simplifies to

$$T_{\mathrm{NO}}^* = \sum_k T_{\mathrm{NO}}(k),$$

where $T_{\mathrm{NO}}(k)$ denotes (9) applied to stratum k alone. Its asymptotic distribution is chi-squared with $df = K(I-1)$.

This statistic also applies when both variables are ordinal and one expects location shifts in the row means, but one does not expect a monotone trend in those row means in every stratum. An analogous model treats X as ordinal and Y as nominal. Then, the X–Y association term has form $u_i v_{jk}$, for ordered row scores and unordered column parameters that differ by stratum.

For Table 1, the test statistic for the most general alternative equals $T_{\mathrm{N}}^* = 10.28$ ($df = 12$) and has a chi-squared asymptotic P-value of 0.59 (The zero column total that occurs in all but the first and last table forces the corresponding fitted values to be 0 and causes a reduction in df from 20 to 12 for this test). The statistics for the nominal-by-ordinal and ordinal-by-ordinal alternatives equal $T_{\mathrm{O}}^* = T_{\mathrm{NO}}^* = 9.67$ ($df = 10$) and have an asymptotic P-value of 0.47.

Each test statistic discussed in this section has df increased by a multiplicative factor of K compared to the corresponding test statistic assuming homogeneous association. This can be a major detriment toward achieving decent power, especially when K is large. When the degree of association heterogeneity is not severe, one is better off using the tests of Section 2. For instance, suppose X and Y are both ordinal and model (10) is plausible. If the partial associations all have the same direction (i.e., all $\beta_k > 0$ or all $\beta_k < 0$), the score statistic T_{O} based on the simpler model (5) is likely to be preferable. That is, the advantage of having a statistic based on $df = 1$ rather than $df = K$ outweighs loss of information from lack of fit of the model on which the statistic is based.

In this regard, the remark of Mantel (1963) in a similar context is instructive: "That a linear regression is being tested does not mean that an assumption of linearity is being made. Rather it is that test of a linear component of regression provides power for detecting any progressive association which may exist." On the other hand, if the partial associations have different direction, with some $\beta_k > 0$ and some $\beta_k < 0$, then T_O may fail to detect the association and the score statistic T_O^* for the heterogeneous model (10) is more appropriate.

When the control variable Z is also ordinal, one can attempt to improve the power by providing further structure for the nature of the interaction across the levels of Z. For instance, when X and Y are ordinal, suppose one can specify a set of scores $\{w_k\}$ such that one expects model (10) to hold with β_k roughly proportional to w_k. This resulting simplification of that model has conditional score statistic for testing conditional independence equal to

$$\frac{\{\sum_k w_k[\sum_i \sum_j u_i v_j n_{ijk} - E(\sum_i \sum_j u_i v_j n_{ijk})]\}^2}{\sum_k w_k^2 \, Var(\sum_i \sum_j u_i v_j n_{ijk})}, \tag{12}$$

where the formula for the null expectation and variance are those given preceding formula (6). This is a single-degree-of-freedom chi-squared statistic that simplifies to T_O when all $w_k = 1$.

4. Approximation of exact *P*-values

For fixed I, J, and K, standard asymptotic chi-squared theory applies to the statistics presented in the previous two sections. In practice, however, for sparse tables such as Table 1, large-sample results are often suspect, particularly as the df value for the statistic increases. We now discuss more precise tests of conditional independence using these statistics.

To test conditional independence (CI) against an alternative hypothesis that corresponds to a more complex model (M), the exact conditional approach eliminates nuisance parameters by using the distribution of the minimal sufficient statistics for M, conditional on the minimal sufficient statistics for CI (see, e.g., Andersen, 1974; Agresti, 1992). The conditional reference set for CI is the set of all tables having the same row totals $\{n_{i+k}\}$ and columns totals $\{n_{+jk}\}$ for the partial tables as the observed data. Denote this set by

$$\Gamma = \{\boldsymbol{x} : x_{i+k} = n_{i+k}, \; x_{+jk} = n_{+jk}, \text{ all } i,j,k\}.$$

The null conditional distribution of the cell counts consists of independent generalized hypergeometric distributions for the various strata, and does not depend on the nuisance parameters. From Birch (1965), the conditional distribution is proportional to $(\prod_i \prod_j \prod_k x_{ijk}!)^{-1}$, defined over all tables in Γ. The exact P-value is the null conditional probability that the relevant test statistic T takes value at least as large as the observed value. The calculation is based on the conditional distribution for the statistic induced by the generalized hypergeometric distributions defined on Γ.

That is, letting

$$\Gamma_t = \{\boldsymbol{x} \in \Gamma : T \geq t\},$$

the exact P-value is

$$P = \left[\sum_{\Gamma_t} \left(\prod_i \prod_j \prod_k x_{ijk}!\right)^{-1}\right] \Big/ \left[\sum_{\Gamma} \left(\prod_i \prod_j \prod_k x_{ijk}!\right)^{-1}\right].$$

All the test statistics in this paper have the form

$$T = \boldsymbol{s}' \, Var(\boldsymbol{s})^{-1} \boldsymbol{s},$$

where $\boldsymbol{s}$ is a score vector. In Section 2, for instance, the score vector has $(I-1)(J-1)$ elements in the nominal case, $(I-1)$ elements in the nominal-by-ordinal case, and 1 element in the ordinal case. For a particular statistic, the covariance matrix in this general expression is the same for each table in Γ. Thus, the calculation of T for each table in Γ primarily involves recalculating the extra sufficient statistic for the association terms in the relevant loglinear model for the alternative hypothesis.

Software for exact tests in three-way tables exists for limited cases, as discussed later in Section 5. Even for cases in which software exists, the reference set Γ is often too large for an exact P-value computation, since it grows exponentially in n and the table dimensions. When the sample size is moderately large but the table has many cells and is sparse, the use of standard asymptotic theory is questionable (particularly for tests having a large df), but exact methods are usually infeasible.

Since software is not available in the generality needed for exact tests using the statistics we presented for $I \times J \times K$ tables, one can approximate the exact conditional result as closely as needed by performing a Monte-Carlo simulation of the exact distribution of the conditional score statistic. Agresti et al. (1979) utilized this method for a variety of tests for two-way tables. The method consists of sampling contingency tables randomly from Γ in proportion to their probabilities, and computing an unbiased point estimate and a narrow confidence interval for the exact P-value.

To illustrate, suppose we want to estimate an exact P-value for some statistic T, and let t_o denote its observed value. We sample M contingency tables with replacement from the reference set Γ, where M is chosen to give the desired degree of accuracy with some fixed probability. For instance, $M = 12\,600$ is sufficient to estimate a P-value that is no greater than 0.050 to within an accuracy of 0.005 with probability 0.99. For the ith table sampled, let $y_i = 1$ if it is in Γ_{t_o}, and let $y_i = 0$, otherwise. The point estimate of the exact P-value is

$$\hat{P} = \frac{1}{M} \sum y_i,$$

the proportion of sampled tables falling in Γ_{t_o}, having standard error $[P(1-P)/M]^{1/2}$.

One generates K independent generalized hypergeometrics in order to generate the K partial tables for each random table in the Monte-Carlo simulations. Kreiner (1987) described this approach for tests of conditional independence. Our algorithm uses a table-generation procedure suggested by Patefield (1981) to generate each of

the K partial tables at each step. Even for large tables or large sample sizes, one can quickly approximate exact P-values as closely as needed for any of the statistics presented in the previous two sections.

For practical applications, we prefer this approximation to others, such as the saddlepoint (Pierce and Peters, 1992; Agresti et al., 1993). For tests of conditional independence, it is available more generally (e.g., for multi-degree-of-freedom statistics for testing vectors of parameters), its accuracy is known to the user, and that accuracy can be set as finely as one requires. Though the process could be speeded up by using importance sampling (as in Mehta et al., 1988), we have not encountered any problems with obtaining precise results quickly using our algorithm in a workstation environment.

5. Example

We illustrate the nearly exact tests using the data in Table 1. Table 2 summarizes results for the asymptotic tests reported in Sections 2 and 3. The data are so sparse that these large-sample approximations are questionable. To estimate P-values for exact conditional tests, we used Monte-Carlo sampling with $M = 100\,000$. This guarantees that P-value estimators fall within 0.004 of the true P-value with probability at least 0.99, and to within 0.002 of the true value with at least this probability when the true P-value is no greater than 0.050. Table 2 also shows results for the nearly exact tests, in terms of 99% confidence intervals for the exact P-values. These intervals are based on inverting results of 0.01-level large-sample tests for a binomial proportion, using the null standard error (i.e., the endpoints are roots of a quadratic equation). The ordinal-by-ordinal analysis refers to the one-sided alternative.

Table 2 illustrates that asymptotic P-values can be rather poor when the df value for the test is large and the sample size is small. The exact P-value of 1.000 for the most general test permitting heterogeneity simply indicates that no table configuration with the given margins can produce a larger Pearson statistic in any one of the strata, which is evident by visual inspection of the partial tables.

Table 2
Results of asymptotic, nearly exact, and exact tests of conditional independence for Table 1

	Test statistic	Degrees of freedom	Asymptotic P-value	Est. exact P-value	Exact P-value
Assuming homogeneous association					
Nominal-by-nominal	1.90	2	0.39	(0.413, 0.421)	—
Nominal-by-ordinal	1.81	1	0.18	(0.210, 0.217)	0.216
Ordinal-by-ordinal	1.81	1	0.09	(0.136, 0.142)	0.140
Permitting three-factor interaction					
Nominal-by-nominal	10.28	12	0.59	(0.99993, 1.000)	1.000
Nominal-by-ordinal	9.67	10	0.47	(0.609, 0.617)	—

For each nearly exact test having an estimated exact P-value reported in Table 2, computations based on 100 000 simulations took between one and three minutes on a Sun SPARCstation 10–30 with 32 MB main memory, 1.2 GB hard disk, and 86.5 MIPS. The computing time is roughly proportional to the size of the table or the sample size. For instance, the same tests conducted on a $4 \times 3 \times 10$ table containing data for two additional dose levels of the drug and having twice as many cells and about twice as large a sample took about twice as long.

We purposely chose an example having only two rows, since software exists for exact analyses for some cases, making comparisons possible. Table 2 also reports these exact results, where available. For $2 \times J \times K$ tables, StatXact (1991) provides an exact test for a statistic that refers to the special case of alternative (7) with $I = 2$. For Table 1, StatXact reports $P = 0.216$ for this analysis, compared to (0.210, 0.217) for the Monte-Carlo approach reported in Table 2. For the one-sided alternative, StatXact reports $P = 0.140$, compared to (0.136, 0.142) for the Monte-Carlo approach reported in Table 2. StatXact also uses Monte-Carlo methods with $2 \times J \times K$ tables when exact inference is impractical.

Software for the general $I \times J \times K$ case is more limited. The DIGRAM software (Kreiner, 1989) provides Monte-Carlo approximation for exact tests against the general alternative, based on the Pearson goodness-of-fit statistic for the model of conditional independence. This is the unconditional score statistic for the alternative of the saturated model. When the stratum sample sizes are identical, the conditional and unconditional score statistics have the same ordering of the sample space, and thus have identical exact P-value. DIGRAM also provides an ordinal test based on a partial association version of the gamma measure of association. That test does not correspond to a loglinear model alternative, but would tend to be powerful for trend alternatives similar to those detected by (5). Yao and Tritchler (1993) used the Pearson statistic for exact tests of conditional independence for $2 \times 2 \times K$ tables. Baglivo et al. (1992, 1993) discussed an algorithm for this exact test for the $I \times J \times K$ case, but did not provide software. Morgan and Blumenstein (1991) described an algorithm for testing the fit of loglinear models using an ordering of tables in Γ based solely on the table probability. The algorithm requires complete enumeration of tables in Γ, and is impractical when n, I, J, or K have moderate size.

When $K = 1$, the exact test that orders tables in Γ by (6) for the linear-by-linear alternative was discussed by Agresti et al. (1990) and Cohen and Sackrowitz (1992), and is available in StatXact (1991). For arbitrary K but $I = J = 2$, widely available software including StatXact provides tests of conditional independence for alternative (1).

6. Nearly exact tests comparing marginal distributions of repeated categorical responses

The tests for the homogeneous association alternative proposed in Section 2 can be used to compare marginal distributions of multivariate responses. Suppose each subject is measured on a categorical response having J categories for each of I

measurements. For instance, each subject might be measured on the same response variable at each of I separate occasions. We refer to the I separate components of the multivariate response as *items*. Let ϕ_{ijk} denote the probability that subject k makes response j for the ith item. Consider the model

$$\log(\phi_{ijk}/\phi_{iJk}) = \alpha_{jk} + \beta_{ij}. \tag{13}$$

For each response category paired with the baseline category J, this logit model is additive in subject and item effects. This is a multinomial version of the *Rasch model* (Rasch, 1961).

This model implies the loglinear model of quasi-symmetry for the J^I cross-classification of responses for the I items. In fact, conditional ML estimates of $\{\beta_{ij}\}$, based on conditioning on sufficient statistics for $\{\alpha_{jk}\}$, are identical to ML estimates of main effect parameters in the quasi-symmetry model (Conaway, 1989). Given that model (13) holds, marginal homogeneity in the J^I table is equivalent to $\beta_{1j} = \cdots = \beta_{Ij}$ for $j = 1,\ldots,J-1$. For this model, this special case corresponds to complete symmetry in the J^I cross-classification.

Among the most common tests of first-order marginal homogeneity in J^I contingency tables are the likelihood-ratio test and score test based on comparing the fits of the complete symmetry and quasi-symmetry models (Caussinus, 1966; Darroch, 1981). They are large-sample chi-squared tests, having $df = (I-1)(J-1)$. For a sample of N subjects, one can apply the score test statistics of Section 2 to the $I \times J \times N$ table $\{n_{ijk}\}$ in which n_{ijk} denotes the number of responses by subject k that fall in category j for item i. Each cell count in this contingency table is a 0 or 1, and $n_{i+k} = 1$ for all i and k.

When the response categories are nominal, one would use statistic (4), with $df = (I-1)(J-1)$, a statistic discussed by Darroch (1981). For $J = 2$, this statistic is algebraically identical to McNemar's statistic when $I = 2$ and Cochran's Q statistic when $I > 2$ (Somes, 1986). When the response categories are ordinal, one could use statistic (9), with $df = (I-1)$, to detect differences among the I marginal mean responses. For equally spaced scores, this provides a test of marginal homogeneity for the special case of (13) in which $\beta_{ij} - \beta_{i,j+1} = \beta_i$; such a common difference for all j corresponds to a constant item effect in an adjacent-categories logit model (Agresti, 1993). If the items themselves are ordered, one could use statistic (6), with $df = 1$, to improve power for detecting a trend in the mean response across the items, corresponding to a linear trend across items in β_i.

These large-sample chi-squared tests may be inadequate if some of the marginal counts are small. Agresti et al. (1992) presented alternative tests for large, highly sparse tables, which rely on Wald statistics using jackknife estimates of covariance matrices. Their approach is also an asymptotic one, and may behave poorly for small samples. In such cases, one could use the nearly exact versions of the tests described in Section 4, conditional on sufficient statistics for nuisance parameters in models such as (13). This approach takes some computational effort, because of the typically large number of strata. It has simple structure, however, and unlike other approaches, has guaranteed accuracy to a chosen level of precision regardless of the sample size.

Table 3
Results of tennis matches for women players in 1989–1990 season

Winner	Loser				
	Seles	Graf	Sabatini	Navratilova	Sanchez
Seles	—	2	1	3	2
Graf	3	—	6	3	7
Sabatini	0	3	—	1	3
Navratilova	3	0	2	—	3
Sanchez	0	1	2	1	—

We illustrate this approach to nearly exact testing of marginal homogeneity by conducting the test for Table 3, which summarizes results of matches among five women tennis players during the 1989–1990 season. For instance, Steffi Graf won 3 of the 5 matches that she and Monica Seles played. The quasi-symmetry model for these data is equivalent to the Bradley–Terry model (Fienberg and Larntz, 1976). The hypothesis of marginal homogeneity for this table is equivalent to identical player parameters in the Bradley–Terry model, or equal chances of a victory for each player in each pair. The likelihood-ratio test based on comparing the symmetry and quasi-symmetry models has a test statistic of 11.5 based on $df = 4$, for an asymptotic P-value of 0.021. Entering the data for the 46 matches as 92 observations in a $2 \times 5 \times 46$ cross-classification of outcome (winner, loser) by players by match results in a value for the score test statistic (4) of 10.6, for an asymptotic P-value of 0.031. The 99% confidence interval for the exact P-value for this test, based on 100 000 simulations, equals (0.024, 0.027). For these data, the asymptotic approaches perform adequately.

7. Comments

For small data sets with relatively few levels of the variables, exact methods can be conservative because of the high degree of discreteness of the test statistic. One could then instead use a mid P-value. The test is no longer 'exact,' but results can be trusted more than asymptotic results. Alternatively, one could use a modified P-value that maintains exactness but reduces conservativeness (Kim and Agresti, 1995). In practice, the conservativeness problem diminishes rapidly as n, I, J, and K increase.

Other approaches exist for approximating results of exact inferences. For instance, Baglivo et al. (1988) suggested a hybrid approach that combines exact and asymptotic techniques. Kolassa and Tanner (1994) used a double-saddlepoint approximation for components of the relevant conditional distribution, and then used Gibbs sampling to approximate the distribution of the statistic of interest. This approach is considerably more complex for the problem of testing conditional independence, and unlike simple Monte-Carlo, convergence to the exact result does not occur as the number of simulations increases. However, for more complex models, Gibbs sampling yields results for cases in which simple Monte-Carlo is difficult to apply. See

Diaconis and Sturmfels (1993) and Forster et al. (1996) for application of Markov chain Monte-Carlo methods to some complex models.

The first author has prepared a FORTRAN program, designed for UNIX work-stations, for conducting Monte-Carlo approximation of exact P-values for tests of conditional independence discussed in this article. This is available from the authors by e-mail or by sending a formatted $3\frac{1}{2}$ inch diskette.

Acknowledgements

This research was partially supported by a grant from the National Institutes of Health. The authors thank Dr. Sudeep Kundu and Merck Research Laboratories for permitting the use of the data given in Table 1 for this paper.

References

Agresti, A., A survey of exact inference for contingency tables, *Statist. Sci.*, **7** (1992) 131–177.

Agresti, A., Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters, *Scand. J. Statist.*, **20** (1993) 63–72.

Agresti, A. and A. Kezouh, Association models for multidimensional cross-classifications of ordinal variables, *Comm. Statist.*, **A 12** (1983) 1261–1276.

Agresti, A., J.B. Lang and C.R. Mehta, Some empirical comparisons of exact, modified exact, and higher-order asymptotic tests of independence for ordered categorical variables, *Comm. Statist. Simulation Comput.*, **22** (1993) 1–18.

Agresti, A., S. Lipsitz and J.B. Lang, Comparing marginal distributions of large, sparse contingency tables, *Comput. Statist. Data Anal.*, **14** (1992) 55–73.

Agresti, A., C.R. Mehta and N.R. Patel, Exact inference for contingency tables with ordered categories, *J. Amer. Statist. Assoc.*, **85** (1990) 453–458.

Agresti, A., D. Wackerly and J. Boyett, Exact conditional tests for cross-classifications: Approximation of attained significance levels, *Psychometrika*, **44** (1979) 75–83.

Andersen, A.H., Multidimensional contingency tables, *Scand. J. Statist.*, **1** (1974) 115–127.

Baglivo, J., D. Olivier and M. Pagano, Methods for the analysis of contingency tables with large and small counts, *J. Amer. Statist. Assoc.*, **83** (1988) 1006–1013.

Baglivo, J., D. Olivier and M. Pagano, Methods for exact goodness-of-fit tests, *J. Amer. Statist. Assoc.*, **87** (1992) 464–469.

Baglivo, J., D. Olivier and M. Pagano, Analysis of discrete data: Rerandomization methods and complexity, *Comput. Statist. Data Anal.*, **16** (1993) 175–184.

Birch, M.W., The detection of partial association II: The general case, *J. Roy. Statist. Soc. Ser. B*, **27** (1965) 111–124.

Caussinus, H., Contribution à l'analyse statistique des tableaux de correlation, *Ann. Fac. Sci. Univ. Toulouse*, **29** (1966) 77–182.

Clogg, C.C., Some models for the analysis of association in multiway cross-classifications having ordered categories, *J. Amer. Statist. Assoc.*, **77** (1982) 803–815.

Cohen, A. and H.B. Sackrowitz, An evaluation of some tests of trend in contingency tables, *J. Amer. Statist. Assoc.*, **87** (1992) 470–475.

Conaway, M., Analysis of repeated categorical measurements with conditional likelihood methods, *J. Amer. Statist. Assoc.*, **84** (1989) 53–62.

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.

Diaconis, P. and B. Sturmfels, Algebraic algorithms for sampling from conditional distributions, Tech. Report No. 430 (Dept. of Statistics, Stanford Univ., 1993).

Fienberg, S.E. and K. Larntz, Loglinear representation for paired and multiple comparison models, *Biometrika*, **63** (1976) 245–254.

Forster, J.J., J.W. McDonald and P.W.F. Smith, Monte Carlo exact conditional tests for log-linear and logistic models, *J. Roy. Statist. Soc. Sec. B*, **58** (1996) 445–453.

Goodman, L.A., Simple models for the analysis of association in cross-classifications having ordered categories, *J. Amer. Statist. Assoc.*, **74** (1979) 537–552.

Kim, D. and A. Agresti, Improved exact inference about conditional association in three-way contingency tables, *J. Amer. Statist. Assoc.*, **90** (1995) 632–639.

Kolassa, J.E. and M.A. Tanner, Approximate conditional inference in exponential families via the Gibbs sampler, *J. Amer. Statist. Assoc.*, **89** (1994) 697–702.

Kreiner, S., Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies, *Scand. J. Statist.*, **14** (1987) 97–112.

Kreiner, S., User Guide to DIGRAM, a program for discrete graphical modeling, Research Report 89/10 (Statistical Research Unit, Univ. of Copenhagen, 1989).

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.

Mantel, N., Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure, *J. Amer. Statist. Assoc.*, **58** (1963) 690–700.

Mehta, C.R., N.R. Patel and P. Senchaudhuri, Importance sampling for estimating exact probabilities in permutational inference, *J. Amer. Statist. Assoc.*, **83** (1988) 999–1005.

Morgan, W.M. and B.A. Blumenstein, Exact conditional tests for hierarchical models in multidimensional contingency tables, *Appl. Statist.*, **40** (1991) 435–442.

Patefield, W.M., An efficient method of generating random $R \times C$ tables with given row and column totals, *J. Roy. Statist. Soc. Ser. C*, **30** (1981) 91–97.

Pierce, D.A. and D. Peters, Practical use of higher order asymptotics for multiparameter exponential families, *J. Roy. Statist. Soc. Ser. B*, **54** (1992) 701–737.

Rasch, G., On general laws and the meaning of measurement in psychology, 321-333 in: J. Neyman (Ed.), *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, Vol. 4 (Univ. of California Press, Berkeley, 1961).

Somes, G.W., The generalized Mantel–Haenszel statistic, *Amer. Statist.*, **40** (1986) 106–108.

StatXact, *StatXact: Statistical Software for Exact Nonparametric Inference*, Version 2 (Cytel Software, Cambridge, MA, 1991).

Yao, Q. and D. Tritchler, An exact analysis of conditional independence in several 2×2 contingency tables, *Biometrics*, **49** (1993) 233–236.