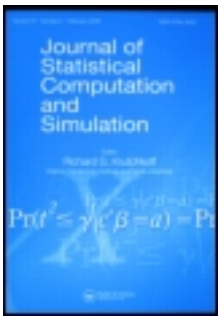

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

A likelihood ratio test of quasi-independence for sparse two-way contingency tables

Balgobin Nandram [a], Dilli Bhatta [a] & Dhiman Bhadra [b]

[a] Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, 01609, USA

[b] Production and Quantitative Methods Area, Indian Institute of Management, Ahmedabad, Gujarat, 380-015, India

Published online: 15 Jul 2013.

To cite this article: Journal of Statistical Computation and Simulation (2013): A likelihood ratio test of quasi-independence for sparse two-way contingency tables, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2013.815190

To link to this article: http://dx.doi.org/10.1080/00949655.2013.815190

Journal of Statistical Computation and Simulation, 2013
http://dx.doi.org/10.1080/00949655.2013.815190

Taylor & Francis
Taylor & Francis Group

A likelihood ratio test of quasi-independence for sparse two-way contingency tables

Balgobin Nandram[a]*, Dilli Bhatta[a] and Dhiman Bhadra[b]

[a]*Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, USA;* [b]*Production and Quantitative Methods Area, Indian Institute of Management, Ahmedabad, Gujarat 380-015, India*

(*Received 20 November 2012; final version received 11 June 2013*)

We consider a likelihood ratio test of independence for large two-way contingency tables having both structural (non-random) and sampling (random) zeros in many cells. The solution of this problem is not available using standard likelihood ratio tests. One way to bypass this problem is to remove the structural zeroes from the table and implement a test on the remaining cells which incorporate the randomness in the sampling zeros; the resulting test is a test of quasi-independence of the two categorical variables. This test is based only on the positive counts in the contingency table and is valid when there is at least one sampling (random) zero. The proposed (likelihood ratio) test is an alternative to the commonly used ad hoc procedures of converting the zero cells to positive ones by adding a small constant. One practical advantage of our procedure is that there is no need to know if a zero cell is structural zero or a sampling zero. We model the positive counts using a truncated multinomial distribution. In fact, we have two truncated multinomial distributions; one for the null hypothesis of independence and the other for the unrestricted parameter space. We use Monte Carlo methods to obtain the maximum likelihood estimators of the parameters and also the *p*-value of our proposed test. To obtain the sampling distribution of the likelihood ratio test statistic, we use bootstrap methods. We discuss many examples, and also empirically compare the power function of the likelihood ratio test relative to those of some well-known test statistics.

Keywords: chi-squared test; maximum likelihood estimators; Monte Carlo methods; quasi-independence; truncated multinomial distribution; zero counts

1. Introduction

Large two-way contingency tables almost always have cells with small and/or zero counts. Such tables frequently occur in survey problems where the categorical variables have many levels. For these tables, the use of the standard chi-squared test and the likelihood ratio test becomes problematic. Here, we assume that at least one of these zero counts is a random zero, and the test of independence becomes a test of quasi-independence of the two categorical variables. In this article, we show how to use the likelihood ratio test for sparse categorical tables in which many cells have observed zero counts.

Bishop et al. [1] considered 'incomplete' tables with zero cell counts known in advance; these are known as structural zeros. There are tables in which zero cells may vanish for larger sample sizes; such cells are called sampling zeros and they are random (i.e. they are not fixed in advance as

*Corresponding author. Email: balnan@wpi.edu

structural zeros). Structural zeros can be dealt with simply by dropping them from the standard test statistics although it is necessary to identify them. Therefore, the resulting test of independence becomes a test of quasi-independence (i.e. the test is run on the remaining cells).

However, sampling zeros must be taken into consideration, and they are more difficult to study because the standard likelihood ratio test does not exist. One way to avoid the difficulties associated with sampling zeros is to implement a test of quasi-independence of the two categorical variables while incorporating the randomness in the sampling zeros. This test is based only on the positive counts in the contingency table, thereby leading naturally to our approach which uses the truncated multinomial distribution for modelling the positive cell counts. Bishop et al. [1, Section 5.2.1] discuss the notion of quasi-independence for two-way contingency tables specifically for the case when there are structural zeros. In this article, we consider a modification of this definition since we do not ignore the sampling zeros.

In this article, we formulate a likelihood ratio test of quasi-independence for sparse categorical tables in which many cells have observed zero counts. Our test is sensible from a practical viewpoint as we do not need to know whether a cell is a structural or a sampling zero. However, our procedure requires at least one zero cell to be a sampling zero. This assumption is required because we need a nonzero probability on all zero cells (structural and sampling) combined into a single cell so that all positive cells have a total probability smaller than 1. The proposed test is based on the positive counts in the contingency table which are modelled using the truncated multinomial distribution. We do not consider the problem in which there is uncertainty about whether a specific cell is a structural zero or a sampling zero. For a mathematical treatment of this problem within the Bayesian paradigm, see Consonni and Pistone.[2] Because we are looking for the exact distribution of the likelihood ratio test statistic, our method allows us to deal simultaneously with the problem of sparse tables (expected cell counts smaller than 5).

There have been many proposals to correct for the sparseness of large tables; Ishii-Kuntz [3] provides a review of many strategies for dealing with zero cells in contingency tables. In practice, the usual choice is to drop these zero cells or add a small constant to all the cells. This latter choice is sensitive to the constant used. For example, adding 0.5 to every cell of the contingency table, as suggested by Goodman,[4] tends to make the sparse cells equally probable, thereby making the chi-squared statistic too conservative especially when the table density is less than 5, and this effect is very severe when there are many cells. See Agresti [5, p. 397] for additional discussion about adding 0.5 to the zero cells. There are other suggestions such as (a) add a small quantity such as 0.2 only to zero cells,[6] (b) add the reciprocal of the number of response categories [7] to the zero cells, (c) define zero divided by zero arbitrarily as zero [8] and (d) increase the sample size sufficiently to remove all zero cells.[9] Finally, Clogg and Eliason [10] suggested replacing sampling zeros by 0.000000001 or even a smaller number and then check the results against those obtained without such an adjustment or, as suggested by Agresti,[5, p. 397] perform a sensitivity analysis. While this is a minor over smoothing, it is still an ad hoc procedure. Recently, Beh and Farver [11] suggested adding 0.05 to the zero cells. In fact, adding 0.05, 0.5 or so to every cell of a table with many zero cells results in severe over smoothing.

An alternative to maximum likelihood estimation and chi-squared goodness-of-fit test is to use Fisher's exact test of independence.[5, Section 3.5] This test conditions on both margins (not naturally fixed though) of the two-way table leading to a hypergeometric distribution under independence for the free cell in a 2×2 table. The test is exact because the probabilities of the hypergeometric distribution can be calculated exactly under the hypothesis of independence. Larger two-way tables need Monte Carlo methods or asymptotic theory [1, Section 10.6] or special algorithms.[12] However, because the hypergeometric distribution is discrete it is not possible to obtain a specific significance level (e.g. .05); a randomization procedure is necessary. In fact, both Fisher's exact test and the chi-squared test are conservative because their actual significance level is below the nominal level.[13,14] Hashemi et al. [15] has a Bayesian approach using Bayes

factors which is conditional on only one margin of a 2×2 table. For a general two-way table and for non-ignorable missing data, Nandram [16] has a Bayesian method to test for independence when both margins are random; see also Nandram and Choi.[17] Fiser's exact test was critcized by Berkson [18] on the grounds that the marginal totals are informative. This leads to smaller variablility and hence smaller p-values. In addition, Fisher's exact test cannot be computed for large (more than 24 cells or so) because the workspace required for the network algorithm [19] on SAS 9.2 is currently prohibitive. Fisher's exact test can be computed with structural zeros but again only for small tables.[20] At the moment it is not possible to compute the p-value of Fisher's exact test for large tables with structural zeros and uncertainty about whether a cell is a structural or sampling zero partly because of the large workspace needed.

There is a growing body of research on general two-way tables, both categorical variables being ordinal. The key interest is to test for positive association of the two categorical variables under stochastic ordering. The intention is to generalize Fisher's exact test. This is a specialized type of literature of importance in its own right. Bartolucci et al. [21] consider both conditional and unconditional tests, Bartolucci and Scaccia [22] developed an exact conditional approach for certain forms of positive association using the multivariate generalized hypergeometric distribution, an extension of Fisher's exact test, and Bartolucci et al. [23] has a Bayesian approach in which the Bayes factor is used to test for independence under stochastic ordering. In this paper, we are considering a test of quasi-independence for two-way categorical tables in which the categories are nominal, not ordinal, and there are both structural and sampling zeros and uncertainty about which cells are structural or sampling zeros. This is itself an ubiquitous problem. In Appendix A we show how to extend our method to two-way tables in which both categorical variables are ordinal.

Log-linear models are commonly used to analyse multi-dimensional tables. Brown and Fuchs [24] have a good discussion of log-linear models for sparse contingency tables with zero counts; see also Baker et al. [25] for other views on structural and sampling zeros. Recently there has been increased activities in the use of log-linear models for multi-dimensional tables with sampling zeros. It has been shown that for certain patterns of sampling zeros maximum likelihood estimators (MLEs) may not exist for the parameters of the log-linear model. For an extensive theoretical discussion, see Fienberg and Rinaldo.[26,27]

There are difficulties in fitting log-linear models to tables with sampling zeros. Because the problem is caused by small samples, asymptotic theory, the backbone of log-linear models, does not help. However, log-linear models can be fitted to tables with structural zeros; see Christensen [28, Chapter 8] for a detailed discussion of this topic and note particularly the very informative discussion about how to deal with sampling zeros. Clogg and Eliason [10] has a very interesting discussion about some common problems in log-linear analysis and they also outlined strategies for dealing with them. For example, Clogg and Eliason [10] pointed out that there are difficulties in finding degrees of freedom when there are zero cells. They describe how to do so by calculating the difference between the effective dimension of the design matrix and the actual number of independent parameters estimated. Because we restrict attention to two-way categorical tables, which are highly applicable, it is not necessary to use log-linear models although we are considering large and sparse two-way tables. In fact, we provide an exact test which avoids asymptotic theory. It is indeed an open problem of how to analyse categorical tables with both structural and sampling zeros and even worst when it is not known with certainty which cells are structural zeros and which are sampling zeros. This is the problem we study for a $r \times c$ categorical table. In Appendix B we show how to extend our method to multi-way categorical tables.

The plan of the rest of the paper is as follows. Section 2 has a preliminary discussion and reviews many different test statistics, some of which are used routinely. In Section 3 we describe the likelihood ratio test for quasi-independence of the two categorical variables. In Section 4 we show how to compute the MLEs and obtain the distributions of the likelihood ratio statistic

using a Monte Carlo method. In Section 5 we present several examples with exact distributions and p-values, and we compare our likelihood ratio test with many other well-known tests. We also study the empirical power function of our likelihood ratio test statistic. Section 6 has concluding remarks.

2. Preliminary discussion

We consider the situation in which there are n individuals who are to be placed in k cells through an independent and identical sampling scheme. Clearly, the numbers of individuals falling in the k cells have a multinomial distribution. It is a standard practice to judge the sparseness of the table by the ratio $\eta = n/k$ with smaller values indicating more sparse tables; η is called the table density. Sparseness occurs in two-way tables when at least one categorical variable has several levels or the sample size, n, is small. In this article, we will consider sparse tables in which all marginal counts are positive.

Let $n_{ij},\ i = 1, \ldots, r,\ j = 1, \ldots, c$, denote the cell counts in a two-way contingency table. The marginal counts are given by $a_i = \sum_{j=1}^{c} n_{ij},\ i = 1, \ldots, r$, and $b_j = \sum_{i=1}^{r} n_{ij},\ j = 1, \ldots, c$. Note that, for a $r \times c$ contingency table, $k = rc$. Let π_{ij} denote the cell probabilities, where $\sum_{i=1}^{r}\sum_{j=1}^{c} \pi_{ij} = 1$, $p_i = \sum_{j=1}^{c} \pi_{ij}$ and $q_j = \sum_{i=1}^{r} \pi_{ij}$. The independence hypothesis states that $\pi_{ij} = p_i q_j,\ i = 1, \ldots, r,\ j = 1, \ldots, c$, where $\sum_{i=1}^{r} p_i = 1$ and $\sum_{j=1}^{c} q_j = 1$. Let $\hat{\lambda}_{ij} = a_i b_j / n$ denote the MLEs of the means of n_{ij} under the hypothesis of independence of the two categorical variables. Then, the Pearson chi-squared statistic is given by

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\lambda}_{ij})^2}{\hat{\lambda}_{ij}}$$

and the likelihood ratio statistic G^2 is given by

$$G^2 = 2 \sum_{ij} n_{ij} \log \left(\frac{n_{ij}}{\hat{\lambda}_{ij}} \right),$$

where, for both statistics, $\hat{\lambda}_{ij}$ are assumed positive. It is well known that Pearson X^2 and the likelihood ratio statistic G^2 have equivalent asymptotic chi-squared distributions with $(r-1)(c-1)$ degrees of freedom; the asymptotic approach is as n goes to infinity with $k = rc$ fixed. However, these results do not hold for sparse contingency tables; see Agresti [5, Section 7.7] and Bishop et al.[1, Chapter 5]

The accuracy of the chi-squared approximation depends on both n and k. The standard recommendation for reasonable performances of X^2 or G^2 is that the expected cell counts should be at least 5. When there are sampling zeros, strictly speaking, the likelihood ratio test cannot be carried out. For the Pearson chi-squared statistic, under independence, the expected cell counts are strictly positive, thereby increasing the value of the Pearson chi-squared statistic and leading to smaller p-values.

Cressie and Read [29] provided normal approximations to the null distributions of X^2 and G^2, and Read [30] studied the chi-squared approximations to X^2 and G^2. Koehler [31] and Koehler and Larntz [32] examined the accuracy of these approximations. The distribution of G^2 is usually poorly approximated by a chi-squared distribution when $\eta = n/k$ is less than 5; X^2 can be better approximated by a chi-squared distribution for smaller sample sizes compared to G^2. However, the chi-squared approximation tends to perform poorly for sparse tables containing both small and moderately large expected cell counts.

Cressie and Read [29] introduced a family of statistics called the power divergence statistics of which X^2 and G^2 are special cases; see also Read and Cressie [33] for an exhaustive discussion

on power divergence statistics. The power divergence statistic is given by

$$P^2 = \frac{2}{\varphi(\varphi+1)} \sum_{ij} n_{ij} \left\{ \left(\frac{n_{ij}}{\hat{\lambda}_{ij}} \right)^{\varphi} - 1 \right\}, -\infty < \varphi < \infty,$$

which is always positive and is defined as limits of P^2 at -1 and 0. Both the chi-squared test statistic, X^2, and the likelihood ratio statistic, G^2, are special cases of P^2 (i.e. $\varphi = 1$ gives the chi-squared test statistic and taking the limit as φ goes to zero gives the likelihood ratio test statistic). This is a very rich class of statistics and it contains many other statistics. One prominent statistic is the Freeman–Tukey statistic ($\varphi = -\frac{1}{2}$) which is

$$F^2 = 4 \sum_{ij} \left(\sqrt{n_{ij}} - \sqrt{\hat{\lambda}_{ij}} \right)^2$$

and a more popular one is the Cressie–Read statistic ($\varphi = 2/3$) which is

$$C^2 = \frac{9}{5} \sum_{ij} n_{ij} \left\{ \left(\frac{n_{ij}}{\hat{\lambda}_{ij}} \right)^{2/3} - 1 \right\}.$$

The idea is that by adjusting the simple chi-squared statistic we still get the same asymptotic chi-squared distribution for every member of the power divergence family of statistics. Cressie and Read [29] found that C^2 is less susceptible to the effects of sparseness than X^2 and G^2; see also Read and Cressie [33, Section 4.5] for recommendations. Even though there are conflicting recommendations regarding the choice of φ that results in the optimal statistic, in almost all cases they recommend a value of φ in $(-1, 2]$ and they state that when the sample size is larger than 10, C^2 is an excellent choice. In our investigations, we will study X^2, G^2, F^2 and C^2 which are all members of the power divergence class of statistics.

Garcia-Perez and Nunez-Anton [34] have recently studied seven members of the power divergence family including X^2 and G^2 using a simulation study. They found that the rate of convergence of the power divergence statistic depends on the parameter indexing the family. Among the seven members studied, they found that X^2 is the best up to a table density as low as 2; a moment correction increases the accuracy of X^2 for table density lower than 2. They showed that G^2 performs poorly and hence did not recommend it.

Another adjustment to the Pearson X^2 is given by Zelterman.[35] The adjustment is the D^2 statistic,

$$D^2 = \sum_{ij} \frac{\{(n_{ij} - \hat{\lambda}_{ij})^2 - n_{ij}\}}{\hat{\lambda}_{ij}}.$$

It is interesting that D^2 is not a member of the class of divergence statistics. Zelterman [35] also showed that when properly normalized, under the null hypothesis of independence, D^2 has an asymptotic standard normal distribution as both n and k approach infinity. Therefore, the asymptotic p-value for the test of independence is easy to obtain. It is true that for such tables, X^2 and D^2 are not equivalent and X^2 will reject the alternative hypothesis more often than D^2. Note that the contribution of $n_{ij} = 0$ to D^2 is the same as that of X^2 (i.e. $\hat{\lambda}_{ij}$ is added to each statistic).

Clearly, it will be better to obtain the exact distribution of the likelihood ratio statistic for the test of independence for sparse contingency tables to avoid the uncertainty associated with X^2, G^2, D^2, F^2 and C^2 and allied adjusted statistics. Our idea is to use the truncated multinomial distribution under no restriction (the whole space) and under the null hypothesis of independence. Letting k_0 denote the number of cells with structural and sampling zeros, the sparseness of the

Table 1. Summary of the characteristics of the six examples (E1–E6).

Example	Dimension	k	n	k_0	M	η	η_0
E1	11×24	264	279	186	22	1.06	3.58
E2	16×5	80	219	24	16	2.74	3.91
E3	6×28	168	539	54	21	3.21	4.73
E4	6×17	102	270	31	13	2.65	3.80
E5	4×4	16	21	4	3	1.31	1.75
E6	28×26	728	129	627	5	0.19	1.28

Note: n is the sample size; k is the number of cells; k_0 is the number of cells with structural and sampling zeros; M is the largest cell; the table density, $\eta_0 = n/(k - k_0)$, is larger than $\eta = n/k$.

table is given by $\eta_0 = n/(k - k_0)$, a considerably smaller sparseness than $\eta = n/k$ because k_0 can be substantial for sparse tables. For example, consider Table 1 on the six examples (E1–E6) that we discuss later. In E1, $n = 279, k = 264$ and $k_0 = 186$, and in E6, $n = 129, k = 728$ and $k_0 = 627$; in E1 and E6, k_0 is relatively large with respect to k. Here, the idea is that the sparseness of the table is reduced considerably and the test is essentially done only on the positive cells with all the structural and sampling zeros treated as a single nonnegative cell.

In Section 4 we present several examples with exact distributions, p-values, and a comparison of our proposed likelihood ratio test with X^2, G^2, D^2, F^2 and C^2. We also compare the empirical power function of our likelihood ratio test statistic with that of X^2, G^2, D^2, F^2 and C^2.

3. Likelihood ratio test

We denote the proposed likelihood ratio test statistic by T^2. In this section, we discuss how to obtain T^2, its exact distribution under the null hypothesis, and its p-value for a sparse $r \times c$ contingency table with sampling zeros. The likelihoods are obtained by assuming truncated multinomial distribution to accommodate the positivity restriction on the cell counts. Our procedure uses Monte Carlo methods to obtain the sampling distribution of the test statistic, and therefore, the p-value.

The null hypothesis of independence is given by $\pi_{ij} = p_i q_j,\ i = 1, \ldots, r,\ j = 1, \ldots, c$, and the alternative hypothesis is that there is at least one cell with $\pi_{ij} \neq p_i q_j$. There is no restriction on the whole space which is the union of these two hypotheses. We assume that the individuals respond independently in the $k = rc$ cells of the two-way table. Therefore, we always have the multinomial distribution albeit with the truncation restriction under the hypotheses. To account for the randomness of the sampling zeros, we assume that the nonzero cell counts are positive random variables. Next, we formulate the null hypothesis of quasi-independence.

Let $\mathcal{N}$ denote the set of positive cells of the two-way table. Also let $\mathcal{R}_i = \{(i,j) : n_{ij} > 0,\ j = 1, \ldots, c\}, i = 1, \ldots, r$ and $\mathcal{C}_j = \{(i,j) : n_{ij} > 0,\ i = 1, \ldots, r\},\ j = 1, \ldots, c$, where $\mathcal{R}_i$ is the set of positive counts in ith row, and $\mathcal{C}_j$ is the set of positive counts in jth column. Throughout inference is conditional on $\mathcal{N}$. It is worth noting that, unlike Fisher's exact test, our likelihood ratio test does not condition on the two margins of the two-way categorical table; it is only conditional on the set of positive cells.

Then the null hypothesis of quasi-independence is given by the restriction,

$$\pi_{ij} = \begin{cases} p_i q_j, & (i,j) \in \mathcal{N}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

In Equation (1), $p_i = \sum_{j \in \mathcal{R}_i} \pi_{ij},\ i = 1, \ldots, r$, and $q_j = \sum_{i \in \mathcal{C}_j} \pi_{ij},\ j = 1, \ldots, c$. Finally, it is worth noting that $\sum_{i=1}^{r} p_i = \sum_{j=1}^{c} q_j = 1$ and under the null hypothesis of quasi-independence,

we have

$$0 < \sum_{(i,j)\in\mathcal{N}} \pi_{ij} = \sum_{(i,j)\in\mathcal{N}} p_i q_j < 1, \sum_{i=1}^{r}\sum_{j=1}^{c} \pi_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} p_i q_j = 1.$$

We now construct the truncated multinomial distributions for the positive cells under the unrestricted parameter space and the restricted parameter space (i.e. the null hypothesis). Under the unrestricted parameter space, we have

$$p(\underset{\sim}{n} \mid \underset{\sim}{\pi}) = n! \prod_{(ij)\in\mathcal{N}} \frac{1}{n_{ij}!} \left\{ \frac{\pi_{ij}}{\sum_{(i,j)\in\mathcal{N}} \pi_{ij}} \right\}^{n_{ij}}, n_{ij} > 0, \sum_{(i,j)\in\mathcal{N}} n_{ij} = n, \tag{2}$$

and under the restricted parameter space, the corresponding distribution is given by

$$p(\underset{\sim}{n} \mid \underset{\sim}{p}, \underset{\sim}{q}) = n! \prod_{(ij)\in\mathcal{N}} \frac{1}{n_{ij}!} \left\{ \frac{p_i q_j}{\sum_{(i,j)\in\mathcal{N}} p_i q_j} \right\}^{n_{ij}}, n_{ij} > 0, \sum_{(i,j)\in\mathcal{N}} n_{ij} = n. \tag{3}$$

Equations (2) and (3) reflect/elaborate our assumption of representing all the 0 cells with a single positive cell. Then, the structural and sampling zeros represent a small number of individuals, say $n_0 \geq 0$. Thus, the effective sample size is $N = n + n_0$, which is a latent variable. It is easy to see this because we can augment the probability mass function of $\underset{\sim}{n}$ to include N or n_0 as follows. From Equation (2) we have

$$p(\underset{\sim}{n}, N \mid \underset{\sim}{\pi}) = (N-1)! n \left\{ \prod_{(ij)\in\mathcal{N}} \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!} \right\} \frac{(1 - \sum_{(ij)\in\mathcal{N}} \pi_{ij})^{N-n}}{(N-n)!}, N \geq n, \tag{4}$$

and from Equation (3) we have

$$p(\underset{\sim}{n}, N \mid \underset{\sim}{p}, \underset{\sim}{q}) = (N-1)! n \left\{ \prod_{(ij)\in\mathcal{N}} \frac{(p_i q_j)^{n_{ij}}}{n_{ij}!} \right\} \frac{\{1 - \sum_{(ij)\in\mathcal{N}} (p_i q_j)\}^{N-n}}{(N-n)!}, N \geq n. \tag{5}$$

Thus, Equations (4) and (5) are presumably convenient to perform an expectation-maximization (EM) algorithm to obtain the MLEs of the π_{ij}. This is a standard data augmentation scheme when the EM algorithm is used.

From Equation (4) or Equation (5), it is easy to deduce that the marginal probability mass function of N is negative binomial. Specifically, from Equation (2), we have, $p(\underset{\sim}{n}, N \mid \underset{\sim}{\pi}) = p(\underset{\sim}{n} \mid N, \underset{\sim}{\pi}) p(N \mid \underset{\sim}{\pi})$, where $N \mid \underset{\sim}{\pi} \sim \mathrm{NB}(n, \sum_{(ij)\in\mathcal{N}} \pi_{ij})$ with *NB* denoting the negative binomial probability mass function, and from Equation (3), we have, $p(\underset{\sim}{n}, N \mid \underset{\sim}{p}, \underset{\sim}{q}) = p(\underset{\sim}{n} \mid N, \underset{\sim}{p}, \underset{\sim}{q}) p(N \mid \underset{\sim}{p}, \underset{\sim}{q})$, where $N \mid \underset{\sim}{p}, \underset{\sim}{q} \sim \mathrm{NB}(n, \sum_{(ij)\in\mathcal{N}} p_i p_j)$. Note that, under this formulation, both $\sum_{(ij)\in\mathcal{N}} \pi_{ij}$ and $\sum_{(ij)\in\mathcal{N}} p_i q_j$ are assumed to be strictly smaller than unity. That is, all zero cells are assumed to have a positive total probability, an assumption that is typically made (e.g. [2]). Clearly, this assumption is sensible if at least one zero cell is a sampling zero; we believe that adding a small positive constant to a structural zero cell is not sensible.

However, note that Equations (4) and (5) are never used in our development. So one does not need to know what n_0 or N is. That is, throughout we work with only Equations (2) and (3).

Since the logarithm of the likelihood ratio test statistic is $T^2 = -2[\ln\{p(\underset{\sim}{n} \mid \underset{\sim}{\hat{p}}, \underset{\sim}{\hat{q}})\} - \ln\{p(\underset{\sim}{n} \mid \underset{\sim}{\hat{\pi}})\}]$, where $\underset{\sim}{\hat{p}}$ and $\underset{\sim}{\hat{q}}$ are the MLEs under Equation (3), and $\underset{\sim}{\hat{\pi}}$ is the maximum likelihood

estimator under Equation (2), we have

$$T^2 = 2 \sum_{(ij)\in\mathcal{N}} n_{ij} \left[\ln \left\{ \frac{\hat{\pi}_{ij}}{\sum_{(i,j)\in\mathcal{N}} \hat{\pi}_{ij}} \right\} - \ln \left\{ \frac{\hat{p}_i \hat{q}_j}{\sum_{(i,j)\in\mathcal{N}} \hat{p}_i \hat{q}_j} \right\} \right]. \tag{6}$$

Thus, T^2 is obtained once $\hat{p}$, $\hat{q}$, and $\hat{\underset{\sim}{\pi}}$ are computed. It is worth noting that T^2 has $(r-1)(c-1)-(k_0-1)$ degrees of freedom (i.e. k_0-1 less than the number without the assumption of sampling zeros), where as defined earlier, k_0 is the number of cells with sampling zeros. We do not rely on asymptotic theory and we provide the exact distribution of T^2. Thus, our likelihood ratio test is an exact test. The challenge to find the distribution of the likelihood ratio test statistic is to obtain the MLEs of the parameters under these scenarios.

4. Computations

The key issue that remains is how to compute the MLEs. Once this is done, it is easy to use a Monte Carlo method to obtain the sampling distribution of T^2 and hence the p-value for the test of independence.

It is difficult to obtain the MLEs of the π_{ij} in model (2) or those of the p_i and q_j in model (3) using numerical optimization. However, one might use the EM algorithm [36] in Equation (4) or Equation (5) to obtain the required MLEs; see Appendix C for a description of the EM algorithm in our context. This is also not easy because the EM algorithm is very sensitive to the starting values in this problem.

Therefore, we consider a Monte Carlo method to obtain the MLEs. Our strategy is to convert the likelihood functions to posterior densities using prior distributions that leave the likelihood function unchanged. Then random samples can be drawn from these posterior densities.

In Equation (2) we take

$$\left(\pi_{ij}, (i,j) \in \mathcal{N}, 1 - \sum_{(i,j)\in\mathcal{N}} \pi_{ij} \right) \sim \text{Dirichlet}(1,\ldots,1).$$

Similarly, in Equation (3) we take $\underset{\sim}{p}$ and $\underset{\sim}{q}$ to be independent with

$$\underset{\sim}{p} \sim \text{Dirichlet}(1,\ldots,1), \underset{\sim}{q} \sim \text{Dirichlet}(1,\ldots,1).$$

Thus, using these 'uniform' priors in Equation (2) or Equation (3) and applying Bayes' theorem, we will get the required posterior densities which are exactly the same as the likelihood functions. This is a standard result in Bayesian statistics.

We can now draw samples of $\underset{\sim}{\pi}$ and $(\underset{\sim}{p}, \underset{\sim}{q})$ from these posterior densities. With these samples, we can then find the posterior mode which is, in fact, the MLE. Under the whole parameter space, the joint posterior density is given by

$$p(\underset{\sim}{\pi} \mid \underset{\sim}{n}) \propto \prod_{(ij)\in\mathcal{N}} \left\{ \frac{\pi_{ij}}{\sum_{(i,j)\in\mathcal{N}} \pi_{ij}} \right\}^{n_{ij}}, \ \pi_{ij} > 0, \quad \sum_{(ij)\in\mathcal{N}} \pi_{ij} < 1, \tag{7}$$

and under the restricted parameter space the joint posterior density is given by

$$p(\underset{\sim}{p}, \underset{\sim}{q} \mid \underset{\sim}{n}) \propto \prod_{(ij)\in\mathcal{N}} \left\{ \frac{p_i q_j}{\sum_{(i,j)\in\mathcal{N}} p_i q_j} \right\}^{n_{ij}}, \ p_i, q_j > 0 \sum_{i=1}^{r} p_i = 1, \quad \sum_{j=1}^{c} q_j = 1. \tag{8}$$

It is not easy to draw samples from Equation (7) or Equation (8) directly because these posterior densities are intractable. One can use approximate Bayesian computation,[37] but this is not

necessary in our case. It turns out that we can simply use Monte Carlo optimization; see Robert and Casella [38, Chapter 5] for details. This latter procedure requires random samples from the posterior density given in Equation (7).

The Monte Carlo optimization is advantageous because it will find the global maximum value in the case when there are several local maxima. This is different from standard numerical optimization routines which can converge to a local maximum value, and this is erroneous. For example, the Nelder–Mead and Newton–Raphson algorithms can converge to a local maximum value, and so these algorithms will not be feasible for our purpose. Moreover, these algorithms may not be efficient when the number of parameters in the likelihood function is large, something that is true in our case; see Table 1.

Fortunately, it is easy to draw samples from the posterior density in Equation (7) indirectly using random draws, not Markov chains (e.g. a Metropolis–Hastings sampler). This is advantageous because it ensures non-repeated sample values with probability one, and no convergence monitoring is required.

We string out the $\pi_{ij}, (i,j) \in \mathcal{N}$, and denote the new vector by $\underset{\sim}{\gamma} = (\gamma_1, \ldots, \gamma_S)'$, where S is the cardinality of $\mathcal{N}$. Similarly, we string out the $n_{ij}, (i,j) \in \mathcal{N}$, and denote the new vector by $\underset{\sim}{t} = (t_1, \ldots, t_S)'$. Then, as in Equation (7),

$$p(\underset{\sim}{\gamma} \mid \underset{\sim}{t}) \propto \prod_{s=1}^{S} \left\{ \frac{\gamma_s}{\sum_{s=1}^{S} \gamma_s} \right\}^{t_s}, \ \gamma_s > 0, \quad \sum_{s=1}^{S} \gamma_s < 1. \tag{9}$$

We need to draw $\underset{\sim}{\gamma}$ from Equation (9) which can be done using random sample generation techniques.

Now, we make the transformation $\tau_s = \gamma_s / \sum_{s=1}^{S} \gamma_s$, $s = 1, \ldots, S-1$, and $\Gamma = \sum_{s=1}^{S} \gamma_s$. Then, we have $\gamma_s = \tau_s \Gamma$, $s = 1, \ldots, S-1$ and $\gamma_S = (1 - \sum_{s=1}^{S-1} \tau_s)\Gamma$. It is now easy to show that the Jacobin of the transformation is Γ^{S-1}. Therefore, the joint posterior density function of $\underset{\sim}{\tau}$ and Γ will be

$$p(\underset{\sim}{\tau}, \Gamma \mid \underset{\sim}{t}) \propto \left(\prod_{s=1}^{S-1} \tau_s^{n_s} \right) \left(1 - \sum_{s=1}^{S-1} \tau_s \right)^{n_S} \Gamma^{n_S + S - 1}. \tag{10}$$

Then, $\underset{\sim}{\tau}$ and Γ in Equation (10), given $\underset{\sim}{t}$, are independent with

$$\left(\tau_1, \ldots, \tau_{S-1}, 1 - \sum_{s=1}^{S-1} \tau_s \right) \mid \underset{\sim}{t} \sim \text{Dirichlet}(n_1 + 1, \ldots, n_S + 1), \Gamma \mid \underset{\sim}{t} \sim \text{Beta}(n_S + S, 1), \tag{11}$$

and it is therefore easy to get samples of τ_s, $s = 1, \ldots, S-1$, and Γ from Equation (11). It is worth noting that the posterior densities of both $\underset{\sim}{\tau}$ and $\underset{\sim}{t}$ are unimodal. Retransforming, we get back samples of γ_s, $s = 1, \ldots, S$, and in turn, $\pi_{ij}, (i,j) \in \mathcal{N}$.

However, it is much more difficult to draw samples from Equation (8). Fortunately, the samples drawn from Equation (7) can be converted to samples of p_i, $i = 1, \ldots, r$, and q_j, $j = 1, \ldots, c$, from Equation (8). This is true because it is important to note that we can move from Equations (2) to (3) via the transformation $\pi_{ij} = p_i q_j, (i,j) \in \mathcal{N}$. That is, $p_i = \sum_{j=1}^{c} \pi_{ij}$, $i = 1, \ldots, r$, and $q_j = \sum_{i=1}^{r} \pi_{ij}$, $j = 1, \ldots, c$; recall that $\pi_{ij} = 0$, $(i,j) \notin \mathcal{N}$. Thus, if we can draw samples from the posterior density (7), we have samples from the posterior density (8) automatically.

We denote the generated samples by $\pi_{ij}^{(h)}, (i,j) \in \mathcal{N}$, $h = 1, \ldots, M$, and $p_i^{(h)}$, $i = 1, \ldots, r$, $q_j^{(h)}$, $j = 1, \ldots, c$, $h = 1, \ldots, M$. In our numerical examples, we have used $M = 5000$, a very conservative choice (i.e. somewhat smaller values of M are mostly adequate). We have computed the likelihood functions under the independence hypothesis and the whole parameter space and have picked the largest one in each situation. Note that, from the manner in which the sample

values are drawn, they are different with probability one. Let $\hat{\pi}_{ij}, (i,j) \in \mathcal{N}$, denote the maximum likelihood estimates corresponding to the whole space and $\hat{p}_i,\ i = 1,\ldots,r,\ \hat{q}_j,\ j = 1,\ldots,c$, denote the corresponding maximum likelihood estimates under the hypothesis of independence. Hence, we have obtained a value of the likelihood ratio statistic T^2 in Equation (6).

To obtain the sampling distribution of T^2, we need to bootstrap the $r \times c$ categorical table a large number of times. For each bootstrap sample (BS), we calculate the value of T^2. We can bootstrap under the hypothesis of independence to get the sampling distribution under independence. To obtain the p-value of the test, we simply calculate the proportion of the bootstrap values which are larger than the observed T^2 value. We can also compute the sampling distribution in the whole parameter space in a similar manner. The sampling distribution of T^2 is obtained using the Parzen–Rosenblatt density estimator on the sample values of T^2. In our case, we generate 1000 bootstrap values of T^2.

Finally, we describe how to bootstrap the observed table. To do so, we maintain the sampling zeros in the observed data, and fill in the positive cells. Note that some small positive cells can become sampling zeros in the simulated data. Thus, the degrees of freedom is also a random variable. Let $n_{ij}^{(o)}, (i,j) \in \mathcal{N}$, denote the observed counts with $\sum_{(i,j)\in\mathcal{N}} n_{ij}^{(o)} = n^{(0)}$. We want to bootstrap $n_{ij}^{(o)}, (i,j) \in \mathcal{N}$; we denote the bootstrapped values by $n_{ij}, (i,j) \in \mathcal{N}$.

Let $u_j = \sum_{i\in\mathcal{C}_j} n_{ij}^{(o)}/n^{(o)},\ j = 1,\ldots,c$. We draw $(a_1,\ldots,a_c) \sim \text{Multinomial}(n^{(o)}, \underline{u})$ until $a_j > 0,\ j = 1,\ldots,c$. To draw samples from the whole space, we define $t_{ij} = n_{ij}^{(o)} / \sum_{i\in\mathcal{C}_j} n_{ij}^{(o)}$, $(i,j) \in \mathcal{N}$. Next, we draw $\underline{n}_j \overset{\text{ind}}{\sim} \text{Multinomial}(a_j, \underline{t}_j), j = 1,\ldots,c$ until $\sum_{j\in\mathcal{R}_i} n_{ij} > 0, i = 1,\ldots,r$. Samples under the independence hypothesis are drawn in a similar manner. We define $t_i = \sum_{j\in\mathcal{R}_i} n_{ij}^{(o)}/n^{(o)},\ i = 1,\ldots,r$, and draw $\underline{n}_j \overset{\text{ind}}{\sim} \text{Multinomial}(a_j, \underline{t})$ until $\sum_{j\in\mathcal{R}_i} n_{ij} > 0,\ i = 1,\ldots,r$. We repeat the process to obtain 1000 draws of the simulated contingency tables in each case.

5. Numerical analysis

In this section, we present the details of the numerical and computational procedures we have carried out for illustrative purposes. Specifically, in Section 5.1 we describe six numerical examples and in Section 5.2 we present a power comparison of X^2, G^2, D^2, F^2, C^2 and T^2.

5.1. *Numerical examples*

We use the following six examples for illustrative purposes and to make comparisons of the different tests. Recall that our likelihood ratio test requires at least one of the zero cells to be a sampling zero and this is true in each of the six examples.

E1: This example is on a multiple response problem that motivates our research. Bilder and Loughin [39] presented a joint table of results from a Kansas farmer survey categorizing four waste storage methods and five sources of veterinary information.

E2: Researchers at the University of Florida classified the stomach contents of alligators into five categories: Fish, invertebrate, reptile, bird and other. We have reformulated the data into two categorical variables by crossing lake, sex and size to form one categorical variable. However, in this example conclusions are problematic because this two-way table comes from a collapsed table. Agresti [5] used these data to illustrate how to fit multivariate logit models.

E3: These data are based on a classification of body mass index and family income for a US state in which the data are obtained from the third National Health and Nutrition Examination Survey.

E4: This example deals with the classification of the number of graduate degrees granted by year and degree type (PhD, AM, MAS, MAF, MAI, MME) in the Department of Mathematical sciences at Worcester Polytechnic Institute. These data were made available by the secretary of the said department.

E5: This example is based on an Inventory Management Programme. It is a standard practice to put the complete stock of 1120 T-shirts into the market on February 1 and inventory is taken every Saturday thereafter for 52 weeks ending on January 31 of the next year. This data set categorizes the number of sales of T-shirts by four sizes and four styles after the first week of sales on 8 February 2003.

E6: This data set summarizes the results of a survey of women employed as mathematicians or statisticians to relate years since bachelors degree and monthly salary. Zelterman [35] used this contingency table as an illustrative example.

These tables are quite large and sparse. We present a summary of the characteristics of the six data sets in Table 1. In E4, we are sure that some zero cells are structural zeros, but this information is not important to perform our likelihood ratio test. Thus, for convenience we treat all six data sets equally and a test of independence of the two categories is needed. It is useful to note that for E6, the table density ($\eta = 0.19$) makes it difficult for the standard Pearson X^2 or G^2 to cope; also $\eta = 1.31$ for E5. However, note that when sampling zeros are treated as a combined positive cell, the table densities are much larger.

Table 2 lists the p-values corresponding to the different tests of the examples. Except for G^2 and T^2, for all examples we have used the complete tables (i.e. all cells are included). In E2 and E4 there are virtually no differences among the tests but for E1, E3, E5 and E6 there are significant differences. In E1 T^2 agrees with X^2, G^2 and D^2 but differs from F^2 and C^2. In E3 T^2 agrees with D^2 and C^2 but not X^2, G^2 and F^2. In E5 T^2 agrees with all tests except F^2. In E6 T^2 agrees with all tests except G^2 and D^2. We note that while E1 and E6 have large numbers of zeros, E2 has just 24 zeros.

We have looked at the p-values for incomplete tables (i.e. without the cells with zero counts) for X^2, F^2, C^2. We can not do so for D^2 because it was developed for sparse complete tables. We noticed important differences in X^2, F^2 and C^2 for the complete and the incomplete tables except for E2 and E4. In E1 the p-values with F^2 and C^2 are now .00002 and .00000. In E3 F^2 gives a p-value of .58534 and C^2 .00044. In E5 F^2 has a p-value of .76705. The differences are more dramatic in E6. Without the zero cells in E6 for X^2, F^2 and C^2 the p-values are all .00000; now it is very different from T^2 but similar to G^2 and D^2. As indicated by Garcia-Perez and Nunez-Anton,[34] X^2 does not perform well when $\eta < 2$. Researchers have considered further adjustments to the Pearson X^2 statistic; see, for example, Davis [40] who obtained an improved approximation to the distribution of X^2 by using its first three moments, but this is still an approximation. The act

Table 2. P-values for the test of independence by example.

Example	X^2	G^2	D^2	F^2	C^2	T^2
E1	.00015	.00000	.00002	.11582	.44235	.00000
E2	.00021	.00000	.00000	.00000	.00030	.00151
E3	.02748	.00043	.11490	.03369	.32231	.21923
E4	.00374	.00000	.00012	.00000	.00412	.04269
E5	.41613	.07370	.55551	.02737	.40498	.81782
E6	.23801	.00000	.00020	.99999	.99999	.96922

Note: X^2 is the Pearson chi-squared statistic, G^2 is the likelihood ratio statistics, D^2 is the adjusted Pearson chi-squared statistic, F^2 is the Freeman–Tukey statistic, C^2 is the Cressie–Read statistic and T^2 is the new likelihood ratio statistic.

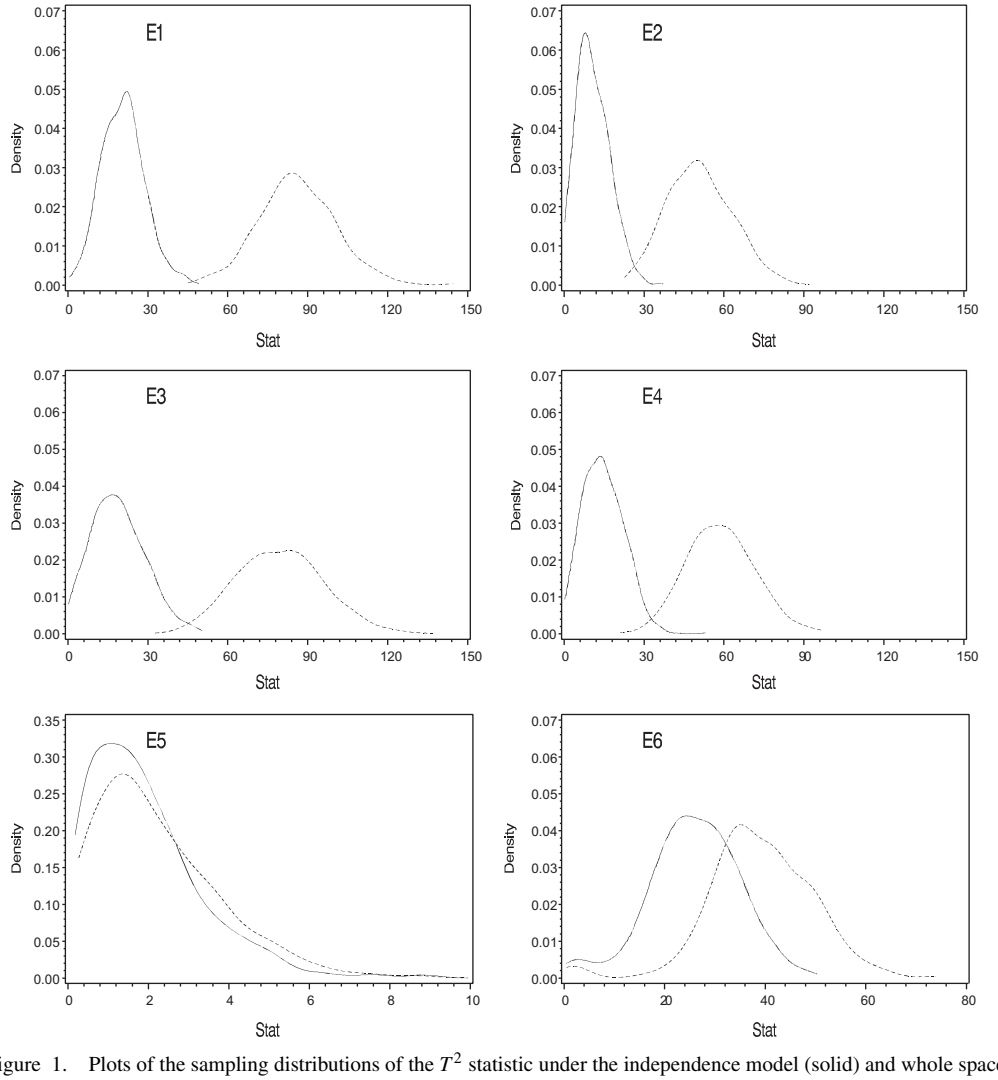

Figure 1. Plots of the sampling distributions of the T^2 statistic under the independence model (solid) and whole space (dotted) by example.

of replacing an incomplete table by a complete table will increase the degrees of freedom and the observed value of the test statistics because expected values are now nonnegative instead of zeros.

Figure 1 shows the sampling distributions of the proposed likelihood ratio test statistic T^2 under the independence hypothesis and the unrestricted parameter space. It is clear that the distributions are not symmetric and hence not normal. Also, it is apparent that T^2 does not follow a chi-squared distribution. We note that in all six examples, the two sampling distributions are very different; specifically the unrestricted distributions are to the right of the restricted distributions. This is very clear for E1, E2, E3, E4. In E5 there are minor differences between the two distributions and in E6 there is significant overlap between the two sampling distributions. As expected, the unrestricted distributions have more spread than the restricted ones. Therefore, it is clear that the proposed likelihood ratio test can discriminate very well between the null hypothesis of independence and the alternative hypothesis.

Table 3. Actual computation times of the three parts of the likelihood ratio test for the six examples.

Example	Dimension	*Real time*				
		n	LR	Boot	C	tt
E1	11×24	279	1.98	7:34.51	0.08	7.610
E2	16×5	219	1.53	5:31.53	0.08	5.552
E3	6×28	539	2.72	10:53.88	0.08	10.945
E4	6×17	270	2.02	6:38.44	0.08	6.676
E5	4×4	21	1.39	2:37.69	0.08	2.653
E6	28×26	129	2.73	9:43.11	0.08	9.765

Note: *n* is the sample size; LR refers to the time (seconds) to compute the observed likelihood ratio test statistc (two MLE's); Boot is the time to draw the 1000 BSs (the entries $a : b.cd$ means a minutes and $b.cd$ seconds, e.g. for E1 it took 7 min and 34.51 s to draw the BSs); *C* is the time (minutes) taken to count the number of bootstrap values which are at least as large as observed test statistic (p-value); and *tt* is the total time (minutes) to complete all the computations which were done on our 850 MHz computer.

We have recorded the time to do the computation using our likelihood ratio test. On our 850 MHz station the computation time to obtain the p-value of the likelihood ratio test has three parts which are (a) computation of the observed likelihood ratio test statistic, (b) drawing the 1000 BSs and (c) counting the number of BSs with the value of the test statistic at least as large as the observed value. In Table 3 we have presented these three times and the total time for all six examples. There are two things to observe. First, the bulk of the computation time is taken up to draw the BSs and the times taken to compute the observed test statistic and to do the counting are negligible. Second, as expected, the time to draw the BSs depend on the size of the contingency table. The total time ranges from 2.7 min for E5 (a small table) to 10.9 min for E3 (a large table).

We have performed a stability analysis on our procedure. There are two sample sizes we need to select. These are the sample size for the stochastic optimization (SO) to obtain the MLEs and the sample size for the BS to do the Monte Carlo integration (counting) for the p-value. It turns out that larger sample sizes are needed for SO than for BS. Thus, we have taken the SO sample size to be 5000 and the BS sample size to be 1000, a reasonable specification. So we discuss stability with respect to the SO sample size. We give an illustration for E5; the same procedure works for all examples. For SO with sample sizes of 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000 the p-values are respectively .923, .927, .834, .779, .797, .812, .790, .807, .805. Clearly, then for SO sample sizes larger than about 3000 there is a high degree of stability. We have found similar patterns for the other five examples.

We have also looked at the EM algorithm more carefully. We observed that the MLEs are very sensitive to the specifications of the starting values of the EM algorithm. We believe that the difficulty arises because there is no real information about N, the latent variable which represents the total sample size. Expectation is taken over N in the EM algorithm. Motivated by the difficulty in using the EM algorithm, we have integrated N (an important innovation) to perform the SO. This provides a stable algorithm.

We have applied Fisher's exact test to our six examples. On SAS 9.2 Fisher's exact test uses the network algorithm [19] which needs large workspace for even moderately large tables like the ones we are discussing. With roughly 24 or more cells (e.g. a 4×6 or a 3×8 contingency table) Fisher's exact test cannot be computed. Also note that SAS does not account for structural zeros; all zeros are treated as sampling zeros. For E5, a 4×4 contingency table with a sample size of 21, SAS gives a p-value of .58170. As expected, because of the conservative nature of Fisher's exact test this p-value should be smaller than the p-value from T^2 which is .81782. For the other five examples, SAS responded,'WARNING: Computing exact p-values for the problem may require much time and memory. Press system interrupt key to terminate exact computations.' We now understand that for large tables even without structural zeros the computation of the p-value for Fisher's exact test is an open problem. One possibility is to develop a Monte Carlo method.

5.2. *Power comparison*

In this section, we use the data from E4 to compare the power functions of the four statistics, X^2, G^2, F^2, C^2, D^2 and T^2. For the power comparison any of the six examples can be used; it does not matter that E4 is a somewhat artificial example.

To describe departures from the null hypothesis of independence, we consider a mixture distribution under the alternative hypothesis. We assume that the distribution under the alternative hypothesis is multinomial with cell probabilities $\pi_{ij} = (1-\omega)\tilde{p}_i\tilde{q}_j + \omega\tilde{\pi}_{ij}$, where $\tilde{p}_i$, $\tilde{q}_j$ and $\tilde{\pi}_{ij}$ are obtained from the observed contingency table. Here, $\omega = 0$ corresponds to the null hypothesis distribution and $\omega = 1$ corresponds to the alternative hypothesis of dependence. Therefore, values of ω close to zero give local alternatives, and larger values of ω give larger departures from the null hypothesis.

Letting v_α denote the critical value of an upper tailed test of size α, the power function is given by

$$P\{V > v_\alpha \mid \underset{\sim}{n} \sim \text{Multinomial}(n, \underset{\sim}{\pi})\}, \tag{12}$$

where V is one of the six test statistics (X^2, G^2, D^2, F^2, C^2 and T^2). Here n is the observed sample size for E4. The data are generated in the same way as described before for calculating the p-values; the only difference being that, in this case, the data are generated from a Multinomial$(n, \underset{\sim}{\pi})$ distribution in Equation (12).

Table 4. Comparison of the power functions for Example 4 of X^2, G^2, D^2, F^2, C^2 and T^2 by the test size, α, and the mixture coefficient, ω, in the distribution under the alternative hypothesis.

α	ω	X^2	G^2	D^2	F^2	C^2	T^2
.050	0.0	$.050_{.002}$	$.050_{.002}$	$.050_{.002}$	$.050_{.002}$	$.050_{.002}$	$.050_{.002}$
	0.1	$.053_{.002}$	$.059_{.002}$	$.063_{.002}$	$.056_{.002}$	$.055_{.002}$	$.057_{.002}$
	0.2	$.075_{.003}$	$.092_{.003}$	$.104_{.003}$	$.086_{.003}$	$.085_{.003}$	$.092_{.003}$
	0.3	$.145_{.004}$	$.182_{.004}$	$.212_{.004}$	$.161_{.004}$	$.177_{.004}$	$.171_{.004}$
	0.4	$.283_{.005}$	$.354_{.005}$	$.404_{.005}$	$.303_{.005}$	$.337_{.005}$	$.334_{.005}$
	0.5	$.512_{.005}$	$.604_{.005}$	$.655_{.005}$	$.529_{.005}$	$.583_{.005}$	$.540_{.005}$
	0.6	$.760_{.004}$	$.835_{.004}$	$.868_{.003}$	$.771_{.004}$	$.820_{.004}$	$.745_{.004}$
	0.7	$.935_{.002}$	$.964_{.002}$	$.974_{.002}$	$.939_{.002}$	$.958_{.002}$	$.901_{.003}$
	0.8	$.991_{.001}$	$.997_{.001}$	$.998_{.000}$	$.992_{.001}$	$.995_{.001}$	$.967_{.002}$
.025	0.0	$.025_{.002}$	$.025_{.002}$	$.025_{.002}$	$.025_{.002}$	$.025_{.002}$	$.025_{.002}$
	0.1	$.028_{.002}$	$.031_{.002}$	$.034_{.002}$	$.030_{.002}$	$.029_{.002}$	$.035_{.002}$
	0.2	$.041_{.002}$	$.054_{.002}$	$.060_{.002}$	$.045_{.002}$	$.047_{.002}$	$.059_{.002}$
	0.3	$.085_{.003}$	$.120_{.003}$	$.138_{.003}$	$.095_{.003}$	$.110_{.003}$	$.114_{.003}$
	0.4	$.192_{.004}$	$.259_{.004}$	$.296_{.005}$	$.202_{.004}$	$.240_{.004}$	$.250_{.004}$
	0.5	$.391_{.005}$	$.497_{.005}$	$.537_{.005}$	$.408_{.005}$	$.470_{.005}$	$.442_{.005}$
	0.6	$.654_{.005}$	$.760_{.004}$	$.797_{.004}$	$.664_{.005}$	$.738_{.004}$	$.663_{.005}$
	0.7	$.883_{.003}$	$.938_{.002}$	$.951_{.002}$	$.892_{.003}$	$.927_{.003}$	$.849_{.004}$
	0.8	$.979_{.001}$	$.992_{.001}$	$.995_{.001}$	$.984_{.001}$	$.990_{.001}$	$.944_{.002}$
.010	0.0	$.010_{.001}$	$.010_{.001}$	$.010_{.001}$	$.010_{.001}$	$.010_{.001}$	$.010_{.001}$
	0.1	$.010_{.001}$	$.013_{.001}$	$.015_{.001}$	$.011_{.001}$	$.012_{.001}$	$.014_{.001}$
	0.2	$.014_{.001}$	$.026_{.002}$	$.031_{.002}$	$.020_{.001}$	$.023_{.002}$	$.025_{.002}$
	0.3	$.033_{.002}$	$.063_{.002}$	$.076_{.003}$	$.044_{.002}$	$.057_{.002}$	$.060_{.002}$
	0.4	$.086_{.003}$	$.166_{.004}$	$.191_{.004}$	$.113_{.003}$	$.149_{.004}$	$.147_{.004}$
	0.5	$.218_{.004}$	$.362_{.005}$	$.413_{.005}$	$.272_{.004}$	$.338_{.005}$	$.300_{.005}$
	0.6	$.464_{.005}$	$.650_{.005}$	$.687_{.005}$	$.527_{.005}$	$.615_{.005}$	$.525_{.005}$
	0.7	$.744_{.004}$	$.887_{.003}$	$.909_{.003}$	$.810_{.004}$	$.864_{.003}$	$.748_{.004}$
	0.8	$.928_{.003}$	$.983_{.001}$	$.987_{.001}$	$.960_{.002}$	$.976_{.002}$	$.890_{.003}$

Note: The distribution under the alternative hypothesis is multinomial with cell probabilities $(1-\omega)\tilde{p}_i\tilde{q}_j + \omega\tilde{\pi}_{ij}$, where $\tilde{p}_i$, $\tilde{q}_j$ and $\tilde{\pi}_{ij}$ are obtained from the observed contingency table. In each entry the notation a_b means that a is the estimate of the power and b is the standard error of the estimate.

The critical value is obtained by taking the $100(1-\alpha)$th percentile point of the test statistics from the data generated for the p-values. We also use a method based on the Parzen–Rosenblatt density estimator to find the critical value; however, the two answers are very similar. Since the power function in (12) is a function of $\omega \in [0, 1]$, we vary ω in this range to study it. For each value of ω, we generate 1000 contingency tables, and for each table we calculate the six test statistics (X^2, G^2, D^2, F^2, C^2 and T^2). Then, we obtain the proportion of values of each of the test statistics exceeding its critical value. We have repeated the entire process 10 times to get 10 critical values (p-values) and 10 estimated powers at each value of ω. We actually have 10 estimated power functions for each test; these curves are very close. However, for each test, our final estimated power function is obtained by averaging the 10 estimated powers at each of the selected values of ω. We also computed the overall standard error by combining the 10 standard errors at each value of ω.

In Table 4, we present the power of each of the six statistics at values of $\omega = .00, .10, \ldots, .80$, for three test sizes, .050, .025, and .010. The highest power function corresponds to D^2; this is followed in rough order by G^2, C^2, T^2, F^2 and X^2. For strong dependence, X^2 tends to have higher

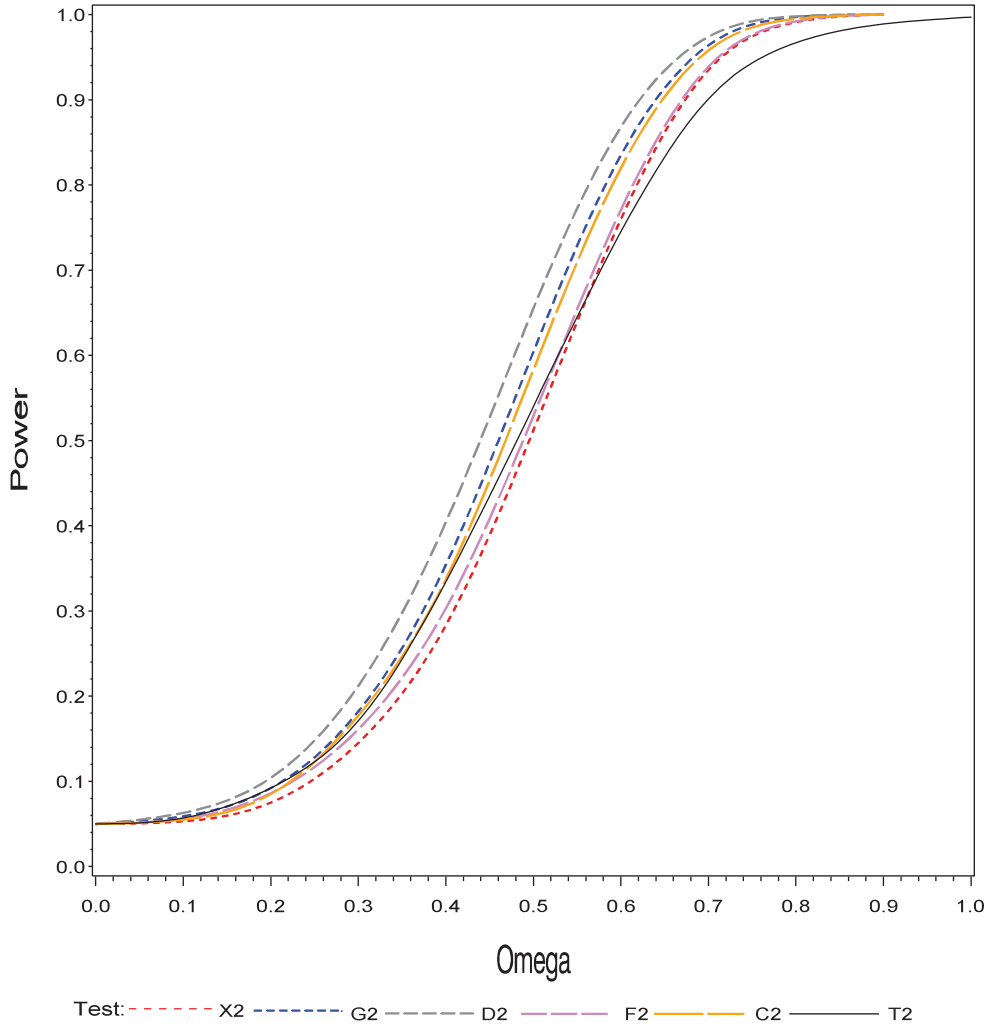

Figure 2. Plots of the estimated power functions of the six tests for E4, $\alpha = .050$.

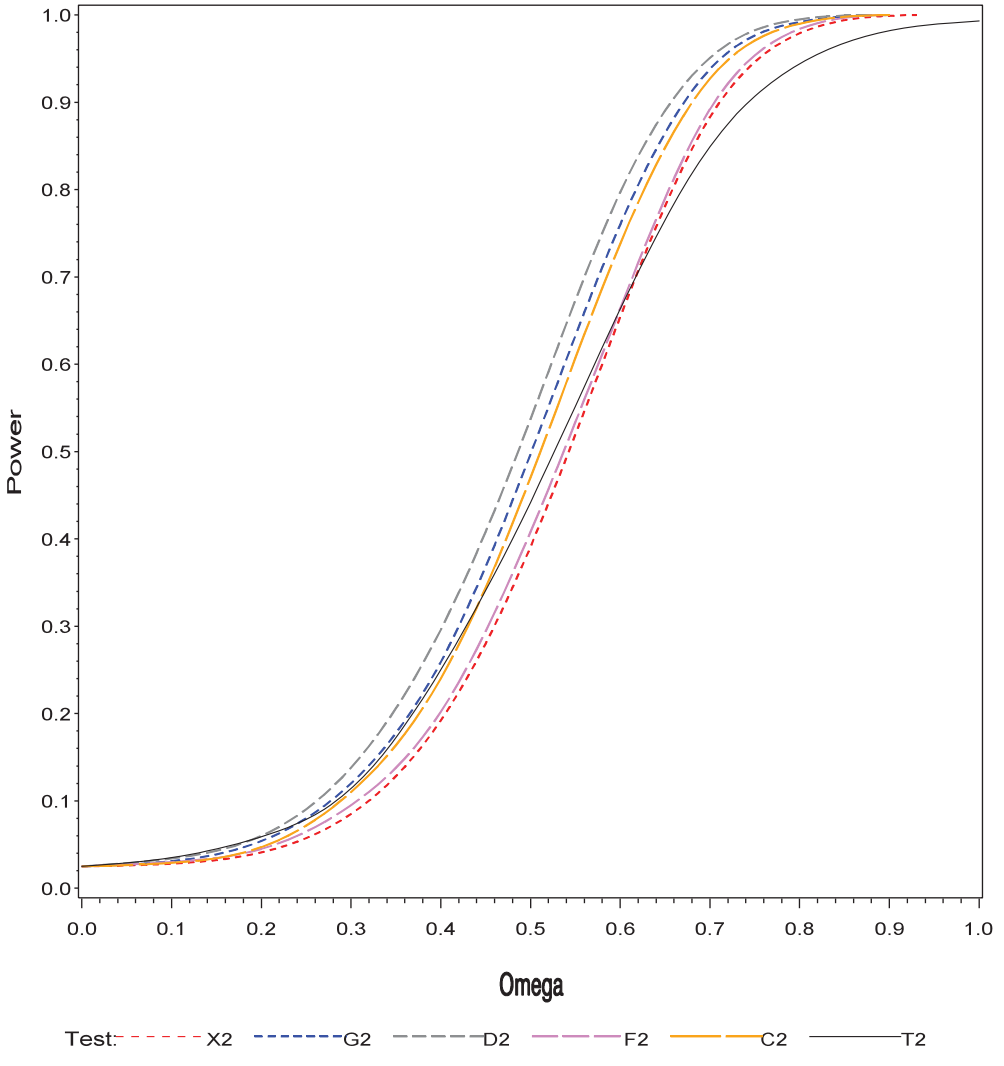

Figure 3. Plots of the estimated power functions of the six tests for E4, $\alpha = .025$.

power than T^2. For weak dependence, the power functions of these tests are very similar, but for at least moderate dependence, D^2, C^2 and G^2 have higher powers. For example, at size .025 for $\omega = 0.2, X^2, G^2, D^2, F^2, C^2$ and T^2 have powers .041, .054, .060, .045, .047 and .059, respectively; for all these cases, the numerical standard errors are smaller than .002. Also at size .025 for $\omega = .6$, X^2, G^2, D^2 and T^2 have powers of .654, .760, .797, .664, .738 and .663, respectively; for all these cases, the numerical standard errors are smaller than .005. As seen in Table 4 similar patterns are observed for sizes $\alpha = .05, .01$.

We have plotted the estimated power functions of the six test statistics for E4 in Figures 2–4. The curves are smooth because of the averaging we have done on the 10 estimated powers. The curve for D^2 is the highest, followed by G^2, C^2, F^2 then T^2 which almost always is higher that the curve for X^2. This is true for all three sizes (.050, .025, .010) we have considered.

The natural question that comes up is, 'Why does the test based on T^2 appear not as the best?' First, the test based on G^2 is inappropriate because an observed zero cell count does not have any contribution regardless of whether it is a sampling zero or a structural zero. That is, G^2 does not

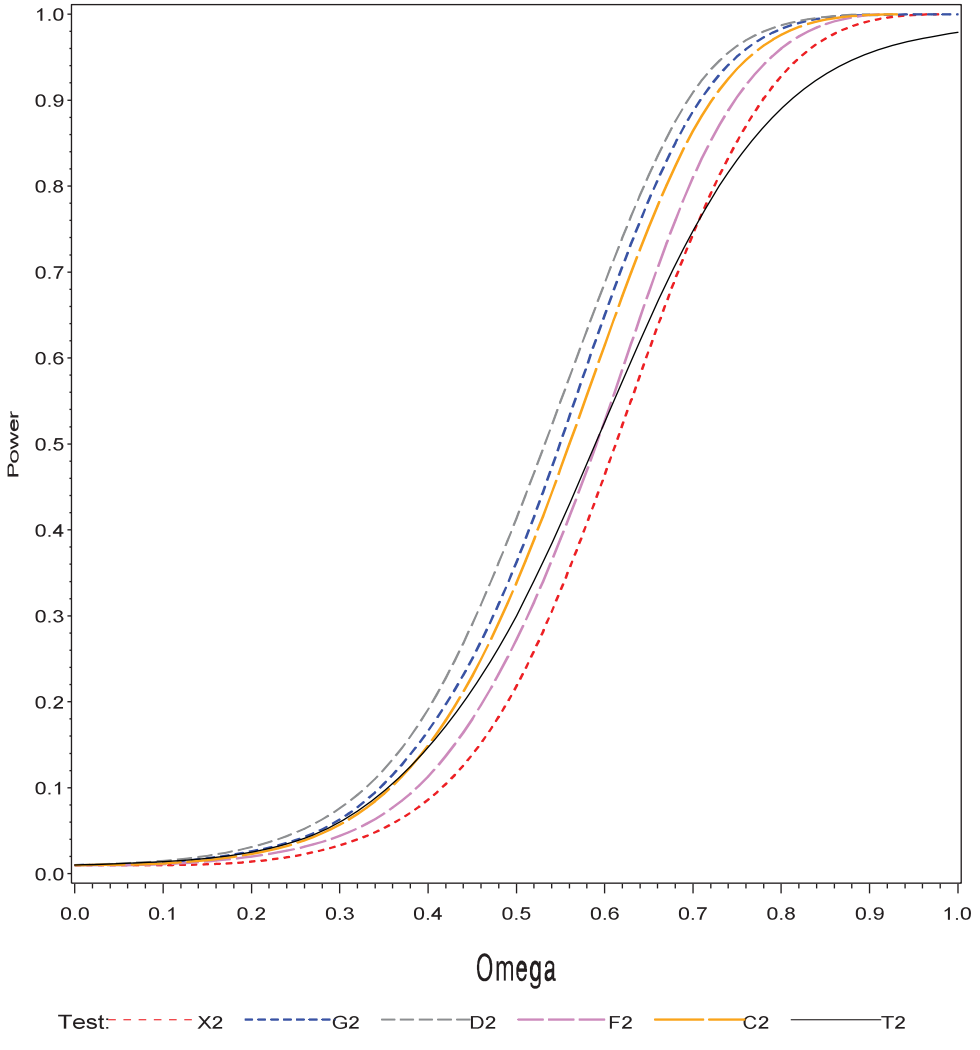

Figure 4. Plots of the estimated power functions of the six tests for E4, $\alpha = .010$.

account for randomness of the sampling zeros. The proposed likelihood ratio test based on T^2 is actually a corrected version of G^2 for contingency tables with both sampling and structural zeros. We are, therefore, left with X^2, F^2, C^2 and D^2; D^2 was constructed specifically for sparse tables.

According to Zelterman,[35] 'Under sequences of local alternative hypotheses the test based on D^2 exhibits moderate power when the X^2 test is biased.' He also stated that

> With a large, sparse multinomial distribution where n and k are both large, X^2 and D^2 will usually behave as normal random variables with means and variances that are unrelated to the chi-squared distribution. In the sparse distribution, X^2 and D^2 are not equivalent and X^2 will accept the null hypothesis too often under certain alternative hypotheses.

It is true that X^2 and D^2 treat the zeros exactly the same (i.e. when the (i,j)th cell is zero, the same value $\hat{\lambda}_{ij}$ under independence is added to X^2 or D^2). The benefit of D^2 is that, when the cell counts are positive, D^2 helps to remove the skewness in X^2, so that normality is reasonable under independence.

Both F^2 and C^2 are reasonable statistics with higher power than T^2. They can be used with both sampling zeros and structural zeros. The trouble here is that the expected value of a zero cell is

always nonnegative (mostly positive) and this increases the degrees of freedom and the values of the test statistics, which can lead to much larger p-values than without the zero cells as in E6. This is also true for the X^2 statistic. It is also true that the structural zeros can be eliminated from the table. However, when there is uncertainty about whether zeros are structural or sampling zeros, none of these tests can be used.

So, although the power comparison is a guide for choosing the optimal test and D^2 has the highest power, we prefer T^2 because it is the one that actually accounts for the uncertainty about sampling zeros and structural zeros and it corrects for the deficiency in G^2. Moreover, the test based on T^2 is a test in which the null hypothesis is that of quasi-independence, not independence as in standard contingency tables. When there are many sampling zeros, X^2, G^2, F^2, C^2 and D^2 tend to give overly strong evidence against the null hypothesis.

Therefore, T^2 is a very useful statistic, and computation involved to find its p-value is worth the effort; it takes less than 12 min to compute the p-value or the power of T^2 on our 850 MHz computer. It takes slightly less time to find the p-values of the other tests when the 'exact' distributions (not asymptotic distributions) are used.

6. Concluding remarks

We have considered a likelihood ratio test for quasi-independence in large two-way categorical tables which are likely to have both structural and sampling zeros. Our procedure requires at least one sampling zero, and it is competitive to some alternatives. One draw back of our procedure is that it is a conditional (on the set of positive cells) test but it is not conditional on the margins as in Fisher's exact test. It is also true that T^2 is an exact test and it does not rely on asymptotic theory.

We have shown how to obtain the exact sampling distribution of the likelihood ratio test statistic under the independence hypothesis, and have therefore obtained its exact p-value. We calculated both the MLEs and the distribution of the likelihood ratio statistic using Monte Carlo methods. While D^2 has higher power than X^2, G^2, F^2, C^2 and T^2, we believe that the test based on T^2 is most appropriate for the problem with many structural and sampling zeros and when there is uncertainty about which cells are structural zeros or sampling zeros (i.e. cells with zero counts are combined into a single positive cell and one does not have to know the count in this cell).

It is true that our procedure resembles a Bayesian method. However, the Bayesian method requires calculation of the Bayes factor [41] which is sensitive to prior specifications. This is also a computationally intensive procedure. An alternative is to use Bayesian estimation procedure which is the main strength of Bayesian statistics. Nandram and Choi [17] has a possible procedure which we plan to extend to two-way contingency tables with structural and sampling zeros.

In many problems, some cells are sampling zeros and others are structural zeros. However, when there is uncertainty about whether a specific cell is a sampling zero or a structural zero, an important practical problem arises. One would then need to incorporate information about which cells are more likely to be sampling zeros. For example, in some tables, one piece of information is that the zeros 'near' the positive cells are likely to be sampling zeros, not structural zeros. Thus, one would need to use a Bayesian approach (e.g. [2]).

Finally, we note that we have considered a simple random sample in this article. We plan to extend this work to more complex sample designs (e.g. two-stage cluster sampling), where the sample schemes must be taken into consideration for a test of independence; p-values obtained under the assumption of simple random sampling will be too small if the data are obtained under cluster sampling. Similar problems exist for stratified random sampling. This is a much more complex problem and it is one of our on-going activities. However, Nandram et al. [42] presents a Bayesian approach using surrogate sampling.

Acknowledgements

We are grateful to the two reviewers whose comments help to improve the manuscript enormously. We are also thankful to the editor for his encouragement.

References

[1] Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis. Cambridge, MA: MIT Press; 1975.
[2] Consonni G, Pistone, G. Algebraic Bayesian analysis of contingency tables with possibly zero-probability cells. Stat Sinica. 2007;17:1355–1370.
[3] Ishii-Kuntz M. Ordinal log-linear models. Sage University series on quantitative applications in the social sciences, Newbury Park, CA: Sage; 1994.
[4] Goodman LA. The multivariate analysis of association in cross-classifications having ordered categories. J Am Stat Ass. 1970;74:320–334.
[5] Agresti A. Categorical data analysis. 2nd ed. New York: Wiley; 2002.
[6] Evers M, Namboodiri NK. On the design matrix strategy in the analysis of categorical data. In: Schuessler KF, editor. Sociological methodology 1979. San Francisco, CA: Jossey Bass; 1979. p. 86–111.
[7] Grizzle JE, Starmer CF, Koch GC. Aanalysis of categorical data linear models. Biometrics. 1969;26:489–504.
[8] Fienberg SE. The analysis of cross-classified categorical data. 2nd ed. Cambridge: MIT Press; 1980.
[9] Knocke D, Burke PJ. Log-linear models. Sage University Paper Series on quantitative applications in the social sciences, Series No. 07-020, Beverly Hills, CA: Sage; 1980.
[10] Clogg CC, Eliason SR. Some common problems in log-linear analysis. Sociol Methods Res. 1987;16(1):8–44.
[11] Beh EJ, Farver, TB. An evaluation of non-iterative methods for estimating the linear-by-linear parameter of ordinal log-linear models. Aust New Zealand J Stat. 2009;51:335–352.
[12] Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J Am Stat Assoc. 1983;78:427–434.
[13] D'Agostino RB, Chase W, Belanger A. The appropriateness of some common procedures for testing the equality of two independent binomial populations. Am Statistician. 1988;42:198–202.
[14] Upton GJG. A comparison of alternative tests for 2×2 comparative trial. J R Stat Soc Ser A. 1982;145:86–105.
[15] Hashemi L, Nandram B, Goldberg R. Bayesian analysis for a single 2×2 table. Stat Med. 1997;16:1311–1328.
[16] Nandram, B. Bayesian inference of the cell probabilities of a two-way categorical table under non-ignorability. Commun Stat – Theory Methods. 2009;38(16):3015–3030.
[17] Nandram B, Choi JW. Alternative tests of independence in two-way categorical tables. J Data Sci. 2007;5:217–237.
[18] Berkson J. In dispraise of fisher's exact test: do the marginal totals of the 2×2 table contain relevant information respecting the table proportions. J Stat Plan Inference. 1978;2:27–42.
[19] Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in $I \times I$ contingency tables. J Am Stat Assoc. 1983;78:427–434.
[20] West LJ, Hankin RKS. Exact tests for two-way contingency tables with structural zeros. J Stat Softw. 2008;28:1–19.
[21] Bartolucci F, Forcina, A, Dardoni V. Positive quadrant dependence and marginal modeling in two-way tables with ordered margins. J Am Stat Assoc. 2001;96:1497–1505.
[22] Bartolucci F, Scaccia L. Testing for positive association in contingency tables with fixed margins. Comput Stat Data Anal. 2004;47:195–210.
[23] Bartolucci F, Scaccia L, Farcomeni A. Bayesian inference through encompassing priors and importance sampling for a class of marginal models for categorical data. Comput Stat Data Anal. 2012;56:4067–4080.
[24] Brown MB, Fuchs C. On maximum likelihood estimation in sparse contingency tables. Comput Stat Data Anal. 1983;1:3–15.
[25] Baker RJ, Clarke MRB, Lane PW. Zero entries in contingency tables. Comput Stat Data Anal. 1985;3:33–45.
[26] Fienberg SE, Rinaldo A. Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. J Stat Plan Inference. 2007;137:3430–3445.
[27] Fienberg SE, Rinaldo A. Maximum likelihood estimation in log-linear models. Ann Stat. 2012;40:996–1023.
[28] Christensen R. Log-linear models and logistic regression. New York: Springer; 1990.
[29] Cressie N, Read TRC. Multinomial goodness-of-fit tests. J R Stat Soc Ser B. 1984;46:440–464.
[30] Read TRC. Small-sample comparisons for power divergence goodness-of-fit statistics. J Am Stat Assoc. 1984;79:929–935.
[31] Koehler KJ. Goodness-of-fit tests for log-linear models in sparse contingency tables. J Am Stat Assoc. 1986;81: 483–493.
[32] Koehler KJ, Larntz, K. An empirical investigation of goodness-of-fit statistics for sparse multinomial. J Am Stat Assoc. 1980;75:483–493.
[33] Read TRC, Cressie N. Goodness-of-fit statistics for discrete multivariate data. New York: Springer; 1988.
[34] Garcia-Perez MA, Nunez-Anton V. Accuracy of the power-divergence statistics for testing independence and homogeneity in two-way contingency tables. Commun Stat – Simul Comput. 2009;38:503–512.
[35] Zelterman, D. Goodness-of-fit tests for large sparse multinomial distributions. J Am Stat Assoc. 1987;82:624–629.
[36] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B. 1977;39:1–22.

[37] Blum MGB. Approximate Bayesian computation: a nonparametric perspective. J Am Stat Assoc. 2010;105: 1178–1187.
[38] Robert CP, Casella G. Monte Carlo statistical methods. New York: Springer; 1999.
[39] Bilder CR, Loughin TM. Testing for marginal independence between two categorical variables with multiple responses. Biometrics. 2004;60:241–248.
[40] Davis CS. A new approximation to the distribution of the Pearson's chi-square. Statist Sinica. 1993;3:189–196.
[41] Kass RE, Raftery AE. Bayes factor. J Am Stat Assoc. 1995;90:773–795.
[42] Nandram B, Bhatta DR, Sedransk J, Bhadra D. A Bayesian test of independence in a two-way contingency table using surrogate sampling. J Stat Plan Inference. 2013;143:1392–1408. Available from http://dx.doi.org/10.1016/j.jspi.2013.03.011.
[43] McCullagh P. Regression models for ordinal data (with discussion). J R Stat Soc Ser B. 1980;42(2):109–142.

Appendix 1. Ordinal $r \times c$ categorical tables

We consider a $r \times c$ categorical table and the case in which both categorical variables are ordinal. The case of one ordinal variable is simpler. We show how to extend our method to two ordinal variables; our method does not work immediately but the overall principle is similar. We use a method for ordinal data, similar to McCullagh [43] which treats the response categories as contiguous intervals on a continuous scale with unknown cut-points.

Denoting the first categorical variable by I and the second by J, we assume that

$$P(I = i) = p_i \quad \text{and} \quad P(J = j \mid I = i) = q_{j\,|\,i},\ i = 1, \ldots, r,\ j = 1, \ldots, c,$$

where the levels of the categorical variables are not necessarily equally spaced or numeric. Let $\gamma_i = \sum_{i'=1}^{i} p_{i'}$ and $\delta_{j\,|\,i} = \sum_{j'=1}^{j} q_{j'\,|\,i},\ i = 1, \ldots, r,\ j = 1, \ldots, c$. The cell counts are $n_{ij},\ i = 1, \ldots, r, j = 1, \ldots, c$, with $\sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = n$, the sample size. We take the general (unrestricted) model to be

$$\begin{aligned} \ln\left(\frac{\gamma_i}{1-\gamma_i}\right) &= \theta_i, \quad i = 1, \ldots, r-1, \\ \ln\left(\frac{\delta_{j\,|\,i}}{1-\delta_{j\,|\,i}}\right) &= \frac{\phi_j - \alpha_i}{\beta_i}, \quad i = 1, \ldots, r,\ j = 1, \ldots, c-1, \end{aligned} \tag{A1}$$

where $-\infty < \alpha_i < \infty, \beta_i > 0$ and for identifiability we take $\alpha_r = 0$ and $\beta_r = 1$. The cut-points are $-\infty = \theta_0 < \theta_1 < \cdots < \theta_{r-1} < \theta_r = \infty$ and $-\infty = \phi_0 < \phi_1 < \cdots < \phi_{c-1} < \phi_c = \infty$.

It follows from Equation (A1) that

$$p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} - \frac{e^{\theta_{i-1}}}{1+e^{\theta_{i-1}}}, \quad i = 1, \ldots, r$$

and

$$q_{j\,|\,i} = \frac{e^{(\phi_j - \alpha_i)/\beta_i}}{1+e^{(\phi_j - \alpha_i)/\beta_i}} - \frac{e^{(\phi_{j-1} - \alpha_i)/\beta_i}}{1+e^{(\phi_{j-1} - \alpha_i)/\beta_i}}, \quad i = 1, \ldots, r, j = 1, \ldots, c.$$

Thus, without any restriction on the two categorical variables, we have

$$\pi_{ij} = p_i q_{j\,|\,i} = \left\{\frac{e^{\theta_i}}{1+e^{\theta_i}} - \frac{e^{\theta_{i-1}}}{1+e^{\theta_{i-1}}}\right\} \left\{\frac{e^{(\phi_j - \alpha_i)/\beta_i}}{1+e^{(\phi_j - \alpha_i)/\beta_i}} - \frac{e^{(\phi_{j-1} - \alpha_i)/\beta_i}}{1+e^{(\phi_{j-1} - \alpha_i)/\beta_i}}\right\} \tag{A2}$$

and under independence in Equation (A2), $q_{j\,|\,i} = q_j$ where q_j is obtained by setting $\alpha_i = 0,\ \beta_i = 1,\ i = 1, \ldots, r$ (i.e. $\pi_{ij} = p_i q_j$ under independence). For completeness we define $\eta_j = \sum_{j'=1}^{j} q_{j'}$ and $\ln\{\eta_j/(1-\eta_j)\} = \phi_j,\ j = 1, \ldots, c$.

Again, letting $\mathcal{N}$ denote the set of positive cells, the likelihood functions, analogous to Equations (2) and (3), are easy to write down. But now under the null hypothesis the likelihood is a function of $\underset{\sim}{\theta}$ and $\underset{\sim}{\phi}$ and without any restriction it is a function of $\underset{\sim}{\theta}$, $\underset{\sim}{\phi}$, $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$. Thus, the likelihood ratio test can be obtained. The only remaining issue is how to obtain the MLEs.

We transform the parameters to the interval $(0, 1)$ and use independent uniform priors on all the parameters. Thus, the joint posterior densities under the two hypotheses are proper. Then, to draw samples, we simply use a Gibbs sampler with grids dividing $(0, 1)$ into 100 subintervals or so. Again, we only need to sample from the joint posterior under the alternative hypothesis to obtain $\underset{\sim}{\theta}$, $\underset{\sim}{\phi}$, $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$, and therefore, $\pi_{ij}, (i, j) \in \mathcal{N}$ using Equation (A2). Then, samples of p_i and q_i are obtained as we have done for the $r \times c$ categorical table. Thus, under the null hypothesis we can generate samples of θ_i and ϕ_j from the respective joint posterior density and, therefore, the MLEs using SO.

In the same spirit, our procedure can be applied to multi-way tables. As an example, for a $r \times c \times \ell$ contingency table, denoting the first categorical variable by I, the second by J and the third by T, we assume that

$$P(I = i) = p_i, P(J = j \mid I = i) = q_{j\mid i} \quad \text{and} \quad P(T = t \mid I = i, J = j) = u_{t\mid i,j},$$

where $i = 1, \ldots, r$, $j = 1, \ldots, c$, $t = 1, \ldots, \ell$, and again the levels of the categorical variables are not necessarily equally spaced or numeric. In principle, one can proceed in a similar manner as for two-way tables, but this will need further research.

Appendix 2. multi-dimensional tables

Our method can be extended to multi-way tables. For simplicity, consider a three-way table ($r \times c \times \ell$). We have cell probabilities π_{ijk}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, $k = 1, \ldots, \ell$ and cell counts n_{ijk}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, $k = 1, \ldots, \ell$, where $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{\ell}\pi_{ijk} = 1$ and $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{\ell} n_{ijk} = n$, the sample size. We want to test the null hypothesis of independence, which is

$$\pi_{ijk} = p_i q_j t_k, \quad i = 1, \ldots, r, j = 1, \ldots, c, k = 1, \ldots, \ell,$$

where $\sum_{i=1}^{r} p_i = 1$, $\sum_{j=1}^{c} q_j = 1$ and $\sum_{k=1}^{\ell} t_k = 1$. Again, we assume that all margins are positive and there may be many zero cells with at least one sampling zero. We also use the standard assumption of multinomial sampling.

Let $\mathcal{N}$ denote the set of positive cells. Then, the null hypothesis of quasi-independence is given by the restriction,

$$\pi_{ijk} = \begin{cases} p_i q_j t_k, & (i,j,k) \in \mathcal{N} \\ 0 & \text{otherwise.} \end{cases} \tag{A3}$$

Also let $\mathcal{R}_i = \{(i,j,k) : n_{ijk} > 0,\ j = 1, \ldots, c,\ k = 1, \ldots, \ell\}$, $i = 1, \ldots, r$, $\mathcal{C}_j = \{(i,j,k) : n_{ijk} > 0,\ i = 1, \ldots, r,\ k = 1, \ldots, \ell\}$, $j = 1, \ldots, c$, and $\mathcal{T}_k = \{(i,j,k) : n_{ijk} > 0,\ i = 1, \ldots, r,\ j = 1, \ldots, c\}$, $k = 1, \ldots, \ell$, where $\mathcal{R}_i$ is the set of positive counts in ith plane ($j \times k$), $\mathcal{C}_j$ is the set of positive counts in jth plane ($i \times k$) and $\mathcal{T}_k$ is the set of positive counts in kth plane ($i \times j$). In Equation (A3) $p_i = \sum_{(j,k)\in\mathcal{R}_i}\pi_{ijk}$, $i = 1, \ldots, r$, $q_j = \sum_{(i,k)\in\mathcal{C}_j}\pi_{ijk}$, $j = 1, \ldots, c$ and $t_k = \sum_{(i,j)\in\mathcal{T}_k}\pi_{ijk}, k = 1, \ldots, \ell$. Again, inference is conditional on $\mathcal{N}$.

Analogous to Equations (2) and (3) we now construct the truncated multinomial distributions for the positive cells under the unrestricted parameter space and the restricted parameter space (i.e. the null hypothesis). Under the unrestricted parameter space, we have

$$p(\underset{\sim}{n} \mid \underset{\sim}{\pi}) = n! \prod_{(ijk)\in\mathcal{N}} \frac{1}{n_{ijk}!} \left\{ \frac{\pi_{ijk}}{\sum_{(i,j,k)\in\mathcal{N}}\pi_{ijk}} \right\}^{n_{ijk}}, \ n_{ijk} > 0, \quad \sum_{(i,j,k)\in\mathcal{N}} n_{ijk} = n, \tag{A4}$$

and under the restricted parameter space, the corresponding distribution is given by

$$p(\underset{\sim}{n} \mid \underset{\sim}{p}, \underset{\sim}{q}) = n! \prod_{(ijk)\in\mathcal{N}} \frac{1}{n_{ijk}!} \left\{ \frac{p_i q_j t_k}{\sum_{(i,j,k)\in\mathcal{N}} p_i q_j t_k} \right\}^{n_{ijk}}, \ n_{ijk} > 0, \quad \sum_{(i,j,k)\in\mathcal{N}} n_{ijk} = n. \tag{A5}$$

We can obtain the likelihood ratio test in a manner similar to the one for the $r \times c$ table. The sampling-based method is used to obtain the MLEs in Equations (A4) and (A5). The bootstrap method, used to obtain the p-value of the test, should work the same as for the $r \times c$ table. Other tests for the three-way table can be performed as well.

Appendix 3. EM algorithm

We can perform the EM algorithm in Equation (4) to obtain the MLEs of the parameters. Here, the observed data are $n_{ij}, (i,j) \in \mathcal{N}$ and the missing data are N, the effective sample size.

Because $N \mid \underset{\sim}{\pi} \sim \text{NB}\{n, \sum_{(i,j)\in\mathcal{N}}\pi_{ij}\}$, we have $E(N \mid \underset{\sim}{\pi}) = n/\sum_{(i,j)\in\mathcal{N}}\pi_{ij}$. This is the E-step.

Now, given $\underset{\sim}{\pi}, N, [\{n_{ij}, (i,j) \in \mathcal{N}\}, N - \sum_{(i,j)\in\mathcal{N}} n_{ij}] \sim \text{Multinomial}[N, \{\pi_{ij}, (ij) \in \mathcal{N}\}, 1 - \sum_{(ij)\in\mathcal{N}}\pi_{ij})]$ and so the MLEs are $\hat{\pi}_{ij} = n_{ij}/N, (ij) \in \mathcal{N}$. This is the M-step.

Thus, it is easy to perform the EM algorithm. The difficulty is that, in this specific problem, the algorithm is sensitive to the starting values. This is partly due to the fact that there is no real information about N; thus we have integrated N out of the likelihood function.