Conditional Inference on Tables With Structural Zeros

Yuguo CHEN

We develop a set of sequential importance sampling (SIS) strategies for sampling nearly uniformly from two-way zero-one or contingency tables with fixed marginal sums and a given set of structural zeros. The SIS procedure samples tables column by column or cell by cell by using appropriate proposal distributions, and enables us to approximate closely the null distributions of a number of test statistics involved in such tables. When structural zeros are on the diagonal or follow certain patterns, more efficient SIS algorithms are developed which guarantee that every generated table is valid. Examples show that our methods can be applied to make conditional inference on zero-one and contingency tables, and are more efficient than other existing Monte Carlo algorithms.

Key Words: Contingency table; Exact test; Monte Carlo method; Sequential importance sampling; Zero-one table.

1. INTRODUCTION

A zero-one table is a matrix in which each entry is either 0 or 1. A contingency table is a matrix in which each element is a nonnegative integer. We refer to an entry as a *structural zero* if it is constrained to be zero. This may happen when it is known a priori that an entry has a zero value based on a certain feature or the underlying structure of the data. See Sections 2 and 8 for examples. Problems of testing hypotheses about two-way zero-one or contingency tables with fixed marginal sums and a given set of structural zeros arise in many different contexts, including ecological studies, educational tests, and social networks. For zero-one tables, the reference distribution for the null hypothesis is often chosen to be the uniform distribution over all tables with given marginal sums and structural zeros (Rasch 1960; Connor and Simberloff 1979; Wasserman and Faust 1994). For contingency tables, the null hypothesis may assume the hypergeometric (for testing quasi-independence, see Goodman 1968) or uniform (for conditional volume tests, see Diaconis and Efron 1985) distribution over all tables with given marginal sums and structural zeros.

Yuguo Chen is Assistant Professor, Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, IL 61820 (E-mail: *yuguo@uiuc.edu*).

Journal of Computational and Graphical Statistics, Volume 16, Number 2, Pages 445–467
DOI: 10.1198/106186007X209226

Making inferences conditional on marginal sums is important for creating a probabilistic basis for a test and removing the effect of nuisance parameters on tests (Lehmann 1986, chap. 4), but it presents challenging computational problems. No good analytic approximations to the null distributions of various test statistics are available due to the complicated interactions among the constraints on marginal sums and structural zeros. Several Markov chain Monte Carlo (MCMC) algorithms have been proposed to approximate the null distribution of any test statistic for zero-one tables (Connor and Simberloff 1979; Roberts and Stone 1990; Rao, Jana, and Bandyopadhyay 1996; Roberts 2000) or contingency tables (Smith, Forster, and McDonald 1996; Aoki and Takemura 2005; Rapallo 2006) with fixed marginal sums and structural zeros. The MCMC approaches for such problems often involve designing complicated Markov moves in order to connect all tables with given constraints, and tend to have a long autocorrelation time.

Snijders (1991) was the first to consider importance sampling in the context of zero-one tables with fixed marginal sums and structural zeros. However, the variation of the importance weights is often large in his method, and it sometimes generates a large proportion of invalid tables. Chen, Diaconis, Holmes, and Liu (2005) used a sequential importance sampling (SIS) approach to generate zero-one and contingency tables with given marginal sums, which is shown to be efficient. The presence of structural zeros, which was not considered by Chen et al. (2005), adds another layer of complexity to the sampling procedure. It requires more careful derivation of the proposal distribution which is ideally close to the target distribution and easy to sample from. The technical details for proving that the sequential sampling procedure always generates valid tables are much harder.

In this article, we extend Chen et al.'s (2005) method to sample tables with fixed marginals and a given set of structural zeros. The SIS approach generates tables column by column or cell by cell by using appropriate proposal distributions, and provides us with a fast, accurate approximation to the null distribution of any test statistic. A byproduct of our approach is to give an estimate of the total number of tables with given marginal constraints and structural zeros, which is another interesting but difficult problem. We compare SIS with other existing Monte Carlo methods on several examples to demonstrate the efficiency of our algorithms. Note that a two-way zero-one table can be thought of as a high-dimensional incomplete contingency table (Kelderman 1984), but we find it more useful to take advantage of the special structure of a zero-one table and think of it as a two-way table, which records for each of a set of units whether or not the unit has certain properties (a 1 denotes that the unit has a given property and a 0 denotes otherwise).

The article is organized as follows. Section 2 provides several motivating examples for conditional inference on tables with structural zeros. Section 3 gives statistical motivations for conditional inference. Section 4 introduces the basic SIS methodology. Section 5 describes how we apply conditional-Poisson sampling to generate zero-one tables. Section 6 proposes a more efficient SIS method that always generates valid tables when the structural zeros follow certain patterns. Section 7 considers the sampling strategy for contingency tables. Section 8 shows some applications and numerical examples, and Section 9 provides concluding remarks.

2. MOTIVATING EXAMPLES

In the analysis of social networks, zero-one tables have been used to represent relational data. Table 1 is the friendship network among 21 high-tech managers in a small manufacturing organization on the west coast of the United States. Each manager was given a list of the names of the other 20 managers, and asked "Who are your friends?" The (i, j)th entry t_{ij} of the table T is 1 if manager i thinks manager j is his/her friend, and 0 otherwise. All entries on the diagonal are structural zeros, denoted by [0] throughout the article to distinguish it from a sampling zero. These data were collected by Krackhardt (1987) and further analyzed by Wasserman and Faust (1994, p. 631) and Roberts (2000).

The row sum r_i is the number of friends that person i perceives he or she has and the column sum c_j is the number of people who think person j is his/her friend. Network analysts are often interested in testing whether the observed relational data are generated from the uniform distribution over all zero-one tables with given marginal sums and a zero diagonal. It was stated by Wasserman and Faust (1994, p. 550) that "This distribution is extremely important in social network analysis, since it can be used to control statistically for both choices made by each actor and choices received." One test statistic of interest is the number of mutual or reciprocal choices

$$M(T) = \sum_{i<j} t_{ij} t_{ji}, \tag{2.1}$$

which can be used to test whether there is a tendency towards mutuality (Snijders 1991; Roberts 2000). To carry out this test, we need to find the distribution of M under the null hypothesis that the observed relational table is a random draw from the uniform distribution over all zero-one tables with given marginal sums and a zero diagonal. The null hypothesis of no tendency towards mutuality is rejected if the observed value of the test statistic M is too large.

Zero-one tables also form an important type of data in educational/psychological tests. Suppose n persons are asked to answer m questions (items). We can construct a zero-one table based on all the answers. A 1 in cell (i, j) means that the ith person answered the jth question correctly, and a 0 means otherwise. Rasch (1960) proposed a simple linear logistic model to measure people's ability. The Rasch model assumes that each person's ability is characterized by a parameter θ_i, each item's difficulty is characterized by a parameter β_j, and

$$P(t_{ij} = 1) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}, \tag{2.2}$$

where t_{ij} is the ith person's answer to the jth question. The responses t_{ij} are assumed to be independent. In some cases, each person is only asked to answer a subset of the available questions. For example, the questions for each person may be randomly drawn from a large problem bank. Sometimes in order to measure the progress that students are making in different grades, the tests are designed so that there are a few overlapping questions for students in different grades (personal communication with Don Burdick and Jack Stenner). We call the (i, j)th entry a structural zero if person i is not asked to answer question j.

Table 1. Friendship relation between 21 high-tech managers.

Manager	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Total
1	[0]	1	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	5
2	1	[0]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	3
3	0	0	[0]	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	2
4	1	1	0	[0]	0	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	6
5	0	1	0	0	[0]	0	0	0	1	0	1	0	0	1	0	0	1	0	1	0	1	7
6	0	1	0	0	0	[0]	1	0	1	0	0	1	0	0	0	0	1	0	0	0	1	6
7	0	0	0	0	0	0	[0]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	1	0	0	0	[0]	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	[0]	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	1	0	1	0	0	1	1	[0]	0	1	0	0	0	1	0	0	0	1	0	7
11	1	1	1	1	1	0	0	1	1	0	[0]	1	1	0	1	0	1	1	1	0	0	13
12	1	0	0	1	0	0	0	0	0	0	0	[0]	0	0	0	0	1	0	0	0	1	4
13	0	0	0	0	1	0	0	0	0	0	1	0	[0]	0	0	0	0	0	0	0	0	2
14	0	0	0	0	0	0	1	0	0	0	0	0	0	[0]	1	0	0	0	0	0	0	2
15	1	0	1	0	1	1	0	0	1	0	1	0	0	1	[0]	0	0	0	1	0	0	8
16	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	[0]	0	0	0	0	0	2
17	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	[0]	0	1	1	1	18
18	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[0]	0	0	0	1
19	1	1	1	0	1	0	0	0	0	0	1	1	0	1	1	0	0	0	[0]	1	0	9
20	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	[0]	0	2
21	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	[0]	4
Total	8	10	5	5	6	2	3	5	6	1	6	8	1	5	4	4	6	4	5	3	5	

The number of items answered correctly by each person (the column sums) are minimal sufficient statistics for the ability parameters and the number of people answering each item correctly (the row sums) are minimal sufficient statistics for the item difficulty parameters. An exact test for the goodness-of-fit of the Rasch model is based on the conditional distribution of the zero-one matrix of responses with fixed marginal sums and structural zeros. It is easy to see that under model (2.2), all the zero-one tables are uniformly distributed conditional on the marginal sums and structural zeros. Thus, implementing the test requires the ability to sample zero-one tables from the null distribution. See Chen and Small (2005) for more discussion on various tests of the Rasch model.

The above two examples use the same reference distribution for the null hypothesis. It is possible to use the Rasch model (2.2) to unify the two zero-one examples which assume that the row and column sums contain sufficient information about the characteristics of the objects in the model. For example, we can interpret θ_i as person i's perception of his/her ability to make friends, and β_j as the friendliness of person j in the social network example. Holland and Leinhardt (1981) proposed an exponential family of probability distributions, denoted by p_1, to study social network problems. The Rasch model and the p_1 distribution with the reciprocity parameter $\rho = 0$ are equivalent conditional on the total number of 1's in the table being fixed. Snijders (1991) points out that the uniformly most powerful unbiased test for the reciprocity parameter ρ in the p_1 distribution requires conditioning on the row and column sums and structural zeros.

3. CONDITIONAL INFERENCE

Exact conditional inference for categorical data eliminates nuisance parameters and does not rely on questionable asymptotic approximations in hypothesis testing. It controls Type I errors exactly. Cox (1988) and Reid (1995) argued that exact conditional inference can make probability calculations more relevant to the data under study. When the subjects in statistical studies are not obtained by a sampling scheme, conditioning on the marginal sums can be seen as a pragmatic way of creating a probabilistic basis for a test (Lehmann 1986, chap. 4). See Agresti's (1992) review article for detailed discussion of conditional inference and other supporting arguments.

For two-way contingency tables without structural zeros, Diaconis and Efron (1985) proposed the conditional volume test to address the question of whether the Pearson χ^2-statistic of a contingency table is "atypical" when the observed table is regarded as a draw from the uniform distribution over tables with the given marginal sums. The p value from the conditional volume test is useful in interpreting Pearson's χ^2 statistic in the test for independence. When a contingency table contains structural zeros, see Table 4 and other examples in Bishop, Fienberg, and Holland (1975, chap. 5), Goodman (1968) introduced quasi-independence as a generalization of the usual model of independence. Following Diaconis and Efron's reasoning, the conditional volume test can be applied to contingency tables with structural zeros as a way to measure how far the table is from quasi-independence. To carry out the test, we need to find the distribution of the χ^2-statistic when tables with fixed marginal sums and structural zeros are uniformly distributed. An exact test is also preferred

in the test of quasi-independence because for sparse contingency tables, the asymptotic χ^2 approximation may not be adequate (Smith, Forster, and McDonald 1996; Aoki and Takemura 2005). The null distribution of this test is hypergeometric over the same set of tables as the conditional volume test.

4. SEQUENTIAL IMPORTANCE SAMPLING

Let $\Sigma_{\mathbf{rc}}(\Omega)$ denote the set of all $m \times n$ zero-one or contingency tables with row sums $\mathbf{r} = (r_1, \ldots, r_m)$, column sums $\mathbf{c} = (c_1, \ldots, c_n)$, and the set of structural zeros Ω. Assume $\Sigma_{\mathbf{rc}}(\Omega)$ is nonempty. The p value for conditional inference on tables, conditional on the marginal sums and structural zeros, can be written as

$$\mu = E_p f(T) = \sum_{T \in \Sigma_{\mathbf{rc}}(\Omega)} f(T)p(T), \tag{4.1}$$

where $p(T)$ is the underlying distribution on $\Sigma_{\mathbf{rc}}(\Omega)$, which is usually uniform or hypergeometric and only known up to a normalizing constant, and $f(T)$ is a function of the test statistic. For example, if we let $f(T) = 1_{\{M(T) \geq M(T_0)\}}$, where T_0 is the observed table and $M(T)$ is defined in (2.1), formula (4.1) gives the p value for the test of mutuality in social networks.

In many cases, sampling from $p(T)$ directly is difficult. The importance sampling approach is to simulate a table $T \in \Sigma_{\mathbf{rc}}(\Omega)$ from a different distribution $q(\cdot)$, where $q(T) > 0$ for all $T \in \Sigma_{\mathbf{rc}}(\Omega)$, and estimate μ by

$$\hat{\mu} = \frac{\sum_{i=1}^N f(T_i)p(T_i)/q(T_i)}{\sum_{i=1}^N p(T_i)/q(T_i)}, \tag{4.2}$$

where $T_1, \ldots, T_N$ are iid samples from $q(T)$. When the underlying distribution $p(T)$ is uniform, $p(T_i)$ in the numerator and denominator of (4.2) can be canceled out. We can also estimate the total number of tables in $\Sigma_{\mathbf{rc}}(\Omega)$ by

$$|\widehat{\Sigma}_{\mathbf{rc}}(\Omega)| = \frac{1}{N} \sum_{i=1}^N \frac{1}{q(T_i)}, \tag{4.3}$$

because $|\Sigma_{\mathbf{rc}}(\Omega)| = \sum_{T \in \Sigma_{\mathbf{rc}}(\Omega)} \frac{1}{q(T)} q(T)$. The underlying distribution on $\Sigma_{\mathbf{rc}}(\Omega)$ corresponding to this case is uniform.

In order to evaluate the efficiency of an importance sampling algorithm, we can look at the number of iid samples from the target distribution that are needed to give the same standard error for $\hat{\mu}$ as N importance samples. A rough approximation for this number is the *effective sample size* $\text{ESS} = N/(1 + cv^2)$ (Kong, Liu, and Wong 1994), where the *coefficient of variation* (cv) is defined as

$$\text{cv}^2 = \frac{\text{var}_q\{p(T)/q(T)\}}{E_q^2\{p(T)/q(T)\}}. \tag{4.4}$$

Accurate estimation generally requires a low cv^2, that is, $q(T)$ must be sufficiently close to $p(T)$. The standard error of $\hat{\mu}$ or $|\widehat{\Sigma}_{\mathbf{rc}}(\Omega)|$ can be simply estimated by further repeated sampling or by the δ-method (Snijders 1991; Chen et al. 2005).

A good proposal distribution is essential to the efficiency of the importance sampling method. However, it is usually not easy to devise good proposal distributions for high-dimensional problems such as sampling tables from a complicated target space $\Sigma_{\mathbf{rc}}(\Omega)$. To overcome this difficulty, Chen et al. (2005) suggested a sequential importance sampling approach based on the following factorization

$$q(T = (t_1, \ldots, t_n)) = q(t_1)q(t_2|t_1)q(t_3|t_2, t_1)\ldots q(t_n|t_{n-1}, \ldots, t_1), \tag{4.5}$$

where $t_1, \ldots, t_n$ denote the configurations of the columns of T. This factorization suggests that instead of trying to find a proposal distribution for the entire table all at once, it may be easier to sample the table sequentially, column by column, and use the partially sampled table and updated constraints to guide the choice of the proposal distribution. This breaks up the problem of constructing a good proposal distribution into manageable pieces. Each column has a much lower dimension compared to the whole table; therefore it is relatively easy to design a good proposal distribution $q(t_k|t_{k-1}, \ldots, t_1)$ to approximate $p(t_k|t_{k-1}, \ldots, t_1)$, $k = 1, \ldots, n$. We describe in the following sections how to use SIS to sample zero-one and contingency tables.

5. SAMPLING ZERO-ONE TABLES WITH STRUCTURAL ZEROS

To avoid triviality, we assume that none of the row or column sums is zero, none of the row sums is n and none of the column sums is m. Suppose there are g_i structural zeros in the ith row and s_j structural zeros in the jth column, $i = 1, \ldots, m$, $j = 1, \ldots, n$, and denote the positions of all structural zeros as

$$\Omega = \{(i, j) : \ (i, j) \quad \text{is a structural zero}\}. \tag{5.1}$$

The first column of the table is sampled conditional on its marginal sum c_1. This is equivalent to putting c_1 ones in $m - s_1$ possible slots. Conditional on the realization of the first column, the row sums of the $m \times (n-1)$ subtable are computed by subtracting the respective numbers in the first column from the original row sums. We then focus on the new $m \times (n-1)$ subtable and sample the first column of the subtable (the second column of the original table) in the same manner. This procedure is repeated recursively until all the columns are sampled.

We modify notation to denote the subtables considered in the sequential sampling procedure. After the first $l-1$ columns of the original table have been sampled and removed from further consideration, denote the updated row sums of the $m \times (n-l+1)$ subtable by $r_j^{(l)}$, $j = 1, \ldots, m$ (note $r_j^{(1)} = r_j$), and denote the updated column sums of the $m \times (n-l+1)$ subtable by $c_j^{(l)}$, $j = 1, \ldots, n-(l-1)$. Here $(c_1^{(l)}, \ldots, c_{n-(l-1)}^{(l)}) = (c_l, \ldots, c_n)$. Let

$$\Omega^{(l)} = \{(i, j-l+1) : \ (i, j) \in \Omega \text{ and } j \geq l\}. \tag{5.2}$$

denote the set of structural zeros in the $m \times (n - l + 1)$ subtable. Let $g_i^{(l)}$ and $s_j^{(l)}$ be the number of structural zeros in the ith row and the jth column of the $m \times (n-l+1)$ subtable.

Chen et al. (2005) suggested the "conditional-Poisson (CP)" sampling method to sample the c_1 nonzero positions for the first column (and subsequently the other columns) when there are no structural zeros. Here we extend their argument to handle structural zeros. We show that the CP distribution is still a good sampling distribution when there are structural zeros, but the parameters of the CP distribution need to be adjusted for structural zeros.

CP sampling is a method for sampling c units from the set $\{1, \ldots, m\}$ without replacement. Let

$$\mathbf{Z} = (Z_1, \ldots, Z_m) \tag{5.3}$$

be independent Bernoulli trials with probability of successes $\mathbf{p} = (p_1, \ldots, p_m)$. Then $S_{\mathbf{Z}} = Z_1 + \cdots + Z_m$ has the *Poisson-binomial* distribution. Here $Z_i = 1$ corresponds to unit i being selected in the sample. The distribution of CP sampling is the conditional distribution of $\mathbf{Z}$ given $S_{\mathbf{Z}}$, which can be written as

$$P(Z_1 = z_1, \ldots, Z_m = z_m \mid S_{\mathbf{Z}} = c) \propto \prod_{k=1}^{m} w_k^{z_k}, \tag{5.4}$$

where $w_i = p_i/(1-p_i)$. We use the drafting sampling scheme proposed by Chen, Dempster, and Liu (1994) and Chen and Liu (1997) to implement the CP sampling method.

The motivation for using CP sampling and the choice of the weights are given in the following theorem, which is an extension of Theorem 1 of Chen et al. (2005).

Theorem 1. *For the uniform distribution over all $m \times n$ zero-one tables with given row sums $r_1, \ldots, r_m$, first column sum c_1, and the set of structural zeros Ω (defined by (5.1)), the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ (defined by (5.4)) given $S_{\mathbf{Z}} = c_1$ with*

$$p_i = \begin{cases} r_i/(n - g_i), & \text{if} \quad (i,1) \notin \Omega, \\ 0, & \text{if} \quad (i,1) \in \Omega, \end{cases} \tag{5.5}$$

where g_i is the number of structural zeros in the ith row.

The proof of Theorem 1 is given in Appendix A (p. 462). When the target distribution is uniform over $\Sigma_{\mathbf{rc}}(\Omega)$, we would expect that the proposal distribution $q(t_1) = P(t_1|r_1, \ldots, r_m, c_1, \Omega)$ is reasonably close to $p(t_1) = P(t_1|r_1, \ldots, r_m, c_1, \ldots, c_n, \Omega)$, the true marginal distribution of the first column t_1 which is analytically intractable. Hence, the theorem suggests that CP sampling should generally have a low cv^2. Suppose the first $l-1$ columns have been sampled, we then generate column l with the CP sampling method using the weights $1_{\{(i,1)\notin\Omega^{(l)}\}} r_i^{(l)}/[n - (l-1) - g_i^{(l)} - r_i^{(l)}]$. During the sampling process, a row is removed from further consideration if its row sum equals zero because that row can be filled by zeros. Similarly, a row can be filled by ones if its row sum equals the number of possible positions left in that row.

Another justification for the use of CP sampling is based on the asymptotic analysis of Bender (1974). Bender gives the following approximation to the number of zero-one matrices in $\Sigma_{\mathbf{rc}}(\Omega)$:

$$|\Sigma_{\mathbf{rc}}(\Omega)| \approx \Delta_{\mathbf{rc}}(\Omega) \equiv \frac{M!}{\prod_{i=1}^{m} r_i! \prod_{j=1}^{n} c_j!} \times \exp\left\{ -\left(\sum_{i=1}^{m} r_i^2 - M\right)\left(\sum_{j=1}^{n} c_j^2 - M\right) \Big/ 2M^2 - \sum_{(i,j)\in\Omega} \frac{r_i c_j}{M} \right\}, \quad (5.6)$$

where $M = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$. This approximation works well for large sparse zero-one matrices. Let $v(i_1, \ldots, i_{c_1})$ be the zero-one vector of length m which has i_kth $(1 \le k \le c_1)$ component equal to 1 and all other components equal to 0. So $v(i_1, \ldots, i_{c_1})$ is a potential candidate for the first column t_1. To avoid violating the requirement of structural zeros, we only consider the configurations $v(i_1, \ldots, i_{c_1})$ such that $(i_k, 1) \notin \Omega$, $k = 1, \ldots, c_1$. Let $\mathbf{r}^{(2)} = (r_1^{(2)}, \ldots, r_m^{(2)})$ and $\mathbf{c}^{(2)} = (c_2, \ldots, c_n)$ be the updated row and column sums after we sample the first column $t_1 = v(i_1, \ldots, i_{c_1})$. Then by approximation (5.6), we have

$$p(t_1 = v(i_1, \ldots, i_{c1})) \approx \frac{\Delta_{\mathbf{r}^{(2)}\mathbf{c}^{(2)}}(\Omega^{(2)})}{\Delta_{\mathbf{rc}}(\Omega)} \propto \prod_{k=1}^{c_1} r_{i_k} \exp\left\{ d r_{i_k} + \sum_{j=2}^{n} \frac{c_j}{M - c_1} 1_{\{(i_k,j)\in\Omega\}} \right\},$$

where $d = \sum_{j=2}^{n} (c_j^2 - c_j)/(M - c_1)^2$. Thus, this approximation also suggests that a good proposal distribution for sampling the first column is the CP distribution, but with weights proportional to $1_{\{(i,1)\notin\Omega\}} r_i \exp\{d r_i + \sum_{j=2}^{n} 1_{\{(i,j)\in\Omega\}} c_j/(M - c_1)\}$.

The CP sampling strategies based on Theorem 1 and (5.6) performed well for all examples we have tested. For the rest of the article, we focus on the CP sampling strategy based on the approximation using Theorem 1. We refer to this strategy as SIS-CP1.

6. MORE EFFICIENT SIS METHOD FOR CERTAIN ZERO-ONE TABLES

The SIS-CP1 algorithm cannot guarantee a valid zero-one table in the end. Sometimes the sampling cannot proceed after generating a few columns. When this happens, we can simply throw away this partially generated table which is equivalent to assigning a zero weight to this bad sample.

A potentially more efficient strategy is to design the sampling procedure more carefully so as to guarantee the existence of subtables with the updated row sums, column sums, and structural zeros. In other words, we need to incorporate the existence conditions of subtables into the sampling scheme so that only valid configurations of each column are sampled. Chen et al. (2005) use this idea to develop efficient sampling algorithms for zero-one tables with given marginal sums. When there are structural zeros, it is difficult to incorporate the existence conditions, because in general 2^{m+n} inequalities need to be checked in order to guarantee the existence of $m \times n$ zero-one tables with fixed marginal sums and a given set of structural zeros (Mirsky 1971, p. 205; Chen 2006). In this scenario, sampling the tables without checking the existence of subtables is usually more efficient, as long as the percentage of invalid tables is not too large.

When there is at most one structural zero in each row and column, Chen (2006) shows that the 2^{m+n} inequalities in Mirsky's theorem can be reduced to mn inequalities, which makes it much easier to check during the sampling process. In the following, we show that the mn inequalities can be incorporated into our sampling scheme so as to guarantee the existence of subtables with the updated marginal sums and structural zeros. This helps us develop a more efficient SIS procedure for sampling from $\Sigma_{\mathbf{rc}}(\Omega)$, when Ω contains at most one structural zero in each row and column. A particular application of this strategy is conditional inference on square tables with a zero diagonal, which arise very often in social network analysis, see Table 1. We state Chen's (2006) theorem below.

Theorem 2. *(Chen 2006) The necessary and sufficient conditions for the existence of an $m \times n$ zero-one matrix with given row sums $r_1, \ldots, r_m$, column sums $c_1, \ldots, c_n$, and the set of structural zeros Ω with at most one structural zero in each row and column, is that*

$$\sum_{i=1}^{k}\sum_{j=1}^{l} a_{u_i^* v_j^*} \geq \sum_{i=1}^{k} r_{u_i^*} - \sum_{j=l+1}^{n} c_{v_j^*}, \quad k = 1, \ldots, m;\ l = 1, \ldots, n, \tag{6.1}$$

where $r_{u_1^*}, \ldots, r_{u_m^*}$ and $c_{v_1^*}, \ldots, c_{v_n^*}$ are the reordered row and column sums so that $r_{u_1^*} \geq \cdots \geq r_{u_m^*}$ and $c_{v_1^*} \geq \cdots \geq c_{v_n^*}$, and

$$a_{u_i^* v_j^*} = \begin{cases} 0, & \text{if} \quad (u_i^*, v_j^*) \in \Omega, \\ 1, & \text{if} \quad (u_i^*, v_j^*) \notin \Omega. \end{cases} \tag{6.2}$$

In order to incorporate the mn inequalities in Theorem 2 into our sampling scheme, we order the row and column sums as follows. We first order the column sums from largest to smallest. Without loss of generality, assume $c_1 \geq \cdots \geq c_n$ are already ordered from largest to smallest. For the ith row with row sum r_i, define $y(i)$ as follows.

$$y(i) = \begin{cases} j, & \text{if } (i, j) \in \Omega, \\ n+1, & \text{if there is no structural zero in the } i\text{th row.} \end{cases}$$

Here $y(i)$ is well defined because there is at most one structural zero in each row. Reorder the rows according to $(r_i, y(i))$, that is,

$$\begin{aligned} &\text{the } i\text{th row should be before the } j\text{th row if} \\ &\text{(i) } r_i > r_j, \text{ or (ii) } r_i = r_j \text{ and } y(i) < y(j). \end{aligned} \tag{6.3}$$

Without loss of generality, assume that $r_1 \geq \cdots \geq r_m$ already satisfy rule (6.3). Therefore in (6.1) and (6.2), $u_i^* = i$, $v_j^* = j$, $i = 1, \ldots, m$; $j = 1, \ldots, n$.

After the first column of the $m \times n$ table is sampled, we know from Theorem 2 that a necessary and sufficient condition for the existence of an $m \times (n-1)$ zero-one table with row sums $r_1^{(2)}, \ldots, r_m^{(2)}$, column sums $c_1^{(2)}, \ldots, c_{n-1}^{(2)}$, and structural zeros $\Omega^{(2)}$ is that

$$\sum_{i=1}^{k}\sum_{j=1}^{l} a_{u_i, j} \geq \sum_{i=1}^{k} r_{u_i}^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)}, \quad k = 1, \ldots, m;\ l = 1, \ldots, n-1, \tag{6.4}$$

where $r^{(2)}_{u_1} \geq \cdots \geq r^{(2)}_{u_m}$ are the reordered row sums, and

$$a_{u_i,j} = \begin{cases} 0, & \text{if} \quad (u_i, j) \in \Omega^{(2)}, \\ 1, & \text{if} \quad (u_i, j) \notin \Omega^{(2)}. \end{cases}$$

It is not easy to check this condition during the sampling process because the ordering of $r^{(2)}_1, \ldots, r^{(2)}_m$ is not known before sampling the first column. This difficulty can be overcome by the alternative condition in the following theorem, whose proof is given in Appendix B (p. 463).

Theorem 3. *Suppose $r_1 \geq \cdots \geq r_m$ and $c_1 \geq \cdots \geq c_n$ are the row and column sums of a zero-one table with the set of structural zeros Ω, where there is at most one structural zero in each row and column. Assume the ordering of row sums also satisfies rule (6.3). Let $(b_1, \ldots, b_m)$ be a zero-one vector such that $\sum_{i=1}^m b_i = c_1$, and $b_i = 0$ if $(i, 1) \in \Omega$, $i = 1, \ldots, m$. Then the necessary and sufficient condition for the existence of an $m \times (n-1)$ subtable with row sums $r^{(2)}_1, \ldots, r^{(2)}_m$ where $r^{(2)}_i = r_i - b_i$, column sums $c^{(2)}_1, \ldots, c^{(2)}_{n-1}$ where $c^{(2)}_j = c_{j+1}$, and structural zeros $\Omega^{(2)}$ defined in (5.2), is that*

$$\sum_{i=1}^{k} \sum_{j=1}^{l} a_{ij} \geq \sum_{i=1}^{k} r^{(2)}_i - \sum_{j=l+1}^{n-1} c^{(2)}_j, \quad k = 1, \ldots, m; \; l = 1, \ldots, n-1, \tag{6.5}$$

where

$$a_{ij} = \begin{cases} 0, & \textit{if} \quad (i, j) \in \Omega^{(2)}, \\ 1, & \textit{if} \quad (i, j) \notin \Omega^{(2)}. \end{cases}$$

This result is not obvious because $r^{(2)}_1, \ldots, r^{(2)}_m$ is not necessarily ordered. For example, for a 3×3 zero-one matrix with row sums 2, 2, 1, column sums 2, 2, 1, and a structural zero at position (3, 3), if the first column is sampled as $(1, 0, 1)'$, then the updated row sums are 1, 2, 0, which are not in decreasing order.

Define $h_k = \min_{0 \leq l \leq n-1} \{\sum_{i=1}^k \sum_{j=1}^l a_{ij} + \sum_{j=l+1}^{n-1} c^{(2)}_j\}$, for $k = 1, \ldots, m$. Condition (6.5) can be restated as follows.

- For each $1 \leq k \leq m$, if $\sum_{i=1}^k r_i > h_k$, then we need to put at least $\sum_{i=1}^k r_i - h_k$ ones at or before the kth row in the first column. We call k a knot for convenience.

We can summarize the above condition by two vectors: $(k[1], k[2], \ldots)$, which records the positions of the knots, and $(v[1], v[2], \ldots)$, which records how many ones we must put at or before those knots. Redundant knots should be removed before implementation. For example, if $v[j] \leq v[i]$ for some $j > i$, then we ignore knot $k[j]$. For the jth column, define $x(j)$ as follows.

$$x(j) = \begin{cases} i, & \text{if } (i, j) \in \Omega, \\ m + 1, & \text{if there is no structural zero in the } j\text{th column.} \end{cases}$$

Here $x(j)$ is well defined because there is at most one structural zero in each column. Using the above condition, we design the following more efficient sampling strategy.

- Choose an integer o_1 uniformly from $[v[1], \min\{k[1] - 1_{\{0<x(1)\le k[1]\}}, c_1\}]$ as the number of ones we put at or before row $k[1]$. Sample the o_1 positions between row 1 and row $k[1]$ using the CP sampling with weights $1_{\{(i,1)\notin\Omega\}} r_i/(n - g_i - r_i)$ (see Section 5).

- Choose an integer o_2 uniformly from $[\max\{v[2] - o_1, 0\}, \min\{k[2] - k[1] - 1_{\{k[1]<x(1)\le k[2]\}}, c_1 - o_1\}]$ as the number of ones we put between row $k[1]$ and row $k[2]$. Sample the o_2 positions between row $k[1]$ and row $k[2]$ using the CP sampling again.

- Continue the procedure until all the knots in column 1 have been considered. After the first column is sampled, we record the probability $q(t_1)$ of getting such a sample, update the row sums, rearrange the updated row sums according to rule (6.3), and repeat the procedure with the second column.

We refer to the above sampling strategy as SIS-CP2. When there is at most one structural zero in each row and column, SIS-CP2 guarantees that we can always generate a valid table in $\Sigma_{\mathbf{rc}}(\Omega)$, and it tends to be more efficient than SIS-CP1. The extra computational cost for checking the existence conditions was negligible.

7. SAMPLING CONTINGENCY TABLES WITH STRUCTURAL ZEROS

Theorem 11.5.1 of Mirsky (1971, p. 205) gives necessary and sufficient conditions for the existence of an integer matrix with prescribed bounds for its entries, row sums, and column sums. We use Mirsky's theorem to derive the necessary and sufficient conditions for the existence of a contingency table with fixed marginal sums and a given set of structural zeros. The following corollary, whose proof is given in Appendix C (p. 465), is a direct consequence of Mirsky's Theorem 11.5.1. For two sets I and J, define $I \times J = \{(i, j) : i \in I, j \in J\}$.

Corollary 1. *Assume $r_1, \ldots, r_m$ and $c_1, \ldots, c_n$ are non-negative integers and $\sum_{i=1}^m r_i = \sum_{j=1}^n c_j$. The necessary and sufficient conditions for the existence of an $m \times n$ contingency table with given row sums $r_1, \ldots, r_m$, column sums $c_1, \ldots, c_n$, and the set of structural zeros Ω, is that for all $I \subset \{1, \ldots, m\}$, $J \subset \{1, \ldots, n\}$ with $I \times J \subset \Omega$,*

$$\sum_{i\in I} r_i \le \sum_{j\notin J} c_j. \tag{7.1}$$

We want to sample contingency tables column by column as we did for zero-one tables. In order to guarantee that every table we generated is a valid table, we will use Corollary 1 to derive the conditions that the entries in each column need to satisfy. Since this is a recursive procedure, the following theorem focuses on the requirements for entries in the first column.

Theorem 4. *Suppose $r_1, \ldots, r_m$ and $c_1, \ldots, c_n$ are the row and column sums of a contingency table with the set of structural zeros Ω. Let $(t_{11}, \ldots, t_{m1})$ be a vector of*

non-negative integers such that $\sum_{i=1}^{m} t_{i1} = c_1$, $t_{i1} = 0$ *if* $(i, 1) \in \Omega$, *and* $t_{i1} \le r_i$, $i = 1, \ldots, m$. *Then the necessary and sufficient condition for the existence of an* $m \times (n-1)$ *subtable with row sums* $r_1 - t_{11}, \ldots, r_m - t_{m1}$, *column sums* $c_2, \ldots, c_n$, *and structural zeros* $\Omega_2 = \Omega \cap \{(i, j) : i = 1, \ldots, m;\ j = 2, \ldots, n\}$, *is that*

$$\sum_{i \in I} t_{i1} \ge \sum_{i \in I} r_i - \sum_{j \notin J} c_j, \tag{7.2}$$

for all $I \subset \{1, \ldots, m\}$, $J \subset \{2, \ldots, n\}$ *with* $I \times J \subset \Omega_2$.

Theorem 4 can be easily proved by replacing r_i by $r_i - t_{i1}$ in Corollary 1. In general, linear programming can be used to solve (7.2) and compute the lower and upper bounds for each entry t_{i1}. Then we can fill in the entries $t_{11}, \ldots, t_{m1}$ sequentially by picking a possible value for each entry at each step. See Chen, Dinwoodie, and Sullivant (2006) for more discussion on SIS for contingency tables and the properties of linear programming.

Here we consider a special case for which the exact bounds for each entry can be written out explicitly. When Ω_2 contains at most one structural zero in each column, inequality (7.2) simplifies to

$$t_{i1} \ge r_i - \sum_{j=2}^{n} c_j 1_{\{(i,j) \notin \Omega\}}, \quad i = 1, \ldots, m. \tag{7.3}$$

Therefore sampling the first column is equivalent to finding an integer vector $(t_{11}, \ldots, t_{m1})$ such that $\sum_{i=1}^{m} t_{i1} = c_1$ and

$$l_{i1} \le t_{i1} \le u_{i1}, \quad i = 1, \ldots, m,$$

where

$$(l_{i1}, u_{i1}) = \begin{cases} (0, 0), & \text{if} \quad (i, 1) \in \Omega, \\ (\max\{0, r_i - \sum_{j=2}^{n} c_j 1_{\{(i,j) \notin \Omega\}}\}, \min\{r_i, c_1\}), & \text{if} \quad (i, 1) \notin \Omega. \end{cases}$$

Finding such a vector $(t_{11}, \ldots, t_{m1})$ can be realized by the following procedure. Suppose we have already chosen $t_{i1} = t_{i1}^*$ for $1 \le i \le k-1$, then the only restriction on t_{k1} is

$$\max\left\{l_{i1}, c_1 - \sum_{i=1}^{k-1} t_{i1}^* - \sum_{i=k+1}^{m} u_{i1}\right\} \le t_{k1} \le \min\left\{u_{i1}, c_1 - \sum_{i=1}^{k-1} t_{i1}^* - \sum_{i=k+1}^{m} l_{i1}\right\}. \tag{7.4}$$

When the underlying distribution is uniform, which is useful for conditional volume tests, we can choose an integer uniformly at random between the lower and upper bounds in (7.4). When the underlying distribution is hypergeometric, which is useful for tests of quasi-independence, we can choose an integer according to a hypergeometric distribution which only takes values between the lower and upper bounds in (7.4). See Chen, Dinwoodie, and Sullivant (2006) for more discussion on the effect of different sampling distributions.

8. APPLICATIONS

We generated zero-one and contingency tables in the following examples by the SIS algorithms proposed in Sections 5, 6, and 7. All examples were coded in C and run on a

Table 2. Performance comparison of three Monte Carlo methods for estimating p values.

Method	Estimated p value μ_1	Time	Number of samples	Number of invalid samples	cv^2
SIS-CP2	0.042 ± 0.002	20 seconds	10,000	0	0.3
Snijders' method	0.039 ± 0.002	17 minutes	150,000	8950	9.7
Rao, *et al.*'s method	0.038 ± 0.002	110 minutes	10,000	0	–

Pentium 4 computer with 2.4 GHz processor. The C code is available at *http://www.stat.uiuc.edu/~yuguo/software/structural0/*.

8.1 ZERO-ONE TABLES IN SOCIAL NETWORKS

We want to test whether there is a tendency towards mutuality in the friendship network among 21 high-tech managers (Table 1). Since Table 1 contains only one structural zero in each row and column, the more efficient SIS-CP2 method can be applied here. In the following, we compare SIS-CP2 with two other methods: the importance sampling algorithm proposed by Snijders (1991) and the MCMC algorithm proposed by Rao et al. (1996).

For the manager data, the observed test statistic M is 23. The SIS-CP2 method estimated the p value to be 0, so did the other two methods. Therefore the data provide strong evidence against the null hypothesis of no tendency towards mutuality. To compare the efficiency of the three algorithms, we considered a hypothetical p-value: $\mu_1 = P(M(T) \geq 18)$, where T follows the null distribution. The simulation results are summarized in Table 2. The number after the "$\pm$" sign is the standard errors.

The results show that SIS-CP2 is about 50 times faster (to produce the same standard error) than Snijders' algorithm, and more than 300 times faster than Rao et al.'s (1996) algorithm for this example. All 10,000 tables generated by the SIS-CP2 algorithm are guaranteed to be valid. The SIS-CP1 algorithm also generated 10,000 tables in about 20 seconds, but it produced 80 bad tables. A long simulation of 1,000,000 samples based on SIS-CP2 gave an estimate of 0.040 for μ_1.

Based on 10,000 samples from SIS-CP2, we estimated the total number of zero-one tables with the same margins as Table 1 (p. 448) and a zero diagonal to be $(1.88 \pm 0.01) \times 10^{45}$. The large number of possible tables also shows that the idea of calculating the p value by enumerating all possible tables is infeasible in practice. To further challenge our method, we tried a 50×50 table with all marginal sums equal to 25 and a zero diagonal. Based on 100 samples using SIS-CP2, we estimated the total number of tables to be $(4.91 \pm 0.17) \times 10^{643}$. The cv^2 in this case was 0.15. This shows that SIS-CP2 is still very efficient even for large tables.

Table 3. Occurrence matrix for Darwin's finch data. Island name code: A = Seymour, B = Baltra, C = Isabella, D = Fernandina, E = Santiago, F=Rábida, G = Pinzón, H = Santa Cruz, I =Santa Fe, J = San Cristóbal, K = Española, L = Floreana, M = Genovesa, N = Marchena, O = Pinta, P = Darwin, Q = Wolf.

	Island																
Finch	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Large ground finch	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Medium ground finch	1	1	1	1	1	1	1	1	1	1	[0]	1	0	1	1	[0]	[0]
Small ground finch	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Sharp-beaked ground finch	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1
Cactus ground finch	1	1	1	0	1	1	1	1	1	1	[0]	1	0	1	1	[0]	[0]
Large cactus ground finch	0	0	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	1	[0]	1	[0]	[0]	0	0
Large tree finch	[0]	[0]	1	1	1	1	1	1	1	0	[0]	1	[0]	1	1	[0]	[0]
Medium tree finch	[0]	[0]	[0]	0	0	[0]	[0]	0	[0]	[0]	[0]	1	[0]	[0]	[0]	[0]	[0]
Small tree finch	[0]	[0]	1	1	1	1	1	1	1	1	[0]	1	[0]	[0]	1	[0]	[0]
Vegetarian finch	[0]	[0]	1	1	1	1	1	1	1	1	[0]	1	[0]	1	1	[0]	[0]
Woodpecker finch	[0]	[0]	1	1	1	[0]	1	1	[0]	1	[0]	0	[0]	[0]	0	[0]	[0]
Mangrove finch	[0]	[0]	1	1	0	[0]	[0]	0	[0]	[0]	[0]	0	[0]	[0]	[0]	[0]	[0]
Warbler finch	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

8.2 TESTING ZERO-ONE TABLES IN ECOLOGY

In ecological studies, zero-one tables have been used to describe the distribution of certain species over a number of locations. A "1" or "0" in cell (i, j) represents the presence or absence, respectively, of species i at location j. Such zero-one tables are often called occurrence matrices. Table 3 represents 13 species of finches inhabiting 17 islands of the Galápagos Islands (an archipelago in the East Pacific). The data are known as "Darwin's finches" because Charles Darwin collected some of these species when he visited the Galápagos. Many other occurrence matrices are reported in Cook and Quinn (1995).

It has been debated for a long time among ecologists whether the distribution of species reflects competition between species (Sanderson 2000). Connor and Simberloff (1979) proposed to test the null hypothesis that the observed zero-one table is a typical sample drawn uniformly from the set of all zero-one tables with the observed row and column sums and a given set of structural zeros. The number of islands each species inhabits (the row sums) and the number of species on each island (the column sums) are kept fixed under the null hypothesis to reflect the fact that some species are naturally more widespread than others and some islands are naturally more accommodating to a wide variety of species than others (Manly 1995; Connor and Simberloff 1979). The structural zeros here represent the assumption that each species is permitted to occur on only a subset of all locations. This assumption was described in Connor and Simberloff (1979) as follows: "Each species is placed only on islands with species numbers in the range for islands which that species is, in fact, observed to inhabit. That is, the 'incidence' range convention is maintained." In other words, species i can occur on island j if $c_j \in [\min\{c_k : t_{ik} = 1,\ k = 1, \ldots, n\},\ \max\{c_k : t_{ik} = 1,\ k = 1, \ldots, n\}]$, and otherwise cell (i, j) is a structural zero. This additional constraint creates 70 structural zeros in Table 3.

This example can also be put in the Rasch model framework. We can interpret θ_i as the survival ability of each species, and β_j as the habitability of each location. For testing whether there is competition between species, Roberts and Stone (1990) suggested the test statistic

$$\bar{S}^2(T) = \frac{1}{m(m-1)} \sum_{i \neq j} s_{ij}^2, \tag{8.1}$$

where m is the number of species, $S = (s_{ij}) = TT'$ and $T = (t_{ij})$ is the occurrence matrix. Note that s_{ij} is the number of islands on which both the ith and jth species are found. If competition for food controls the distribution of species, then each pair of species will either share a lot of islands, which means they eat different food, or share very few islands, which means they compete with each other for food. Thus, s_{ij} will be either large or small under the alternative hypothesis. It is easy to show that $\sum_{i \neq j} s_{ij}$ is a constant for all tables with given marginal sums, so $\sum_{i \neq j} s_{ij}^2$ tends to have a large value if all the s_{ij} are very uneven. Therefore the competition between species will generally lead to a large value of $\bar{S}^2$, which means the null hypothesis is rejected if the observed $\bar{S}^2(T_0)$ is too large. More test statistics were discussed by Connor and Simberloff (1979), Wilson (1987), Manly (1995), Sanderson, Moulton, and Selfridge (1998), and Sanderson (2000). Our method applies to all of these test statistics.

Since many rows and columns in the finch data contain more than one structural zero, SIS-CP2 and Rao et al.'s method cannot be applied here. We reordered the columns so that the column sums are decreasing. Based on 10,000 sampled tables using SIS-CP1, which took about three seconds, we estimated that the p value is 0.033 ± 0.003. Thus, there is strong evidence against the null hypothesis that there is no competition between species. The null distribution of the test statistic (8.1) in the form of a histogram (computed using the weighted samples) is given in Figure 1. Among the 10,000 samples, there are only 15 invalid tables. Using these samples, we also estimated the total number of zero-one tables with given margins and structural zeros to be $(1.04 \pm 0.02) \times 10^9$.

Snijders' method took about three seconds to generate 4,000 tables, of which 2,501 (62.5%) are invalid tables. Snijders' method estimated a p value of 0.043 ± 0.010. This shows that the SIS-CP1 algorithm is 11 times faster than the Snijders' algorithm for this example. A long simulation of 1,000,000 samples based on SIS-CP1 gave an estimate of 0.036 for the p value.

8.3 Conditional Volume Test for Contingency Tables

Table 4 describes the distribution of genital display in a colony of six squirrel monkeys, labeled as R, S, T, U, V, and W. Ploog (1967) collected the data in order to study the group structure and dynamics of monkey colonies. As pointed out by Ploog, genital display is believed to be a social signal stimulus, which may mean demanding, self-assertion, courting and desiring closer contact, and its appearance may depend on the social situation. There is an active participant and a passive participant in each display. The diagonal cells are structural zeros because a monkey never displays toward itself.

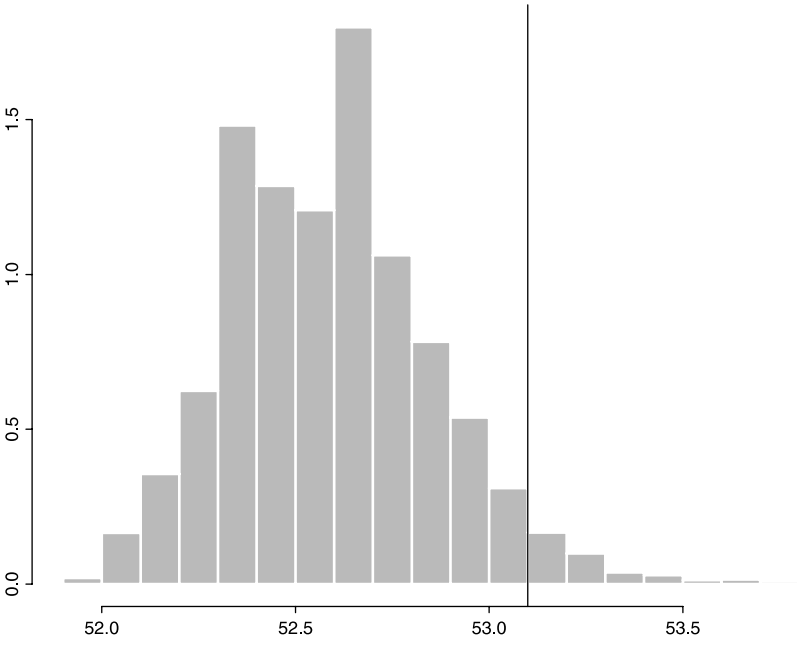

Figure 1. Approximated null distribution of the test statistic $\bar{S}^2(T)$ based on 10,000 weighted samples. The vertical line indicates the observed $\bar{S}^2(T_0) = 53.1$.

Fienberg (1980, p. 145) applied a test of quasi-independence on the data in Table 4. The null hypothesis was rejected at a small significance level which leads to the conclusion that squirrel monkeys tend to choose specific members of the colony to display themselves. When the null hypothesis of quasi-independence is rejected, one may want to know what distribution actually generated the data and how far the table is from quasi-independence. We can use the conditional volume test to partially answer these questions. See Section 3 and Diaconis and Efron (1985).

With 1,000,000 samples produced by our SIS method (see Section 7), which took about three seconds, we estimated the p value for the conditional volume test to be 0.9290 ± 0.0006. Therefore the observed value of the χ^2-statistic is not unusual if the underlying distribution is

Table 4. Distribution of genital display in a colony of six squirrel monkeys.

	Passive Participant					
Active participant	R	S	T	U	V	W
R	[0]	1	5	8	9	0
S	29	[0]	14	46	4	0
T	0	0	[0]	0	0	0
U	2	3	1	[0]	38	2
V	0	0	0	0	[0]	1
W	9	25	4	6	13	[0]

uniform. Such information may be helpful for assessing the social structure in the colony of squirrel monkeys. The MCMC algorithm proposed in Aoki and Takemura (2005) generated 2,000,000 samples (with 500,000 samples as burn-in) in three seconds and estimated the p value to be 0.93 ± 0.01. Thus, SIS is about 270 times faster (to produce the same standard error) than the MCMC algorithm for this example. One million SIS samples estimated the total number of tables to be $(8.76 \pm 0.03) \times 10^{12}$. The cv^2 of the importance weights is 8.6.

9. DISCUSSION

We extended Chen et al.'s (2005) approach to sample zero-one or contingency tables with fixed marginal sums and a given set of structural zeros. The SIS strategies we developed enable us to approximate closely the null distributions of various test statistics about these tables, as well as to obtain an accurate estimate of the total number of tables satisfying the constraints. Our method compares favorably with other existing Monte Carlo algorithms. MCMC methods usually need to design a particular set of Markov bases for each configuration of structural zeros, which makes the algorithms hard to implement besides the thorny convergence issues. The available MCMC approaches typically have a very long autocorrelation time, especially for large tables because the Markov chain is making "small" moves at each step. SIS generates independent tables and can take care of structural zeros easily by directly setting those entries to be zero.

Different orderings of the row sums and column sums sometimes can affect the efficiency of the SIS algorithms. In the finch data example, we found that ordering the column sums from largest to smallest works best. Another option is to sample rows instead of columns. See Chen, et al. (2005) for more discussion. The techniques developed here may also be applied to the case when it is desirable to fix the observed values (not necessarily zero) of certain entries. For example, Smith, Forster, and McDonald (1996) suggest a test of quasi-independence that fixes the entries on the diagonal.

A. PROOF OF THEOREM 1

The proof is similar to the proof of Theorem 1 of Chen et al. (2005). The following algorithm generates tables uniformly from all $m \times n$ zero-one tables with given row sums $r_1, \ldots, r_m$, first column sum c_1, and the set Ω of structural zeros defined in (5.1).

Algorithm:

1. Choose r_i $(i = 1, \ldots, m)$ positions that are not in Ω uniformly from the ith row and put 1's in. The choices of positions are independent across different rows.

2. Accept those tables with given first column sum c_1.

At Step 1, the first cell at the ith row will be chosen to put 1 in with probability

$$p_i = \begin{cases} \binom{n-1-g_1}{r_1-1} \Big/ \binom{n-g_1}{r_1} = r_1/(n-g_1), & \text{if} \quad (i,1) \notin \Omega, \\ 0, & \text{if} \quad (i,1) \in \Omega, \end{cases}$$

where g_1 is the total number of structural zeros in the first row. After Step 1, the marginal distribution of the first column is the same as the distribution of $\mathbf{Z}$ (defined by (5.3)) with

$p_i = r_i/(n - g_1)$. After Step 2, the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ (defined by (5.4)) given $S_{\mathbf{Z}} = c_1$ with $p_i = r_i/(n - g_1)$, because the tables whose first column sum is not c_1 are rejected at Step 2.

B. PROOF OF THEOREM 3

By using the theorem in Mirsky (1971, p. 205), Chen (2006) shows that the necessary and sufficient conditions for the existence of an $m \times (n-1)$ zero-one matrix with given row sums $r_1^{(2)}, \ldots, r_m^{(2)}$, column sums $c_1^{(2)}, \ldots, c_{n-1}^{(2)}$, and the set of structural zeros $\Omega^{(2)}$, is that for all $I \subset \{1, \ldots, m\}$, $J \subset \{1, \ldots, n-1\}$,

$$\sum_{i \in I, j \in J} a_{ij} \geq \sum_{i \in I} r_i^{(2)} - \sum_{j \notin J} c_j^{(2)}. \tag{B.1}$$

This implies that (B.1) holds for $I = \{1, \ldots, k\}$ and $J = \{1, \ldots, l\}$, $k = 1, \ldots, m$; $l = 1, \ldots, n-1$. Therefore (6.5) is a necessary condition.

To prove that (6.5) is also a sufficient condition, it is enough to show that (6.5) implies

$$\sum_{i=1}^{k} \sum_{j=1}^{l} a_{u_i, j} \geq \sum_{i=1}^{k} r_{u_i}^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)}, \quad k = 1, \ldots, m;\ l = 1, \ldots, n-1, \tag{B.2}$$

where $r_{u_1}^{(2)}, \ldots, r_{u_m}^{(2)}$ are the reordered row sums such that $r_{u_1}^{(2)} \geq \cdots \geq r_{u_m}^{(2)}$, because (B.2) is a necessary and sufficient condition for the existence of an $m \times (n-1)$ zero-one table with given row sums $r_1^{(2)}, \ldots, r_m^{(2)}$, column sums $c_1^{(2)}, \ldots, c_{n-1}^{(2)}$, and the set of structural zeros $\Omega^{(2)}$ (see Theorem 2). Notice that $r_1 \geq \cdots \geq r_m$ and $r_i^{(2)} = r_i - b_i$, where b_i is 0 or 1, $i = 1, \ldots, m$, therefore, for every set of neighboring rows $(r_k^{(2)}, r_{k+1}^{(2)})$, either $r_k^{(2)} \geq r_{k+1}^{(2)}$ or $r_k^{(2)} + 1 = r_{k+1}^{(2)}$. We show in the following that for every set of neighboring rows $(r_k^{(2)}, r_{k+1}^{(2)})$ with $r_k^{(2)} + 1 = r_{k+1}^{(2)}$, we can switch these two rows while maintaining property (6.5). If this is proved, then after finite number of such switches, we can change the current ordering of rows $r_1^{(2)}, \ldots, r_m^{(2)}$ to the ordering $r_{u_1}^{(2)}, \ldots, r_{u_m}^{(2)}$ required by (B.2), while maintaining property (6.5), and the theorem is thus proved.

Suppose $(r_k^{(2)}, r_{k+1}^{(2)})$ are two neighboring rows with $r_k^{(2)} + 1 = r_{k+1}^{(2)}$, then (6.5) implies that

$$\sum_{i=1}^{k-1} \sum_{j=1}^{l} a_{ij} \geq \sum_{i=1}^{k-1} r_i^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)}, \quad l = 1, \ldots, n-1, \tag{B.3}$$

$$\sum_{i=1}^{k-1} \sum_{j=1}^{l} a_{ij} + \sum_{j=1}^{l} a_{kj} \geq \sum_{i=1}^{k-1} r_i^{(2)} + r_k^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)}, \quad l = 1, \ldots, n-1, \tag{B.4}$$

$$\sum_{i=1}^{k+1} \sum_{j=1}^{l} a_{ij} \geq \sum_{i=1}^{k+1} r_i^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)}, \quad l = 1, \ldots, n-1. \tag{B.5}$$

To show that we can switch the two rows (r_k, r_{k+1}) while maintaining property (6.5), it is enough to prove that

$$\sum_{i=1}^{k-1}\sum_{j=1}^{l} a_{ij} + \sum_{j=1}^{l} a_{k+1,j} \geq \sum_{i=1}^{k-1} r_i^{(2)} + r_{k+1}^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)}, \quad l = 1, \ldots, n-1. \tag{B.6}$$

We will prove (B.6) by contradiction. Suppose (B.6) does not hold, that is,

$$\sum_{i=1}^{k-1}\sum_{j=1}^{l} a_{ij} + \sum_{j=1}^{l} a_{k+1,j} < \sum_{i=1}^{k-1} r_i^{(2)} + r_{k+1}^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)}, \tag{B.7}$$

for some $1 \leq l \leq n-1$.

Then (B.7), (B.4), and the fact that $r_k^{(2)} + 1 = r_{k+1}^{(2)}$ lead to

$$\sum_{j=1}^{l} a_{k+1,j} \leq \sum_{j=1}^{l} a_{kj}. \tag{B.8}$$

According to rule (6.3),

$$y(k) \leq y(k+1). \tag{B.9}$$

Since $b_k = 1$, we have

$$y(k) > 1. \tag{B.10}$$

Based on (B.9) and (B.10), we immediately have

$$\sum_{j=1}^{l} a_{k+1,j} \geq \sum_{j=1}^{l} a_{kj}. \tag{B.11}$$

So (B.8) and (B.11) imply

$$\sum_{j=1}^{l} a_{k+1,j} = \sum_{j=1}^{l} a_{kj}, \tag{B.12}$$

Combining (B.12), (B.7), (B.4), and the fact that $r_k^{(2)} + 1 = r_{k+1}^{(2)}$, we have

$$\sum_{i=1}^{k-1}\sum_{j=1}^{l} a_{ij} + \sum_{j=1}^{l} a_{kj} = \sum_{i=1}^{k-1} r_i^{(2)} + r_k^{(2)} - \sum_{j=l+1}^{n} c_j^{(2)}, \text{ for some } 1 \leq l \leq n-1, \tag{B.13}$$

Therefore

$$\sum_{j=1}^{l} a_{kj} \leq r_k^{(2)}, \tag{B.14}$$

based on (B.13) and (B.3). Combining (B.14) and (B.7), we have

$$\begin{aligned}\sum_{i=1}^{k+1}\sum_{j=1}^{l} a_{ij} &= \sum_{i=1}^{k-1}\sum_{j=1}^{l} a_{ij} + \sum_{j=1}^{l} a_{k+1,j} + \sum_{j=1}^{l} a_{kj} \\ &< \sum_{i=1}^{k-1} r_i^{(2)} + r_{k+1}^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)} + r_k^{(2)} \\ &= \sum_{i=1}^{k+1} r_i^{(2)} - \sum_{j=l+1}^{n-1} c_j^{(2)}\end{aligned}$$

which contradicts (B.5). The theorem is thus proved.

C. PROOF OF COROLLARY 1

In Theorem 11.5.1 of Mirsky (1971, p. 205), let the lower and upper bounds for the ith row sum be r_i, $i = 1, \ldots, m$, let the lower and upper bounds for the jth column sum be c_j, $j = 1, \ldots, n$, and define

$$c_{ij} = \begin{cases} 0, & \text{if} \quad (i, j) \in \Omega, \\ \sum_{d=1}^{n} r_d + 1, & \text{if} \quad (i, j) \notin \Omega. \end{cases} \tag{C.1}$$

Then necessity of (7.1) is obvious from Mirsky's Theorem 11.5.1. To prove sufficiency, we need to show that (7.1) can guarantee condition (4) in Mirsky's Theorem 11.5.1, that is,

$$\sum_{i\in I, j\in J} c_{ij} \geq \max\left\{\sum_{i\in I} r_i - \sum_{j\notin J} c_j,\ \sum_{j\in J} c_j - \sum_{i\notin I} r_i\right\} \tag{C.2}$$

for all $I \subset \{1, \ldots, m\}$, $J \subset \{1, \ldots, n\}$. Note from $\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$ that

$$\sum_{i\in I} r_i - \sum_{j\notin J} c_j = \sum_{j\in J} c_j - \sum_{i\notin I} r_i,$$

therefore the right hand side of (C.2) can be simplified to $\sum_{i\in I} r_i - \sum_{j\notin J} c_j$. If there are $i' \in I$, $j' \in J$ such that $(i', j') \notin \Omega$, then $c_{i'j'} = \sum_{d=1}^{n} r_d + 1$ alone will be larger than the right hand side of (C.2). Thus we only need to focus on I and J satisfying $I \times J \subset \Omega$ and the corollary follows immediately.

ACKNOWLEDGMENTS

The author thanks the Editor and reviewers for many helpful suggestions and Zhenglei Gao for implementing Rao et al.'s algorithm on the manager data. This research was partly supported by the National Science Foundation grants DMS-0203762 and DMS-0503981.

[Received December 2004. Revised September 2006.]

REFERENCES

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131–153.

Aoki, S., and Takemura, A. (2005), "Markov Chain Monte Carlo Exact Tests for Incomplete Two-Way Contingency Tables," *Journal of Statistical Computation and Simulation*, 75, 787–812.

Bender, E. A. (1974), "The Asymptotic Number of Non-negative Integer Matrices with Given Row and Column Sums," *Discrete Mathematics*, 10, 217–223.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: M.I.T. Press.

Chen, S. X., and Liu, J. S. (1997), "Statistical Applications of the Poisson-Binomial and Conditional Bernoulli Distributions," *Statistica Sinica*, 7 875–892.

Chen, X. H., Dempster, A. P., and Liu, J. S. (1994), "Weighted Finite Population Sampling to Maximize Entropy," *Biometrika*, 81, 457–469.

Chen, Y. (2006), "Simple Existence Conditions for Zero-One Matrices With at Most One Structural Zero in Each Row and Column," *Discrete Mathematics*, 306, 2870–2877.

Chen, Y., Diaconis, P., Holmes, S. P., and Liu, J. S. (2005), "Sequential Monte Carlo Methods for Statistical Analysis of Tables," *Journal of the American Statistical Association*, 100, 109–120.

Chen, Y., Dinwoodie, I. H., and Sullivant, S. (2006), "Sequential Importance Sampling for Multiway Contingency Tables," *The Annals of Statistics*, 34, 523–545.

Chen, Y., and Small, D. (2005), "Exact Tests for the Rasch Model via Sequential Importance Sampling," *Psychometrika*, 70, 11–30.

Connor, E. F., and Simberloff, D. (1979), "The Assembly of Species Communities: Chance or Competition?" *Ecology*, 60, 1132–1140.

Cook, R. R., and Quinn, J. F. (1995), "The Influence of Colonization in Nested Species Subsets," *Oecologia*, 102, 413–424.

Cox, D. R. (1988), "Some Aspects of Conditional and Asymptotic Inference," *Sankhya*, Series A, 50, 314–337.

Diaconis, P., and Efron, B. (1985), "Testing for Independence in a Two-Way Table: New Interpretations of the Chi-Square Statistic," *The Annals of Statistics*, 13, 845–874.

Fienberg, S. E. (1980), *The Analysis of Cross-Classified Categorical Data* (2nd ed.), Cambridge, MA: M.I.T. Press.

Goodman, L. A. (1968), "The Analysis of Cross-Classified Data: Independence, Quasi-independence and Interactions in Contingency Tables With or Without Missing Entries," *Journal of the American Statistical Association*, 63, 1091–1131.

Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association*, 76, 33–50.

Kelderman, H. (1984), "Loglinear Rasch Model Tests," *Psychometrika*, 49, 223–245.

Kong, A., Liu, J. S., and Wong, W. H. (1994), "Sequential Imputations and Bayesian Missing Data Problems," *Journal of the American Statistical Association*, 89, 278–288.

Krackhardt, D. (1987), "Cognitive Social Structures," *Social Networks*, 9, 109–134.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, New York: Wiley.

Manly, B. F. (1995), "A Note on the Analysis of Species Co-occurrences," *Ecology*, 76, 1109–1115.

Mirsky, L. (1971), *Transversal Theory*, New York: Academic Press.

Ploog, D. W. (1967), "The Behavior of Squirrel Monkeys (*Saimiri sciureus*) as Revealed by Sociometry, Bioacoustics, and Brain Stimulation," in *Social Communication Among Primates*, ed. S. Altmann, Chicago: University of Chicago Press, 149–184.

Rao, A. R., Jana, R., and Bandyopadhyay, S. (1996), "A Markov Chain Monte Carlo Method for Generating Random (0,1)-Matrices with Given Marginals," *Sankhya*, Series A, 58, 225–242.

Rapallo, F. (2006), "Markov Bases and Structural Zeros," *Journal of Symbolic Computation*, 41, 164–172.

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Danish Institute for Educational Research.

Reid, N. (1995), "The Roles of Conditioning in Inference," *Statistical Science*, 10, 138–157.

Roberts, A., and Stone, L (1990), "Island-Sharing by Archipelago Species," *Oecologia*, 83, 560–567.

Roberts, J. M., Jr. (2000), "Simple Methods for Simulating Sociomatrices with Given Marginal Totals," *Social Networks*, 22, 273–283.

Sanderson, J.G. (2000), "Testing Ecological Patterns," *American Scientist*, 88, 332–339.

Sanderson, J. G., Moulton, M. P., and Selfridge, R. G. (1998), "Null Matrices and the Analysis of Species Co-occurrences," *Oecologia*, 116, 275–283.

Smith, P. W. F., Forster, J. J., and McDonald, J. W. (1996), "Monte Carlo Exact Tests for Square Contingency Tables," *Journal of the Royal Statistical Society*, Series A, 159, 309–321.

Snijders, T. A. B. (1991), "Enumeration and Simulation Methods for 0-1 Matrices With Given Marginals," *Psychometrika*, 56, 397–417.

Wasserman, S., and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.

Wilson, J. B. (1987), "Methods for Detecting Non-randomness in Species Co-occurrences: A Contribution," *Oecologia*, 73, 579–582.