Sampling and Counting Zero-One Tables For Fixed-Margin Matrices

Abdulkadir Çelikkanat

June 5, 2016

1 Introduction

Problems of testing hypotheses about zero-one 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.

A good analytic approximations to the null distributions of various test statistics are harder 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. Snijders (1991) was the first to consider importance sampling in the context of zero-one tables with fixed marginal sums and structural zeros. Chen et al. [1] proposed a method to sample tables with fixed marginals, and then extended it for tables given set of structural zeros [3]. In this project, I will consider this SIS method for two different data samples.

Definition A *zero-one table* is a matrix in which each entry is either 0 or 1. An entry is referred as a *structural zero* if it is constrained to be zero

	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	14
Medium ground finch	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	13
Small ground finch	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	14
Sharp-beaked ground finch	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1	10
Cactus ground finch	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	0	12
Large cactus ground finch	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2
Large tree finch	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	0	10
Medium tree finch	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Small tree finch	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	10
Vegeterian finch	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	11
Woodpecker finch	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	6
Mangrove finch	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Warbler finch	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
	4	4	11	10	10	8	9	10	8	9	3	10	4	7	9	3	3	

Table 1: Occurrence Matrix for Darwin's Finch Data

Sequential Importance Sampling

Given the row sums $\mathbf{p} = (p_1, p_2, ..., p_m)$, column sums $\mathbf{q} = (q_1, q_2, ..., q_n)$, and let $\sum_{pq}$ be the set of all mxn zero-one tables with row sum $\mathbf{p}$, column sum $\mathbf{q}$.

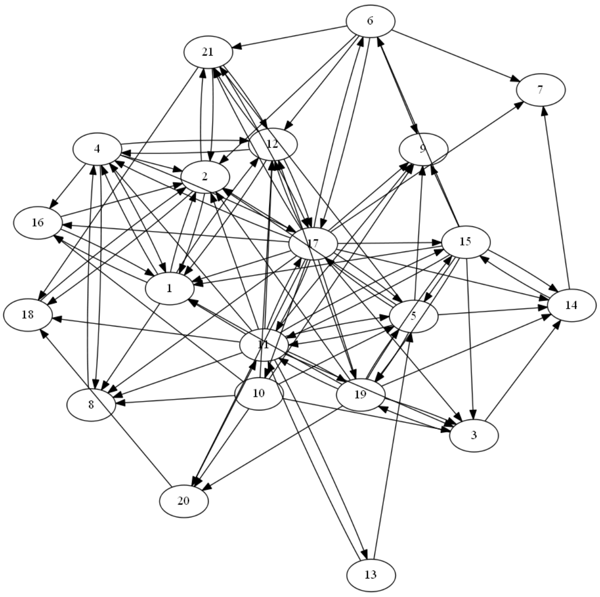


Figure 1: Friendship relation between 21 high-tech managers.

Let p(T) = $1/|\Sigma_{pq}|$ be the uniform distribution over $\sum_{pq}$. Let q be a proposal distribution where q(T)¿0 for all $T \in \Sigma_{pq}$, then we have

$$E_q\left[\frac{1}{q(T)}\right] = \sum_{T\in\Sigma_{pq}} \frac{1}{q(T)} q(T) = |\Sigma_{pq}|.$$

Therefore, we can estimate the number of zero-one tables $|\Sigma_{pq}|$ by

$$|\widehat{\Sigma_{pq}}| = \frac{1}{N}\sum_{1}^{N}\frac{1}{q(T_i)}$$

where N independent identically distributed samples T_1,...,T_N drawn from $q(T)$.

In order to measure the overall efficiency of an importance sampling algorithm, the *effective sample size (ESS)* can be computed, which is defined as

$$ESS = \frac{N}{1+cv^2} \qquad\qquad cv^2 = \frac{var_q\{p(T)/q(T)\}}{E_q^2\{p(T)/q(T)\}}$$

When cv^2 is smaller, the two distributions becomes closer to each other.

Sampling Zero-One Tables

The main problem in importance sampling is to choose a good proposal distribution q(.).
Note that

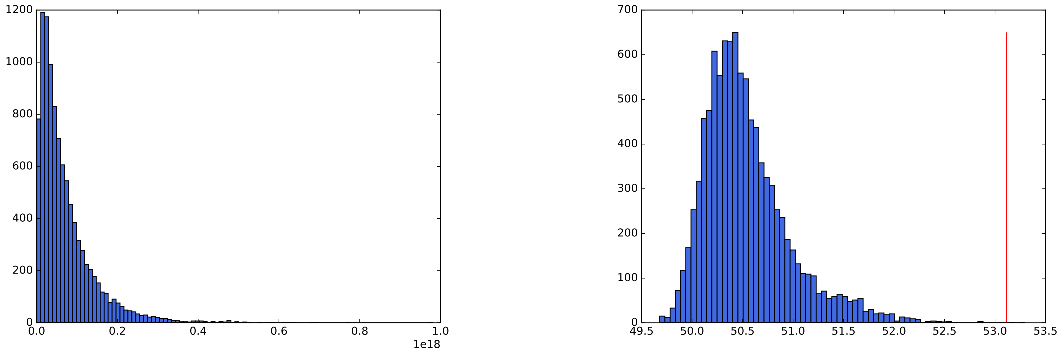
$$q(T = (c_1, c_2, ..., c_n)) = q(c_1)q(c_2|c_1)q(c_3|c_2, c_1)...q(c_n|c_{n-1}, ..., c_1)$$

where $c_1, c_2, ..., c_n$ denote the configurations of the columns of **T**. This factorization suggests a method to generate a table sequentially, column by column. More precisely, the first column of the table is sampled conditional on its marginal sum q_1. The row sums are updated conditional on the realization of the first column, then the second column is sampled conditional on the column sum q_2. The recursive structure of the methods leads us to *sequential importance sampling.*

Sampling From the Conditional Poisson Distribution

Let

$$Z = (Z_1, ..., Z_m)$$



(a) Histogram of 10,000 importance weights (b) Approximated Null Distribution of the Test Statistic

be independent Bernoulli trials with probability of successes $\mathbf{p} = (p_1,...,p_m)$. Then the random variable

$$S_Z = Z_1 + ... + Z_m$$

is said to follow the *Poisson-binomial distribution*.
The conditional distribution of $\mathbf{Z}$ given S_Z is called *conditional poisson (CP)*. If we say $w_i = p_i/(1-p_i)$, then we get

$$P(Z_1 = z_1, ..., Z_m = z_m | S_z = c) \propto \prod_{k=1}^{m} w_k^{z_k}$$

One of the five methods given in [2] to sample from conditional-poisson distribution, the *drafting sampling method*, is adopted for this problem. Let A_k (k=0,...,c) be the set of selected units after k draws without replacement. Hence, $A_0 = \emptyset$, and A_c is the final sample obtained. At the kth step of the drafting sampling (k=1,...,c), a unit $j \in \{1, ..., m\}/A_{k-1}$ is chosen into the sample with probability

$$P(j, \{1, ..., m\}/A_{k-1}) = \frac{w_j R(c-k, \{1, ..., m\}/A_{k-1})}{(c-k+1) R(c-k+1, \{1, ..., m\}/A_{k-1})}$$

where

$$R(s, A) = \sum_{B \subset A, |B|=s} \left(\prod_{i \in B} w_i \right)$$

Justification of the Conditional-Poisson Sampling

Theorem For the uniform distribution over all mxn zero-one tables with given row sums $p_1, ...p_m$ and the first column sum q_1, the marginal distribution of the first column is the same as the conditional distribution of $\mathbf{Z}$ given $S_Z = q_1$ with $p_i = r_i/n$

Since the desired true marginal distribution for the first column c_1 is $p(c_1) = P(c_1 | p_1, ..., p_m, q_1, ..., q_n)$, it is natural to let the proposal distribution of t_1, $q(t_1) = P(c_1 | p_1, ..., p_m, q_1)$, which is exactly conditional-poisson distribution with $p_i = r_i/n$. After sampling the first l-1 columns, then remaining number of columns $n-(l-1)$, and row sums $r_i^{(l)}$ are updated. Further, the column l can be generated with the CP sampling technique with the weights $r_i/[n-(l-1)-r_i^{(l)}]$

Applications and Simulations

The sequential importance sampling (SIS) procedure described in the previous sections are firstly tested over the $12x12$ zero-one tables of row and column sums equal to 2.
The estimated number of the tables is found as $(2.14025 \pm 0.02179)x10^{16}$ for 10,000 samples, where the exact answer is 21,959,547,410,077,200. The value of cv^2 is 0.03744, which is close to the value 0.04 given in the work of Chen et al [1].

For testing whether there is competition between species, Roberts and Stone(1990) suggested the test statistics

$$\widehat{S^2} = \frac{1}{m(m-1)} \sum_{i \neq j} s_{ij}^2$$

where m is the number of species, S=(s_{ij})=AA^T, and A is the occurrence matrix. For the finch data **??**, the observed statistic is 53.1. The estimated p value of this statistic for 10,000 samples is found as 0.0002 which is in the interval $[4 - 2.8, 4 + 2.8] * 10^{-4}$ given in [1]. The estimated total number of zero-one tables is $7.1908x10^{16}$ where the correct answer is 67,149,106,137,567,626

In the following example, a tendency towards mutuality in the friendship network among 21 high-tech managers (table 1) will be checked with the test statistic given by Snijders(1991).

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

For 10,000 samples, the estimated number of tables with the same margin in figure 1 and zero diagonal is computed as $(5.2315 \pm 0.0572)x10^{45}$, p-value as 0. The algorithm produced 62 bad samples, but it is still efficient, cv^2 is computed as 0.187

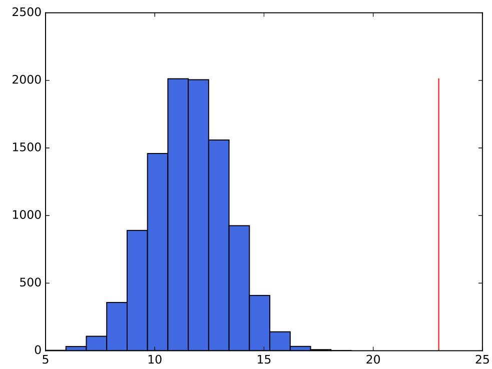


Figure 3: Approximated Null Distribution of the Test Statistic for manager data

References

[1] Diaconis P. Holmes S. P. Chen, Y. and J. S Liu. Sequential monte carlo methods for statistical analysis of tables. (100):109–120, 2005.

[2] X. H.; A. P. Dempster; J. S. Liu Chen. Weighted finite population sampling to maximize entropy. (81):457, 1994.

[3] Y Chen. Conditional inference on tables with structural zeros. (16):445–467, 2007.