Published in Statistics and Probability Letters in 2021:
https://doi.org/10.1016/j.spl.2021.109047

Limiting properties of an equiprobable sampling scheme for 0-1 matrices

Louis-Paul Rivest

Department of Mathematics and Statistics, Université Laval, Québec, G1V 0A6 Canada

Abstract

A sampling scheme that selects a random 0-1 matrix of size $N \times M$ uniformly in the set of 0-1 matrices with predetermined row and column totals is investigated. The limits, as M goes to ∞ and N is fixed, of the column relative frequencies is derived. The limiting values give a sampling design for a population of N units that generalizes the conditional Poisson sampling design introduced by Hajek. A method to calculate the joint selection probabilities for this new design using the known single inclusion probabilities is presented. Numerical examples show that the limiting theorem provides good approximations to the fixed M column probabilities of a random matrix.

Keywords: Conditional Poisson sampling, Entropy, Null model analysis, Occupancy data, Weak law of large number

2010 MSC: 62D05, 60F05

1. Introduction

Let $\mathbf{Z}$ be an $N \times M$ matrix whose entries $\{Z_{ij} : i = 1, \ldots, N; j = 1, \ldots M\}$ are equal to either 0 or 1. We are interested in matrices with fixed row totals, $Z_{i\bullet} = m_i,\ i = 1, \ldots, N$ and fixed column totals $Z_{\bullet j} = n_j,\ j = 1, \ldots, M$ where $\{m_i\}$ and $\{n_j\}$ are positive integers that satisfy $\sum_i m_i = \sum_j m_j$. This work

*Corresponding author: Louis-Paul Rivest, Department of Mathematics and Statistics, Université Laval, Quebec Canada, G1V 0A6: Louis-Paul.Rivest@mat.ulaval.ca

[1] R functions with documentation, and supplementary material are available as annexes in the electronic version of the manuscript.

investigates limiting properties, as M goes to ∞, of random 0-1 matrices $\mathbf{Z}$ that are selected uniformly in the set of matrices with fixed margins.

Mathematical properties or random 0-1 matrices are investigated in Barvinok [1]. They are found in ecology: the rows are species and the columns are sites and they enter in Monte Carlo tests of association between species [5]. In survey sampling, these matrices are used to sample populations with a matrix structure, for instance when N sites can be sampled over M days [8].

Several simulation algorithms are available. Besag and Clifford [3] propose a Markov chain defined on the set of feasible matrices whose stationary distribution is the uniform on that set, see also [10] for a recent discussion for this approach. In the numerical section we used this so-called " trial swap" algorithm implemented in the R-package `vegan` [6]. Algorithms based on importance sampling [4] and rejective methods [8] have also been considered.

A "Weak Law of Large Number" for the column frequencies of a random matrix, as M goes to ∞, is proved in Section 2. The limit involves a generalization of the conditional Poisson sampling design. Section 3 gives a method to calculate the limiting probabilities using the known totals for the matrix margins while Section 4 investigates the small M accuracy of the limits.

2. A limit theorem for random 0-1 matrices

2.1. A generalization of the conditional Poisson sampling design

In survey sampling, the conditional Poisson sampling (CPS) design for selecting a sample of size n in a population of size N, with selection probabilities $\alpha_i,\ i = 1, \ldots, N$ satisfying $\sum_{i=1}^{N} \alpha_i = n$, gives sample $\omega = (\omega_1, \ldots, \omega_N)$, where $\omega_i = 1$ if unit i is sampled and 0 otherwise, a probability equal to

$$p_\omega = \frac{\exp(\sum_{i=1}^{N} \lambda_i \omega_i)}{\sum_{(n)} \exp(\sum_{i=1}^{N} \lambda_i \varpi_i)} \quad \text{if} \quad \sum_{i=1}^{N} \omega_i = n, \tag{1}$$

where index (n) means that the sum is over the $\binom{N}{n}$ possible samples $\varpi = (\varpi_1, \ldots, \varpi_N)$ of size n, see Tillé [9, Section 5.6]. The parameters $\{\lambda_i\}$ are

selected in such a way that the single inclusion probabilities are equal to α_i, that is $\sum_{(n)} \omega_i p_\omega = \alpha_i,\ i = 1, \ldots, N$. This design maximizes the entropy, $-\sum_{(n)} p_\omega \log(p_\omega)$, for given inclusion probabilities $\{\alpha_i\}$.

Consider a generalization of the CPS design where the sample size is random. The probability that it is equal to n is $q_n, n = 1, \ldots, N$ where $\sum_{n=1}^N q_n = 1$ and the selection probabilities $\alpha_i,\ i = 1, \ldots, n$ satisfy $\sum_{i=1}^N \alpha_i = \sum_{n=1}^N n q_n$. The set of possible samples Ω, has size $\sum \binom{N}{n}$ where the sum is on the n for which $q_n > 0$. The entropy is $\mathcal{I}(\mathbf{p}) = -\sum_{\omega \in \Omega} p_\omega \log(p_\omega)$, where $\mathbf{p}$ is the vector for the p_ω's; it needs to be maximized under the constraints

$$\sum_{\omega \in \Omega} \omega_i p_\omega = \alpha_i, i = 1, \ldots, N \quad \text{and} \quad \sum_{\omega \in \Omega} \psi_n(\omega) p_\omega = q_n, n = 1, \ldots, N, \tag{2}$$

where $\psi_n(\omega) = 1$ if $\sum \omega_i = n$ and 0 otherwise. The solution to this problem is given in the next proposition.

Proposition 1. *The following properties hold*

1. *The vector* $\mathbf{p}^*$ *defined by*

$$p^*_\omega = \sum_{n=1}^N q_n \psi_n(\omega) \frac{\exp(\sum_{i=1}^N \lambda_i \omega_i)}{\sum_{(n)} \exp(\sum_{i=1}^N \lambda_i \varpi_i)} \quad \omega \in \Omega, \tag{3}$$

   where the λ_i*'s are evaluated by solving* $\sum_\Omega \omega_i p^*_\omega = \alpha_i, i = 1, \ldots, N$*, maximises the entropy* $\mathcal{I}(\mathbf{p})$ *under the constraints (2).*

2. *If* $\mathbf{p}$ *satisfies (2) and* $||\mathbf{p}^* - \mathbf{p}|| = \sqrt{\sum_{\omega \in \Omega} (p^*_\omega - p_\omega)^2}$ *then* $\mathcal{I}(\mathbf{p}^*) - \mathcal{I}(\mathbf{p}) > ||\mathbf{p}^* - \mathbf{p}||^2/2$.

Proof: The proof for 1. mimics that in Tillé [9, Section 5.6]. As the entropy is a concave function, its maximum is determined using the following Lagrangian,

$$\mathcal{L}(\lambda, \gamma, p_\omega) = \sum_{i=1}^N \lambda_i (\sum_\omega \omega_i p_\omega - p_i) + \sum_{n=1}^N \gamma_n (\sum_\omega \psi_n(\omega) p_\omega - q_n) - \sum_\omega p_\omega \log(p_\omega).$$

The partial derivative with respect to p_ω is

$$\frac{\partial \mathcal{L}(\lambda, \gamma, p_\omega)}{\partial p_\omega} = \sum_{i=1}^{N} \lambda_i \omega_i + \sum_{n=1}^{N} \psi_n(\omega)\gamma_n - 1 - \log(p_\omega), \quad \omega \in \Omega.$$

Equating this expression to 0 gives $p^*_\omega = \exp\left\{\sum_{i=1}^N \lambda_i\omega_i + \sum_{n=1}^N \gamma_n\psi_n(\omega)\right\}$, where the parameters $\{\lambda_i\}$ and $\{\gamma_n\}$ are determined by the constraints (2). The condition involving q_n implies that $\sum_{\omega\in\Omega} \psi_n(\omega)\exp(\sum_{i=1}^N \lambda_i\omega_i + \gamma_n) = q_n$. This gives the following expression that leads to formula (3) for p^*_ω,

$$\gamma_n = \log(q_n) - \log\{\sum_{(n)} \exp(\sum_{i=1}^{N} \lambda_i \varpi_i)\}. \tag{4}$$

To prove 2., consider $\mathcal{I}_0(\mathbf{p}) = -\sum_{\omega\in\Omega} p_\omega \log(p_\omega) + \sum_{\omega\in\Omega}(\mathbf{p}^*_\omega - \mathbf{p}_\omega)^2/2$. This is a concave function defined on the set of probability vectors satisfying (2) as its matrix of second order partial derivatives, the diagonal matrix with entries $\{-1/p_\omega + 1 :\ \omega \in \Omega\}$, is negative definite. Thus $\mathcal{I}_0(\mathbf{p})$ is concave and the proof that it reaches its maximum at $\mathbf{p} = \mathbf{p}^*$ is similar to that for $\mathcal{I}(\mathbf{p})$ presented in 1. Therefore $\mathcal{I}_0(\mathbf{p}^*) > \mathcal{I}_0(\mathbf{p})$; this proves the result.

□

The parameters $\{\lambda_i,\ i = 1, \ldots, N\}$ of the maximum entropy design defined by the probabilities $\{\alpha_i,\ i = 1, \ldots, N\}$ and $\{q_n,\ n = 1, \ldots, N\}$ are over-determined. They can be calculated by setting $\lambda_1 = 0$ and by solving simultaneously the following $N - 1$ equations:

$$\alpha_i = \sum_{n=1}^{N} q_n \frac{\sum_{(n)} \omega_i \exp(\sum \omega_i\lambda_i)}{\sum_{(n)} \exp(\sum \omega_i\lambda_i)},\ i = 2, \ldots, N. \tag{5}$$

The parameters γ_n are then given by (4).

2.2. Limiting properties of the uniform sampling design for matrices

Consider a 0-1 matrix $\mathbf{Z}$ of size $N \times M$ with fixed row and column totals, $Z_{i\bullet} = m_i,\ i = 1, \ldots, N$ and $Z_{\bullet j} = n_j,\ j = 1, \ldots, M$ for positive integers $\{m_i\}$

and $\{n_j\}$ satisfying $\sum m_i = \sum n_j$. An $N \times 1$ vector of sample indicators $\omega \in \Omega$ gives a possible column for $\mathbf{Z}$ where Ω contains 0-1 vectors that sum to one of the column totals $\{n_j\}$. Let x_ω be the number of occurrences of column ω among the M columns of $\mathbf{Z}$. The frequencies $\{x_\omega : \omega \in \Omega\}$ are a tool to describe the matrices $\mathbf{Z}$. The constraints on the margins are similar to (2):

$$\sum_{\omega \in \Omega} \omega_i x_\omega = m_i \quad i = 1, \ldots, N \quad \text{and} \quad \sum_{\omega \in \Omega} \psi_n(\omega) x_\omega = c_n \quad n = 1, \ldots, N, \tag{6}$$

where $c_n,\ n = 1, \ldots, N$, are the frequencies of column total n among $\{n_j : j = 1, \ldots, M\}$ and $\psi_n(\omega)$ is equal to 1 if $\sum \omega_i = n$ and 0 otherwise as defined in Section 2.1. These equations imply that $\sum_{\omega \in \Omega} x_\omega = \sum_{n=1}^{N} c_n = M$ and the number of entries equal to 1 in $\mathbf{Z}$, is $\sum_{i,j} Z_{ij} = \sum_{n=1}^{N} n c_n = \sum_{i=1}^{N} m_i$.

Given a set of frequencies $\{x_\omega : \omega \in \Omega\}$ satisfying the constraints (6) one can construct several matrices $\mathbf{Z}$ as $\{x_\omega\}$ is invariant to permutations of the columns of $\mathbf{Z}$ within subsets with the same column total. Indeed the number of different matrices $\mathbf{Z}$ associated with a set of frequencies $\{x_\omega : \omega \in \Omega\}$ is $\prod c_n! / \{\prod_{\omega \in \Omega} x_\omega!\}$. Summing these expressions for all the possible sets $\{x_\omega\}$ gives the following expression for $\mathcal{N}_M$, the total number of matrices $\mathbf{Z}$ whose row and column totals are $\{Z_{i\bullet} = m_i\}$ and $\{Z_{\bullet j} = n_j\}$, namely

$$\mathcal{N}_M = \sum_{\{z_\varpi\} \in \mathcal{X}_M} \prod c_n! / \{ \prod_{\varpi \in \Omega} z_\varpi! \}, \tag{7}$$

where $\mathcal{X}_M$ is the ensemble of the sets $\{x_\omega : \omega \in \Omega\}$ satisfying (6).

Suppose now that $\mathbf{Z}$ is selected at random. Define the random vector $\{X_\omega : \omega \in \Omega\}$ as the number of occurrences of ω among the columns of $\mathbf{Z}$ and $\widehat{\mathbf{p}}_M = \{\hat{p}_{M\omega} = X_\omega / M : \omega \in \Omega\}$ the vector of relative frequencies. The probability function of the random vector $\{X_\omega : \omega \in \Omega\}$ is

$$P_M[X_\omega = x_\omega : \{x_\omega\} \in \mathcal{X}_M] = \frac{1}{\mathcal{N}_M} \frac{\prod c_n!}{\prod_{\omega \in \Omega} x_\omega!}. \tag{8}$$

Equation (8) defines a discrete probability function on the set $\mathcal{X}_M$.

To exemplify the notation used in this section, consider the following 4×6 matrix, belonging to the family studied in Section 4.1,

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The row totals of $\mathbf{Z}$ are $m_1 = 4, m_2 = m_3 = m_4 = 2$ while the column totals are $n_1 = n_2 = n_4 = n_5 = 1$ and $n_3 = n_6 = 3$ and the non null frequencies of the column totals are $c_1 = 4$ and $c_3 = 2$. The set Ω of possible columns has size 8: it contains all the subsets of size 1 and 3 in a set of size 4. The non null column frequencies for $\mathbf{Z}$ are $x_{(1,0,0,0)} = 2, x_{(0,0,0,1)} = x_{(1,1,1,0)} = x_{(0,0,1,0)} = x_{(1,1,0,1)} = 1$ and the corresponding values of $\hat{p}_{M\omega}$ for $\mathbf{Z}$ are obtained by dividing the frequencies by $M = 6$. The generalized CPS design associated to this matrix has selection probabilities $\alpha_1 = m_1/6 = 2/3, \alpha_2 = \alpha_3 = \alpha_4 = 1/3$ and the respective probabilities for the sample size are $q_1 = c_1/6 = 2/3$ and $q_3 = c_3/6 = 1/3$; the probabilities $p^*_{M\omega}$ are evaluated in Section 4.1. For instance $p^*_{M(1,0,0,0)} = 0.3619$ as compared to the estimate $\hat{p}_{M(1,0,0,0)} = 0.3333$ derived from $\mathbf{Z}$. The number of matrices with the same marginal totals as $\mathbf{Z}$ is $\mathcal{N}_M = 115$. A formula for this calculation is in the Supplementary Material.

The convergence in probability of $\widehat{\mathbf{p}}_M$ as M goes to ∞ is now investigated.

Proposition 2. *If, as M goes to ∞, the sequences $= m_i/M,\ i = 1, \ldots, N$ and c_n/M, for the possible sample sizes, converge to positive limits, then*

$$||\widehat{\mathbf{p}}_M - \mathbf{p}^*_M|| \to 0 \quad \textit{in probability},$$

where $\mathbf{p}^*_M$ is the vector of probabilities for the generalized CPS design defined by (3) with $\alpha_i = m_i/M,\ i = 1, \ldots, N$ and $q_n = c_n/M,\ n = 1, \ldots, N$.

Proof: To prove this result we need to show that for any fixed $\epsilon > 0$,

$$P_M\{\{X_\omega\} : ||\widehat{\mathbf{p}}_M - \mathbf{p}^*_M|| > \epsilon\} \tag{9}$$

converges to 0 where P_M is defined by (8). To simplify the presentation, this proof evaluates (8) as a ratio of sums involving terms such as $M!/(\prod x_\omega!)$. By Stirling's approximation the following inequalities are valid for any $x > 0$, $e\sqrt{x}x^x e^{-x} \geq x! \geq \sqrt{2\pi x}x^x e^{-x}$. For a set $\{x_\omega\}$ satisfying (6), this gives

$$\begin{aligned}
\frac{M!}{\prod x_\omega!} &\approx \exp\left(\frac{\log M}{2} + M\log M - M - \frac{\sum_\omega \log x_\omega}{2} - \sum_\omega x_\omega \log(x_\omega) + \sum_\omega x_\omega\right) \\
&\approx \exp\left\{\frac{\log M}{2} - M\sum_\omega p_\omega \log(p_\omega) - \frac{\sum_\omega \log x_\omega}{2}\right\} \\
&\approx \exp\left\{\frac{\log M}{2} + M\mathcal{I}(\mathbf{p}^*_M) - \frac{M}{2}||\mathbf{p}^*_M - \mathbf{p}_M||^2 - \frac{\sum_\omega \log x_\omega}{2}\right\}
\end{aligned}$$

where $\approx$ means that the ratio of the left to the right hand side belongs to an interval (C_0, C_1) for some positive constant $C_1 > C_0$ independent of M and $\mathbf{p}_M = \{x_\omega/M,\ \omega \in \Omega\}$.

Consider the vector $v^{(M)} = \{M\mathbf{p}^*_{M\omega} : \omega \in \Omega\}$. Define $\{x^*_\omega : \omega \in \Omega\}$ a vector of integers satisfying (6) such that $\sqrt{\sum_\omega (p^*_{M\omega} - x^*_\omega/M)^2}$ is $O(1/M)$. Such a vector exists; it can be constructed from $v^{(M)}$ using the following algorithm:

1. First consider the floors of $v^{(M)}$, the vector of the largest integers smaller than the corresponding entries of $v^{(M)}$. This defines a 0-1 matrix $\mathbf{Z}_0$ with $M - M_0$ columns where $M_0 < 2^N$ is bounded. $\mathbf{Z}_0$ is missing a $N \times M_0$ matrix to satisfy the constraints (6). This $N \times M_0$ matrix is defined in terms of its row and column totals. If these totals satisfy the existence condition of Gale and Ryser, see Barvinok [2], the $N \times M_0$ completion to $\mathbf{Z}_0$ exists and the difference between $\{x^*_\omega\}$ for the completed matrix and $\{M\mathbf{p}^*_M\}$ is bounded or $O(1)$.

2. If a completion does not exist, reapply 1. to a modified $v^{(M)}$ where the entries $v^{(M)}_\omega$ for the largest column total $nm = \sum \omega_i$ are rounded and

sum to c_{nm}. With this modified $v^{(M)}$ the completion of $\mathbf{Z}_0$ do not have columns that sum to nm; it is more likely to satisfy the Gale and Ryser condition as its column totals are less variable. If the margins for the completion do no satisfy Gale and Ryser condition then one rounds the entries of $v^{(M)}$ with the second largest column total and so on.

Since the entropy is differentiable $|\{\mathcal{I}(\{x^*_\omega/M\}) - \mathcal{I}(\mathbf{p}^*_M)|$ is $O(1/M)$ and $M!/\prod x^*_\omega!$ is then $\exp\{M\mathcal{I}(\mathbf{p}^*_M) + O(\log M)\}$. Now (9) is a fraction whose numerator is a sum of $M!/(\prod x_\omega!)$ on the sets $\{x_\omega\}$ for which $||\mathbf{p}^*_M - \mathbf{p}_M|| > \epsilon$. As the size of these sets is less than 2^N and each entry can take a maximum of M values, the number of possible sets in the sum is less than $\exp\{2^N \log(M))\}$ and in view of the bound derived above for each term, the sum is $\exp\left[O\{\log(M)\} + M\mathcal{I}(\mathbf{p}^*_M) - M\epsilon^2/2\right]$. Considering the approximation for $M!/\prod x^*_\omega!$, (9) is $O\{\exp(-M\epsilon^2/4)\}$ and goes to 0 as M goes to ∞.

□

The convergence in Proposition 2 is fast as $P_M\{\{x_\omega\} : ||\widehat{\mathbf{p}}_M - \mathbf{p}^*_M|| > \epsilon\}$ is $O\{\exp(-M\epsilon^2/4)\}$ as compared to $O\{1/(M\epsilon^2)\}$ for the standard Law of Large Number. Consider an $N \times M$ matrix $\mathbf{Z}$ uniformly distributed among those with fixed margins and let $\mathcal{Z}$ be a column selected at random among the M columns of $\mathbf{Z}$. $\mathcal{Z}$ identifies a sample of rows selected in a population of N rows. The underlying sampling design gives a probability of $E\{\hat{p}_{M\omega}\}$ to sample ω. When M is large this probability is nearly equal to $p^*_{M\omega}$. Thus the uniform selection of a matrix $\mathbf{Z}$ provides an approximate method to select a sample according the generalized CPS design of Section 2.1. One simply selects, at random, one of the columns of $\mathbf{Z}$. The special case of equal column totals is interesting since it provides a new, approximate, method to select a CPS sample. For single inclusion probabilities $\{\alpha_i\}$, it uses a random $N \times M$ matrix $\mathbf{Z}$ with row totals $Z_{i\bullet} = m_i$ and constant column totals $Z_{\bullet j} = n$. The integers $\{m_i\}$ satisfy $m_i \approx M\alpha_i$ and $\sum_i m_i = nM$; thus the final selection probabilities $\{m_i/M\}$ might differ slightly from $\{\alpha_i\}$. Finally note that Barvinok [1] is related to Proposition 2 as he studies limiting properties of a randomly selected $\mathbf{Z}$ when

both M and N go to ∞.

3. A Poisson regression model to evaluate the limiting probabilities

This section suggests a method to calculate the limiting probabilities $\mathbf{p}_M^*$ using the known totals for the rows and the columns of $\mathbf{Z}$. It uses a Poisson regression model for the discrete random variables $\{X_\omega : \Omega \in \Omega\}$ that is related to the uniform sampling scheme investigated in Section 2.2. This model postulates that the variables $\{X_\omega\}$ have independent Poisson distributions with means $\{\mu_\omega\}$ that depends on parameters $\{\lambda_i : i = 1, \ldots, N\}$ and $\{\gamma_n : n = 1, \ldots, N\}$, as follows

$$\log \mu_\omega = \sum_i \omega_i \lambda_i + \sum_n \psi_n(\sum \omega_i)\gamma_n \quad \omega \in \Omega. \tag{10}$$

Model (10) is a reparametrization of the lower bound model of [7] to estimate population sizes in capture recapture experiments with heterogeneous capture probabilities. Under this model the sufficient statistics are $m_i = \sum_\omega \omega_i X_\omega$, and $c_n = \sum_\omega \psi_n(\sum \omega_i) X_\omega$, $i, n = 1, \ldots, N$ and the conditional distribution of the data $\{X_\omega\}$ given the sufficient statistics $\{m_i, c_n\}$ gives a probability mass proportional to $1/\prod x_\omega!$ to a vector $\{x_\omega\}$ satisfying (6). Thus this conditional distribution is given by (8) and the uniform selection algorithm of Section 2.2 draws from the conditional distribution of $\{X_\omega : \Omega \in \Omega\}$ given $\{m_i, c_n\}$.

Another interesting feature of model (10) is that maximizing its likelihood gives the parameters $\{\lambda_i\}$ for the generalized CPS design, see (3), involved in the limiting distribution of Proposition 2. The log-likelihood for model (10) is, up to a constant, equal to

$$\mathcal{L}(\lambda, \gamma) = \sum_{n=1}^N \gamma_n c_n + \sum_{i=1}^N \lambda_i m_i - \sum_{n=1}^N \sum_{\omega \in \Omega} \psi_n(\sum \omega_i) \exp(\gamma_n + \sum_{i=1}^N \omega_i \lambda_i). \tag{11}$$

As mentioned in the discussion of (5) we set $\lambda_1 = 0$. The next proposition show that maximizing (11) amounts to solving (5).

Proposition 3. *For fixed values* $\{\lambda_1, \ldots, \lambda_N\}$*, the function* $\mathcal{L}(\lambda, \gamma)$ *is maximum at* γ_n *given by (4) with* $q_n = c_n/M$*. In addition the score equations to estimate the parameters* $\{\lambda_i\}$ *are given by (5) with* $\alpha_i = m_i/M$ *and* $q_n = c_n/M$*.*

Proof: For fixed $\{\lambda_i\}$, the value of $\{\gamma_n\}$ that maximizes $\mathcal{L}$ is obtained by setting the partial derivatives with respect to γ_n equal to 0. This gives (4). Now

$$\max_{\gamma} \mathcal{L}(\lambda_2, ..., \lambda_N, , \gamma_1, \ldots, \gamma_N) = \sum_{n=1}^{N} c_n \log \left\{ \frac{c_n}{\sum_{(n)} \exp(\sum \omega_i \lambda_i)} \right\} + \sum_{i=1}^{N} \lambda_i m_i$$

up to a constant and setting the partial derivatives, with respect λ_i, of $\max_\gamma \mathcal{L}$ gives equation (5). □

In the special case of the CPS design, with $c_n = M$ for some value of n, methods to evaluate the CPS parameters are discussed in Tillé [9, p.80]. The approach by maximization in Proposition 3 appears to be new. The implementation of this method is very simple. First create a pseudo data set $\{x_\omega : \omega \in \Omega\}$ that meets (6) and fit (10) using a standard program for generalized linear model. The generalized CPS probabilities $p^*_{M\omega}$ are given by $\hat{x}_\omega/M$ where $\hat{x}_\omega$ is the predicted value for ω. The only drawback of this approach is that it does not work when N is large, say larger than 15. Indeed memory problems are encountered with standard scoring algorithms for Poisson regression, when the size, $(2^N - 1) \times 1$, of the dependent vector $\{x_\omega : \omega \in \Omega\}$ is large.

4. Investigations of the speed of convergence in Proposition 2

This section compares the $E(\widehat{\mathbf{p}}_M)$ to its limiting approximation $\mathbf{p}^*_M$ in two examples; one evaluates explicitly $E(\widehat{\mathbf{p}}_M)$ in matrices with $N = 4$ rows while the other uses a data set with $N = 12$ species and $M = 17$ sites.

4.1. An example with N=4

This example considers $4 \times M$ matrices where $M = 3k$ for a positive integer k. The row totals are $Z_{1\bullet} = 2k$ and $Z_{2\bullet} = Z_{3\bullet} = Z_{4\bullet} = k$ while the column

totals $Z_{\bullet j}$ are 1 with relative frequency 2/3 or 3 with a relative frequency of 1/3. The special case $k = 2$ is discussed in Section 2.2. There are 8 vectors in Ω and the probability function (7) can be evaluated numerically, details are given in the Supplementary Material. Table 1 gives the expectations and the standard deviations of some relative frequencies $\hat{p}_{M\omega}$.

These matrices have $q_1 = 2/3, q_3 = 1/3$, $\alpha_1 = 2/3$ and $\alpha_2 = \alpha_3 = \alpha_4 = 1/3$. The equation for $i = 2$ in (5), with $z = \exp(\lambda_2)$, is $2z/(3+9z)+(z+2)/(z+3) = 1/3$. The solution is $z = (\sqrt{17} - 3)/4$ and $p^*_{M(1,1,1,0)} = 1/\{3(z + 3)\} = 0.1016$ while $p^*_{M(0,1,1,1)} = z/\{3(z + 3)\} = 0.0285$ and $p^*_{M(1,0,0,0)} = 2/(3 + 9z) = 0.3619$. These values are equal, up to 10^{-3}, to the $M = 150$ expectations in Table 1.

Table 1: Expectation (E) and standard deviation (SD) of $\hat{p}_{M\omega}$ in $4 \times M$ random matrices $\mathbf{Z}$ with column totals $Z_{\bullet j}$ equal to 3 (relative frequency 1/3) or 1 (relative frequency 2/3) and row totals $(2M, M, M, M)/3$

	$M = 3$		$M = 30$		$M = 150$	
Prob	E	SD	E	SD	E	SD
$\hat{p}_{M(1,1,1,0)}$	0.0952	0.1506	0.1012	0.0363	0.1015	0.0157
$\hat{p}_{M(0,1,1,1)}$	0.0476	0.1166	0.0297	0.0283	0.0287	0.0123
$\hat{p}_{M(1,0,0,0)}$	0.3810	0.1166	0.3630	0.0283	0.3621	0.0123

In Table 1, the most important relative discrepancy between the fixed M and the limiting values occurs for the smallest probability, that is for $(0, 1, 1, 1)$. Note also that the single inclusion probabilities, that is the probabilities that entries of the random matrix $\mathbf{Z}$ are equal to 1, depend on M. For instance, in a column j with $Z_{\bullet j} = 1$, $P_M(Z_{1j} = 1) = E(\hat{p}_{M(1,0,0,0)})/(2/3)$ varies with M. This is not so when $\mathbf{Z}$ has equal column totals since both P_M and the limiting CPS design give a probability of $Z_{i\bullet}/M$ to $Z_{ij} = 1$.

4.2. The Galapagos Island Finch data

To illustrate the potential application of Proposition 2 to null model analysis, consider the data presented in Table 1 of Chen et al. [4], minus the data for the Warbler finch, seen at all sites. We get a 12×17 0-1 data matrix $\mathbf{Z}_{obs}$ where the columns correspond to different islands in the Galapagos. The pairwise association between species is of interest as finches with similar morphological charac-

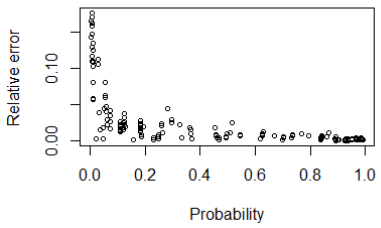

Figure 1: First order moments.

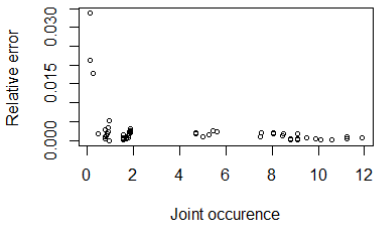

Figure 2: Joint occurences.

teristics are in competition and might not be present on the same islands. This is investigated statistically by comparing the joint occurrences found in the observed matrix, $\mathbf{Z}_{obs}\mathbf{Z}_{obs}^{\top}$, to Monte Carlo joint occurrences simulated by sampling uniformly among 12×17 0-1 matrices with row and column sums equal to those of the data matrix, respectively given by $(1, 2, 2, 6, 10, 10, 10, 11, 12, 13, 14, 14)$ and $(2, 2, 3, 3, 3, 3, 6, 7, 7, 8, 8, 8, 8, 9, 9, 9, 10)$.

The goal of this section is to compare moments of statistics calculated using randomly generated matrices $\mathbf{Z}$ to those derived using the limiting distribution of Proposition 2. The limiting values $\mathbf{p}_M^*$ are given by $\hat{x}_\omega/M, \omega \in \Omega$ where $\{\hat{x}_\omega\}$ are the predicted value when (10) is fitted to $\mathbf{Z}_{obs}$. To compare the first two moments, $E_M(\mathbf{Z})$ and $E_M(\mathbf{Z}\mathbf{Z}^{\top})$ were evaluated using 10^5 matrices simulated using the swap algorithm in [6], details are in the Supplementary Material. The (i, j) entry of the limiting value $E_M^*(\mathbf{Z})$ is obtained as $E_M^*(\mathbf{Z})_{ij} = \sum_\omega \omega_i \psi_{n_j}(\sum \omega_k)\hat{x}_\omega/c_j, 12 \geq i > j > 0$, where n_j is the total for column j and c_j is the number of columns whose total is n_j. For second order moments, $E_M^*(\mathbf{Z}\mathbf{Z}^{\top})_{ij} = \sum_\omega \omega_i\omega_j\hat{x}_\omega, 12 \geq i > j > 0$. Figures 1 and 2 show the relative differences between the simulated and the limiting values for the first two moments.

The limiting probabilities are smaller than the fixed M probabilities when they are small; a similar result was found in section 4.1. This explains the largest errors found in Figure 1. For joint occurrences, the relative errors are essentially zero except for finches with very small occurrence probabilities. Thus, in this

example where $M = 17$ is small, the moments of $\mathbf{Z}$ are well approximated by the limiting values derived from Proposition 2.

[1] Barvinok, A., 2010. On the number of matrices and a random matrix with prescribed row and column sums and 0–1 entries. Advances in Mathematics 224 (1), 316–339.

[2] Barvinok, A., 2012. Matrices with prescribed row and column sums. Linear Algebra and its Applications 436 (4), 820–844.

[3] Besag, J., Clifford, P., 1989. Generalized monte carlo significance tests. Biometrika 76 (4), 633–642.

[4] Chen, Y., Diaconis, P., Holmes, S. P., Liu, J. S., 2005. Sequential monte carlo methods for statistical analysis of tables. Journal of the American Statistical Association 100 (469), 109–120.

[5] Gotelli, N. J., Graves, G. R., 1996. Null models in ecology. Smithsonian Institution Press, University of Michigan USA.

[6] Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., Ohara, R., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., et al., 2013. Package vegan. Community ecology package, version 2 (9).

[7] Rivest, L.-P., Baillargeon, S., 2007. Applications and extensions of Chao's moment estimator for the size of a closed population. Biometrics 63, 999–1006.

[8] Rivest, L.-P., Ebouele, S. E., 2020. Sampling a two dimensional matrix. Computational Statistics & Data Analysis, 106971.

[9] Tillé, Y., 2006. Sampling Algorithms. New York: Springer Science&Business Media.

[10] Wang, G., 2020. A fast mcmc algorithm for the uniform sampling of binary matrices with fixed margins. Electronic Journal of Statistics 14 (1), 1690–1706.

Supplementary Material for " A limit theorem for an equiprobable sampling scheme for 0-1 matrices"

Louis-Paul Rivest

Department of Mathematics and Statistics, Université Laval, 1045 avenue de la médecine, Quebec, QC, G1V 0A6 Canada

4.1 An example with N=4

For $M = 3$ there are 4 basic matrices with row totals $(2, 1, 1, 1)$ and column totals $(3, 1, 1)$:

$$\mathbf{Z}_1 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} ; \mathbf{Z}_2 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} ; \mathbf{Z}_3 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} ; \mathbf{Z}_4 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

In the last three matrices the 2 columns with a row sum of 1 are different: permuting them yields a total of 7 matrices. In this example there are 8 possible values for ω, 4 that sums of 1, $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, and $(0, 0, 0, 1)$, and 4 that sums to 3, $(1, 1, 1, 0)$, $(1, 1, 0, 1)$, $(1, 0, 1, 1)$, $(0, 1, 1, 1)$. For an arbitrary value of k the constraints on row and column totals can be expressed as

- $x_{1,1,1,0} + x_{1,1,0,1} + x_{1,0,1,1} + x_{0,1,1,1} = k$, $x_{1,0,0,0} + x_{0,1,0,0} + x_{0,0,1,0} + x_{0,0,0,1} = 2k$;
- $x_{1,1,1,0} + x_{1,1,0,1} + x_{1,0,1,1} + x_{1,0,0,0} = 2k$, $x_{1,1,1,0} + x_{1,1,0,1} + x_{0,1,1,1} + x_{0,1,0,0} = k$,
- $x_{1,1,1,0} + x_{1,0,1,1} + x_{0,1,1,1} + x_{0,0,0,1} = k$, $x_{1,0,1,1} + x_{1,1,0,1} + x_{0,1,1,1} + x_{0,0,0,1} = k$.

The space of solutions to the constraints has dimension 3. It is indexed by non negative integers i, j, ℓ such that $i + j + \ell \leq k$ as follows $i = x_{1,1,1,0} = x_{0,0,0,1}$, $j = x_{1,1,0,1} = x_{0,0,1,0}$, $\ell = x_{1,0,1,1} = x_{0,1,0,0}$ as $x_{0,1,1,1} = k - i - j - \ell$ and $x_{1,0,0,0} = 2k - i - j - \ell$ and

$$\mathcal{N}_M = \sum_{i+j+\ell \leq k} \frac{(2k)!(k!)}{i!^2 j!^2 \ell!^2 (k-i-j-\ell)!(2k-i-j-\ell)!}.$$

When $k = 1$ it is easily seen that $\mathcal{N}_3 = 7$ as found earlier. For fixed M,

$$E\{\hat{p}_{M(1,1,1,0)}\} = \frac{1}{\mathcal{N}_M} \sum_{i+j+\ell \leq k} \frac{(i/M)(2k)!(k!)^2}{i!^2 j!^2 \ell!^2 (k-i-j-\ell)!(2k-i-j-\ell)!}.$$

The second moment is calculated in a similar way.

Some R-code for Section 4.1

R code to calculate the M=150 column of Table 1:

```
k=50
M<-3*k
res<-numeric(0)
for (i in (0:k)){
  for (j in (0:(k-i))){
    for (ll in (0:(k-i-j))){
    vec1<-factorial(i)^2*factorial(j)^2*factorial(ll)^2*factorial(k-i-j-ll)
    res<-rbind(res,c(i,j,ll, (vec1*factorial(2*k-i-j-ll))^(-1)))
    }
  }
}
res[,4]<-res[,4]/sum(res[,4])
matvp<-cbind(res[,1:3],(k-res[,1]-res[,2]-res[,3]),
(2*k-res[,1]-res[,2]-res[,3]))/M

esp<-t(res[,4])%*%matvp
stesp<-sqrt(t(res[,4])%*%(matvp-outer(rep(1,dim(res)[1]),as.vector(esp)))^2)
xx<-rbind(esp,stesp)
colnames(xx)<-c("1110","1101","1011","0111","1000")
rownames(xx)<-c("E","SD")
xx
```

4.2 Analysis of the Galapagos Island data

This section uses the function `closedpCI.t` of the R-package `Rcapture` to calculate the limiting value $\mathbf{p}_M^*$. It also use the package `vegan` to simulate uniformly distributed matrices. The R-code for the plots in Figures 1 and 2 is also given.

```
data<-c(0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
        0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1,
        1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
        0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
        0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0,
        0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0,
        0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

finch<-matrix(data,12,17,byrow=TRUE)
require(Rcapture)
library(Rcapture)
xx<-closedpCI.t(t(finch),m="Mth", h="LB",h.control=list(neg=FALSE))
fit<-cbind(histpos.t(12),xx$fit$fitted.values)
# For all possible vectors omega, given in the first 12 columns of fit
# column 13 gives the fitted value \hat x_\omega.
#
# Simulation of 10^5 matrices Z with row and column sums equal to that of finch
  library(vegan)
  nm <- nullmodel(finch, "tswap")
  fin.sampl<-simulate(nm, burnin=2000, nsim=100000, thin=2000)
  #
  #  Expected matrices of first order selection Prob
  #
res2<-matrix(0,12,17)
nj<-colSums(finch)
qn<-table(nj)
for (j in (1:17)){
totc<-nj[j]
qtotc<-qn[row.names(qn)==totc]
```

```
indi<-rowSums(fit[,1:12])==totc
res2[,j]<-as.vector(t(fit[,13]*indi)%*%fit[,1:12])/qtotc
}
#
#  Expected matrices of joint occurences
#
res3<-matrix(0,12,12)
for (i in (1:12)){
  for(j in (1:12)){
    res3[i,j]<-sum(fit[,13]*fit[,i]*fit[,j])
  }
}
 #
  # Fixed M first order expectation
  #
  res4<- apply(mpa.sampl,c(1,2),mean)
#
  # Fixed M joint occurences
 #
 res5<-matrix(0,12,12)
for (i in (1:100000)){
  res5<-res5+mpa.sampl[,,i]%*%t(mpa.sampl[,,i])
}
 res5<-res5/100000
 #
 #Plot of the first order relative differences
 #
  plot(as.vector(res4), as.vector(abs((res2-res4)/res2)),
  xlab="Probability", ylab="Relative error ", type="p", pch=1, cex=0.75)
res3v<-res5v<-numeric(0)
#Rfin<-rowsums(finch)
for (i in (1:11)){
  res3v<-c(res3v,res3[i,(i+1):12])
  res5v<-c(res5v,res5[i,(i+1):12])
}
 #
 #Plot of the relative differences for joint occurences
 #
plot(res5v, abs(res3v-res5v)/res5v, xlab="Joint occurence",
 ylab="Relative error ", type="p", pch=1, cex=0.75)
```