# A Practical Algorithm for Exact Inference on Tables

Jeffrey W. Miller* Matthew T. Harrison*

**Abstract**
We describe a dynamic programming algorithm for exact uniform generation of two-way zero-one or contingency tables with fixed margin sums. Monte Carlo samples generated by the algorithm are useful for a variety of statistical tests about these tables. The algorithm is practical for moderately-sized tables and some large, sparse tables. Exact sampling is preferable over existing methods, such as Markov chain Monte Carlo (MCMC) or Sequential Importance Sampling (SIS), because little is known about the convergence rates of these latter methods. The algorithm also computes the exact number of tables with the specified margin sums. We illustrate the method on a variety of published datasets.

**Key Words:** zero-one tables, contingency tables, exact sampling, dynamic programming

## 1. Introduction

Tables indicating the co-occurrence of elements from two sets arise frequently in the statistical analysis of scientific data. Ecology provides for rich source of examples, such as Table 1, indicating the presence of 26 mammalian species in 28 mountain ranges in the American Southwest. To begin to introduce the general problem, we illustrate this particular example. The following questions naturally spring to mind when one is presented with this data. What factors control which species inhabit which mountains? Do relationships among species influence this distribution? Does this data exhibit "structure" or is it simply "random"? Ecologists seeking to answer these questions formulate models of species distribution, and conduct statistical hypothesis tests to evaluate the plausibility of a model given the observed data. Such a test involves three ingredients: (1) a null distribution (a probability distribution on tables, formalizing the notion of randomness), (2) a test statistic (a function on tables, chosen to be indicative of the plausibility of a model), and (3) a dataset (an observed table). The output is a $p$-value for the test statistic evaluated on the dataset.

For instance, for the montane mammals in Table 1, Patterson and Atmar [16] proposed a model in which most of the species historically inhabited a region spanning several mountain ranges and the low-lying areas between, but climate changes caused the populations to recede into the mountains and become extinct in some areas. They suggest that this would generate a distribution in which the set of species found in one mountain range tends to be a subset of those found on another. This led them to consider the following *nested subset statistic*, equal to the number of species-habitat pairs such that the species does not occur in that habitat but does occur in a less-populated habitat:

$$S_n = \sum_i \sum_{j:c_j > m_i} (1 - a_{ij}),$$

where $\mathbf{A} = (a_{ij})$ is a binary matrix with species as rows and habitats as columns (such as Table 1), $c_j = \sum_i a_{ij}$, and $m_i = \min\{c_j : a_{ij} = 1\}$. The standard null distribution

*Division of Applied Mathematics, Brown University, 182 George Street, Providence, RI 02912

**Table 1**: 26 Mammalian Species in 28 Mountain Ranges in the American Southwest [16]

| | *Range* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Species* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| A | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| B | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| C | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| D | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| E | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| F | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| G | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| H | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| J | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| K | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| L | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| M | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| N | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| O | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| U | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| V | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| W | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Z | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

NOTE: See [16] for species and range code. A “1” means the species is present, a “0” means it is absent. The number of binary matrices with row and column sums as in this table is 2,663,296,694,330,271,332,856,672,902,543,209,853,700. This was computed in 32 minutes using Algorithm 3.2 below.

used in ecological studies of this kind is the uniform distribution over binary matrices with given row and column sums. So, in our example using Table 1, this would be the set of binary matrices with row sums (26, 26, 25, 22, 22, 18, 12, 12, 12, 11, 10, 10, 8, 8, 8, 7, 6, 6, 5, 5, 4, 4, 3, 3, 1, 1) and column sums (26, 24, 23, 21, 19, 13, 13, 12, 11, 10, 10, 9, 9, 7, 7, 7, 7, 7, 7, 6, 6, 5, 5, 4, 3, 2, 1, 1). Patterson and Atmar used an approximation to the uniform distribution in which the entries of each column are drawn proportionally to the row sums, conditioned on the column sum (the row sums were not constrained in their approximation). They draw a number of samples from this distribution, estimate a $p$-value of $9 \times 10^{-20}$ for the nested subset statistic on Table 1, and conclude that the data does exhibit significantly more nestedness than one would expect under the null.

However, there is a lingering question: did the use of an approximate distribution significantly affect the result? In fact, the situation described above is typical – the combinatorial problem that arises from constraining the row and column sums makes it difficult to sample exactly from the desired uniform distribution. As a result, on all but the most trivial matrices, researchers have resorted to approximate methods, such as Markov chain Monte Carlo, Sequential Importance Sampling, and heuristic methods such as the one described above. With all these approximate methods, the nagging question remains: is the approximation good enough? In this work, we circumvent this question, by providing an efficient algorithm for sampling exactly from the uniform distribution over binary matrices with given row and column sums (provided the matrix is moderately-sized, as in the example above). As a result, one can obtain exact confidence intervals for statistics estimated by Monte Carlo sampling. Further, our algorithm computes the exact number of such matrices. In addition to binary matrices, the algorithm can also handle non-negative integer matrices, however, here we will focus primarily on binary matrices.

In this paper, we give a recursive formula for exactly counting the number of matrices (binary or non-negative integer) with given row and column sums, and describe an algorithm that uses this result to exactly sample from such matrices. The distinguishing characteristic of the method is exactness and efficiency. The algorithm provably outperforms all existing (exact) algorithms, and although it can be exponentially slow in general, it scales polynomially for certain special cases (such as, for bounded margins). Precise claims and proofs of efficiency will be given in forthcoming work.

The difficulty of the problem arises from the complicated constraints imposed by fixing the row and column sums, combined with the fact that the number of matrices grows rapidly with the size of the matrix. To get a sense of the problem, consider the following trivial example: if the row and column sums are $(2, 2, 1, 1)$, $(3, 2, 1)$, then there are $8$ binary matrices and $24$ non-negative. The $8$ binary matrices are below.

$$\begin{array}{ccc} 1&1&0\\ 1&1&0\\ 1&0&0\\ 0&0&1 \end{array} \quad \begin{array}{ccc} 1&1&0\\ 1&1&0\\ 0&0&1\\ 1&0&0 \end{array} \quad \begin{array}{ccc} 1&1&0\\ 1&0&1\\ 1&0&0\\ 0&1&0 \end{array} \quad \begin{array}{ccc} 1&1&0\\ 1&0&1\\ 0&1&0\\ 1&0&0 \end{array}$$

$$\begin{array}{ccc} 1&1&0\\ 0&1&1\\ 1&0&0\\ 1&0&0 \end{array} \quad \begin{array}{ccc} 1&0&1\\ 1&1&0\\ 1&0&0\\ 0&1&0 \end{array} \quad \begin{array}{ccc} 1&0&1\\ 1&1&0\\ 0&1&0\\ 1&0&0 \end{array} \quad \begin{array}{ccc} 0&1&1\\ 1&1&0\\ 1&0&0\\ 1&0&0 \end{array}$$

On the other hand, there are 2,663,296,694,330,271,332,856,672,902,543,209,853,700 ($\approx 2.7 \times 10^{39}$) binary matrices with row and column sums as in Table 1. This number was computed using our algorithm, and we know of no other existing algorithm capable of handling this example exactly.

## 2. Review

This section contains a brief account of previous work on this problem. The majority of researchers tackling this problem have employed methods falling mainly into five categories: exhaustive methods, heuristic methods, Markov chain Monte Carlo (MCMC), asymptotic formulae, and Sequential Importance Sampling (SIS). Exhaustive methods, in which every matrix in the set is visited, can be used for matrices of trivial size (smaller than, say, $5 \times 5$) (as in [20]), but these are intractable even for modestly-sized matrices such as Table 1, since the number of matrices grows rapidly. Heuristic methods consist of approximations (as in the example of the previous section [16], as well as [6, 12, 5]) which are chosen because they are heuristically similar to the desired distribution, are easy to analyze, and are trivial to compute – however, they are typically known to be incorrect, and come with no guarantees on the quality of the estimate. Asymptotic formulae for the number of matrices are a more sophisticated approach, and major advances in this direction have been made in the last decade (see [2, 3, 13, 15]) – the idea here is to find a closed-form expression (for the number of matrices) that converges as rapidly as possible to the true value as the size of the matrix grows – thus, these have the advantages of easy computation and theoretical guarantees for very large matrices, but they are not applicable for modestly-sized matrices. Furthermore, it is often unclear how to adapt formulae for the number of matrices to obtain $p$-values for complicated test statistics. MCMC is the most common approach (see [1, 4, 14, 18]), having the virtues of being simple to implement, always returning a matrix with the correct row and column sums, and being guaranteed to provide uniform samples in the limit – however, current theory does not provide adequate estimates of the mixing time, so one never really knows how close the samples are to being uniform. SIS techniques have been applied to this problem recently, and appear to provide an improvement over MCMC in terms of the efficiency (see [4]) – however, as with MCMC, current theory does not provide adequate guarantees on convergence rates.

There have been other algorithms presented for finding the exact number of matrices (allowing for non-regular margins), such as [8, 17, 21], however, they appear to be too inefficient to handle many of the matrices that arise in applications, such as Table 1, (although the method described in [8] can handle Table 3, the smallest table we consider here). Further, the capacity for exact sampling has not been demonstrated with these methods (although it should be possible with [8]).

There is a gap in the capabilities of the sampling methods described above. When one has a medium-sized matrix (say, between $5 \times 5$ and $50 \times 50$) and requires theoretical guarantees, then existing methods are inadequate – the matrix is too large for exhaustive methods, and the desired guarantees on accuracy are too stringent for heuristic methods, MCMC, or SIS. Our method fills this gap by providing exact results for matrices of this size.

## 3. Theory

Since the main purpose of this paper is to describe statistical applications of our algorithm, we will simply state the primary mathematical results (a forthcoming paper will provide proofs of these claims, as well as precise statements and proofs regarding efficiency).

Let $N(\mathbf{p}, \mathbf{q})$ be the number of $m \times n$ binary matrices with margins (row and column sums) $\mathbf{p} = (p_1, \dots, p_m) \in \mathbb{N}^m$, $\mathbf{q} = (q_1, \dots, q_n) \in \mathbb{N}^n$ respectively; let $M(\mathbf{p}, \mathbf{q})$ be the corresponding number of $\mathbb{N}$-valued matrices. Let $L : \mathbb{R}^n \to \mathbb{R}^n$ denote the left-shift map: $L\mathbf{r} = (r_2, \dots, r_n, 0)$. Given $\mathbf{r}, \mathbf{s} \in \mathbb{N}^n$ (where $\mathbb{N} := \{0, 1, 2, \dots\}$), let $\mathbf{r}\backslash\mathbf{s} := \mathbf{r} - \mathbf{s} + L\mathbf{s}$, (which may be read as "$\mathbf{r}$ *reduce* $\mathbf{s}$"), let

$$\binom{\mathbf{r}}{\mathbf{s}} := \binom{r_1}{s_1} \cdots \binom{r_n}{s_n},$$

and let $\bar{\mathbf{r}}$ denote the vector of counts, $\bar{\mathbf{r}} := (\bar{r}_1, \bar{r}_2, \dots)$ where $\bar{r}_i := \#\{j : r_j = i\}$. Since the numbers $N(\mathbf{p}, \mathbf{q})$ and $M(\mathbf{p}, \mathbf{q})$ are fixed under permutations of the row sums $\mathbf{p}$ and column sums $\mathbf{q}$, then we may define $\bar{N}(\mathbf{p}, \bar{\mathbf{q}}) := N(\mathbf{p}, \mathbf{q})$ and $\bar{M}(\mathbf{p}, \bar{\mathbf{q}}) := M(\mathbf{p}, \mathbf{q})$

without ambiguity. We say $\mathbf{r}$ *is dominated by* $\mathbf{s}$, and write $\mathbf{r} \leq \mathbf{s}$, if $r_i \leq s_i$ for all $i$. Let $C_n(k) := \{\mathbf{r} \in \mathbb{N}^n : \sum_i r_i = k\}$ be the $n$-part compositions of $k$, and let $C^{\mathbf{s}}(k) := \{\mathbf{r} \in C_n(k) : \mathbf{r} \leq \mathbf{s}\}$ be those compositions dominated by $\mathbf{s}$. We can now state our main results.

**Theorem 3.1 (Recursions)** *The number of matrices with margins* $(\mathbf{p}, \mathbf{q}) \in \mathbb{N}^m \times \mathbb{N}^n$ *is given by*

(1) $\bar{N}(\mathbf{p}, \mathbf{r}) = \displaystyle\sum_{\mathbf{s} \in C^{\mathbf{r}}(p_1)} \binom{\mathbf{r}}{\mathbf{s}} \bar{N}(L\mathbf{p}, \mathbf{r} \backslash \mathbf{s})$ *for binary matrices, and*

(2) $\bar{M}(\mathbf{p}, \mathbf{r}) = \displaystyle\sum_{\mathbf{s} \in C^{\mathbf{r}+L\mathbf{s}}(p_1)} \binom{\mathbf{r} + L\mathbf{s}}{\mathbf{s}} \bar{M}(L\mathbf{p}, \mathbf{r} \backslash \mathbf{s})$ *for* $\mathbb{N}$*-valued matrices,*

*where* $\mathbf{r} = \bar{\mathbf{q}}$*, and in* (2)*, we sum over all* $\mathbf{s}$ *such that* $\mathbf{s} \in C^{\mathbf{r}+L\mathbf{s}}(p_1)$*.*

Proofs will be given in a forthcoming paper. The Gale-Ryser conditions [11, 19] simplify computation of the sum in (1) by providing a necessary and sufficient condition for there to exist a binary matrix with margins $(\mathbf{p}, \mathbf{q})$: if $q_i' := \#\{j : q_j \geq i\}$ and $p_1 \geq \cdots \geq p_m$, then $N(\mathbf{p}, \mathbf{q}) \neq 0$ if and only if $\sum_{i=1}^{j} p_i \leq \sum_{i=1}^{j} q_i'$ for all $j < m$ and $\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i$. This is easily translated into a similar condition in terms of $(\mathbf{p}, \bar{\mathbf{q}})$ and $\bar{N}(\mathbf{p}, \bar{\mathbf{q}})$. The following recursive procedure can be used to compute either $N(\mathbf{p}, \mathbf{q})$ or $M(\mathbf{p}, \mathbf{q})$.

**Algorithm 3.2 (Enumeration)**
*Input:* $(\mathbf{p}, \bar{\mathbf{q}})$*, where* $(\mathbf{p}, \mathbf{q}) \in \mathbb{N}^m \times \mathbb{N}^n$ *are row and column sums such that* $\sum_i p_i = \sum_i q_i$*.*
*Output:* $N(\mathbf{p}, \mathbf{q})$ *(or* $M(\mathbf{p}, \mathbf{q})$*), the number of binary (or* $\mathbb{N}$*-valued) matrices.*
*Storage: Lookup table of cached results, initialized with* $\bar{N}(\mathbf{0}, \mathbf{0}) = 1$ *(or* $\bar{M}(\mathbf{0}, \mathbf{0}) = 1$*).*
(1) *If* $\bar{N}(\mathbf{p}, \bar{\mathbf{q}})$ *is in the lookup table, return the result.*
(2) *In the binary case, if Gale-Ryser gives* $\bar{N}(\mathbf{p}, \bar{\mathbf{q}}) = 0$*, cache the result and return 0.*
(3) *Evaluate the sum in Theorem 3.1, recursing to step* (1) *for each term.*
(4) *Cache the result and return it.*

It is possible to move the Gale-Ryser check directly into the loop, providing a substantial improvement in efficiency, and we use this in the results described below (a description of how this is done will be detailed in forthcoming work). Algorithm 3.2 traverses a directed acyclic graph in which each node represents a distinct set of input arguments to the algorithm, such as $(\mathbf{p}, \bar{\mathbf{q}})$. Node $(\mathbf{u}, \bar{\mathbf{v}})$ is the child of node $(\mathbf{p}, \bar{\mathbf{q}})$ if the algorithm is called (recursively) with arguments $(\mathbf{u}, \bar{\mathbf{v}})$ while executing a call with arguments $(\mathbf{p}, \bar{\mathbf{q}})$. If the initial input arguments are $\mathbf{p}, \bar{\mathbf{q}}$, then all nodes are descendents of node $(\mathbf{p}, \bar{\mathbf{q}})$. Meanwhile, all nodes are ancestors of node $(\mathbf{0}, \mathbf{0})$. Note the correspondence between the children of a node $(\mathbf{u}, \bar{\mathbf{v}})$ and the compositions $\mathbf{s} \in C^{\bar{\mathbf{v}}}(u_1)$ in the binary case, and $\mathbf{s} \in C^{\bar{\mathbf{v}}+L\mathbf{s}}(u_1)$ in the $\mathbb{N}$-valued case, under which $\mathbf{s}$ corresponds with the child $(L\mathbf{u}, \bar{\mathbf{v}} \backslash \mathbf{s})$. We also associate with each node its *count*: the number of matrices with the corresponding margins.

As an additional benefit of caching the counts in a lookup table (as in Algorithm 3.2), once the enumeration is complete we obtain an efficient algorithm for uniform sampling from the set of $(\mathbf{p}, \mathbf{q})$ matrices (binary or $\mathbb{N}$-valued). It is straightforward to prove that since the counts are exact, the following algorithm yields a sample from the uniform distribution.

**Algorithm 3.3 (Sampling)**
*Input:*
· *Row and column sums* $\mathbf{p}, \mathbf{q} \in \mathbb{N}^m \times \mathbb{N}^n$ *such that* $\sum_i p_i = \sum_i q_i$*.*

```
· Lookup table of counts generated by Algorithm 3.2 on input $\mathbf{p}, \bar{\mathbf{q}}$.
Output: A binary (or $\mathbb{N}$-valued) matrix with margins $(\mathbf{p}, \mathbf{q})$, drawn uniformly at random.
(1) Initialize $(\mathbf{u}, \mathbf{v}) \leftarrow (\mathbf{p}, \mathbf{q})$.
(2) If $(\mathbf{u}, \mathbf{v}) = (\mathbf{0}, \mathbf{0})$, exit.
(3) Choose a child $(L\mathbf{u}, \bar{\mathbf{v}}\backslash\mathbf{s})$ of $(\mathbf{u}, \bar{\mathbf{v}})$ with probability proportional to its count times the number of corresponding rows (that is, the $\mathbf{r} \in C^{\mathbf{v}}(u_1)$ such that $\overline{\mathbf{v} - \mathbf{r}} = \bar{\mathbf{v}}\backslash\mathbf{s}$.)
(4) Choose a row uniformly among the corresponding rows.
(5) $(\mathbf{u}, \mathbf{v}) \leftarrow (L\mathbf{u}, \mathbf{v} - \mathbf{r})$.
(6) Goto (2).
```

In step (3), there are $\binom{\bar{\mathbf{v}}}{\mathbf{s}}$ corresponding rows $\mathbf{r}$ in the binary case, and $\binom{\bar{\mathbf{v}}+L\mathbf{s}}{\mathbf{s}}$ in the $\mathbb{N}$-valued case. In step (4), in the binary case of course we only choose among $\mathbf{r} \in \{0, 1\}^n$.

## 4. Applications

In the remainder of the paper, we describe specific cases in which our method has a clear advantage over published results. We focus on binary matrices coming from ecology since our method seems particularly well-suited to problems arising in that field. First, a brief discussion of the scientific problem is given, and then we look at three examples in detail.

### 4.1 Chance or Competition?

In 1975, Jared Diamond detailed an influential theory of species distribution involving a set of "assembly rules", accounting for the effects of competition [9]. In 1979, Connor and Simberloff criticized Diamond's theory with a statistical analysis indicating that by simply conditioning on the number of habitats per species, and species per habitats (conditioning on the margins of the co-occurrence matrix), one obtained values of statistics comparable to those used by Diamond to support his theory [6]. Diamond made rebuttals [12, 10], to which Connor and Simberloff replied [7], sparking a lively debate within the ecology community. As illustrated by the example in the introduction of this paper, researchers continue to propose statistics for testing theories of species distribution, as well as methods for sampling (in order to estimate $p$-values for those statistics). Due to the difficulty of sampling from the desired probability distribution (the uniform distribution over binary matrices with given row and column sums), researchers have used only approximate methods on all but the most trivial cases. However, with the method given in this paper, one can obtain exact results in many non-trivial cases.

### 4.2 Montane Mammals in the American Southwest

For our first example, we apply the techniques of the previous section to the dataset given in Table 1. We compute the number of matrices (exactly), and use exact sampling to estimate the $p$-value for the statistic $S_n$ (defined in the introduction). We find that there are 2,663,296,694,330,271,332,856,672,902,543,209,853,700 ($\approx 2.7 \times 10^{39}$) binary matrices with row and column sums as in Table 1. This was computed in 32 minutes using Algorithm 3.2. We know of no other algorithm capable of exactly enumerating matrices of this size.

In Table 2, we compare the $p$-value found by Patterson and Atmar (using the heuristic method described in the introduction) with the value we estimated from 1 million exact

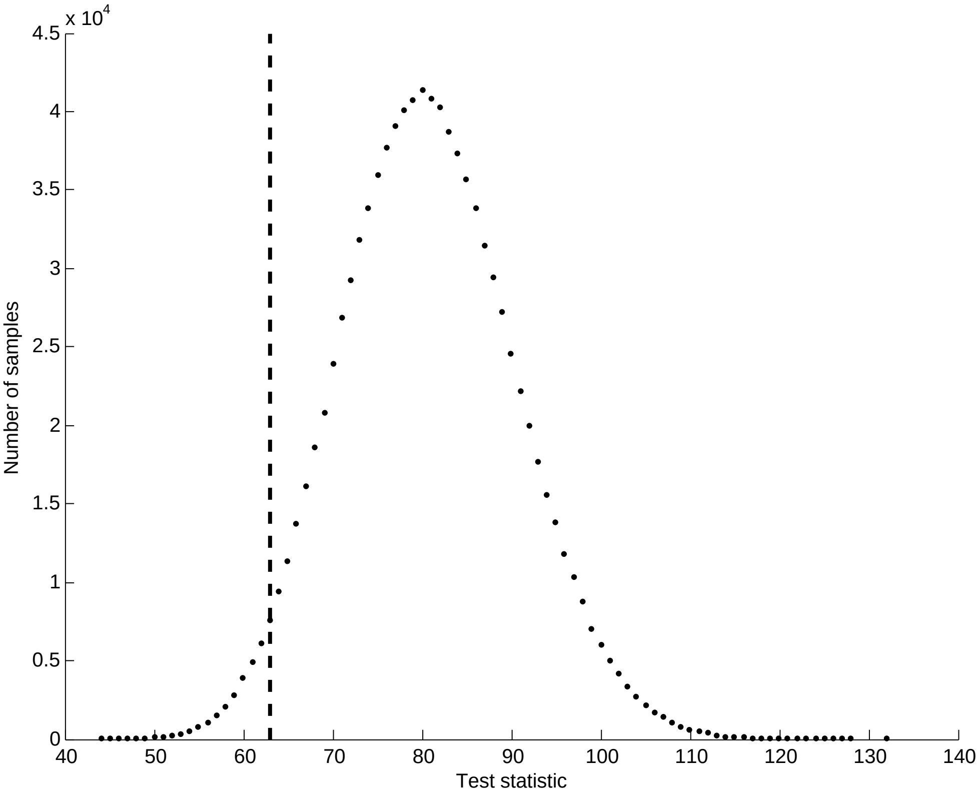

**Figure 1**: Histogram of $S_n$, for 1 million exact samples from the uniform distribution over matrices with margins as in Table 1. The dashed line indicates the value of the statistic for Table 1.

**Table 2**: Results for Montane Mammal Data, using $S_n$

| Method | # samples | $p$-value | mean of $S_n$ | std. dev. | min | max |
|---|---|---|---|---|---|---|
| Exact | 1,000,000 | $0.0322 \pm .00018$ | 80.71 | 9.697 | 44 | 132 |
| Heuristic | 1,000 | $9 \times 10^{-20}$ | 227.9 | 18.135 | 180 | 287 |

samples, obtained at a rate of about 0.009 seconds per sample (using Algorithm 3.3). There are large discrepancies between our results and those reported by Patterson and Atmar. First, we estimate a $p$-value of $0.0322 \pm .00018$ (where the number following the ± sign is the standard error), much larger than their $9 \times 10^{-20}$ (see Figure 1). The value of the test statistic on the original data is $S_n = 63$, considerably larger than the smallest value obtained on their 1000 Monte Carlo samples: $S_n = 180$. (It appears that the authors used the standard deviation of their samples to estimate the $p$-value.) Recall that in their approximation, the entries of each column are drawn proportionally to the row sums, conditioned on the column sum, and that the row sums are not constrained. Apparently, omitting the constraint on the row sums has a drastic effect on the nested subset statistic. As a result, the analysis they performed considerably underestimates the $p$-value.

**Table 3**: 13 Species of Finch in 17 of the Galápagos Islands [4]

| | *Island* | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Species* | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
| 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 4 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 10 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 11 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

NOTE: See [4] for species and range code.

### 4.3 Darwin's Finches

Our next example involves a table indicating the occurrence of 13 species of finch on 17 of the Galápagos Islands (see Table 3). It comes equipped with the colorful name of "Darwin's finches" because Charles Darwin's development of the theory of evolution was initiated by his observations of these birds. The row and column sums of the matrix are (14, 13, 14, 10, 12, 2, 10, 1, 10, 11, 6, 2, 17) and (4, 4, 11, 10, 10, 8, 9, 10, 8, 9, 3, 10, 4, 7, 9, 3, 3), respectively. This table is the subject of an analysis by Chen, Diaconis, Holmes, and Liu [4], in which they use SIS and MCMC to estimate the $p$-value for the statistic:

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

where $m$ is the number of species, $\mathbf{C} = (c_{ij}) = \mathbf{A}\mathbf{A}^T$, and $\mathbf{A}$ is the occurrence matrix.

**Table 4**: Results for Darwin's Finch Data, using $\bar{S}^2$

| Method | # samples | $p$-value |
|---|---|---|
| Exact | 1,000,000 | $(4.67 \pm .22) \times 10^{-4}$ |
| SIS | 1,000,000 | $(3.96 \pm .36) \times 10^{-4}$ |
| MCMC | 15,000,000 | $(3.56 \pm .68) \times 10^{-4}$ |

The results of Chen et al. are reported in Table 4, alongside our results using exact sampling. Our results largely confirm the conclusion of Chen et al. – the $p$-value is small, leading one to reject the null hypothesis (see Figure 2). The computation time for our method is not significantly larger: for this dataset, enumerating the number of matrices takes 1.4 seconds, and sampling takes about 0.002 seconds per sample, while their SIS method takes about 0.001 seconds per sample. (However, on larger matrices, their method

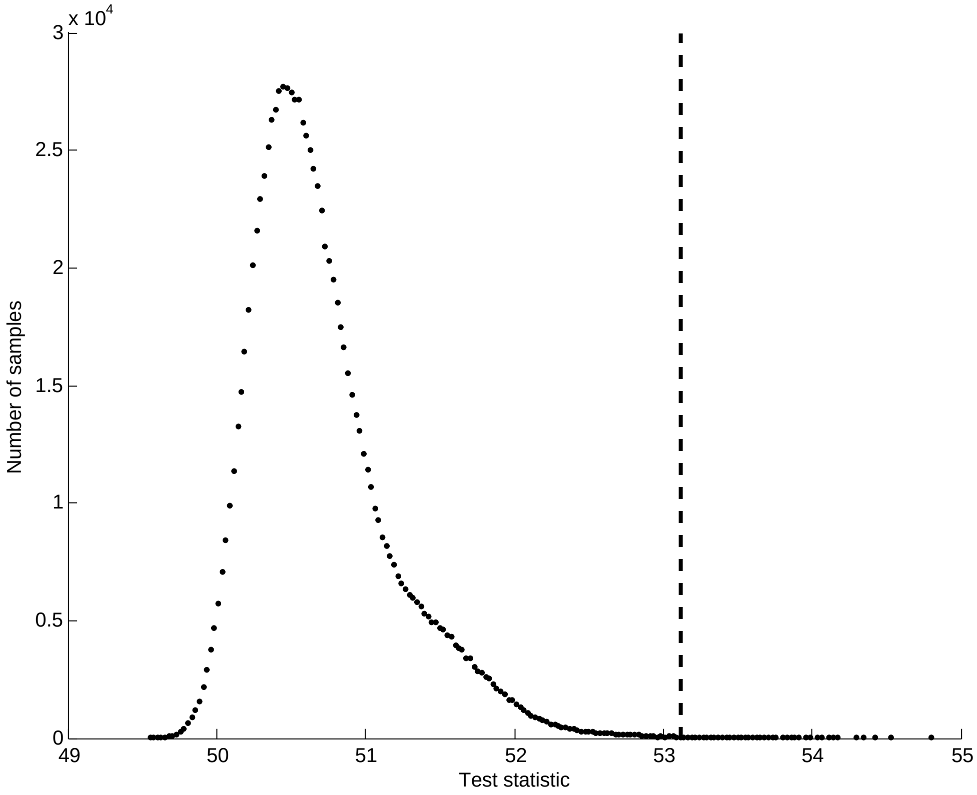

**Figure 2**: Histogram of $\bar{S}^2$, for 1 million exact samples from the uniform distribution over matrices with margins as in Table 3. The dashed line indicates the value of the statistic for Table 3.

should be significantly faster than ours, due to the overhead incurred by enumeration.) The technique of Chen et al. also yields an estimate of $6.7150 \times 10^{16}$ for the number of matrices. We compute the exact number of matrices to be 67,149,106,137,567,626, and note that this agrees with the exact number reported by Chen et al. (obtained by other means).

We would like to emphasize again, however, that in comparison with MCMC and SIS the appeal of our method is not its speed, but rather its exactness. There are no guarantees that the estimates coming from MCMC or SIS have adequately converged. Meanwhile, we are drawing exact i.i.d. samples from the null distribution, which provides many guarantees. For example, an exact 95% confidence interval for the $p$-value is $(4.26, 5.11) \times 10^{-4}$.

### 4.4 Island Lizards in the Gulf of California

Our third and last example is an occurrence table for 20 lizard species on 25 islands in the Gulf of California (see Table 5). The row and column sums of the matrix are (7, 23, 6, 6, 2, 18, 8, 8, 1, 22, 2, 2, 9, 2, 1, 18, 9, 3, 3, 1) and (13, 4, 9, 4, 3, 2, 3, 4, 4, 5, 2, 5, 2, 10, 10, 10, 7, 6, 6, 3, 3, 11, 8, 11, 6), respectively.

Manly [14] performs an analysis on this table in which he uses MCMC to estimate a $p$-value for the statistic:

$$S_d = \frac{1}{m^2} \sum_{i,j} (c_{ij} - d_{ij})^2,$$

**Table 5**: 20 Lizard Species on 25 Islands in the Gulf of California [14]

| | *Island* | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Species* | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y |
| 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 9 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| 11 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 13 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 14 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 17 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 20 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

NOTE: See [14] for species and range code.

**Table 6**: Results for Island Lizards, using $S_d$

| Method | # samples | $p$-value |
|---|---|---|
| Exact | 1,000,000 | $(1.34 \pm .12) \times 10^{-4}$ |
| MCMC | 1,000,000 | $(5.0 \pm .4) \times 10^{-4}$ |

where $m$ is the number of species, (as before, $\mathbf{C} = (c_{ij}) = \mathbf{A}\mathbf{A}^T$ where $\mathbf{A}$ is the occurrence matrix), and $d_{ij}$ is the expected number of times that the species $i$ and $j$ occur on the same island under the null distribution. Since $d_{ij}$ is not known exactly, it is estimated by sampling from the distribution, and this defines an approximation to the statistic. We used 1 million samples to obtain such an estimate. Our results essentially confirm the conclusion reached by Manly – that the data is atypical with respect to this statistic (see Figure 3) – however, we arrive at a somewhat smaller $p$-value (see Table 6). We find the exact number of matrices to be 55,838,420,515,731,001,979,319,625,577,023,858,901,579,264 ($\approx 5.6 \times 10^{43}$), a computation requiring 11 minutes. Following this enumeration, exact samples were obtained at a rate of 0.005 seconds per sample.

The preceding examples demonstrate that the method outlined in this paper is capable of providing exact results (exact counting and exact sampling) in real-world problems involving binary matrices with fixed margins, for which exact results have previously appeared to be intractable.

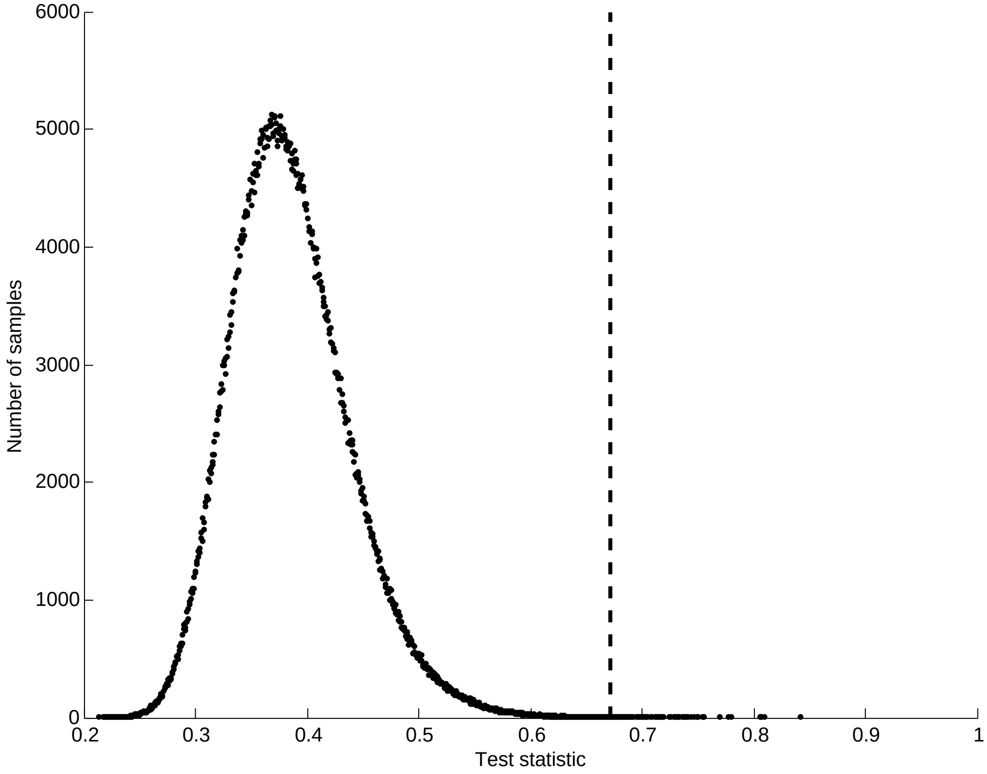

**Figure 3**: Histogram of $S_d$, for 1 million exact samples from the uniform distribution over matrices with margins as in Table 5. Note: It appears that Manly sometimes uses $mS_d$, rather than $S_d$. The dashed line indicates the value of the statistic for Table 5.

## References

[1] R. A. Brualdi. Matrices of zeros and ones with fixed row and column sum vectors. *Linear Algebra and its Applications*, 33:159–231, October 1980.

[2] E. Canfield, C. Greenhill, and B. D. McKay. Asymptotic enumeration of dense 0-1 matrices with specified line sums. *Journal of Combinatorial Theory, Series A*, 115(1):32–66, 2008.

[3] E. R. Canfield and B. D. McKay. Asymptotic enumeration of dense 0-1 matrices with equal row sums and equal column sums. *Electronic Journal of Combinatorics*, 12(R29), 2005.

[4] Y. Chen, P. Diaconis, S. P. Holmes, and J. S. Liu. Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100, 2005.

[5] B. D. Coleman, M. A. Mares, M. R. Willig, and Y. Hsieh. Randomness, area, and species richness. *Ecology*, 63(4):1121–1133, August 1982.

[6] E. F. Connor and D. Simberloff. The assembly of species communities: Chance or competition? *Ecology*, 60(6):1132–1140, December 1979.

[7] E. F. Connor and D. Simberloff. Interspecific competition and species co-occurrence patterns on islands: Null models and the evaluation of evidence. *Oikos*, 41(3):455–465, December 1983.

[8] P. Diaconis and A. Gangolli. Rectangular arrays with fixed margins. In *Discrete Probability and Algorithms*, pages 15–41. Springer-Verlag, New York, 1995.

[9] J. M. Diamond. Assembly of species communities. In *Ecology and evolution of communities*, pages 342–444. Harvard University press, Cambridge, 1975.

[10] J. M. Diamond and M. E. Gilpin. Examination of the "null" model of Connor and Simberloff for species co-occurrences on islands. *Oecologia*, 52(1):64–74, January 1982.

[11] D. Gale. A theorem on flows in networks. *Pacific Journal of Mathematics*, 7:1073–1082, 1957.

[12] M. E. Gilpin and J. M. Diamond. Factors contributing to non-randomness in species co-occurrences on islands. *Oecologia*, 52(1):75–84, January 1982.

[13] C. Greenhill, B. D. McKay, and X. Wang. Asymptotic enumeration of sparse 0-1 matrices with irregular row and column sums. *Journal of Combinatorial Theory, Series A*, 113(2):291–324, 2006.

[14] B. F. J. Manly. A note on the analysis of species co-occurrences. *Ecology*, 76:1109–1115, 1995.

[15] B. D. McKay and X. Wang. Asymptotic enumeration of 0-1 matrices with equal row sums and equal column sums. *Linear Algebra and its Applications*, 373:273–288, 2003.

[16] B. Patterson and W. Atmar. Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biological Journal of the Linnean Society*, 28:65–82, 1986.

[17] B. R. Pérez-Salvador, S. de-los Cobos-Silva, M. A. Gutiérrez-Ándrade, and A. Torres-Chazaro. A reduced formula for the precise number of (0,1)-matrices in A(R,S). *Discrete Mathematics*, 256(1-2):361–372, Sept 2002.

[18] A. Roberts and L. Stone. Island-sharing by archipelago species. *Oecologia*, 83(4):560–567, July 1990.

[19] H. Ryser. Combinatorial properties of matrices of zeros and ones. *Canadian Journal of Mathematics*, 9:371–377, 1957.

[20] T. A. B. Snijders. Enumeration and simulation methods for 0-1 matrices with given marginals. *Psychometrika*, 56(3):397–417, September 1991.

[21] B. Wang and F. Zhang. On the precise number of (0, 1)-matrices in U(R,S). *Discrete Mathematics*, 187:211–220, 1998.