Sampling and Counting Contingency Tables Using Markov Chains

Jonathan Connell

March 14, 2011

Abstract

In this paper we present an overview of contingency tables, provide an introduction to the problems of almost uniform sampling and approximate counting, and show recent results achieved through the use of Markov chains. We focus specifically on contingency tables with two rows, since as of this time little progress has been made in achieving reasonable bounds on arbitrarily sized contingency tables.

1 Introduction

Contingency tables are $m \times n$ matrices $[T[i,j]]$ with row sums $r_1, r_2, ..., r_m$, and column sums $c_1, c_2, ..., c_n$ where $r_i = \sum_{j=1}^{n} T[i,j]$ and $c_j = \sum_{i=1}^{m} T[i,j]$. Contingency tables are commonly used by statisticians to store data in such a way as to make the data easy to analyze with respect to two variables or characteristics. Each entry T[i, j] in the contingency table is always a non-negative integer and represents the number of elements sampled that have both characteristics i and j. To be certain that every element sampled has one and only one location in the table, contingency tables require row and column categories to be both *exclusive* (No element can belong to two different rows or columns.) and *exhaustive* (Row and column categories must cover every possibility.).

As an example, imagine that we wish to conduct a study whose goal is to determine whether or not there is a link between hair color and eye color in humans. Once a random sample of the population had been surveyed, we would create a contingency table with row values corresponding to eye color and column values corresponding to hair color (or vice versa) (See Fig. 1 for example). Then for each person surveyed we would find their corresponding cell in the

table and increment its value by 1. In doing this, we would provide ourselves with a way to view our data that is both easy to construct and easy to analyze.

Eye Color	Hair Color				Total
	Black	Brunette	Red	Blond	
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

Figure 1: Example of a contingency table comparing hair and eye color. From [2]

2 Sampling

When sample surveys are conducted and the data is put into a contingency table it is often of value to compare the data surveyed with randomly generated contingency tables which have the same row and column sums. Diaconis and Saloff-Coste [3] showed a Markov Chain capable of generating a contingency table uniformly at random from the pool of contingency tables with equal row and column sums when the number of rows is two. We will henceforth refer to this Markov chain as the *Diaconis chain*.

For the analysis of the Diaconis chain let us consider contingency tables $[T[i,j]]$ with row sums $r_i = \sum_{j=1}^{n} T[i,j]$ and column sums $c_j = \sum_{i=1}^{2} T[i,j]$. We refer to $\Sigma_{r,c}$ as the set of all contingency tables with row sum set $r = \{r_1, r_2\}$ and column sum set $c = \{c_1, c_2, ..., c_n\}$, and $X \in \Sigma_{r,c}$ as a state in the Diaconis chain.

To determine the next state X' in the Diaconis chain let us say that with probability $1/2$ $X' = X$. Likewise, with probability $1/2$ we will select two unique values j_1 and j_2 uniformly at random such that $1 \leq j_1 < j_2 \leq n$. Using these two values, let us consider the following 2×2 subgraph S:

$$S = \begin{bmatrix} T[1,j_1] & T[1,j_2] \\ T[2,j_1] & T[2,j_2] \end{bmatrix}$$

Let us now uniformly at random choose to either add $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ or $\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$ to S. If doing so would result in an entry in S that is less than 0, we refrain from making this change in favor of doing nothing, as having a negative entry in S would invalidate the contingency table, resulting in an $X' \notin \Sigma_{r,c}$. Otherwise, we make this change to S. Having made the change, we point out that X' differs from X in exactly four entries, but that the row sums and column sums have not changed from X to X'. Hence, from any transition from one state to another, we are assured of having a valid $X' \in \Sigma_{r,c}$ regardless of what action is taken.

Using a coupling argument, Hernek [4] was able to show that the Diaconis chain rapidly mixes with a mixing time quadratic in the number of columns n, and $N = \sum_{i=1}^{2} r_i = \sum_{j=1}^{n} c_j$, the total number of elements in the table. Dyer and Greenhill [1] later gave a modified version of the Diaconis chain that reduced the mixing time to $O(n^2 log(N/\epsilon))$, where ϵ is an accepted error range on how close to uniformly random our chain is.

A rapidly mixing Markov chain has not yet been found that can sample almost uniformly at random from an arbitrarily sized contingency table with given row and column sums.

3 Counting

Another tool that statisticians use when analyzing contingency tables is to count (exactly or approximately) the number of contingency tables that have given row and column sums. In other words, if we were to sample uniformly at random, how many different unique tables do we have the possibility of getting?

To better illustrate the problem of exact counting, we will show how to exactly count the number of 2×2 contingency tables given specific row and column sums as shown by Dyer and Greenhill [1]:

First, let us say that we are given a 2×2 table $T^c_{a,b} = \Sigma_{(a,c-a),(b,c-b)}$ where $0 < a, b < c$. Given row sums $(a, c-a)$ and column sums $(b, c-b)$ we can write our matrix as follows:

$$T^c_{a,b} = \begin{bmatrix} i & (a-i) \\ (b-i) & (c+i-a-b) \end{bmatrix}$$

By expressing our matrix in this form, we can have all entry values represented as a function of an integer i as long as $\max\{0, a+b-c\} \leq i \leq \min\{a, b\}$ (to ensure that no entry in the table

can ever be below 0). Now the problem of counting the number of 2×2 contingency tables that have our given row and column sums has been reduced to counting how many values for i exist within the given criteria, which is

$$\begin{aligned} &\min\{a, b\} + 1 && \text{if } a + b \leq c \\ &c - \max\{a, b\} + 1 && \text{if } a + b > c \end{aligned}$$

While 2×2 contingency tables are very easy to solve, we know counting the number of contingency tables with m rows and n columns is $\#P$-complete, even when m or n (but not both) is 2. It is therefore of great interest whether or not we can closely approximate the number of contingency tables with given row and column sums. It was unknown for some time whether or not this could be done in polynomial time for even tables with only two rows, but Dyer and Greenhill [1] showed a *fully polynomial randomized approximation scheme* (FPRAS) for reducing the problem of approximately counting two-rowed contingency tables with given row and column sums to almost uniform sampling.

This FPRAS for counting $|\Sigma_{r,c}|$, the number of contingency tables with given row and column sums r and c respectively, is shown to run in time which is polynomial in n, $\log(N)$, ϵ^{-1}, and $\log(\delta^{-1})$ where n is the number of columns, N is the sum of all entries in the contingency table, ϵ is the amount of error we are willing to accept when approximating $|\Sigma_{r,c}|$, and $1 - \delta$ is a lower bound on the probability that our approximation A is within ϵ of $|\Sigma_{r,c}|$. In short,

$$\text{Prob}[(1 - \epsilon)|\Sigma_{r,c}| \leq A \leq (1 + \epsilon)|\Sigma_{r,c}|] \geq 1 - \delta.$$

Given that we have a polynomial time randomized approximation scheme for reducing the problem of approximately counting two-rowed contingency tables to the problem of almost uniform sampling, the result we gave in Section 2 that almost uniform sampling from two-rowed contingency tables can be done in polynomial time becomes even more useful and meaningful, as it now allows us to not only sample almost uniformly at random, but shows us that we can count two-rowed contingency tables approximately in polynomial time.

References

[1] M. Dyer and C. Greenhill, Polynomial-time counting and sampling of two-rowed contingency tables, *Theoretical Computer Science*, 246, 2000, pp. 265-278.

[2] P. Diaconis, and B. Efron, Testing for Independence in a Two-Way Table: New Interpretations of the Chi-Square Statistic, The Annals of Statistics, 1985, Vol. 13, No. 3, 845-874.

[3] P. Diaconis and L. Saloff-Coste, Random walk on contingency tables with fixed row and column sums, Tech. rep., Department of Mathematics, Harvard University, (1995).

[4] D. Hernek, Random generation of $2 \times n$ contingency tables, Random Structures and Algorithms, 13 (1998), pp. 7179.

[5] J. Mount, Application of Convex Sampling to Optimization and Contingency Table Generation/Counting, Ph.D. Thesis, Carnegie Mellon University, May 1995.

[6] M. Cryan, M. Dyer, L. Goldberg, M. Jerrum, and R. Martin, Rapidly Mixing Markov Chains for Sampling Contingency Tables with a Constant Number of Rows, Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002, pp. 711-720