Approximate Counting Scheme for $m \times n$ Contingency Tables

Shuji Kijima and Tomomi Matsui

METR 2003-01 JANUARY 2003

Approximate Counting Scheme for $m \times n$ Contingency Tables

Shuji Kijima and Tomomi Matsui

Department of Mathematical Informatics,
Graduate School of Information Science and Technology,
University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan
{kijima, tomomi}@misojiro.t.u-tokyo.ac.jp

Abstract. In this paper, we propose a new counting scheme for $m \times n$ contingency tables. Our scheme is a modification of Dyer and Greenhill's scheme for two rowed contingency tables [5]. We can estimate not only the sizes of error, but also the sizes of the bias of the number of tables obtained by our scheme, on the assumption that we have an approximate sampler.

1 Introduction

A contingency table is a matrix of nonnegative integers with prescribed positive row and column sums. Contingency tables are used in statistics to store data from sample surveys. The problem of exactly counting the number of contingency tables with fixed row and column sums is known to be #P-complete, even when there are only two rows [6].

For a #P-complete problem, randomized approximation is often useful. A fully-polynomial randomized approximation scheme (fpras) is a randomized algorithm, which outputs an approximate solution Z satisfying $0 < \forall\varepsilon < 1$, $0 < \forall\delta < 1$,

$$\Pr\left[(1-\varepsilon)A \le Z \le (1+\varepsilon)A\right] \ge 1-\delta,$$

where A is the exact solution, and whose running time is bounded by a polynomial time of the input size of the problem (row and column sums), ε and δ [7].

Dyer and Greenhill introduced an fpras based on Markov chain Monte Carlo method to count the number of contingency tables with two rows [5]. In 2002, Cryan and Dyer proposed another fpras for contingency tables with constant number of rows [1]. Their scheme is a hybrid algorithm of exact counting and the calculation of convex body.

In this survey, we propose an approximate counting scheme for $m \times n$ tables based on Monte Carlo method, which is an extension and modification of Dyer and Greenhill's algorithm. When we have an approximate uniform sampler for $m \times n$ tables, we can bound the sizes of error in randomized fashion in the same

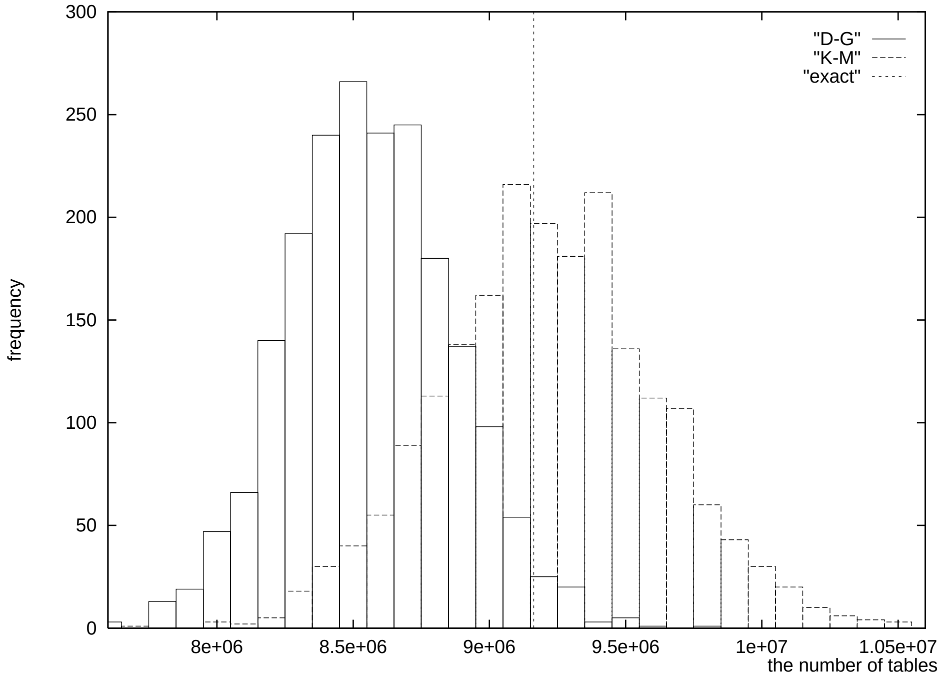

Fig. 1. The histogram

way as Dyer and Greenhill's scheme. If there exists a polynomial time approximate sampler, then we have an fpras for counting $m \times n$ tables. The advantage of our scheme comparing to Dyer and Greenhill's is that we can bound the bias of expectation of estimator. According to our result, the expectation of our approximate solution depends mainly on the difference between the distribution functions of approximate sampler and exact uniform one, and little depends on the Monte Carlo methods.

Fig. 1 shows results of our algorithm and Dyer and Greenhill's algorithm. Fig. 1 is the histogram of obtained solutions (approximate number of tables) of the common instance for two thousands executions of each algorithm. Approximate solutions by Dyer and Greenhill's (abbreviated by “D-G” in the figure) have unignorable bias from the exact number (denoted by “exact”), while solutions obtained by our algorithm (denoted by “K-M”) have little bias. Here we note that, the instance is 2×8 table given row sums $\boldsymbol{r} = (42, 38)$, and column sums $\boldsymbol{s} = (10, \ldots, 10)$, and the exact number of tables is 9,162,736.

Our scheme is different from Dyer and Greenhill's in two points. First, the reduction process of our scheme is deterministic comparing to the fact that Dyer and Greenhill's process is decided probabilistically after sampling. Thus we can parallelize our algorithm easily. Second, we don't use U/M but $(U + 1)/(M + 1)$

as estimator, where U is the number of samples and M is the sampling times. This is an ordinary method in statistics for gaining unbiased estimator.

In Section 2, we propose our approximate counting scheme. In Section 3, we prove some theorems related to the number of tables. In Section 4, we estimate the accuracy of our estimator using the theorem proposed in Section 3. Finally we sum up our discussions in Section 5.

2 Approximate counting algorithm

We denote the set of integers (non-negative integers, positive integers) by $\mathbb{Z}$ ($\mathbb{Z}_+$, $\mathbb{Z}_{++}$), respectively. For integers $m, n \geq 2$, let $\boldsymbol{r} = (r_1, \ldots, r_m) \in \mathbb{Z}_{++}^m$ and $\boldsymbol{s} = (s_1, \ldots, s_n) \in \mathbb{Z}_{++}^n$ be two positive integer partitions of a given positive integer $N \in \mathbb{Z}_{++}$. The set $\varSigma_{\boldsymbol{r},\boldsymbol{s}}$ of $m \times n$ contingency tables with row and column sums $(\boldsymbol{r}, \boldsymbol{s})$ is defined by

$$\varSigma_{\boldsymbol{r},\boldsymbol{s}} \overset{\text{def.}}{=} \left\{ X \in \mathbb{Z}_+^{m \times n} \middle| \begin{array}{l} \sum_{j=1}^n X_{ij} = r_i \, (1 \leq \forall i \leq m), \\ \sum_{i=1}^m X_{ij} = s_j \, (1 \leq \forall j \leq n) \end{array} \right\},$$

where X_{ij} is the value in the cell indexed by ith row and jth column. We define the subset $\varOmega_i \subset \varSigma_{\boldsymbol{r},\boldsymbol{s}}$ $(i = 1, \ldots, m)$ such as $\varOmega_i = \{X \in \varSigma_{\boldsymbol{r},\boldsymbol{s}} \mid X_{in} \geq \lceil s_n/m \rceil\}$. We define

$$\begin{aligned} \tilde{\boldsymbol{r}} &\overset{\text{def.}}{=} (r_1, \ldots, r_k - \lceil s_n/m \rceil, \ldots, r_m)\,, \\ \tilde{\boldsymbol{s}} &\overset{\text{def.}}{=} \begin{cases} (s_1, \ldots, s_{n-1}, \lfloor \frac{m-1}{m} s_n \rfloor)\,, & \text{if } s_n > 1\,, \\ (s_1, \ldots, s_{n-1})\,, & \text{if } s_n = 1\,, \end{cases} \end{aligned}$$

where $k \in \{1, \ldots, m\}$ is an index satisfying $r_k = \max\{r_1, \ldots, r_m\}$. Clearly, $|\varSigma_{\tilde{\boldsymbol{r}},\tilde{\boldsymbol{s}}}| = |\varOmega_k|$. If we know $\rho = |\varOmega_k|/|\varSigma_{\boldsymbol{r},\boldsymbol{s}}|$ and $|\varSigma_{\tilde{\boldsymbol{r}},\tilde{\boldsymbol{s}}}|$, we can calculate the number of tables $|\varSigma_{\boldsymbol{r},\boldsymbol{s}}|$ by $|\varSigma_{\boldsymbol{r},\boldsymbol{s}}| = |\varSigma_{\tilde{\boldsymbol{r}},\tilde{\boldsymbol{s}}}|/\rho$. When we have a uniform sampler on $\varSigma_{\boldsymbol{r},\boldsymbol{s}}$, Monte Carlo method is available to estimate the value of ρ. In fact, we generate M tables by the sampler, and suppose that U samples of M tables are in $\varOmega_k$, then we estimate the value of ρ as $(U+1)/(M+1)$. Then, even if each of M tables is not in $\varOmega_k$, we have positive estimator $1/(M+1)$ of ρ. By applying above procedure recursively, we can reduce the original problem to the problem of counting 2×2 contingency table, which is solvable in constant time. Let R be the number of calls of reduction procedure required to reduce the original problem to 2×2 table. For $i = 1, \ldots, R$, we put $Z_i = (U_i + 1)/(M + 1)$ where U_i is the number of samples in $\varOmega$ generated in ith reduction procedure. We denote the number of obtained 2×2 tables by σ. Finally we estimate $|\varSigma_{\boldsymbol{r},\boldsymbol{s}}|$ as

$$Z = \sigma \prod_{i=1}^{R} (Z_i)^{-1}.$$

The entire algorithm is sketched in Fig. 2.

```
Step 1.
  while $\max\{m, n\} \geq 3$
  do $\rightarrow$
    transpose rows and columns appropriately and assume that $m \leq n$
    sample $M$ tables according to an almost uniform distibution on $\varSigma_{\boldsymbol{r},\boldsymbol{s}}$
    set $U_i$ be the number of samples in $\varOmega_k$ where $r_k = \max\{r_1, \dots, r_m\}$
    set $Z_i = (U_i + 1)/(M + 1)$
    set $\boldsymbol{r} = \tilde{\boldsymbol{r}}$, $\boldsymbol{s} = \tilde{\boldsymbol{s}}$
    return Step 1.
Step 2.
  let $\sigma = \min\{r_1, r_2, s_1, s_2\} + 1$
  and $Z = \sigma \prod_i (Z_i)^{-1}$
```

Fig. 2. approximate counting algorithm for $m \times n$ tables

3 Properties of the number of contingency tables

In our scheme, we define the subset $\varOmega_k$ of $\varSigma_{\boldsymbol{r},\boldsymbol{s}}$ in order to reduce the size of problem. There are two important points in this definition. One is the choice of the row index k, and the other is the value $\lceil s_n/m \rceil$ reduced from the kth element r_k of row sum.

Now, we have two exclusive purposes. We need polynomial time algorithm, so we need to reduce row and column sums effectively. On the other hand we need the ratio ρ to be enough large, since the approximate solution is sensitive for the error of the estimator of ρ if ρ is too small.

We lead the following theorem for this purpose.

Theorem 1 *If an index k satisfies $r_k = \max\{r_1, \dots, r_m\}$, then $|\varOmega_k|/|\varSigma_{\boldsymbol{r},\boldsymbol{s}}| \geq 1/m$.*

With this theorem, we can bound the sizes of the error and bias of approximate solution of our scheme. It is discussed in next section. In the rest of this section, we show Theorem 1.

First, we show the following lemma.

Lemma 2 *Let the vectors $\boldsymbol{r}, \boldsymbol{r}' \in \mathbb{Z}_+^2$ and $\boldsymbol{s} \in \mathbb{Z}_+^n$ satisfy that $|r_1 - r_2| \leq |r'_1 - r'_2|$, $\sum_{i=1}^m r_i = \sum_{i=1}^m r'_i = \sum_{j=1}^n s_j = N$. Then the pair of sets $\varSigma_{\boldsymbol{r},\boldsymbol{s}}$ and $\varSigma_{\boldsymbol{r}',\boldsymbol{s}}$ satisfies $|\varSigma_{\boldsymbol{r},\boldsymbol{s}}| \geq |\varSigma_{\boldsymbol{r}',\boldsymbol{s}}|$.*

Proof: Without loss of generality, we may assume that $r'_1 > r_1 \geq r_2 > r'_2$. Let $P_n(r_2)$ be the number of tables in $\varSigma_{\boldsymbol{r},\boldsymbol{s}}$ where $\boldsymbol{r} = (r_1, r_2)$, and $P_n(u) = 0$ for $u \notin \{0, \dots, N_n\}$, i.e.,

$$P_n(u) = \begin{cases} |\varSigma_{(N_n - u, u),\boldsymbol{s}}| & (\text{if } u \in \{0, \dots, N_n\}), \\ 0 & (\text{otherwise}). \end{cases}$$

Clearly, we only need to prove that

$$P_n(u_1) \geq P_n(u_2), \qquad \lfloor N_n/2 \rfloor \geq \forall u_1 > \forall u_2 > 0,\ u_1, u_2 \in \mathbb{Z}.$$

We prove the above inequality by induction on n.

1. Case $n = 2$. Without loss of generality, we may assume the column sums $\boldsymbol{s} = (s_1, s_2)$ satisfies $s_1 \geq s_2$.
   (a) If $u_1 > u_2 > s_2$ then $P_2(u_1) = P_2(u_2)$ since $P_2(u_1) = s_2 + 1$ and $P_2(u_2) = s_2 + 1$.
   (b) If $u_1 > s_2 \geq u_2$ then $P_2(u_1) \geq P_2(u_2)$ since $P_2(u_1) = s_2 + 1$ and $P_2(u_2) = u_2 + 1$.
   (c) If $s_2 \geq u_1 > u_2$ then $P_2(u_1) > P_2(u_2)$ since $P_2(u_1) = u_1 + 1$ and $P_2(u_2) = u_2 + 1$.
2. Now, consider the case that $n = k + 1$. We add s_{k+1} as $k + 1$st column sum, so that $N_{k+1} = N_k + s_k + 1$
   (a) When $u_1 \leq \lfloor N_k/2 \rfloor$,

$$\begin{aligned} P_{k+1}(u_1) &= P_k(u_1) + P_k(u_1 - 1) + \cdots + P_k(u_1 - s_{k+1}), \\ P_{k+1}(u_2) &= P_k(u_2) + P_k(u_2 - 1) + \cdots + P_k(u_2 - s_{k+1}). \end{aligned}$$

   From the assumption of induction, we have

$$P_k(u_1 - \xi) \geq P_k(u_2 - \xi), \qquad \forall \xi \in [0, s_{k+1}] \cap \mathbb{Z},$$

   and so $P_{k+1}(u_1) \geq P_{K+1}(u_2)$.
   (b) When $\lfloor N_k/2 \rfloor < u_1 \leq \lfloor N_{k+1}/2 \rfloor$,

$$\begin{aligned} P_{k+1}(u_1) = P_k(u_1) + P_k(u_1 - 1) + \cdots + P_k(\lfloor N_k/2 \rfloor + 1) \\ + P_k(\lceil N_k/2 \rceil) + \cdots + P_k(u_1 - s_{k+1}). \end{aligned}$$

   Since $N_k + s_{K+1} = N_k \geq 2u_1$, $N_k - u_1 \geq u_1 - s_{k+1}$ holds. From the definition of $P_n(u)$, $P_k(u) = P_k(N_k - u_1)$ and

$$P_k(u_1) = P_k(N_k - u_1) \geq P_k(u_1 - s_{k+1}).$$

   The property obtained in Case (a) implies that

$$\forall u \leq \lfloor N_k/2 \rfloor, \qquad P_{k+1}(\lfloor N_k/2 \rfloor) \geq P_{k+1}(u).$$

   Thus we only need to consider the case that $\lfloor N_k/2 \rfloor \leq u_2 < u_1 \leq \lfloor N_{k+1}/2 \rfloor$. Then it is easy to see that

$$\begin{aligned} P_{k+1}(u_1) - P_{k+1}(u_2) &= \{P_k(u_1) + \cdots + P_k(u_1 - s_{k+1})\} \\ &\quad - \{P_k(u_2) + \cdots + P_k(u_2 - s_{k+1})\} \\ &\geq (u_1 - u_2)\{P_n(k_1 - s_{k+1}) - P_k(u_1 - s_{k+1} - 1)\} \\ &\geq 0, \end{aligned}$$

   and we obtained the desired result for the case that $n = k + 1$.

We extend Lemma 2 to $m \times n$ tables.

Lemma 3 *Let $\boldsymbol{r}, \boldsymbol{r}' \in \mathbb{Z}_+^m$ and $\boldsymbol{s} \in \mathbb{Z}_+^n$ satisfy that $|r_1 - r_2| \leq |r'_1 - r'_2|$, $r_i = r'_i$ $(3 \leq i \leq m)$, and $\sum_{i=1}^m r_i = \sum_{i=1}^m r'_i = \sum_{j=1}^n s_j = N$. Then the pair of sets $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ and $\Sigma_{\boldsymbol{r}',\boldsymbol{s}}$ satisfies $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \geq |\Sigma_{\boldsymbol{r}',\boldsymbol{s}}|$.*

Proof: For an $m \times n$ table $X \in \Sigma_{\boldsymbol{r},\boldsymbol{s}}$, we define the $2 \times n$ table $\overline{X}$ and the $(m-2) \times n$ table $\underline{X}$ by

$$\begin{aligned} \overline{X_{ij}} &\overset{\text{def.}}{=} X_{ij} && (i = 1, 2, \quad j = 1, \dots n), \\ \underline{X_{ij}} &\overset{\text{def.}}{=} X_{i+2\,j} && (i = 1, \dots, m-2, \quad j = 1, \dots n). \end{aligned}$$

We call each as *upper separated table*, *lower separated table*. Let $\underline{\boldsymbol{r}}$ be

$$\underline{\boldsymbol{r}} = (\underline{r_1}, \dots, \underline{r_{m-2}}) = (r_3, \dots, r_n).$$

We define the set

$$\Gamma \overset{\text{def.}}{=} \left\{ \underline{\boldsymbol{s}} \in \mathbb{Z}_+^n \,\middle|\, 0 \le \underline{s}_j \le s_j,\ \sum_{j=1}^{n} \underline{s_j} = \sum_{i=1}^{m-2} \underline{r_i} \right\},$$

and the set Λ which is the union of the set of $(m-2) \times n$ tables $\Sigma_{\underline{\boldsymbol{r}},\underline{\boldsymbol{s}}}$ as

$$\Lambda \overset{\text{def.}}{=} \bigcup_{\underline{\boldsymbol{s}} \in \Gamma} \Sigma_{\underline{\boldsymbol{r}},\underline{\boldsymbol{s}}}.$$

Now, given $(m-2) \times n$ table $Y \in \Lambda$, let Ξ_Y be the set of tables $X \in \Sigma_{\boldsymbol{r},\boldsymbol{s}}$ which has Y as lower separated table, i.e., $\Xi_Y = \{X \in \Sigma_{\boldsymbol{r},\boldsymbol{s}} | \underline{X} = Y \in \Lambda\}$. Clearly,

$$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| = \sum_{Y \in \Lambda} |\Xi_Y|. \tag{1}$$

From now on, we discuss the size $|\Xi_Y|$. For arbitrary table $Y \in \Lambda$, let $\underline{\boldsymbol{r}}$, $\underline{\boldsymbol{s}}$ be the vectors satisfying $Y \in \Sigma_{\underline{\boldsymbol{r}},\underline{\boldsymbol{s}}}$. Then (1) implies that

$$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| = \sum_{Y \in \Lambda} |\Xi_Y| = \sum_{Y \in \Lambda} |\Sigma_{\overline{\boldsymbol{r}},\overline{\boldsymbol{s}}}|. \tag{2}$$

Now, we consider $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ and $\Sigma_{\boldsymbol{r}',\boldsymbol{s}'}$ again. Since two vectors $\boldsymbol{r}$ and $\boldsymbol{r}'$ satisfy $r_i = r'_i$ $(i = 3, \dots, m)$, for any $Y \in \Lambda$, we can define $\Xi'_Y = \{X \in \Sigma_{\boldsymbol{r}',\boldsymbol{s}} | \underline{X'} = Y \in \Lambda\}$. Equations (2) imply that $|\Xi_Y| = |\Sigma_{\overline{\boldsymbol{r}},\overline{\boldsymbol{s}}}|$ and $|\Xi'_Y| = |\Sigma_{\overline{\boldsymbol{r}},\overline{\boldsymbol{s}}}|$, where $\overline{\boldsymbol{r}'} \overset{\text{def.}}{=} (r'_1, r'_2)$. From the assumption $|r'_1 - r'_2| \ge |r_1 - r_2|$ and Lemma 1,

$$|\Xi_Y| = |\Sigma_{\overline{\boldsymbol{r}},\overline{\boldsymbol{s}}}| \ge |\Sigma_{\overline{\boldsymbol{r}'},\overline{\boldsymbol{s}}}| = |\Xi'_Y|.$$

Thus

$$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| = \sum_{Y \in \Lambda} |\Xi_Y| \ge \sum_{Y \in \Lambda} |\Xi'_Y| |\Sigma_{\boldsymbol{r}',\boldsymbol{s}}|. \qquad \square$$

Now, consider the point $\boldsymbol{y} = (y_1, \dots, y_m)$ in $\mathbb{R}^m$. Put $N^- \overset{\text{def.}}{=} \sum_{j=1}^{n-1} s_j$, and we define the hyperplane

$$S \overset{\text{def.}}{=} \left\{ \boldsymbol{y} \in \mathbb{R}^m \,\middle|\, \sum_{i=1}^{m} y_i = N^- \right\}.$$

And for $\boldsymbol{r} = (r_1, \ldots, r_m)$, we define

$$S^+ \overset{\text{def.}}{=} \{\boldsymbol{y} \in S | y_i \geq 0, \ i = 1, \ldots, m\}, \tag{3}$$

$$S_0 \overset{\text{def.}}{=} \{\boldsymbol{y} \in S | r_i - s_n \leq y_i \leq r_i, \ i = 1, \ldots, m\}. \tag{4}$$

Now let $\boldsymbol{x} \in \mathbb{R}^m$ be $\boldsymbol{x} \overset{\text{def.}}{=} \boldsymbol{r} - \boldsymbol{y}$. For any $\boldsymbol{x}$, a table $X \in \Sigma_{\boldsymbol{r},\boldsymbol{s}}$ satisfying $x_i = X_{in}$ $(i = 1, \ldots, m)$ exists if and only if

$$\begin{aligned} 0 \leq x_i \leq \min\{r_i, s_n\} \qquad & (i = 1, \ldots, m), \\ \sum_{i=1}^m x_i = s_n, \qquad & \boldsymbol{x} \in \mathbb{Z}^m. \end{aligned} \tag{5}$$

We can rewrite these conditions as

$$\begin{aligned} \max\{0, r_1 - s_n\} \leq y_i \leq r_i \qquad & (i = 1, \ldots, m), \\ \sum_{i=1}^m y_i = N^-, \qquad & \boldsymbol{y} \in \mathbb{Z}^m. \end{aligned} \tag{6}$$

These conditions are equivalent to $\boldsymbol{y} \in S^+ \cap S_0 \cap \mathbb{Z}^m$.

For any $k, l \in \{1, \ldots, m\}$ such that $k \neq l$, we define the hyperplane H_{ij} by

$$H_{ij} \overset{\text{def.}}{=} \{\boldsymbol{y} \in \mathbb{R}^m | - y_i + y_j = -r_i + r_j\}.$$

We define S_i by

$$S_i \overset{\text{def.}}{=} \{\boldsymbol{y} \in S_0 | r_i - y_i = \max\{r_j - y_j | l = 1, \ldots, m\}\}.$$

For any $\boldsymbol{y} \in S$ we define the symmetric point with H_{kl} by

$$h_{kl}(\boldsymbol{y}) \overset{\text{def.}}{=} (y_1^*, \ldots, y_m^*), \qquad y_i^* \overset{\text{def.}}{=} \begin{cases} r_i - r_j + y_j & (i, j = k, l, \ i \neq j), \\ y_i & (\text{otherwise}). \end{cases}$$

Clearly, there exists a bijection between the point sets S_i and S_j.

For a given point $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathbb{Z}_+^m$, we define the function

$$q(\boldsymbol{y}) \overset{\text{def.}}{=} \begin{cases} |\Sigma_{\boldsymbol{y},\boldsymbol{s}^-}| & (\boldsymbol{y} \in \mathbb{Z}_+^m \cap S_0 \cap S^+), \\ 0 & (\text{otherwise}), \end{cases} \tag{7}$$

where $\boldsymbol{s}^- \overset{\text{def.}}{=} (s_1, \ldots, s_{n-1})$.

Lemma 4 *Let the index $k \in \{1, \ldots, m\}$ satisfy $r_k = \max\{r_1, \ldots, r_m\}$. Suppose that $l \in \{1, \ldots, m\}$, $l \neq k$ and $\boldsymbol{y}^* = h_{kl}(\boldsymbol{y})$ is the symmetrical point of $\boldsymbol{y} \in S_k$ associated with H_{kl}, then $q(\boldsymbol{y}) \geq q(\boldsymbol{y}^*)$.*

Proof: From the definition of $q(\boldsymbol{y})$, $q(\boldsymbol{y}) = |\Sigma_{\boldsymbol{y},\boldsymbol{s}^-}|$ and $q(\boldsymbol{y}^*) = |\Sigma_{\boldsymbol{y}^*,\boldsymbol{s}^-}|$. Now, $\boldsymbol{y}$ and $\boldsymbol{y}^*$ is different at kth and lth elements, and so

$$\begin{aligned} |y_k^* - y_l^*| &= |(r_k - r_l + y_l) - (r_l - r_k + y_k)| = |(r_k - y_k) - (r_l - y_l) + (r_k - r_l)| \\ &= |(r_k - y_k) - (r_l - y_l)| + (r_k - r_l) = |(r_k - r_l) + (y_k - y_l)| + (r_k - r_l) \\ &\geq |y_k - y_l| - (r_k - r_l) + (r_k - r_l) = |y_k - y_l|. \end{aligned}$$

From Lemma 3, we have

$$|\varSigma_{\boldsymbol{y},\boldsymbol{s}^-}| \geq |\varSigma_{\boldsymbol{y}^*,\boldsymbol{s}^-}|$$

and from the definition of the function q

$$q(\boldsymbol{y}) \geq q(\boldsymbol{y}^*). \qquad \square$$

We define the function $Q_k \overset{\text{def.}}{=} \sum_{\boldsymbol{x}\in S_k} q(\boldsymbol{x})$ for $k = 1, \ldots, m$.

Lemma 5 *If an index k satisfies $r_k = \max\{r_1, \ldots, r_m\}$, then for any index i, $Q_k \geq Q_i$.*

Proof: From Lemma 4, $q(\boldsymbol{y}) \geq q(\boldsymbol{y}^*)$ for any $\boldsymbol{y} \in S_k$. There exists a bijection between the point sets S_k and S_l, and so $|S_k \cap \mathbb{Z}^m| = |S_l \cap \mathbb{Z}^m|$. Thus

$$Q_k = \sum_{\boldsymbol{y}\in\mathbb{Z}^m\cap S_k} q(\boldsymbol{y}) \geq \sum_{\boldsymbol{y}\in\mathbb{Z}^m\cap S_l} q(\boldsymbol{y}^*) = Q_i. \qquad \square$$

With these Lemmas, we can show Theorem 1.
Proof: We define the set of contingency tables $\varUpsilon_{\boldsymbol{x}}$ by

$$\varUpsilon_{\boldsymbol{x}} \overset{\text{def.}}{=} \{X \in \varSigma_{\boldsymbol{r},\boldsymbol{s}} | \boldsymbol{x} = (x_1, \ldots, x_m),\ X_{in} = x_i\ (i = 1, \ldots, m)\}.$$

Let $\boldsymbol{y} = \boldsymbol{r} - \boldsymbol{x}$, $\boldsymbol{s}^- = (s_1, \ldots, s_{m-1})$. It means that for $X \in \varUpsilon_{\boldsymbol{x}}$, $\sum_{j=1}^{n-1} X_{ij} = y_j$, $\sum_{i=1}^{m} X_{ij} = s_j$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n-1$. Therefore $|\varUpsilon_{\boldsymbol{x}}| = |\varSigma_{\boldsymbol{y},\boldsymbol{s}^-}|$. The condition that $|\varUpsilon_{\boldsymbol{x}}| \neq \emptyset$ is described as (5) and equivalent to (6). Also, using S^+, S_0 defined as (3) and (4), we can rewrite the condition as

$$\boldsymbol{y} \in \mathbb{Z}^m \cap S_0 \cap S^+.$$

Now we have,

$$|\varSigma_{\boldsymbol{r},\boldsymbol{s}}| = \sum_{\boldsymbol{x}} |\varUpsilon_{\boldsymbol{x}}| = \sum_{\boldsymbol{y}\in\mathbb{Z}^m\cap S_0\cap S^+} |\varSigma_{\boldsymbol{y},\boldsymbol{s}^-}| = \sum_{\boldsymbol{y}\in S_0} q(\boldsymbol{y}),$$

where the function $q(\boldsymbol{y})$ is defined by (7). When $\boldsymbol{y} \in S_k$, $\boldsymbol{x} = \boldsymbol{r} - \boldsymbol{y}$ satisfies the condition $x_k = \max\{x_1, \ldots, x_m\}$. Then $x_k \geq \lceil s_n/m \rceil$. Hence it is clear that

$$|\varOmega_k| \geq \sum_{\boldsymbol{y}\in S_k} q(\boldsymbol{y}) = Q_k.$$

With Lemma 5,

$$|\varSigma_{\boldsymbol{r},\boldsymbol{s}}| \leq \sum_{i=1}^{m} \sum_{\boldsymbol{y}\in S_i} q(\boldsymbol{y}) = \sum_{i=1}^{m} Q_i \quad \leq \ mQ_k,$$

therefore

$$|\varOmega_K| \geq Q_k \geq \frac{1}{m} |\varSigma_{\boldsymbol{r},\boldsymbol{s}}|. \qquad \square$$

So, if we choose an index k such as $r_k = \max\{r_1, \ldots, r_m\}$, then we obtain an algorithm such that the ratio ρ in each step attains $\rho \geq 1/m$.

4 Bound the error and bias

In this section, we discuss the sizes of error and bias of the estimator obtained by our scheme. For any pair of distributions ν and ν' defined on the finite discrete state space Ω, we define the total variation distance $\mathrm{d_{TV}}(\nu, \nu')$ by

$$\mathrm{d_{TV}}(\nu, \nu') \overset{\text{def.}}{=} \frac{1}{2} \sum_{X \in \Omega} |\nu(X) - \nu'(X)|.$$

Let π be the uniform distribution on $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$. Now, supposing that we have an approximate sampler on $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$, whose probability distribution function ν satisfies that for arbitrary nonnegative $\varepsilon < 1$, $\mathrm{d_{TV}}(\pi, \nu) \leq \varepsilon/6mR$, where we can gain an estimator Z for $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ after R times reduction. If we set $M = 108mR^2\varepsilon^{-2}\ln(2R/\delta)$, then we obtain next two theorems about the sizes of error and bias.

Theorem 6 *The estimator Z satisfies*

$$\Pr\left[(1-\varepsilon)|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \leq Z \leq (1+\varepsilon)|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|\right] \geq 1 - \delta \ .$$

This theorem is proved in a similar way with [6, 5] as follows;

1. for $1 \leq i \leq R$, $\hat{\rho}_i \overset{\text{def.}}{=} \mathrm{E}\left[U_i/M\right] \geq 1/m - \varepsilon/6mR$,
2. $|\rho_i - \hat{\rho}_i| \leq \dfrac{\varepsilon}{6R - \varepsilon}\hat{\rho}_i$,
3. $\Pr\left[|Z_i - \hat{\rho}_i| > \dfrac{\varepsilon}{6R-\varepsilon}\hat{\rho}_i\right] \leq \dfrac{\delta}{R}$,
4. with the probability higher than $1 - \delta$, $|(Z_1 \cdots Z_R)^{-1} - (\rho_1 \cdots \rho_R)^{-1}| \leq \varepsilon(\rho_1 \cdots \rho_R)^{-1}$.

Thus, our scheme is an fpras if we have a polynomial time approximate uniform sampler. What is more, we can bound the sizes of the bias of $\mathrm{E}\left[Z\right]$.

Theorem 7 *The estimator Z satisfies*

$$\frac{|\mathrm{E}\left[Z\right] - \Sigma_{\boldsymbol{r},\boldsymbol{s}}|}{|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|} \leq \frac{\varepsilon}{4} + \mathrm{e}^{-90R^3\varepsilon^{-2}\ln(2R/\delta)} \leq \left(\frac{1}{4} + \frac{1}{10^{27}}\right)\varepsilon \ .$$

Proof: Since $Z_1, \ldots, Z_R$ are independent,

$$\mathrm{E}\left[Z\right] = \sigma \mathrm{E}\left[\prod_{i=1}^{R} \frac{1}{Z_i}\right] = \sigma \prod_{i=1}^{R} \mathrm{E}\left[\frac{1}{Z_i}\right].$$

Now, we have

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{Z_i}\right] &= \sum_{U_i=0}^{M} \frac{M+1}{U_i+1}\binom{M}{U_i}\hat{\rho}_i^{\,U_i}(1-\hat{\rho}_i)^{M-U_i} \\
&= \frac{1}{\hat{\rho}_i}\sum_{U_i=0}^{M}\binom{M+1}{U_i+1}\hat{\rho}_i^{\,U_i+1}(1-\hat{\rho}_i)^{M-U_i} \\
&= \frac{1}{\hat{\rho}_i}\left\{1 - (1-\hat{\rho}_i)^{M+1}\right\}.
\end{aligned}$$

Then the equality

$$\mathrm{E}\left[Z\right] = \sigma \prod_i \frac{1}{\hat{\rho}_i} \left\{1 - (1 - \hat{\rho}_i)^{M+1}\right\} \tag{8}$$

holds. From (8), $||\Sigma_{\boldsymbol{r},\boldsymbol{s}}| - \mathrm{E}\left[Z\right]|$ satisfies

$$\begin{aligned}
||\Sigma_{\boldsymbol{r},\boldsymbol{s}}| - \mathrm{E}\left[Z\right]| &= \left|\sigma \prod_i \frac{1}{\rho_i} - \sigma \prod_i \frac{1}{\hat{\rho}_i}\left\{1 - (1 - \hat{\rho}_i)^{M+1}\right\}\right| \\
&\leq \left|\sigma \prod_i \frac{1}{\rho_i} - \sigma \prod_i \frac{1}{\hat{\rho}_i}\right| \left|\left\{1 - (1 - \hat{\rho}_i)^{M+1}\right\}\right| + \left|\sigma \prod_i \frac{1}{\rho_i}(1 - \hat{\rho}_i)^{M+1}\right|,
\end{aligned}$$

and

$$\begin{aligned}
&\left|\sigma \prod_i \frac{1}{\rho_i} - \sigma \prod_i \frac{1}{\hat{\rho}_i}\right| = \sigma \prod_i \frac{1}{\rho_i}\left|1 - \prod_i \frac{\rho_i}{\hat{\rho}_i}\right| \leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \left\{\left(1 + \frac{\varepsilon}{6R - \varepsilon}\right)^R - 1\right\} \\
&\quad \leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \left\{\exp\left(\frac{\varepsilon R}{6R - \varepsilon}\right) - 1\right\} \leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \frac{\varepsilon R}{6R - \varepsilon - \varepsilon R} \leq \frac{\varepsilon}{4} |\Sigma_{\boldsymbol{r},\boldsymbol{s}}|,
\end{aligned}$$

and so

$$||\Sigma_{\boldsymbol{r},\boldsymbol{s}}| - \mathrm{E}\left[Z\right]| \leq \frac{\varepsilon}{4} |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| + |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \prod_i (1 - \hat{\rho}_i)^{M+1}.$$

Lastly, we estimate the magnitude of $\prod_i (1 - \hat{\rho}_i)^{M+1}$ by

$$\begin{aligned}
\prod_i (1 - \hat{\rho}_i)^{M+1} &\leq \left\{1 - \left(\frac{1}{m} - \frac{\varepsilon}{6mR}\right)\right\}^{RM} = \left\{1 - \left(\frac{1 - \frac{\varepsilon}{6R}}{m}\right)\right\}^{RM} \\
&\leq \left(1 - \frac{5}{6m}\right)^{RM} \leq \left(1 - \frac{5}{6m}\right)^{R\, 108 m R^2 \varepsilon^{-2} \ln(2R/\delta)} \\
&\leq \left(\mathrm{e}^{-1}\right)^{\frac{5}{6m} 108 m R^3 \varepsilon^{-2} \ln(2R/\delta)} \\
&= \mathrm{e}^{-90 R^3 \varepsilon^{-2} \ln(2R/\delta)}.
\end{aligned}$$

Thus the bias becomes as follows

$$\begin{aligned}
||\Sigma_{\boldsymbol{r},\boldsymbol{s}}| - \mathrm{E}\left[Z\right]| &\leq \frac{\varepsilon}{4} |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| + \mathrm{e}^{-90 R^3 \varepsilon^{-2} \ln(2R/\delta)} |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \\
&\leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \left(\frac{\varepsilon}{4} + \mathrm{e}^{-90 R^3 \varepsilon^{-2} \ln(2R/\delta)}\right). \qquad \square
\end{aligned}$$

Then, we can see that the bias mainly depends on the difference between the distribution function of approximate sampler and exact uniform one, and little depends on the Monte Carlo method.

5 Conclusion and future work

We proposed a new approximate counting scheme for $m \times n$ contingency tables. We consider the properties of the number of contingency tables, and bound the bias of our estimator. We show a computational result of our algorithm and Dyer and Greenhill's, for two rowed contingency tables.

When we have a polynomial time approximate samples sampler, our scheme becomes an fpras. In 2002, Cryan et al. showed that heat bath Markov chain for contingency tables with constant number of rows is rapidly mixing [2]. Of course, employing their results, our scheme becomes an fpras for contingency tables with constant number of rows. The existence of polynomial time approximate uniform sampler for $m \times n$ contingency tables is still an open problem.

References

1. M. Cryan and M. Dyer, "A polynomial-time algorithm to approximately count contingency tables when the number of rows is constant," *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)* (2002), pp. 240–249.
2. M. Cryan, M. Dyer, L. A. Goldberg, M. Jerrum, and R. Martin, "Rapidly mixing Markov chains for sampling contingency tables with constant number of rows," *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science (FOCS)* (2002), pp. 711–720.
3. P. Diaconis and L. Saloff-Coste, "Random walk on contingency tables with fixed row and column sums," *Tech. rep.*, Department of Mathmatics, Harvard University, 1995.
4. M. Dyer, A. Frieze, and R. Kannan, "A random polynomial time algorithm for approximating the volume of convex bodies," *Journal of the ACM*, 38 (1991), pp. 1–17.
5. M. Dyer and C. Greenhill, "Polynomial-time counting and sampling of two-rowed contingency tables," *Theoretical Computer Sciences*, 246 (2000), pp. 265–278.
6. M. Dyer, R. Kannan, and J. Mount, "Sampling contingency tables," *Random Structures and Algorithms*, 10 (1997), pp. 487–506.
7. M. Jerrum and A. Sinclair, "The Markov chain Monte Carlo method: an approach to approximate counting and integration," In *Approximation Algorithm for NP-hard Problems* (Dorit Hochbaum, ed.), PWS, 1996, pp. 482–520.
8. L. G. Valiant, "The complexity of computing the permanent," *Theoretical Computer Science* 8 (1979), pp. 189–201.