卒業論文

2元分割表の個数数え上げ問題

指導教官　松井 知己 助教授

東京大学 工学部 計数工学科

来嶋 秀治

2002年2月12日

目次

1 序論 **1**

1.1 2 元分割表の個数数え上げ問題 1

1.2 本論文の構成 1

2 数え上げアルゴリズム **2**

2.1 MCMC 法による数え上げアルゴリズム 2

2.1.1 数え上げアルゴリズム 2

2.1.2 モンテカルロ法による数え上げ 3

2.1.3 マルコフ連鎖による一様サンプリング 4

2.2 計算量と精度 4

2.3 数え上げアルゴリズムの改良 5

2.4 推定量の偏りについて 8

3 計算機実験 **11**

3.1 計算機実験について 11

3.2 計算結果と考察 12

3.2.1 推定量の偏りに関する実験 12

3.2.2 マルコフ連鎖の推移回数に関する実験 13

3.2.3 サンプル数 M と精度に関する実験 15

3.2.4 表値の合計 N と誤差の関係。 15

3.2.5 列数 n と精度、および計算時間に関する実験 15

3.2.6 計算時間 16

4 結論 **17**

4.1 結論 17

4.2 課題 17

A マルコフ連鎖の既約性と非周期性 **18**

B 理論的精度保証の証明 **19**

謝辞 **23**

参考文献 **24**

概要

本論文では、マルコフ連鎖モンテカルロ法 (MCMC 法) を用いて 2 行 n 列の 2 元分割表の個数の近似解を求めるアルゴリズムについて議論する。2000 年に Dyer and Greenhill によって提案されたアルゴリズムでは列数、表中の値の合計の入力サイズおよび誤差の逆数の多項式時間で理論的に精度が保証された解が得られる。このアルゴリズムを実装し、その実際の性能について確かめた。また、計算機実験によりこのアルゴリズムによる推定量が偏っていることが確認された。これに対し、偏りの大きさの上限を理論的に見積もることができる新しいアルゴリズムを提案した。また、計算機実験を行い、このアルゴリズムで求められる推定量の偏りが既存のアルゴリズムに比べ小さくなることを確認した。

第 1 章

序論

2 元分割表は、正の整数からなる行和と列和を持ち、表値として非負整数を取る表 (行列) である。2 元分割表は医療統計の分野などで統計データを扱うのに用いられている。2 元分割表の行と列の相関を検定する際、表値の合計が十分大きくないと、カイ二乗分布の当てはまりの悪さが問題となるため、正確法が用いられる。正確法を行う上で、周辺和が与えられている 2 元分割表の個数を知ることは重要である。

1.1 2 元分割表の個数数え上げ問題

行和と列和が与えられている m 行 n 列の 2 元分割表の個数を数え上げる問題は、m 本の等式制約を持つ n 変数の整数計画問題の許容解の個数を求める問題と等価である。この問題は $m = 2$ の時でさえ解くのが難しいことが知られている。

そこで、マルコフ連鎖モンテカルロ法 (MCMC 法) を用いてこの問題の近似解を求める。

MCMC 法とは、ある値の算出にモンテカルロ法を用い、そのサンプリングにマルコフ連鎖を用いる方法である。サンプリングにマルコフ連鎖を用いる理由は、状態の総数が不明であり、多すぎるため、一様サンプリングが困難だからである。定常分布として、一様分布をもつマルコフ連鎖を構築し、任意の初期状態からマルコフ連鎖の推移を十分繰り返して得られる状態をサンプルとして用いる。

本論文では、従来のアルゴリズムに比べ、推定量の偏りが小さくなるような新しいアルゴリズムを提案する。そして、このアルゴリズムを用いて得られた推定量の偏りの大きさの上限を理論的に保証する。また、計算機実験を行い、この理論的精度を保証するパラメータの値について検証する。

1.2 本論文の構成

本論文は、2 行 n 列の 2 元分割表の個数数え上げ問題を MCMC 法で解き、個数の近似解として推定量を得ることを目的とする。この推定量は精度と偏りの大きさを理論的に保証することができる。

以下に本論文の構成を述べる。

第 2 章において、分割表の数え上げアルゴリズムを紹介し、理論的な推定量の精度保証をする。また、推定量の偏りが小さくなるような新しいアルゴリズムを提案し、そのアルゴリズムで得られる推定量の偏りの大きさの上限を理論的に保証する。第 3 章において、第 2 章で提示した近似解法の計算機実験の結果を記し、理論との比較、考察を行う。最後に第 4 章で、結論と今後の課題を述べる。

第 2 章

数え上げアルゴリズム

本章では、MCMC 法を用いて 2 行 n 列分割表の個数を近似的に求めるアルゴリズムについて述べる。このアルゴリズムは Diaconis and Gangolli [1] および Hernek [3] のアルゴリズムを Dyer and Greenhill [2] が改良したものである。このアルゴリズムで求まる近似解の理論的な精度保証を行う。次に、既存のアルゴリズムに比べ、推定量の真の値からの偏りが小さくなるようなアルゴリズムを新たに提案する。さらに、新しいアルゴリズムで得られる推定量の偏りの大きさの上限を理論的に保証する。したがって、このアルゴリズムで、理論的に精度と偏りの大きさが保証された近似解を求めることができる。

2.1 MCMC 法による数え上げアルゴリズム

2.1.1 数え上げアルゴリズム

$\boldsymbol{r} = (r_1, r_2),\ \boldsymbol{s} = (s_1, \ldots, s_n)$ は自然数の要素からなり、$\sum_{i=1}^{2} r_i = \sum_{j=1}^{n} s_j = N$ (但し N は自然数) を満たすベクトルとする。各行の行和が $\boldsymbol{r}$、列和が $\boldsymbol{s}$ で表し、非負整数値を表値にとる 2 行 n 列の 2 元分割表全体の集合 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ を、

$$\Sigma_{\boldsymbol{r},\boldsymbol{s}} = \left\{ X \in \mathbb{N}_0^{2\times n} : \sum_{j=1}^{n} X_{ij} = r_i \quad (1 \le i \le 2)\,, \quad \sum_{i=1}^{2} X_{ij} = s_j \quad (1 \le j \le n) \right\} \tag{2.1}$$

で表す。但し、$\mathbb{N}_0$ は非負整数値全体の集合を表す。

この $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ に対して、集合 $\Omega_l\,(l = 0, 1, 2)$ を

$$\Omega_l = \begin{cases} \Sigma_{\boldsymbol{r},\boldsymbol{s}}\,, & l = 0\,, \\ \{X \in \Sigma_{\boldsymbol{r},\boldsymbol{s}} : X_{ln} \ge \lceil s_n/2 \rceil\}\,, & l = 1, 2\,, \end{cases} \tag{2.2}$$

で定義する。但し、$X_{i,j}$ は分割表 X の i 行 j 列の表値を表す。

また、$J \in \{1, 2\}$ に対し、$\boldsymbol{r'}$、$\boldsymbol{s'}$ を、

$$\boldsymbol{r'} = \begin{cases} (r_1 - \lceil s_n/2 \rceil\,,\, r_2), & \text{if } J = 1, \\ (r_1\,,\, r_2 - \lceil s_n/2 \rceil), & \text{if } J = 2, \end{cases} \tag{2.3}$$

$$\boldsymbol{s'} = \begin{cases} (s_1, \ldots, s_{n-1}, \lfloor s_n/2 \rfloor), & \text{if } s_n > 1, \\ (s_1, \ldots, s_{n-1}), & \text{if } s_n = 1, \end{cases} \tag{2.4}$$

で定義する。この時、Ω_J に含まれる分割表 X は全て $X_{Jn} \ge \lceil s_n/2 \rceil$ であるから、

$$|\Omega_J| = |\Sigma_{\boldsymbol{r'},\boldsymbol{s'}}| \tag{2.5}$$

が成り立つ。したがって、

$$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| = \left(\frac{|\Omega_J|}{|\Omega_0|}\right)^{-1} |\Sigma_{\boldsymbol{r}',\boldsymbol{s}'}|\,, \qquad 0 < \frac{|\Omega_J|}{|\Omega_0|} \leq 1, \tag{2.6}$$

となる。

以上の記号を用いて、状態 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ に対する操作 $\mathcal{T}$ を次のように定義する：

操作 $\mathcal{T}$

i) 状態 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ に対して適当に J を決める。

ii) この J に対して、$\dfrac{|\Omega_J|}{|\Omega_0|}$, $\boldsymbol{r}'$, $\boldsymbol{s}'$ を求める。

iii) 状態を $\Sigma_{\boldsymbol{r}',\boldsymbol{s}'}$ に更新する。□

さて、集合 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ の要素数 $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ を求めるアルゴリズムを考える。

まず、$\boldsymbol{s}$ が 2 次元ならば、2 行 2 列の分割表の個数 $\sigma = |\Sigma_{(r_1,r_2),(s_1,s_2)}|$ は

$$\sigma = 1 + \min(r_1, r_2, s_1, s_2) \tag{2.7}$$

で簡単に得られる。

次に、$\boldsymbol{s}$ が 3 次元以上の場合、$\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ に対して、操作 $\mathcal{T}$ を $R = \sum_{q=3}^{n} \lfloor 1 + \log_2(s_q) \rfloor$ 回繰り返すと、2 行 2 列の分割表の集合 $\Sigma_{(r_1,r_2),(s_1,s_2)}$ が得られる。この時、i 回目の操作における $|\Omega_J|/|\Omega_0|$ の値を ρ_i $(i = 1, 2, \ldots, R)$ とすると、$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ は

$$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| = \sigma(\rho_1 \cdots \rho_R)^{-1} \tag{2.8}$$

で求めることができる。

2.1.2 モンテカルロ法による数え上げ

現実には上記のアルゴリズムにおいて $|\Omega_0|$、$|\Omega_J|$ を求めることは困難であるので $\rho_i = |\Omega_J|/|\Omega_0|$ は簡単には求まらない。したがって、状態 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ からモンテカルロ法で ρ_i の推定量 Z_i を求めることを考える。

操作 $\hat{\mathcal{T}}$　操作 $\mathcal{T}$ において、状態 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ から一様サンプリングを行い、M 個のサンプルを得る。このサンプル全体の集合を $\mathcal{S}$ と置き、集合 $\mathcal{U}_l$ $(l = 1, 2)$ を

$$\mathcal{U}_l = \{X : X \in \Omega_l \cap S\} \tag{2.9}$$

で定める。この時、$J \in \{1, 2\}$ を

$$J = \begin{cases} 1 & \text{if} \quad |\mathcal{U}_1| \geq |\mathcal{U}_2|, \\ 2 & \text{otherwise} \end{cases} \tag{2.10}$$

で定義すると、$|\mathcal{U}_J|/M$ は $|\Omega_J|/|\Omega_0|$ の推定量として見なすことができる。□

$\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ に操作 $\hat{\mathcal{T}}$ を $R = \sum_{q=3}^{n} \lfloor 1 + \log_2(s_q) \rfloor$ 回繰り返せば、ρ_i の推定量 Z_i $(1 \leq i \leq R)$ を得られる。この Z_i を用いると、$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ の推定量

$$Z = \sigma(Z_1 Z_2 \cdots Z_R)^{-1} \tag{2.11}$$

を求めることができる。

2.1.3　マルコフ連鎖による一様サンプリング

上記のモンテカルロ法では、$\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ からの一様サンプリングを仮定しているが、実際には $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ のサイズがわからないので、一様サンプリングは困難である。したがって、状態 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ 内をマルコフ連鎖で推移させ、近似的な一様サンプリングを行うことを考える。

状態空間 $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ 上にマルコフ連鎖 $\mathcal{M}$ を生成し、時刻 t におけるマルコフ連鎖 $\mathcal{M}$ の状態 X から次の時刻 $t+1$ の $\mathcal{M}$ の状態 X' への推移を以下のように決定する：

i) $1 \leq j_1 < j_2 \leq n$ を一様分布からランダムに選ぶ。

ii) 時刻 t における状態 $X \in \Sigma_{\boldsymbol{r},\boldsymbol{s}}$ の i 行 j 列の値を X_{ij} で表す時、

$$\begin{aligned} b_i &= X_{ij_1} + X_{ij_2} \qquad i \in \{1,2\}, \\ c_k &= X_{1j_k} + X_{2j_k} \qquad k \in \{1,2\} \end{aligned}$$

とする。行和 $\boldsymbol{b} = (b_1, b_2)$、列和 $\boldsymbol{c} = (c_1, c_2)$ が与えられた非負整数値を表値に持つ 2 行 2 列の分割表の集合を $\mathcal{Y}$ とする。

iii) $Y \in \mathcal{Y}$ を確率 $1/|\mathcal{Y}|$ で決める。

iv) 時刻 $t+1$ における状態 $X' \in \Sigma_{\boldsymbol{r},\boldsymbol{s}}$ を

$$X'_{ij} = \begin{cases} Y_{ik}, & j = j_k, \\ X_{ij}, & \text{otherwise} \end{cases}$$

とする。 □

このマルコフ連鎖は既約で非周期的であることがわかる (付録 A)。また、Dyer and Greenhill はこのマルコフ連鎖は急速に混交して $\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ 上の一様分布とみなせる定常分布を取り、混交に必要な時間 $\tau(\epsilon)$ は $\forall \epsilon > 0$ に対して、

$$\tau(\epsilon) \leq \frac{n(n-1)}{2} \ln\left(N\epsilon^{-1}\right) \tag{2.12}$$

で押さえられることを証明している [2]。従って、任意の初期状態からマルコフ連鎖 $\mathcal{M}$ を $T \geq \tau(\epsilon)$ 回推移して得られたサンプル X は誤差 ϵ の範囲で十分一様なサンプルと見なすことができる。こうして近似的に一様なサンプリングが得られる。

2.2　計算量と精度

このアルゴリズムで求まる近似解の計算量と精度の関係を考える。適当な初期状態からマルコフ連鎖 $\mathcal{M}$ を T 回推移するごとにサンプルを 1 つ取るとする。サンプリング数 M、マルコフ連鎖の推移回数 T を

$$M = \left\lceil 150\mathrm{e}^2 R^2 \epsilon^{-2} \ln\left(3R\delta^{-1}\right) \right\rceil , \tag{2.13}$$

$$T = \left\lceil \tau\left(\epsilon/(15R\mathrm{e}^2)\right) \right\rceil \tag{2.14}$$

$$= \left\lceil \frac{n(n-1)}{2} \ln\left(15R\mathrm{e}^2 N\epsilon^{-1}\right) \right\rceil \tag{2.15}$$

とおくと、$\Sigma_{\boldsymbol{r},\boldsymbol{s}}$ の推定量 Z の値を求めるのに必要なマルコフ連鎖の推移回数の総数は

$$RMT \tag{2.16}$$

となる。但し、前に定義したように $R = \sum_{q=3}^{n} \lfloor 1 + \log_2(s_q) \rfloor$ であり、操作 $\hat{T}$ を再帰的に呼び出す回数である。この時、推定量 Z の精度として

$$\mathrm{Prob}\left[(1-\epsilon)|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \le Z \le (1+\epsilon)|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \right] \ge 1-\delta \tag{2.17}$$

が理論的に保証される (付録 B)。

また、

$$\begin{aligned} R &= \sum_{q=3}^{n} \lfloor 1 + \log_2(s_q) \rfloor & (2.18) \\ &\le n \log_2 N & (2.19) \end{aligned}$$

であることから 、計算量は $n,\ \log(N),\ \epsilon^{-1},\ \log(\delta^{-1})$ の多項式で押さえることができる。

2.3 数え上げアルゴリズムの改良

Dyer and Greenhill [2] の提案するアルゴリズムでは、$\rho_i = 1/2$ の時常に $Z_i \ge 1/2$ となり、推定量 Z は真の値 $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ に比べ小さくなる。また、ρ_i が $1/2$ に近い時、操作 $\hat{T}$ において、J として $|\Omega_l| < |\Omega_{\bar{l}}|$ $(l, \bar{l} \in \{1,2\}, \bar{l} \ne l)$ となる l が採択されることがある。この時本来 $\rho_i = |\Omega_l|/|\Omega_0| < 1/2$ なのに $Z_i \ge 1/2$ となるので、推定量は真の値より小さくなる。さらに、更新された分割表の集合の個数 $|\Sigma_{\boldsymbol{r}',\boldsymbol{s}'}| = |\Omega_l|$ は本来採択されるはずであった $|\Omega_{\bar{l}}|$ より小さくなるので、推定量は真値よりも小さくなる。

この点を改良し、推定量の真の値からの偏りを小さくするために、本論文では、Dyer and Greenhill のアルゴリズムに 2 つの改良を行う。ひとつ目は J の定め方を $r_J \ge r_{\bar{J}}$ $(J, \bar{J} \in \{1,2\},\ \bar{J} \ne J)$ で定めるとすることである。こうすることで J はサンプルによらず一定に定まり、$Z_i - \rho_i$ も正負両側に分布する。また、後に述べる定理 1 から常に $\rho_i \ge 1/2 > 1/(2\mathrm{e}^2)$ となるので、精度保証の理論と矛盾しない。ふたつ目は ρ_i の推定量 Z_i を $(U_i+1)/(M+1)$ とすることである。但し、U_i は i 回目の操作 $\hat{T}$ で求まる $|\mathcal{U}_J|$ のことである。これは、次の節で偏りの大きさの上限を理論的に保証するためである。直感的な説明としては、$U_i = 0$ の時、$1/Z_i$ が定義できなくなるのを防ぎ、また、$U_i = M$ の時、$Z_i > 1$ とならないようにするためである。

ここでは、一つ目の改良で行われた J のとり方が、$\rho_i \ge 1/(2\mathrm{e}^2)$ という条件を満たしていることを示す。

定理 1　行和ベクトル $\boldsymbol{r} = (r_1, r_2)$ について、$r_1 > r_2$ のとき $|\Omega_1| \ge |\Omega_2|$ が成り立つ。

これを証明するために以下の補題 1 を示す。

補題 1　自然数ベクトル $\boldsymbol{s} = (s_1, \ldots, s_n)$ が与えられているとする。自然数 $N_n = s_1 + \cdots + s_n$ を定義する。行和が $\boldsymbol{r} = (r_1, r_2)$、列和が $\boldsymbol{s}$ となる 2 行 n 列分割表の集合の要素数を $\mathrm{P}_n[r_2]$ で表すことにする。また、$u \notin [0, N_n],\ u \in \mathbb{Z}$ を満たす u については、$\mathrm{P}_n[u] = 0$ と定義しておく。すなわち $u \in \mathbb{Z}$ に対して、

$$\mathrm{P}_n[u] = \begin{cases} |\Sigma_{(N_n-u,u),\boldsymbol{s}}|, & \text{if } 0 \le u \le N_n, \\ 0, & \text{otherwise}, \end{cases} \tag{2.20}$$

である。この時、$\lfloor N_n/2 \rfloor \geq \forall u_1 > \forall u_2 > 0$ に対して、

$$\mathrm{P}_n\left[u_1\right] \geq \mathrm{P}_n\left[u_2\right] \tag{2.21}$$

が成り立つ。

証明　帰納法で示す。

i) $k = 2$ の場合。2 行 2 列の分割表について、$s_1 \geq s_2$ としても一般性を失わない。

(a) $u_1 > u_2 > s_2$ の時は、$\mathrm{P}_2\left[u_1\right] = s_2 + 1$ かつ $\mathrm{P}_2\left[u_2\right] = s_2 + 1$ より $\mathrm{P}_2\left[u_1\right] = \mathrm{P}_2\left[u_2\right]$ である。

(b) $u_1 > s_2 \geq u_2$ の時は、$\mathrm{P}_2\left[u_1\right] = s_2 + 1$ かつ $\mathrm{P}_2\left[u_2\right] = u_2 + 1$ より $\mathrm{P}_2\left[u_1\right] \geq \mathrm{P}_2\left[u_2\right]$ である。

(c) $s_2 \geq u_1 > u_2$ の時は、$\mathrm{P}_2\left[u_1\right] = u_1 + 1$ かつ $\mathrm{P}_2\left[u_2\right] = u_2 + 1$ より $\mathrm{P}_2\left[u_1\right] > \mathrm{P}_2\left[u_2\right]$ である。

よって、2 行 2 列の時成り立つ。

ii) 2 行 k 列で仮定が成り立っているとする。$k + 1$ 列目に s_{k+1} を加え、$N_{k+1} = N_k + s_{k+1}$ とする。$u_2 < u_1$ に対して、

(a) $u_1 \leq \lfloor N_k/2 \rfloor$ の時

$$\begin{aligned}
\mathrm{P}_{k+1}\left[u_1\right] &= \mathrm{P}_k\left[u_1\right] + \mathrm{P}_k\left[u_1 - 1\right] + \cdots + \mathrm{P}_k\left[u_1 - s_{k+1}\right], \\
\mathrm{P}_{k+1}\left[u_2\right] &= \mathrm{P}_k\left[u_2\right] + \mathrm{P}_k\left[u_2 - 1\right] + \cdots + \mathrm{P}_k\left[u_2 - s_{k+1}\right],
\end{aligned}$$

である。仮定より、

$$\mathrm{P}_k\left[u_1 - \xi\right] \geq \mathrm{P}_k\left[u_2 - \xi\right], \qquad \forall \xi \in [0, s_{k+1}] \cap \mathbb{Z},$$

が成り立ち、$\mathrm{P}_{k+1}\left[u_1\right] \geq \mathrm{P}_{k+1}\left[u_2\right]$ が得られる。

(b) $\lfloor N_k/2 \rfloor < u_1 \leq \lfloor N_{k+1}/2 \rfloor$ の時

$$\begin{aligned}
\mathrm{P}_{k+1}\left[u_1\right] = \mathrm{P}_k\left[u_1\right] + \mathrm{P}_k\left[u_1 - 1\right] + \cdots + \mathrm{P}_k\left[\lfloor N_k/2 \rfloor + 1\right] \\
+ \mathrm{P}_k\left[\lfloor N_k/2 \rfloor\right] + \cdots + \mathrm{P}_k\left[u_1 - s_{k+1}\right]
\end{aligned}$$

である。ここで、$N_k + s_{k+1} = N_{k+1} \geq 2u_1$ より、$N_k - u_1 \geq u_1 - s_{k+1}$ が成り立つ。
$\mathrm{P}_k\left[u\right]$ の定義より、$\mathrm{P}_k\left[u\right] = \mathrm{P}_k\left[N_k - u\right]$ であることに注意すると、

$$\begin{aligned}
\mathrm{P}_k\left[u_1\right] &= \mathrm{P}_k\left[N_k - u_1\right] \\
&\geq \mathrm{P}_k\left[u_1 - s_{k+1}\right]
\end{aligned}$$

である。また、(a) より、$0 \leq \forall u \leq \lfloor N_k/2 \rfloor$ について、

$$\mathrm{P}_{k+1}\left[\lfloor N_k/2 \rfloor\right] \geq \mathrm{P}_{k+1}\left[u\right]$$

がわかるので、$\lfloor N_k/2 \rfloor \leq u_2 < u_1 \leq \lfloor N_{k+1}/2 \rfloor$ についてのみ考えれば良い。$u_1 > u_2$ より、

$$\begin{aligned}
\mathrm{P}_{k+1}\left[u_1\right] - \mathrm{P}_{k+1}\left[u_2\right] &= \left\{\mathrm{P}_k\left[u_1\right] + \cdots + \mathrm{P}_k\left[u_1 - s_{k+1}\right]\right\} - \left\{\mathrm{P}_k\left[u_2\right] + \cdots + \mathrm{P}_k\left[u_2 - s_{k+1}\right]\right\} \\
&\geq (u_1 - u_2)(\mathrm{P}_n\left[k_1 - s_{k+1}\right] - \mathrm{P}_k\left[u_1 - s_{k+1} - 1\right]) \\
&\geq 0
\end{aligned}$$

となる。よって $k + 1$ 列にも成り立つ。

iii)　i)、ii) より題意は帰納的に示された。□

以上の補題 1 を使って、定理 1 の証明を行う。

証明　列和 $\boldsymbol{s}$、行和 $\boldsymbol{r}$、表値の合計 $N' = N + s_n$ が与えられた 2 行 n 列　$(n \geq 3)$ の分割表を考える。すべての $X \in \Sigma_{\boldsymbol{r},\boldsymbol{s}}$ の内、$X_{2\,n} = t$ となる X の個数を $\mathrm{Q}\,[t]$　$(t \in \mathbb{Z})$ で表すとする。但し、$t > \min\{s_n, r_2\}$ については $\mathrm{Q}\,[t] = 0$ と定義しておく。この時、

$$\mathrm{Q}\,[t] = \mathrm{P}_{n-1}\,[r_2 - t] \qquad (t \geq \max\{0, s_n - r_1\}) \tag{2.22}$$

が成り立つ。

i）$r_2 < \lfloor s_n/2 \rfloor$ の時
$X_{2n} < \lfloor s_n/2 \rfloor$ より $|\Omega_2| = 0$ なので仮定は成り立つ。

ii）$r_2 \geq \lfloor s_n/2 \rfloor$ かつ　$r_1 < s_n$ の時

$$\begin{aligned}
|\Omega_1| &= \mathrm{Q}\,[s_n - r_1] + \mathrm{Q}\,[s_n - r_1 + 1] + \cdots + \mathrm{Q}\,\left[\left\lceil \tfrac{s_n-1}{2} \right\rceil\right] \\
&= \mathrm{P}_{n-1}\,[r_2 - s_n + r_1] + \mathrm{P}_{n-1}\,[r_2 - s_n + r_1 - 1] + \cdots + \mathrm{P}_{n-1}\left[r_2 - \left\lceil \tfrac{s_n-1}{2} \right\rceil\right], \\
|\Omega_2| &= \mathrm{Q}\left[\left\lfloor \tfrac{s_n+1}{2} \right\rfloor\right] + \cdots + \mathrm{Q}\,[r_2] \\
&= \mathrm{P}_{n-1}\left[r_2 - \left\lfloor \tfrac{s_n+1}{2} \right\rfloor\right] + \cdots + \mathrm{P}_{n-1}\,[0]
\end{aligned}$$

である。ここで、$\mathrm{P}_n\,[u] = \mathrm{P}_n\,[N - u]$ より、

$$\begin{aligned}
|\Omega_1| &= \mathrm{P}_{n-1}\,[N - (r_2 - s_n + r_1)] + \mathrm{P}_{n-1}\,[N - (r_2 - s_n + r_1 - 1)] \\
&\quad + \cdots + \mathrm{P}_{n-1}\left[N - (r_2 - \left\lceil \tfrac{s_n-1}{2} \right\rceil)\right] \\
&= \mathrm{P}_{n-1}\,[0] + \mathrm{P}_{n-1}\,[1] + \cdots + \mathrm{P}_{n-1}\left[r_2 - \left\lfloor \tfrac{s_n+1}{2} \right\rfloor\right] \\
&\quad + \cdots + \mathrm{P}_{n-1}\left[r_1 - \left\lfloor \tfrac{s_n+1}{2} \right\rfloor\right] \\
&\geq |\Omega_2|
\end{aligned}$$

となる。

iii）$r_2 \geq \lfloor s_n/2 \rfloor$ かつ $r_1 \geq s_n$ の時

$$\begin{aligned}
|\Omega_1| &= \mathrm{Q}\,[0] + \mathrm{Q}\,[1] + \cdots + \mathrm{Q}\,\left[\left\lceil \tfrac{s_n-1}{2} \right\rceil\right] \\
&= \mathrm{P}_{n-1}\,[r_2] + \mathrm{P}_{n-1}\,[r_2 - 1] + \cdots + \mathrm{P}_{n-1}\left[r_2 - \left\lceil \tfrac{s_n-1}{2} \right\rceil\right], \\
|\Omega_2| &= \mathrm{Q}\left[\left\lfloor \tfrac{s_n+1}{2} \right\rfloor\right] + \cdots + \mathrm{Q}\,[s_n] \\
&= \mathrm{P}_{n-1}\left[r_2 - \left\lfloor \tfrac{s_n+1}{2} \right\rfloor\right] + \cdots + \mathrm{P}_{n-1}\,[r_2 - s_n]
\end{aligned}$$

である。したがって、

$$\begin{aligned}
\left|\left\{\mathrm{Q}\,[0], \ldots, \mathrm{Q}\,\left[\left\lceil \tfrac{s_n-1}{2} \right\rceil\right]\right\}\right| &= \left\lceil \tfrac{s_n-1}{2} \right\rceil - 0 + 1 \\
&= 1 + \left\lceil \tfrac{s_n-1}{2} \right\rceil, \\
\left|\left\{\mathrm{Q}\left[\left\lfloor \tfrac{s_n+1}{2} \right\rfloor\right], \ldots, \mathrm{Q}\,[s_n]\right\}\right| &= s_n - \left\lfloor \tfrac{s_n+1}{2} \right\rfloor + 1 \\
&= 1 + \left\lceil \tfrac{s_n-1}{2} \right\rceil
\end{aligned}$$

であることがわかる。ここで、

$$\begin{cases}
\theta_k &= \left\lfloor \left| \frac{N}{2} - r_2 \right| \right\rfloor, \\
\theta_{k-1} &= \left\lfloor \left| \frac{N}{2} - r_2 - 1 \right| \right\rfloor, \\
\vdots & \\
\theta_1 &= \left\lfloor \left| \frac{N}{2} - r_2 - \left\lceil \frac{s_n-1}{2} \right\rceil \right| \right\rfloor,
\end{cases}$$

$$\left\{\begin{array}{lcl}\psi_1 &=& \left\lfloor \left| \frac{N}{2} - r_2 - \left\lfloor \frac{s_n+1}{2} \right\rfloor \right| \right\rfloor, \\ \psi_2 &=& \left\lfloor \left| \frac{N}{2} - r_2 - \left\lfloor \frac{s_n+1}{2} \right\rfloor - 1 \right| \right\rfloor, \\ \vdots & & \\ \psi_k &=& \left\lfloor \left| \frac{N}{2} - r_2 - s_n \right| \right\rfloor \end{array}\right.$$

を定義すると、

$$\begin{aligned} |\Omega_1| &= \mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \theta_k\right] + \mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \theta_{k-1}\right] + \cdots + \mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \theta_1\right], \\ |\Omega_2| &= \mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \psi_1\right] + \mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \psi_2\right] + \cdots + \mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \psi_k\right] \end{aligned}$$

で表される。また、

$$r_2 \leq \left\lfloor \frac{N + s_n}{2} \right\rfloor$$

より、

$$r_2 - \left\lceil \frac{s_n - 1}{2} \right\rceil \leq \left\lfloor \frac{N}{2} \right\rfloor$$

が導かれるので、$\forall v\ (1 \leq v \leq k)$ に対して、

$$\theta_v \leq \psi_v$$

が成り立つ。このことと補題 1 から

$$\mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \theta_v\right] \geq \mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \psi_v\right]$$

が言え、

$$|\Omega_1| - |\Omega_2| = \sum_{v=1}^{k} \left(\mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \theta_v\right] - \mathrm{P}_{n-1}\left[\left\lfloor \tfrac{N}{2} \right\rfloor - \psi_v\right] \right) \geq 0$$

が成り立つ。よって題意は示された。□

2.4 推定量の偏りについて

この節では、新しいアルゴリズムで得られる推定量の真の値からの偏りについて議論する。Dyer and Greenhill [2] のアルゴリズムでは、推定量 Z が真の値に比べ、小さくなることが予想されるが、この偏りの大きさを理論的に見積もることは困難である。しかし、今回提案したアルゴリズムでは $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ の推定量である Z は偏りの大きさを理論的に見積もることができる。

命題　今回用いたアルゴリズムで求まる Z の期待値 $\mathrm{Exp}\,[Z]$ は真の値 $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ に対して、

$$0 \leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| - \mathrm{Exp}\,[Z] \leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}|\,(1/2)^{M+1}\,R \tag{2.23}$$

を満たす。

証明　$Z_i = (U_i + 1)/(M + 1)$ を ρ_i の推定量とした時の $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ の推定量 Z の期待値は、各 Z_i が独立であることに注意して、

$$\begin{aligned}\mathrm{Exp}\,[Z] &= \sigma \mathrm{Exp}\left[\prod_i \frac{1}{Z_i}\right] \\ &= \sigma \prod_i \mathrm{Exp}\left[\frac{1}{Z_i}\right]\end{aligned}$$

となる。ここで、

$$\begin{aligned}\mathrm{Exp}\left[\frac{1}{Z_i}\right] &= \sum_{U_i=0}^{M} \frac{M+1}{U_i+1}\binom{M}{U_i}\rho_i^{U_i}(1-\rho_i)^{M-U_i} \\ &= \sum_{U_i=0}^{M} \frac{M+1}{U_i+1}\frac{M!}{U_i!(M-U_i)!}\rho_i^{U_i}(1-\rho_i)^{M-U_i} \\ &= \frac{1}{\rho_i}\sum_{U_i=0}^{M} \frac{(M+1)!}{(U_i+1)!(M-U_i)!}\rho_i^{U_i+1}(1-\rho_i)^{M-U_i} \\ &= \frac{1}{\rho_i}\sum_{U_i=0}^{M} \binom{M+1}{U_i+1}\rho_i^{U_i+1}(1-\rho_i)^{M-U_i} \\ &= \frac{1}{\rho_i}\left\{\sum_{U_i'=1}^{M+1} \binom{M+1}{U_i'}\rho_i^{U_i'}(1-\rho_i)^{M+1-U_i'} - (1-\rho_i)^{M+1}\right\} \\ &= \frac{1}{\rho_i}\left\{1-(1-\rho_i)^{M+1}\right\}\end{aligned}$$

であるから、

$$\mathrm{Exp}\,[Z] = \sigma \prod_i \frac{1}{\rho_i}\left\{1-(1-\rho_i)^{M+1}\right\} \tag{2.24}$$

である。まず、

$$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| - \mathrm{Exp}\,[Z] \geq 0$$

は明らかである。次に、

$$\begin{aligned}|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| - \mathrm{Exp}\,[Z] &= \sigma\prod_i \frac{1}{\rho_i} - \sigma\prod_i \frac{1}{\rho_i}\left\{1-(1-\rho_i)^{M+1}\right\} \\ &= \sigma\prod_i \frac{1}{\rho_i}\left[1-\prod_i\left\{1-(1-\rho_i)^{M+1}\right\}\right] \\ &\leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}|\left[1-\prod_i\left\{1-(1/2)^{M+1}\right\}\right] \\ &= |\Sigma_{\boldsymbol{r},\boldsymbol{s}}|\left[1-\left\{1-(1/2)^{M+1}\right\}^R\right] \\ &= |\Sigma_{\boldsymbol{r},\boldsymbol{s}}|\left[1-\left\{1-(1/2)^{M+1}\right\}\right] \\ &\qquad \left[1+\left\{1-(1/2)^{M+1}\right\}+\cdots+\left\{1-(1/2)^{M+1}\right\}^{R-1}\right] \\ &\leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}|\,(1/2)^{M+1}\,R\end{aligned}$$

が成り立つ。したがって、

$$0 \leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| - \mathrm{Exp}\left[Z\right] \leq |\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \left(1/2\right)^{M+1} R$$

となり、題意は示される。□

この結果から、推定量 Z の期待値 $\mathrm{Exp}\left[Z\right]$ は真の値 $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ に比べて小さくなり、偏りの大きさは $|\Sigma_{\boldsymbol{r},\boldsymbol{s}}| \left(1/2\right)^{M+1} R$ で押さえられることが理論的に示された。

第 3 章

計算機実験

この章では、前述の理論的な数え上げアルゴリズムに対して、推定量の偏り、精度、および計算時間の観点から計算機実験を行った結果と考察を述べる。この実験から計算に必要なパラメータの実現可能な値を考察する。

3.1 計算機実験について

2 行 n 列の分割表について実験を行った。MCMC 法で得られた近似解について、真の値との誤差を調べる。

今回の実験の目的は、

- 従来のアルゴリズムと改良したアルゴリズムの偏りの比較
- マルコフ連鎖の推移回数と精度の関係
- サンプル数 M と精度の関係
- 表値の合計 N と精度の関係
- 列数 n と精度の関係

について調べることである。

実験において変動するパラメータは

- 列数 n
- 表値の合計 N
- 行和ベクトル $\boldsymbol{r}$
- 列和ベクトル $\boldsymbol{s}$
- サンプリングの個数 M
- モンテカルロ法の反復の回数 R
- マルコフ連鎖の推移回数 T
- マルコフ連鎖の初期状態

である。この内、行和ベクトル $\boldsymbol{r}$ と列和ベクトル $\boldsymbol{s}$ に関しては列数 n、表値の合計 N を与え、ランダムに発生させる。具体的には、$\boldsymbol{r}, \boldsymbol{s}$ はそれぞれ自然数 N をランダムに 2 分割、n 分割して生成した。また、R は $\boldsymbol{r}, \boldsymbol{s}$ に従属したパラメータであり、直接的に作為を及ぼすことはできない。

このアルゴリズムで解を求めるにはマルコフ連鎖を T 回推移して得られるサンプルを M 個抽出し、これを R 回反復しなければならない。すなわち計算量

$$RMT$$

が必要であるが、理論的精度を得るには正の整数 ϵ, δ に対して、

$$\begin{aligned} M &= \left\lceil 150\mathrm{e}^2 R^2 \epsilon^{-2} \ln(3R\delta^{-1}) \right\rceil \\ T &= \left\lceil \frac{n(n-1)}{2} \ln(N\epsilon^{-1}) \right\rceil \end{aligned}$$

となる。$n = 8, N = 1000$ の時、$R \simeq 40$ であり、$\epsilon = 0.01, \delta = 0.1$ に対して、$M \simeq 1.2 \times 10^{11},\ T \simeq 600$ であり、計算量 $RMT \simeq 2.9 \times 10^{15}$ となってしまう。これは実用時間内で実現不可能である。ここで、理論的精度が保証するのは、最悪ケースであることを鑑み、予備実験の結果 $n = 8, N = 1000$ 程度の時、$M \simeq 100000,\ T \simeq 200$ でも 10^{-2} 程度の精度が得られることが経験的にわかったので、$M,\ T$ は実験者があらかじめ指定することにした。

マルコフ連鎖の初期状態については、理論的には任意の初期状態からはじめても T 回の推移で近似的に一様サンプリングが得られる。しかし、初期状態をうまく取ればより少ない推移で一様サンプリングが得られると予想される。このことに関する実験および考察は次節の「マルコフ連鎖の推移回数に関する実験」の小節で行うが、一般的には初期値変動型を用いた。初期値変動型とは、各ステップにおいて、1 個目のサンプリングを行う時のみ表値を実験者が与え、2 個目以降のサンプリング時のマルコフ連鎖の初期値はその前の回のサンプリングで得られた表値を使う方法である。

また、真の値との比較を行っているが、実際には真の値として 10^{-5} 程度の丸め誤差を含んだ値を使っている。しかし、MCMC 法で求まる値が今回の実験ではせいぜい 10^{-3} 精度であることから、これを用いることとした。

以下の実験において特に断りのない場合、$T = 200,\ N = 1000,\ M = 10^5$ とする。また、実験結果は定めたパラメータ T, M, N, n に対して、各試行ごとに、$\boldsymbol{r}, \boldsymbol{s}$ をランダムに定め、試行を 10 回繰り返して得られた値の平均である。

実験には、CPU：Celeron 733MHz(Intel 社)、メモリ：128MB の PC を用いた。

3.2 計算結果と考察

3.2.1 推定量の偏りに関する実験

まず、Dyer and Greenhill [2] のアルゴリズムを用いたプログラムと今回提案したアルゴリズムについて誤差の絶対値の平均と誤差の算術平均に関して比較した。この実験に関しては $n = 6$ であり、列和の決め方も $s_k = \mathrm{RAND}[19] (1 \leq k \leq n)$ とした。$\mathrm{RAND}[m]$ とは、一様分散にしたがって 1 から m までの自然数を返す関数である。従って N は決まった値を取らず、$N \simeq 60$ である。また、$T = 200$ とした。

この結果を表 3.1 に示す。誤差の絶対値の平均に関しては、両アルゴリズムの値は同程度であるが、Dyer のアルゴリズムででは誤差の算術平均が大きく負に偏っているのに比べて、改良したアルゴリズムではほとんど偏りが見られなかった。

表 3.1　Dyer のアルゴリズムと改良したアルゴリズムの偏りの比較

		$M = 10^3$	$M = 10^4$	$M = 10^5$	$M = 10^6$
Dyer	ϵ	-0.080826	-0.021568	-0.007275	-0.001221
	$\|\epsilon\|$	0.105371	0.022995	0.009105	0.001964
改良	ϵ	0.024739	-0.003787	-0.000864	-0.000368
	$\|\epsilon\|$	0.072223	0.021146	0.006729	0.002719

Dyer のアルゴリズムで作ったプログラムでは、M が十分大きくない時、実際に $\rho_i = |\Omega_J|/|\Omega_0| < 1/2$ となる J が選ばれる「逆転現象」が起きており、この時の誤差は大きく負に偏っていた。また、r_1 と r_2 の差があまり大きくない時、各ステップで理論比 $\rho_i = 1/2$ となることが多く、この時も誤差は負に偏ることが多かった。

3.2.2　マルコフ連鎖の推移回数に関する実験

この小節はマルコフ連鎖の推移回数と誤差についての実験である。ここで用いたプログラムは 2 種類に分けられる。

初期値固定型　：　各サンプリングにおけるマルコフ連鎖の初期状態を固定したもの。

初期値変動型　：　各サンプリングにおけるマルコフ連鎖の初期状態は 1 個目のサンプリング時のみ与え、2 個目以降のサンプリングにおけるマルコフ連鎖の初期状態を、その前に得られたサンプルに設定したもの。

初期状態の与え方は以下のとおりである：

一般性を失うことなく $r_1 \geq r_2$ とする。

i) X_{11} から順に $X_{1\,k} = s_k,\ (1 \leq k \leq n)$ を与えていく。

ii) $\sum_k X_{1\,k} = r_1$ となったら、今度は列和を超えないように $X_{2\,k}$ に数値を与えていく。□

初期状態を与えられた 2 種類のプログラムそれぞれに対して、マルコフ連鎖の推移回数 T を変化させ、誤差の変化を調べた。

実験は $n = 8,\ n = 16$ について行い、サンプル数も $M = 10^4,\ M = 10^5$ の 2 とおりについて行った。

結果を表 3.2 、表 3.3 に示す。

この実験で得られた結果について考察する。まず、M を固定して見ると、T を大きくしても各 M の値によって精度に限度が見られる。例えば、$n = 8,\ M = 10^4$ の時、固定型にしても変動型にしても精度は 5% であり、$n = 16,\ M = 10^5$ の時、精度は 2% で T を大きくしても精度は良くならない。

次に T を固定してみると、固定型の場合、T の値が十分大きくない時、M の値を大きくしても、あまり精度は変わらない。それに対して変動型を用いると、T の値が同じでも M を大きくすると精度が上がる。

理論的混交時間 T は、$n = 8$ の時 $\epsilon = 0.05$ に対して $R = 45$ として $T = 514$、$n = 16$ の時 $\epsilon = 0.1$ に対して $R = 90$ として $T = 1657$、となる。実験の結果から、混交時間を推定すると、$n = 8$ の場合、平均ステップ数 $\hat{R}$ は 38.8 回であるが、$\epsilon = 0.05$ に対する混交時間 $T = 50$ で精度が安定する。$n = 16$ についても平均ステップ数 $\hat{R}$ は 87.8 回で、$\epsilon = 0.1$ に対する推定混交時間は $T = 200$ である。

表 3.2　マルコフ連鎖の初期状態と推移回数による誤差変化（列数 8）。

$n = 8$

	固定型		変動型	
	$M = 10^4$	$M = 10^5$	$M = 10^4$	$M = 10^5$
$T = 1$	–	–	0.091462	0.044455
$T = 2$	–	–	0.049411	0.022905
$T = 3$	–	–	0.052106	0.024237
$T = 5$	–	–	0.043893	0.010891
$T = 10$	714.955036	97.319204	0.047183	0.010905
$T = 20$	0.736416	1.312236	0.052326	0.014693
$T = 30$	0.137628	0.139983	0.064169	0.010777
$T = 50$	0.041060	0.058587	0.044688	0.013775
$T = 70$	0.050492	0.026839	0.056945	0.009063
$T = 100$	0.047324	0.011368	0.056145	0.017895
$T = 200$	0.040457	0.018711	0.054939	0.013230
$T = 400$	0.034374	0.013048	0.049772	0.013131
$T = 800$	0.068829	0.009813	0.041833	0.017043

表 3.3　マルコフ連鎖の初期状態と推移回数による誤差変化（列数 16）。

$n = 16$

	固定型		変動型	
	$M = 10^4$	$M = 10^5$	$M = 10^4$	$M = 10^5$
$T = 1$	–	–	0.323820	0.046637
$T = 2$	–	–	0.139531	0.043179
$T = 3$	–	–	0.115960	0.037651
$T = 5$	–	–	0.113963	0.016320
$T = 10$	–	–	0.075256	0.029963
$T = 20$	10619.7	9826.1	0.087001	0.024318
$T = 30$	11.278630	71.912392	0.064620	0.016104
$T = 50$	1.122720	0.482482	0.079872	0.021812
$T = 70$	0.181155	0.337703	0.067719	0.018466
$T = 100$	0.287543	0.137110	0.089454	0.016573
$T = 200$	0.097839	0.017608	0.071761	0.018947
$T = 400$	0.057786	0.019117	0.072482	0.017024
$T = 800$	0.063889	0.020238	0.091764	0.015658

3.2.3　サンプル数 M と精度に関する実験

上の実験より、$n = 8, 16$ の時、$T = 200$、初期値変動型で十分に 2 桁精度が求まると推測される。この小節では $n = 8, 16$ に対して、M を変化させて精度の変化を調べた。結果は表 3.4 のとおりである。

表 3.4　サンプリング個数 M と誤差の関係

	$M = 10^3$	$M = 10^4$	$M = 10^5$	$M = 10^6$
$n = 8$	0.130641	0.054939	0.013230	0.004218
$n = 16$	0.203034	0.071761	0.018947	0.004853

実験の結果から、ϵ^{-1} が $\sqrt{M}$ に比例することが読み取れる。

3.2.4　表値の合計 N と誤差の関係。

$n = 8,\ M = 10^5$ を固定し、表値の合計 N を変化させ N と誤差の変化の関係を調べた (表 3.5)。

表 3.5　表値の合計 N と誤差、反復回数 R、計算時間の関係

合計値 N	平均誤差 $\lvert\epsilon\rvert$	平均反復回数 R	平均計算時間 [秒]
$N = 10$	0.004895	7.4	147.3
$N = 100$	0.009944	23.5	471.3
$N = 1000$	0.013230	43.8	885.0
$N = 10000$	0.025660	63.4	1283.8

この結果より、ϵ^{-1} は R に比例していると言える。

3.2.5　列数 n と精度、および計算時間に関する実験

$N = 1000,\ M = 10^5$ を固定して、$n = 8, 16, 32, 48$ について精度を調べた。結果を表 3.6 に記す。

表 3.6　列数 n と誤差、反復回数 R、計算時間の関係

列数 n	平均誤差 $\lvert\epsilon\rvert$	平均反復回数 R	平均計算時間 [秒]
$n = 8$	0.013230	43.8	885.0
$n = 16$	0.018947	87.8	1731.1
$n = 32$	0.034496	156.4	3018.2
$n = 48$	0.037718	213.4	4090.3

ϵ^{-1} は R に比例している様子がうかがえる。また、n に比べ N がある程度大きいので、R は n に比例する様子が見られる。

3.2.6 計算時間

以上の実験において、計算時間は $R,\ M,\ T$ にほぼ完全に比例した。$R = 45,\ M = 10^5,\ T = 200$ の時計算にかかった時間は 913.8[s] であった。

第 4 章

結論

行和、列和の与えられた 2 行 n 列の分割表の個数を MCMC 法を用いて近似的に求めるアルゴリズムを扱った。既存のアルゴリズムに比べ、真の値からの偏りが小さくなるような新しいアルゴリズムを提案した。このアルゴリズムを用いると真の値に対する推定量の偏りが理論的に保証されることを示した。また、計算機実験を行い、新しいアルゴリズムで得られる推定量の偏りが従来のアルゴリズムに比べ小さくなることを確認し、アルゴリズム中の各パラメータと精度の関係を考察した。

4.1 結論

Dyer and Greenhill [2] のアルゴリズムに比べ、偏りが小さくなるようなアルゴリズムを作ることができた。このアルゴリズムによって得られる推定量の期待値は、反復回数 R、モンテカルロ法のサンプリング量 M に対して、$(1/2)^{M+1}R|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ ほど真の値から小さくなることを示すことができた。但し、$|\Sigma_{\boldsymbol{r},\boldsymbol{s}}|$ は条件を満たす分割表全体の個数を表す。

また、計算機実験の結果から、マルコフ連鎖の理論的推移回数は多すぎると推測される。さらに、初期値の取り方を工夫すれば、目標とする精度の推定量を得るために必要なマルコフ連鎖の推移回数より少ない回数で目標とする精度を達成することができることがわかった。モンテカルロ法については、今回の実験程度のサイズの分割表については、サンプリング量 M は誤差 ϵ の -2 乗に比例し、ϵ は反復回数 R に比例する様子が観察された。しかし、理論の与えるサンプリング量 M は目標とする精度 ϵ に対して大きすぎることがわかった。今回の理論的精度保証は最悪の場合を想定したものであり、一般的な場合にはオーバーサンプリングであると言える。

4.2 課題

今回の計算機実験の結果、精度の保証された推定量を得るために必要な計算量はもっと少ないことが予想される。必要計算量の上限を小さくするような精度保証の方法について考える余地がある。

また、今回扱った分割表はは 2 行のものに限定している。現在、3 行以上の分割表の個数数え上げ問題については、理論的に精度の保証された推定量を多項式時間で求めるアルゴリズムが見つかっていない。今後、3 行以上の分割表に対するモンテカルロ法のアルゴリズム、および、一様分布を定常分布に持つようなマルコフ連鎖を構築し、理論的に精度を保証する必要がある。

付録 A

マルコフ連鎖の既約性と非周期性

ここでは、2.1.3 で述べたマルコフ連鎖 $\mathcal{M}$ が既約で非周期であることを証明する。

- 既約性について

  帰納法で示す。$\forall X, \forall Y \in \mathcal{M}$ について、

  i) X と Y の表値が 2 列でのみ異なり、他の $n-2$ 列の表値が一致している時
  異なる列が 1 列目と 2 列目としても一般性を失わない。この時、

  $$\sum_{j=1,2} X_{ij} = \sum_{j=1,2} Y_{ij} = b_i \,, \qquad i = 1, 2 \,,$$
  $$\sum_{i=1,2} X_{ij} = \sum_{i=1,2} Y_{ij} = c_j \,, \qquad j = 1, 2$$

  より、$\mathcal{M}$ の定義から X と Y は相互到達可能である。

  ii) k 列以下の列において表値が異なる $\forall X$ と $\forall Z$ が相互到達可能の時
  X と $k+1$ 列異なる Y について考える。一般性を失うことなく、$1 \sim k+1$ 列が異なるとする。l を

  $$|Y_{1l} - X_{1l}| = \min\{\, |Y_{11} - X_{11}|, |Y_{12} - X_{12}|, \ldots, |Y_{1k+1} - X_{1k+1}| \,\}$$

  でひとつ決める。l に対して

  $$(X_{1l} - Y_{1l})(X_{1l'} - Y_{1l'}) < 0$$

  となる l' $(l' \neq l, 1 \leq l' \leq k+1)$ を選ぶと、

  $$Z_{ij} = \begin{cases} X_{ij} \,, & j = l \,, \\ Y_{ij} - (X_{il} - Y_{il}) \,, & j = l', \\ Y_{ij} \,, & \text{otherwise} \end{cases}$$

  なる Z が存在して Y と Z は相互到達可能である。この時、X と Z は k または $k-1$ 列間においてのみ異なる分割表となるので、仮定より、X と Z は相互到達可能である。従って、$k+1$ 列間においてのみ異なる $\forall X$ と $\forall Y$ についても相互到達可能である。

  iii) i)、ii) より帰納的に $\forall X, \forall Y \in \mathcal{M}$ について相互到達可能。
  よって題意は示された。

- 非周期性について

  マルコフ連鎖 $\mathcal{M}$ の定義から、$\mathcal{M}$ における任意の状態 X は周期 1 の自己ループを持つ。従って非周期的である。

付録 B

理論的精度保証の証明

2.2 で与えられた理論的精度保証を証明する 。以下の証明で $\hat{\rho}_i = \mathrm{Exp}\,[Z_i]$ とする。

i) for $1 \leq i \leq R, |\rho_i - \hat{\rho}_i| \leq \epsilon/(15Re^2)$

$T = \tau(\epsilon/(15Re^2))$ としているので、$|\rho_i - \hat{\rho}_i| \leq \epsilon/(15Re^2)$ は明らか。

ii) for $1 \leq i \leq R, \mathrm{Prob}\left[\hat{\rho}_i \geq 1/(2e^2)\right] \leq 1 - \delta/(3R)$

$\mathcal{U}_J$ として、$M/2$ 回以上出た方を選んでいるので、$\mathrm{Prob}\left[\hat{\rho}_i \geq 1/(2e^2)\right]$ が $M/2$ 回以上出る確率を考える。

すなわち表の出る確率が $1/(2e^2)$ 未満のコインを M 回以上表の出る確率の上界は

$$\begin{aligned}
&\sum_{k=\frac{M}{2}}^{M} \binom{M}{k} \left(\frac{1}{2e^2}\right)^k \left(1 - \frac{1}{2e^2}\right)^{M-k} \leq \sum_{k=\frac{M}{2}}^{M} \binom{M}{k} \left(\frac{1}{2e^2}\right)^{\frac{M}{2}} \\
&= \left(\frac{1}{2e^2}\right)^{\frac{M}{2}} \frac{2^M}{2} \leq \left(\frac{4}{2e^2}\right)^{\frac{M}{2}} = \left(\frac{2}{e^2}\right)^{\frac{M}{2}} \leq \left(\frac{2}{2e^2}\right)^{\frac{M}{2}} \\
&= \left(\frac{1}{e}\right)^{\frac{M}{2}} = \left(\frac{1}{e}\right)^{\frac{150}{2} e^2 R^2 \epsilon^{-2} \ln(\frac{3R}{\delta})} \\
&\leq \left(\frac{1}{e}\right)^{\ln(\frac{3R}{\delta})} = \frac{\delta}{3R}
\end{aligned}$$

ゆえに $\mathrm{Prob}\left[\hat{\rho}_i \geq 1/(2e^2)\right] \leq 1 - \delta/(3R)$

iii) $\hat{\rho}_i \geq (1/2e^2)$ ならば $|\rho_i - \hat{\rho}_i| \leq (\epsilon/5R)\hat{\rho}_i$

i) より

$$\begin{aligned}
|\rho_i - \hat{\rho}_i| &\leq \frac{\epsilon}{15Re^2} = \left(\frac{\epsilon}{5R}\right)\left(\frac{1}{3e^2}\right) \\
&\leq \frac{\epsilon}{5R}\left(\frac{1}{2e^2}\right) \leq \frac{\epsilon}{5R}\hat{\rho}_i
\end{aligned}$$

iv) $\rho_i \geq \hat{1}/(2e^2)$ ならば $\mathrm{Prob}\left[|Z_i - \hat{\rho}_i| > (\epsilon/(5R))\hat{\rho}_i\right] \leq 2\delta/3R$

公式

$$\mathrm{Prob}\left[\left|\frac{\hat{m}}{m} - p\right| \geq \lambda p\right] \leq 2\, e^{\frac{-\lambda^2 mp}{3}}$$

が知られているので、これに

$$\frac{\hat{m}}{m} = Z_i, m = M, \lambda = \frac{\epsilon}{5R}, p = \hat{\rho}_i$$

を代入すると、

$$\begin{aligned}
&\mathrm{Prob}\left[|Z_i - \hat{\rho}_i| > \left(\frac{\epsilon}{5R}\right)\hat{\rho}_i\right] \leq 2\ \mathrm{e}^{-\left(\frac{\epsilon}{5R}\right)^2 \frac{1}{3}150e^2R^2\epsilon^{-2}\ln\frac{3R}{\delta}\hat{\rho}_i} \\
&= 2\ \mathrm{e}^{-\frac{\epsilon^2}{25R^2}\frac{150}{3}\mathrm{e}^2R^2\epsilon^{-2}\ln\frac{3R}{\delta}\hat{\rho}_i} \\
&= 2\ \mathrm{e}^{-2\mathrm{e}^2\hat{\rho}_i\ln\frac{3R}{\delta}} \\
&\leq 2\ \mathrm{e}^{-\ln\frac{3R}{\delta}} = \frac{2\delta}{3R}
\end{aligned}$$

v) $1-\delta$ の確率で、$|(Z_1\cdots Z_R)^{-1} - (\rho_1\cdots\rho_R)^{-1}| \leq \epsilon(\rho_1\cdots\rho_R)^{-1}$

iii) より

$$\begin{aligned}
|\rho_i - \hat{\rho}_i| &\leq \frac{\epsilon}{5R}\hat{\rho}_i \\
-\frac{\epsilon}{5R}\ \hat{\rho}_i \leq \rho_i - \hat{\rho}_i &\leq \frac{\epsilon}{5R}\ \hat{\rho}_i \\
\left(1-\frac{\epsilon}{5R}\right)\hat{\rho}_i \leq \rho_i &\leq \left(1+\frac{\epsilon}{5R}\right)\hat{\rho}_i
\end{aligned}$$

また、iv) より、$1-\delta/(3R)$ の確率で、

$$\begin{aligned}
|Z_i - \hat{\rho}_i| &\leq \frac{\epsilon}{5R}\hat{\rho}_i \\
-\frac{\epsilon}{5R}\hat{\rho}_i \leq Z_i - \hat{\rho}_i &\leq \frac{\epsilon}{5R}\hat{\rho}_i \\
\left(1-\frac{\epsilon}{5R}\right)\hat{\rho}_i \leq Z_i &\leq \left(1+\frac{\epsilon}{5R}\right)\hat{\rho}_i \\
\left(1+\frac{\epsilon}{5R}\right)^{-1} Z_i \leq \hat{\rho}_i &\leq \left(1-\frac{\epsilon}{5R}\right)^{-1} Z_i
\end{aligned}$$

これらより、

$$\begin{aligned}
\left(1-\frac{\epsilon}{5R}\right)\left(1+\frac{\epsilon}{5R}\right)^{-1} Z_i \leq \rho_i &\leq \left(1+\frac{\epsilon}{5R}\right)\left(1-\frac{\epsilon}{5R}\right)^{-1} Z_i \\
\left(1-\frac{\epsilon}{5R}\right)\left(1+\frac{\epsilon}{5R}\right)^{-1} \leq \frac{\rho_i}{Z_i} &\leq \left(1+\frac{\epsilon}{5R}\right)\left(1-\frac{\epsilon}{5R}\right)^{-1} \\
\left(1-\frac{\epsilon}{5R}\right)^{R}\left(1+\frac{\epsilon}{5R}\right)^{-R} \leq \frac{\rho_1\cdots\rho_R}{Z_1\cdots Z_R} &\leq \left(1+\frac{\epsilon}{5R}\right)^{R}\left(1-\frac{\epsilon}{5R}\right)^{-R}
\end{aligned}$$

$$\text{(右辺)} = \left(1+\frac{\epsilon}{5R}\right)^{R}\left(1-\frac{\epsilon}{5R}\right)^{-R} = \left(\frac{5R+1}{5R-\epsilon}\right)^{R} = \left(1+\frac{2\epsilon}{5R-\epsilon}\right)^{R}$$

$$
\begin{aligned}
&\leq \left(1+\frac{2\epsilon}{5R}\right)^{R} = \left\{\left(1+\frac{2\epsilon}{5R-1}\right)^{\frac{5R-1}{2\epsilon}}\right\}^{\frac{2\epsilon}{5R-1}R} \\
&\leq \mathrm{e}^{\frac{2R}{5R-1}\epsilon} \\
&= 1+\frac{2R}{5R-1}\epsilon+\frac{1}{2!}\left(\frac{2R}{5R-1}\epsilon\right)^{2}+\frac{1}{3!}\left(\frac{2R}{5R-1}\epsilon\right)^{3}+\cdots \\
&= 1+\frac{2R}{5R-1}\epsilon\left\{1+\frac{1}{2!}\left(\frac{2R}{5R-1}\epsilon\right)+\frac{1}{3!}\left(\frac{2R}{5R-1}\epsilon\right)^{2}+\cdots\right. \\
&\leq 1+\frac{2R}{5R-1}\epsilon\left\{1+\left(\frac{2R}{5R-1}\epsilon\right)+\left(\frac{2R}{5R-1}\epsilon\right)^{2}+\cdots\right. \\
&= 1+\frac{2R}{5R-1}\,\epsilon\,\frac{1}{1-\frac{2R}{5R-1}} \\
&= 1+\frac{2R}{5R-1}\,\epsilon\,\frac{5R-1}{3R-1} = 1+\frac{2R}{3R-1}\,\epsilon \\
&\leq 1+\epsilon
\end{aligned}
$$

$$
\begin{aligned}
(\text{左辺}) &= \left(1-\frac{\epsilon}{5R}\right)^{R}\left(1+\frac{\epsilon}{5R}\right)^{-R} \\
&= \left\{\left(1-\frac{\epsilon}{5R}\right)^{-R}\left(1+\frac{\epsilon}{5R}\right)^{R}\right\}^{-1} = (\text{右辺})^{-1} \\
&\geq \frac{1}{1+\epsilon} \geq 1-\epsilon
\end{aligned}
$$

以上より、

$$
\begin{gathered}
1-\epsilon \leq \frac{\rho_1\cdots\rho_R}{Z_1\cdots Z_R} \leq 1+\epsilon \\
(1-\epsilon)(\rho_1\cdots\rho_R)^{-1} \leq (Z_1\cdots Z_R) \leq (1+\epsilon)(\rho_1\cdots\rho_R)^{-1} \\
|(Z_1\cdots Z_R)^{-1}-(\rho_1\cdots\rho_R)^{-1}| \leq \epsilon(\rho_1\cdots\rho_R)^{-1}
\end{gathered}
$$

ii) より、$1-\dfrac{\delta}{3R}$ 以上の確率で $\hat{\rho}_i \geq \dfrac{1}{2e^2}$

iv) より、$\hat{\rho}_i \geq \dfrac{1}{2e^2}$ が成り立てば、$1-\dfrac{2\delta}{3R}$ 以上の確率で $|Z_i-\hat{\rho}_i| \leq \dfrac{\epsilon}{5R}\hat{\rho}_i$

$$
\therefore \left(1-\frac{2\delta}{3R}\right)\left(1-\frac{\delta}{3R}\right) \text{ 以上の確率で } |Z_i-\hat{\rho}_i| \leq \frac{\epsilon}{5R}\hat{\rho}_i
$$

$$
\begin{aligned}
\left(1-\frac{2\delta}{3R}\right)\left(1-\frac{\delta}{3R}\right) &= 1-\frac{\delta}{3R}-\frac{2\delta}{3R}+\frac{2\delta^2}{9R^2} \\
&\geq 1-\frac{\delta}{R}
\end{aligned}
$$

つまり、

$$\mathrm{Prob}\left[|Z_i - \hat{\rho}_i| \leq \frac{\epsilon}{5R}\hat{\rho}_i\right] \geq 1 - \frac{\delta}{R}$$

$$\begin{aligned}
\therefore \quad & \mathrm{Prob}\left[1 - \epsilon \leq \frac{\rho_1 \cdots \rho_R}{Z_1 \cdots Z_R} \leq 1 + \epsilon\right] \geq \left(1 - \frac{\delta}{R}\right)^R \\
& = 1 + \binom{R}{1}\left(-\frac{\delta}{R}\right) + \binom{R}{2}\left(-\frac{\delta}{R}\right)^2 + \binom{R}{3}\left(-\frac{\delta}{R}\right)^3 + \cdots \\
& 1 - \delta + \left(-\frac{\delta}{R}\right)^2\left\{\binom{R}{2} - \binom{R}{3}\frac{\delta}{R}\right\} + \left(-\frac{\delta}{R}\right)^4\left\{\binom{R}{4} - \binom{R}{5}\frac{\delta}{R}\right\} + \cdots \\
& = 1 - \delta + \sum_{k=1}^{\lceil\frac{R}{2}\rceil - 1}\left(-\frac{\delta}{R}\right)^{2k}\left\{\binom{R}{2k} - \binom{R}{2k+1}\left(\frac{\delta}{R}\right)\right\} \\
& = 1 - \delta + \sum_{k=1}^{\lceil\frac{R}{2}\rceil - 1}\left(-\frac{\delta}{R}\right)^{2k}\frac{R!}{(R-2k-1)!(2k)!}\left(\frac{1}{R-2k} - \frac{1}{2k+1}\frac{\delta}{R}\right) \\
& \geq 1 - \delta
\end{aligned}$$

ここで、

$$\begin{aligned}
\frac{\sigma}{\rho_1 \cdots \rho_R} &= |\Sigma_{\boldsymbol{r},\boldsymbol{s}}|, \\
\frac{\sigma}{Z_1 \cdots Z_R} &= Z
\end{aligned}$$

とおくと、

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

謝辞

この研究をするにあたって、テーマと方向性を示唆し、研究の進め方、論文の書き方にいたるまで細かく指導していただいた松井知己助教授にお礼申し上げます。また、数理第 2 研の松浦史郎助手はじめ研究室の諸先輩方には TEX の使い方、プログラミングの組み方等について助言をいただき心から感謝します。

参考文献

[1] P. Diaconis and L. Saloff-Coste, *"Random walk on contingency tables with fixed row and column sums,"* Tech. Report, Department of Mathematics, Harvard University, 1995.

[2] M. Dyer and C. Greenhill, *"Polynomial-time counting and sampling of two-rowed contingency tables,"* Theoretical Computer Sciences, 246(2000), pp. 265–278.

[3] D. Hernek, *"Random generation of* $2 \times n$ *contingency tables,"* Random Struct. Algorithms, 13(1998), pp. 71–79.