Choice of the Dirichlet parameter to estimate measures for square contingency tables

中川 智之 [1], 桃﨑 智隆 [2], 長 光司 [2], 富澤 貞男 [1]
1. 東京理科大学 理工学部 情報科学科 *
2. 東京理科大学 理工学研究科 情報科学科 *
Tomoyuki Nakagawa, Tomotaka Momozaki, Koji Cho, Sadao Tomizawa
Department of Information Sciences, Tokyo University of Science

1 導入

本稿は, Momozaki *et al.* [4] の要約を与え, Momozaki *et al.* [4] で与えられた尺度の推定量を対称性の尺度に適用し, 既存手法との比較を行う. 本稿では, 確率変数 (X, Y) が同一カテゴリを持つ 2 元の正方分割表を考える. また, X が i カテゴリ, Y が j カテゴリに入るセル確率を $p_{ij} = \Pr(X = i, Y = j)$ で表現する. 正方分割表では, 主対角成分にデータが集まりやすく独立性が成り立たない場合が多く, 独立性よりも対称性に興味がある場合が多い. Bowker [1] は, 対称モデル $p_{ij} = p_{ji}$ $(\forall i, j)$ を定義し, 検定を与えた. 実際には, 対称モデルは多くの場合に成り立たない. 検定は, “モデルが成り立つかどうか” を確かめることは可能であるが, “モデルからの違いはどの程度なのか” を測ることが出来ない. このような問題に対してはモデルからの隔たりを測る尺度を用いることが重要である. 尺度は, “モデルからの違いはどの程度なのか” を定量的に測ることが可能であり, 2 つの分割表のどちらがモデルから隔たっているかを比べることも可能である. これまで様々なモデルからの隔たりを測る尺度も多くの研究で提案されている ([3, 5, 7, 9, 10, 12] など). 特に, Tomizawa *et al.* [12] と Tahata *et al.* [9] は対称モデルからの隔たりを測る尺度を提案している. 一方で, 尺度は未知のセル確率によって構成されており, 推定が必要であ

* 〒278-8510 千葉県野田市山崎 2641

る．多くの研究では, 標本比率を代入して推定し, デルタ法を用いて区間推定を行なっている．しかしながら, このような方法はサンプルサイズが十分でない場合に大きなバイアスが起こる可能性がある．そこで, [6, 8, 11] などで高次のバイアス補正法が提案されている．この方法は高次のバイアスを補正するが, 推定した尺度の値域の外に値を取ることがあるため, 尺度の性質として相応しくない．

Momozaki *et al.* [4] では, ベイズ推定量を用いて平均二乗誤差 (MSE) を最小にする Dirichlet parameter を選ぶことでこれらの問題点を解決している．本稿では, Momozaki *et al.* [4] の方法の概要を説明し, Tomizawa *et al.* [12] と Tahata *et al.* [9] の対称モデルからの隔たりを測る尺度に適用した結果を与える．また既存の推定方法と比較を行う．

2　ベイズ推定量を用いた尺度の推定

本節では, Momozaki *et al.* [4] の結果を正方分割表に限って紹介する．本質的には, 正方分割表でなく一般の分割表において, 同様の結果が得られている．まず分割表の (i,j) セルに入る度数を n_{ij} とすると, 分割表は $\{n_{ij}\}$ は $(n,\boldsymbol{p})$ の多項分布 $\mathrm{Mult}(n,\boldsymbol{p})$ に従うとみなすことができる．ここで, $n=\sum_{i,j} n_{ij}$, $\boldsymbol{p}=(p_{11},p_{12},\ldots,p_{r^2})^\top$: $r^2\times 1$ は全てのセル確率を並べたベクトル ("$\top$": 転置) とする．多項分布のセル確率のベイズ推定には, 共役事前分布である Dirichlet 分布が用いられることが多い．ここでは, 次の対称な Dirichlet 分布を用いることを考える．

$$\pi(\boldsymbol{p}|\alpha)=\frac{\Gamma(r^2\alpha)}{(\Gamma(\alpha))^{r^2}}\prod_{i,j}p_{ij}^{\alpha-1}$$

このとき, ベイズ推定量は次で与えられる．

$$\hat{p}_{ij}^{(\alpha)}=\frac{n_{ij}+\alpha}{n+r^2\alpha}=\frac{n}{n+r^2\alpha}\frac{n_{ij}}{n}+\frac{r^2\alpha}{n+r^2\alpha}\frac{1}{r^2}$$

ここで, $\alpha=1$ が一様事前分布, $\alpha=1/2$ が Jefferys 事前分布になることがわかっている．また, Fienberg and Holland [2] では, セル確率の推定量の MSE を最小にする α を与えている．しかしながら, これらの α は尺度の推定に関しては最適であるかは分からないため, そのまま plug-in をしても尺度の推定に対するバイアスや MSE を改善できるとは言えない．そこで Momozaki *et al.* [4] ではベイズ推定量を plug-in した尺度の推定量の MSE を最小にする α の決定を行った．

まず尺度はセル確率によって構成されているため, ここでは $f(\boldsymbol{p})$ という関数で表現する．またベイズ推定量を plug-in した尺度の推定量は $f(\hat{\boldsymbol{p}}^{(\alpha)})$ で与えられる．このとき, Momozaki *et al.* [4] では以下の定理を与えている．

定理 2.1 (Momozaki *et al.* [4])**.**

$$\tilde{\alpha} = \arg\min_{\alpha} \lim_{n\to\infty} n^2 \mathrm{MSE}[f(\hat{\boldsymbol{p}}^{(\alpha)})] = \frac{A_1}{A_2}$$

ここで, $\boldsymbol{c} = r^2\boldsymbol{p} - \mathbf{1}$, A_1, A_2 は以下で与える.

$$\begin{aligned}
A_2 =& \mathrm{tr}\left[\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}^\top}\right)\boldsymbol{c}\boldsymbol{c}^\top\right], \\
A_1 =& \frac{1}{2}\boldsymbol{c}^\top\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)\mathrm{tr}\left[\left(\frac{\partial^2 f(\boldsymbol{p})}{\partial \boldsymbol{p}\partial \boldsymbol{p}^\top}\right)(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\right] \\
&+ \mathrm{tr}\left[\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)\boldsymbol{c}^\top\left(\frac{\partial^2 f(\boldsymbol{p})}{\partial \boldsymbol{p}\partial \boldsymbol{p}^\top}\right)(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\right] \\
&+ r^2\ \mathrm{tr}\left[\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}^\top}\right)(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\right]
\end{aligned}$$

この結果より, α が MSE に n^2 で初めて影響が出ていることがわかる. 実際に推定量として用いる場合は, A_1 と A_2 はともに未知パラメータ $\boldsymbol{p}$ に依っているので, 本稿は Momozaki *et al.* [4] と同様に A_1 と A_2 に標本比率 $\hat{\boldsymbol{p}}$ を代入し, $\hat{\alpha}$ とする. このように構成した推定量 $f(\hat{p}^{(\hat{\alpha})})$ は尺度の値域から外れることはなく, MSE を漸近的には改善している.

3 対称性に関する尺度

本節では, 数値実験で用いる 2 つの対称性からの隔たりを測る尺度を紹介する. 以降は $p_{ij} + p_{ji} \neq 0$ とする. まず Tomizawa *et al.* [12] で与えられた尺度は以下で与えられる.

定義 3.1 (Tomizawa *et al.* [12] の対称性からの隔たりを測る尺度)**.**

$$f(\boldsymbol{p}) = \Phi^{(\lambda)} = \sum_{i<j}(p_{ij}^* + p_{ji}^*)\phi_{ij}^{(\lambda)} \quad for\ \lambda > -1$$

ここで, $\delta = \sum_{i\neq j} p_{ij}$, $p_{ij}^* = p_{ij}/\delta$ とし, $p_{ij}^c = p_{ij}/(p_{ij} + p_{ji})$, $\phi_{ij}^{(\lambda)}$ は以下で与えられる.

$$\phi_{ij}^{(\lambda)} = 1 - \frac{\lambda 2^\lambda}{2^\lambda - 1}H_{ij}^{(\lambda)}, \quad H_{ij}^{(\lambda)} = \frac{1}{\lambda}\left[1 - (p_{ij}^c)^{\lambda+1} - (p_{ji}^c)^{\lambda+1}\right]$$

またこの尺度は (i)$0 \leq \Phi^{(\lambda)} \leq 1$, (ii)$\Phi^{(\lambda)} = 0 \Leftrightarrow p_{ij} = p_{ji}$, (iii)$\Phi^{(\lambda)} = 1 \Leftrightarrow p_{ij}^c = 0 (p_{ji}^c = 1)$ または $p_{ji}^c = 0 (p_{ij}^c = 1)$ $(i < j)$ が成り立つことがわかっている [12].

同様に Tahata *et al.* [9] で与えられた尺度は以下で与えられる.

定義 3.2 (Tahata *et al.* [9] の対称性からの隔たりを測る尺度)**.**

$$\varphi = \frac{4}{\pi} \sum_{i=1}^{r-1} \sum_{j=i+1}^{r} (p_{ij}^* + p_{ji}^*) \left(\theta_{ij} - \frac{\pi}{4} \right)$$

ここで, $\delta = \sum_{i \neq j} p_{ij}$, $p_{ij}^* = p_{ij}/\delta$ とし, θ_{ij} は以下で与えられる.

$$\theta_{ij} = \arccos \left(\frac{p_{ij}}{\sqrt{p_{ij}^2 + p_{ji}^2}} \right)$$

またこの尺度は (i)$-1 \leq \varphi \leq 1$, (ii)$p_{ij} = p_{ji} \Rightarrow \varphi = 0$, (iii)$\varphi = 1 \Leftrightarrow p_{ij}^c = 0 (p_{ji}^c = 1)$, (iv)$(i < j)$, $\varphi = -1 \Leftrightarrow p_{ji}^c = 0 (p_{ij}^c = 1)$, $(i < j)$ が成り立つことがわかっている [9].

どちらの尺度も対称モデルが成り立つときは, 0 を取ることがわかる. 注意として, $\Phi^{(\lambda)} = 0$ は対称モデルと同値であるが, $\varphi = 0$ は対称モデルとは限らない. (Tahata *et al.* [9] では, $\varphi = 0$ の構造を Average Symmetry と呼んでいる.) 一方で, φ は $\Phi^{(\lambda)}$ では区別できなかった隔たりの向きを区別することができる (詳細は [9] を参照).

4 数値実験

この節では, 提案推定量 ("New") と既存手法: MLE の plug-in 型 ("Samp.Prop"), Improved Estimator("Improved"), ベイズ型推定量: Fienberg and Holland [2] の plug-in 型 ("FBM"), Uniform prior ($\alpha = 1$), Jeffreys prior ($\alpha = 0.5$) の 6 つの推定量の良さをバイアスと MSE の観点から検証する. 各確率構造から 10000 回のモンテカルロ・シミュレーションを用いて比較を行う. また $\gamma := n/r^2 = 1, 2, \ldots, 10$ をセル数に対するサンプル数の比としてこれを動かして確かめる. 全てグラフは横軸を γ として表示している.

4.1 Tomizawa et al. [12] の尺度

まず, Tomizawa et al. [12] の尺度に関して, 表 1 の確率構造を考える. 表 1(a) は尺度の値が 0 に近く対称性に近い場合の確率構造になっている. 一方で, 表 1(b) は尺度の値が 1 に近く対称性に遠い場合の確率構造になっている.

表 1　人工確率表

(a) $\Phi^{(1)} = 0.099$

0.100	0.060	0.038	0.071
0.038	0.100	0.061	0.026
0.068	0.051	0.100	0.031
0.029	0.066	0.061	0.100

(b) $\Phi^{(1)} = 0.800$

0.100	0.089	0.004	0.102
0.005	0.100	0.094	0.007
0.084	0.002	0.100	0.111
0.009	0.088	0.005	0.100

表 1(a) と表 1(b) のバイアスと MSE の結果をそれぞれ図 1 と図 2 に示す. 図 1 から表 1(a) の確率構造の場合は FHM や Uniform が一番良い. これは表 1(a) は対称性に近いので, 対称性に近い事前分布の方が良くなる. 一方で, 図 2 から表 1(b) の確率構造の場合はが Samp.Prop が一番良い. これは表 1(b) は対称性から離れおり, 対称性に近い事前分布の方が悪影響していることがわかる. 提案手法は, 基本的にはどの手法でも安定して推定できている.

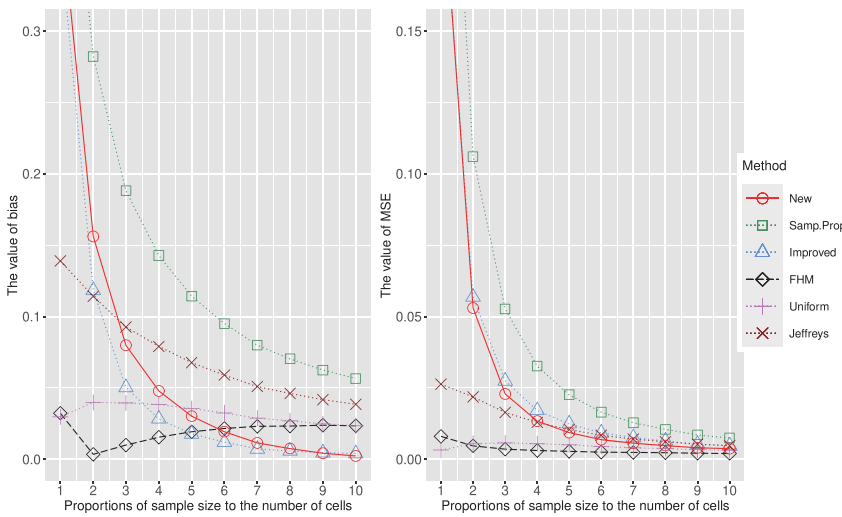

図 1　表 1(a) に対する数値実験の結果

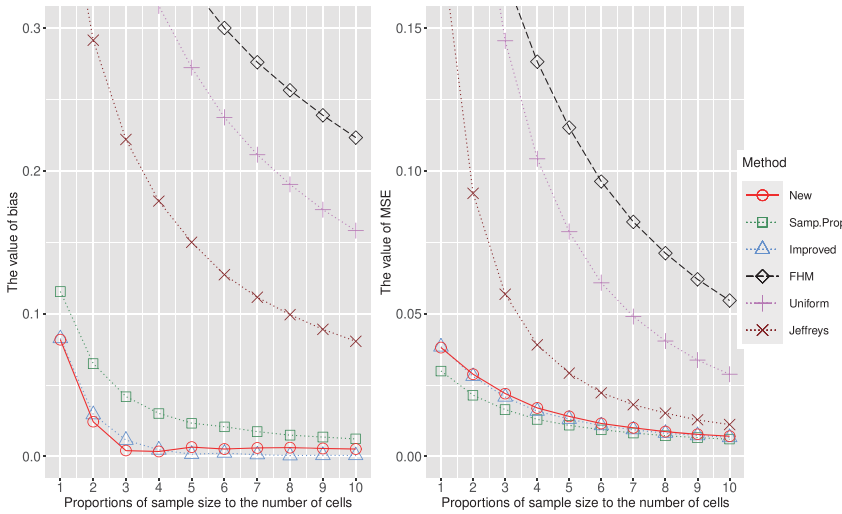

図 2　表 1(b) に対する数値実験の結果

4.2　Tahata *et al.* [9] の尺度

次に Tahata *et al.* [9] の尺度に関して, 表 2 の確率構造を考える. 表 2(a) は尺度の値が 1 に近く, 対称性から遠い, 右上に確率が集中した確率構造になっている. 一方で, 表 2(b) は尺度の値が 0 に近く対称性に近い場合の確率構造になっている.

表 2　人工確率表

(a) $\varphi = -0.761$

0.100	0.094	0.082	0.071
0.018	0.100	0.089	0.081
0.012	0.021	0.100	0.088
0.007	0.014	0.023	0.100

(b) $\varphi = -0.030$

0.100	0.049	0.044	0.052
0.045	0.100	0.054	0.047
0.044	0.042	0.100	0.061
0.059	0.048	0.055	0.100

表 2(a) と表 2(b) のバイアスと MSE の結果をそれぞれ図 3 と図 4 に示す. 図 3 から表 2(a) の確率構造の場合はが Samp.Prop が一番良い. これは表 2(a) は対称性から離れており, 対称性に近い事前分布の方が悪影響していることがわかる. 一方で, 図 4 から表

2(b) の確率構造の場合は FHM が一番良い．これは表 2(a) は対称性に近いので, 対称性に近い事前分布の方が良くなる．提案手法は, 基本的にはどの手法でも安定して推定できている．Samp.Prop と Improved がほぼ同じ値であり, Tahata *et al.* [9] の尺度ではバイアス補正がうまくいっていないことが分かる．

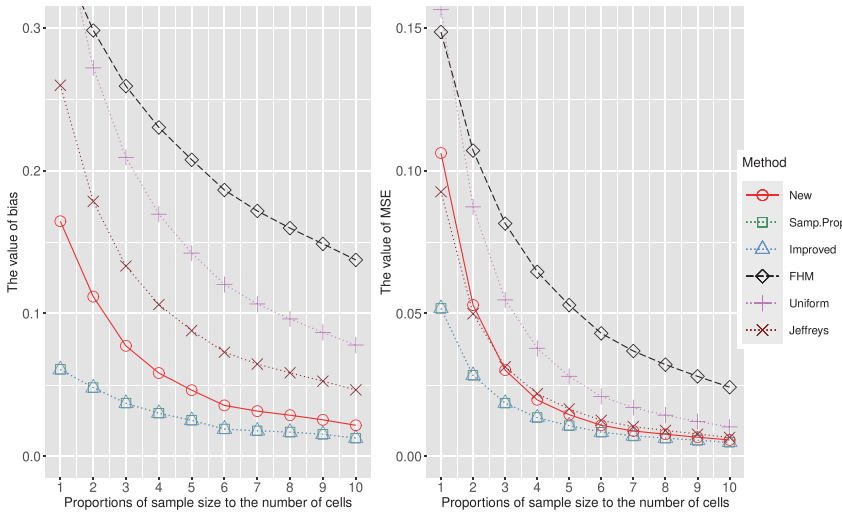

図 3　表 2(a) に対する数値実験の結果

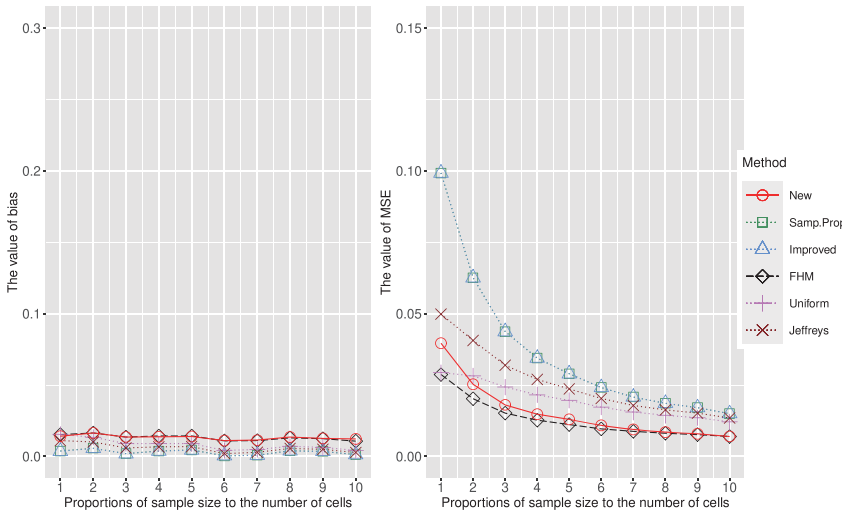

図 4　表 2(b) に対する数値実験の結果

5　まとめ

本稿では, Momozaki *et al.* [4] の Dirichlet parameter の選択方法を概要を与え, Tomizawa *et al.* [12] と Tahata *et al.* [9] の対称モデルからの隔たりを測る尺度に適用した結果を与えた. Momozaki *et al.* [4] の提案手法は, どのような尺度の値でも安定した推定を与えることがわかった. 一方で, サンプルサイズがすごく少ない場合 ($\gamma = 1$) は推定精度にバラツキがあり, 今後の課題である. またベイズ法を用いているので, 尺度の事後分布を導出することも今後の課題である.

謝辞

RIMS 共同研究「ベイズ法と統計的推測」にて貴重な講演の機会を頂き, ありがとうございます. また本研究を行うにあたって, 有益なコメントを頂いた東京理科大学の田畑教授に感謝申し上げます. 最後に, 本研究は JSPS 科研費 JP19K14597, JP20K03756 の助成を受けたものです.

参考文献

[1] Albert H Bowker. A Test for Symmetry in Contingency Tables. *Journal of the American Statistical Association*, 43:572–574, 1948.

[2] Stephen E Fienberg and Paul W Holland. Simultaneous Estimation of Multinomial Cell Probabilities. *Journal of the American Statistical Association*, 68:683–691, 1973.

[3] Leo A Goodman and William H Kruskal. Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, 49:732–764, 1954.

[4] Tomotaka Momozaki, Koji Cho, Tomoyuki Nakagawa, and Sadao Tomizawa. Estimation of measures for two-way contingency tables using the bayesian estimators. *arXiv preprint arXiv:2109.09339*, 2021.

[5] Tomotaka Momozaki, Tomoyuki Nakagawa, Aki Ishii, Yusuke Saigusa, and Sadao Tomizawa. Two-dimensional index of departure from the symmetry model for square contingency tables with nominal categories. *Symmetry*, 13(11):2031, 2021.

[6] Tomoyuki Nakagawa, Ryoma Namba, Kiyotaka Iki, and Sadao Tomizawa. Improved approximate unbiased estimators of the measure of departure from partial symmetry for square contingency tables. *SUT Journal of Mathematics*, 57(2):167–183, 2021.

[7] Yusuke Saigusa, Tomomasa Takada, Aki Ishii, Tomoyuki Nakagawa, and Sadao Tomizawa. Measure of departure from cumulative local symmetry for square contingency tables having ordered categories. *Biometrical Letters*, 57(1):23–35, 2020.

[8] Kouji Tahata, Ryota Tomisato, and Sadao Tomizawa. An Improved Approximate Unbiased Estimator of Log-Odds Ratio for 2×2 Contingency Tables. *Advances and Applications in Statistics*, 9:1–12, 01 2008.

[9] Kouji Tahata, Kouji Yamamoto, Noriyuki Nagatani, and Sadao Tomizawa. A measure of departure from average symmetry for square contingency tables with ordered categories. *Austrian Journal of Statistics*, 38(2):101–108, 2009.

[10] Henri Theil. On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, 76(1):103–154, 1970.

[11] Sadao Tomizawa, Nobuko Miyamato, and Noriaki Ohba. Improved Approximate

Unbiased Estimators of Measure of Asymmetry for Square Contingency Tables. *Advances and Applications in Statistics*, 7:47–63, 01 2007.

[12] Sadao Tomizawa, Takashi Seo, and Hideharu Yamamoto. Power-Divergence-Type Measure of Departure from Symmetry for Square Contingency Tables that Have Nominal Categories. *Journal of Applied Statistics*, 25:387–398, 1998.