正方分割表における非対称モデルとその性質

東京理科大学　田畑　耕治
Kouji Tahata
Tokyo University of Science
E-mail address: kouji_tahata@rs.tus.ac.jp

東京理科大学大学院　郡　優介
Yusuke Kori
Tokyo University of Science

1　はじめに

行と列が順序のある同じ分類からなる正方分割表（順序カテゴリ正方分割表）は，ある個体から異なる二つの時点で得られる順序カテゴリカルデータを表形式にまとめることによって得られる．他にもマッチドペアデータなどから，順序カテゴリ正方分割表が得られる．このような正方分割表は，医学・薬学，心理学，政治学など様々な分野に見られる．順序カテゴリ正方分割表の一つの特徴として，観測度数の多くが分割表の主対角またはその周辺に集中する．例えば，表 1 は Stuart (1953) により解析された 1943 年から 1946 年までに英国の王立軍需工場で働いていた年齢 30 才から 39 才までの女性 7477 名の左右裸眼視力データである．この例では，実に全体の約 70%の観測度数が分割表の主対角部分に集中している．したがって，正方分割表の解析においては分類間の統計的独立性よりもむしろ対称性の解析に関心がある．正方分割表における対称性のモデリングについては，例えば，Kateri and Papaioannou (1997)，Kateri and Agresti (2007)，Tahata (2020, 2022) などがある．

表 1: 英国人女性の左右裸眼視力データ (Stuart 1953)

Right eye grade	Left eye grade				Total
	Highest (1)	Second (2)	Third (3)	Lowest (4)	
Highest (1)	1520	266	124	66	1976
Second (2)	234	1512	432	78	2256
Third (3)	117	362	1772	205	2456
Lowest (4)	36	82	179	492	789
Total	1907	2222	2507	841	7477

正方分割表の解析において，Tahata (2020) は f-divergence に基づく非対称モデルを導入した．このモデルを本稿では，$\mathrm{AS}_k[f]$ モデルと記す．この $\mathrm{AS}_k[f]$ モデルは，Kateri and

Papaioannou (1997) が提案した f-divergence に基づく準対称（QS[f]）モデルや Kateri and Agresti (2007) が提案した f-divergence に基づく順序準対称（OQS[f]）モデルを特別な場合に含む．本稿では，$\mathrm{AS}_k[f]$ モデルの同値表現を与える．また，この同値表現を用いて，対称性からの隔たりを表すパラメータの推定量の漸近分布を Lang (2004) の結果を利用して導出する．

正方分割表において，種々の対称性（または非対称性）のモデルが提案されている．しかし，それらモデルの適合度検定を実行するためには，専門的な知識を必要とする．そのような背景から，Lawal (2001, 2004) や Lawal and Sundheim (2002) は，non-standard log-linear model を用いて SAS や SPSS に対称性のモデルの適合度検定を実装した．また，R 言語を用いた実装については Kateri (2014) や Tan (2017) で紹介されているが，それらを参考に $\mathrm{AS}_k[f]$ モデルの適合度検定を実装することは難しい．したがって，本稿では，新たに $\mathrm{AS}_k[f]$ モデルの同値表現を与えることによって，$\mathrm{AS}_k[f]$ モデルの適合度検定を R 言語を用いて実装する．この実装によって解析したい分割表データ，関数 f，スコア $\{u_i\}$，非対称パラメータ数 k を指定するだけで $\mathrm{AS}_k[f]$ モデルの適合度検定が実行出来るようになるため，分割表解析の専門知識がないユーザでも簡単に利用することができる．

2　モデルとその性質

順序カテゴリ $r \times r$ 正方分割表の (i, j) セル確率を π_{ij} とする $(i = 1, \ldots, r; j = 1, \ldots, r)$．また，行と列に既知のスコア $u_1 < \cdots < u_r$ を割り当てる．

任意に与えられた k $(k = 1, \ldots, r-1)$ に対して，$\mathrm{AS}_k[f]$ モデルは次のように定義される：

$$F(2\pi_{ij}^c) = \sum_{h=1}^{k} u_i^h \alpha_h + \gamma_{ij} \quad (i = 1, \ldots, r; j = 1, \ldots, r) \tag{1}$$

ただし，$\gamma_{ij} = \gamma_{ji}$，$\pi_{ij}^c = \pi_{ij}/(\pi_{ij} + \pi_{ji})$，$F(t) = f'(t)$ である (Tahata 2020)．ここに，f は $(0, \infty)$ で二階微分可能な狭義凸関数で $f(1) = 0$ を満たし，$f(0) = \lim_{t \to 0} f(t)$，$0 \cdot f(0/0) = 0$，$0 \cdot f(a/0) = a \lim_{t \to \infty}(f(t)/t)$ とする．$\mathrm{AS}_k[f]$ モデルの導出については，田畑 (2020) を参照されたい．$k = r - 1$ のときの $\mathrm{AS}_k[f]$ モデルは QS[f] モデルであり，$k = 1$ のときの $\mathrm{AS}_k[f]$ モデルは OQS[f] モデルである．

$\mathrm{AS}_{r-1}[f]$ モデルは，次のようにも表せる：

$$F(2\pi_{ij}^c) = \sum_{h=1}^{r-1} \prod_{s=1}^{h} (u_i - u_s)\alpha_h^* + \gamma_{ij} \quad (i = 1, \ldots, r; j = 1, \ldots, r)$$

ただし，$\gamma_{ij} = \gamma_{ji}$ である．これは，$k = r - 1$ のときの (1) の同値表現である．また，$\alpha_1^* = \cdots = \alpha_{r-1}^* = 0$ のとき，このモデルは対称（S）モデル（すなわち，$\pi_{ij} = \pi_{ji},\ i < j$）である (Bowker 1948)．

$j = 2, 3, \ldots, r$ に対して，

$$\alpha^*_{j-1} = \frac{F(2\pi^c_{1j}) - F(2\pi^c_{j1}) + \sum_{h=1}^{j-2} \left\{ \prod_{i=1}^{h} (u_j - u_i) \right\} \alpha^*_h}{-\prod_{i=1}^{j-1} (u_j - u_i)} \tag{2}$$

が成り立つ．このとき，(2) の右辺にあるパラメータ α^*_h は，セル確率 π_{ij} の関数として表せることに注意する．つまり，(2) はセル確率 π_{ij} の関数として $\alpha^*_1, \ldots, \alpha^*_{r-1}$ を扱えることを示している．

順序カテゴリ $r \times r$ 正方分割表の (i, j) セル度数を n_{ij} とし $(i = 1, \ldots, r; j = 1, \ldots, r)$，$n = \sum_i \sum_j n_{ij}$ とする．(n_{ij}) が多項分布 Multi$(n, \boldsymbol{\pi})$ からの標本とする．ただし，$\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1r}, \ldots, \pi_{r1}, \ldots, \pi_{rr})^\top$ である．また，n_{ij} の期待度数を m_{ij} とする．ここに，$m_{ij} = n\pi_{ij}$ である．Lang (2004) は，期待度数ベクトル $\boldsymbol{m} = (m_{11}, \ldots, m_{1r}, \ldots, m_{r1}, \ldots, m_{rr})^\top$ と制約関数 $\boldsymbol{h}$ を用いて表現される Multinomial-Poisson Homogeneous（MPH）モデルを導入し，MPH モデルにおける最尤法による推定とその大標本における性質を示した．

AS$_k[f]$ モデルは次のようにも表現可能である：

$$\boldsymbol{h}(\boldsymbol{\pi}) = \boldsymbol{0}$$

ただし，$\boldsymbol{h}(\boldsymbol{\pi}) = (h_{1,k+2}(\boldsymbol{\pi}), \ldots, h_{1r}(\boldsymbol{\pi}), h_{23}(\boldsymbol{\pi}), \ldots, h_{2r}(\boldsymbol{\pi}), \ldots, h_{r-1,r}(\boldsymbol{\pi}))^\top$ であり

$$h_{ij}(\boldsymbol{\pi}) = F(2\pi^c_{ij}) - F(2\pi^c_{ji}) - \sum_{l=1}^{k} \left\{ \prod_{s=1}^{l} (u_i - u_s) - \prod_{t=1}^{l} (u_j - u_t) \right\} \alpha^*_l$$

である．ここに，(2) よりパラメータ α^*_l $(l = 1, \ldots, k)$ はセル確率 $\boldsymbol{\pi}$ の関数であり，$d_k = r(r-1)/2 - k$ とするとき，$\boldsymbol{h}(\boldsymbol{\pi})$ は $d_k \times 1$ ベクトルである．したがって，AS$_k[f]$ モデルは MPH モデルの特別な場合として含まれることがわかる．

セル確率 π_{ij} とセル期待度数 m_{ij} の AS$_k[f]$ モデルの下での最尤推定量をそれぞれ $\hat{\pi}_{ij}$ と $\hat{m}_{ij}$ とする．すなわち，$\hat{m}_{ij} = n\hat{\pi}_{ij}$ である．最尤推定値は，(対数) 尤度方程式を Newton-Raphson 法などを用いて解くことによって得られる．ここで，

$$\boldsymbol{\alpha}^*(\boldsymbol{\pi}) = (\alpha^*_1(\boldsymbol{\pi}), \ldots, \alpha^*_k(\boldsymbol{\pi}))^\top$$

とおく．このとき，$\boldsymbol{\alpha}^*(\hat{\boldsymbol{\pi}})$ の漸近分布を考える．ただし，$\hat{\boldsymbol{\pi}} = (\hat{\pi}_{11}, \ldots, \hat{\pi}_{1r}, \ldots, \hat{\pi}_{r1}, \ldots, \hat{\pi}_{rr})^\top$ である．ここに，関数 $\alpha^*_l(\boldsymbol{\pi})$ $(l = 1, \ldots, k)$ は (2) によって定義されることに注意する．Lang (2004) の Theorem 3 より，

$$\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \underset{d}{\to} N(\boldsymbol{0}, \boldsymbol{\Sigma})$$

ただし，$\boldsymbol{\Sigma} = \boldsymbol{D} - \boldsymbol{\pi}\boldsymbol{\pi}^\top - \boldsymbol{D}\boldsymbol{H}(\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H})^{-1}\boldsymbol{H}^\top \boldsymbol{D}$ である．ここに，$\boldsymbol{D}$ は $\boldsymbol{\pi}$ の第 i 番目の成分を第 i 番目の対角成分にもつ対角行列，$\boldsymbol{H} = \partial \boldsymbol{h}^\top(\boldsymbol{\pi})/\partial \boldsymbol{\pi}$ である．

デルタ法を用いることにより，

$$\sqrt{n}(\boldsymbol{\alpha}^*(\hat{\boldsymbol{\pi}}) - \boldsymbol{\alpha}^*(\boldsymbol{\pi})) \underset{d}{\to} N(\mathbf{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$$

ただし，$\boldsymbol{A} = \partial\boldsymbol{\alpha}^*(\boldsymbol{\pi})/\partial\boldsymbol{\pi}^\top$ である．$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$ のプラグイン推定量を $\widehat{\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top}$ とする．また，$\widehat{\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top}$ の第 l 番目の対角成分を $\widehat{v(\alpha_l^*)}$ とする $(l = 1, \ldots, k)$．このとき，信頼係数 $(1-p)\times 100\%$ の α_l^* に対する近似信頼区間は，

$$\left[\ \alpha_l^*(\hat{\boldsymbol{\pi}}) - z_{p/2}\sqrt{\frac{\widehat{v(\alpha_l^*)}}{n}},\ \alpha_l^*(\hat{\boldsymbol{\pi}}) + z_{p/2}\sqrt{\frac{\widehat{v(\alpha_l^*)}}{n}}\ \right]$$

で与えられる．ただし，z_p を標準正規分布の上側 $100p\%$ 点とする．

3　実装と適用例

本研究では，関数 DisplayASkfResult を R 言語で実装した．この関数により $\mathrm{AS}_k[f]$ モデルの適合度検定に関連する種々の情報を得ることができる．ソースコードは，

https://github.com/icy-mountain/MasterResearch

から入手可能である．表 1 を用いて関数 DisplayASkfResult の利用方法を解説する．

表 1 の分割表データを R のコンソールに次のように入力する．

```
> visiondata.women <- c(1520,266,124,66,234,1512,432,78,117,362,1772,205,36,82,179,
492)
```

関数 DisplayASkfResult は，$k = 0$ とすると S モデルの適合度検定を実行できる．

```
> DisplayASkfResult(freq=visiondata.women,f="(1-t)^2",name="t",score=c(1,2,3,4),k=0)
**********result**********
k: 0
f: (1-t)^2
df: 6
G2: 19.25
pValue: 0.003763
```

ここに，df はカイ二乗分布の自由度，G2 は尤度比カイ二乗統計量の値，pValue は検定の p 値である．この他にも S モデルの下での最尤推定値などが出力されるが，ここでは省略する．この結果から，表 1 のデータに対して S モデルの当てはまりが悪いことを確認できる．

次に，S モデルを拡張したモデルである $\mathrm{AS}_k[f]$ モデルの当てはめを考える．例えば，$f(t) = (1-t)^2$，$(u_1, u_2, u_3, u_4) = (1, 2, 3, 4)$，$k = 1$ とした $\mathrm{AS}_k[f]$ モデルの適合度検定は次のように実行できる．

```
> DisplayASkfResult(freq=visiondata.women,f="(1-t)^2",name="t",score=c(1,2,3,4),k=1)
**********result**********
k: 1
f: (1-t)^2
df: 5
G2: 7.271
pValue: 0.2013
alpha_stars:
             Estimate  Std.Error  Confidential.Interval
alpha_star1   -0.21426   0.06137   [  -0.33455,  -0.09397] *
(*) means interval excluding 0.
```

この結果から，表 1 のデータに対して $f(t) = (1-t)^2$ とした $\mathrm{AS}_1[f]$ モデルの当てはまりが良いことを確認できる．さらに，$k = 2$ とした $\mathrm{AS}_k[f]$ モデルの適合度検定は次のように実行できる．

```
> DisplayASkfResult(freq=visiondata.women,f="(1-t)^2",name="t",score=c(1,2,3,4),k=2)
**********result**********
k: 2
f: (1-t)^2
df: 4
G2: 7.267
pValue: 0.1224
alpha_stars:
             Estimate  Std.Error  Confidential.Interval
alpha_star1    -0.2201    0.1213   [  -0.45789,   0.01776]
alpha_star2   0.003263  0.056041   [   -0.1066,    0.1131]
(*) means interval excluding 0.
```

この結果から，表 1 のデータに対して $f(t) = (1-t)^2$ とした $\mathrm{AS}_2[f]$ モデルの当てはまりも良いことを確認できる．$\mathrm{AS}_1[f]$ モデルと $\mathrm{AS}_2[f]$ モデルは包含関係にあることから，$\mathrm{AS}_2[f]$ モデルが成り立つことを仮定した下での $\mathrm{AS}_1[f]$ モデルの条件付き適合度検定を利用することができる．$7.271 - 7.267 = 0.004 < 3.84$ より条件付き検定の結果から $\mathrm{AS}_1[f]$ モデルが表 1 のデータに対して適切なモデルであると考えられる．つまり，条件付き確率の差に関して，

$$\hat{\pi}^c_{ij} - \hat{\pi}^c_{ji} = (i-j)\frac{-0.21426}{4} \quad (i < j)$$

と推測される．$i < j$ に対して，$i - j < 0$ であることから $\hat{\pi}^c_{ij} - \hat{\pi}^c_{ji} > 0$ である．また，分割表の主対角から離れるほど（$i - j$ が小さくなるほど），条件付き確率の差は主対角からの距離 $|i-j|$ に依存して大きくなると推測される．以上のことから，表 1 の英国人女性の左右裸眼視力において，左目の視力よりも右目の視力の方が良い傾向にあると推測される．

謝辞

本研究はJSPS科研費 JP20K03756，JP20H00576，及び京都大学数理解析研究所の助成を受けたものです.

参考文献

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association* **43**, 572–574.

Kateri, M. (2014). *Contingency Table Analysis. Methods and Implementation Using R.* Birkhäuser, Basel, Switzerland.

Kateri, M. and Agresti, A. (2007). A class of ordinal quasi-symmetry models for square contingency tables. *Statistics and Probability Letters* **77**, 598–603.

Kateri, M. and Papaioannou, T. (1997). Asymmetry models for contingency tables. *Journal of the American Statistical Association* **92**, 1124–1131.

Lang, J. B. (2004). Multinomial-poisson homogeneous models for contingency tables. *The Annals of Statistics* **32**, 340–383.

Lawal, H. B. (2001). Modeling symmetry models in square contingency tables with ordered categories. *Journal of Statistical Computation and Simulation* **71**, 59–83.

Lawal, H. B. (2004). Using a glm to decompose the symmetry model in square contingency tables with ordered categories. *Journal of Applied Statistics* **31**, 279–303.

Lawal, H. B. and Sundheim, R. A. (2002). Generating factor variables for asymmetry, non-independence and skew-symmetry models in square contingency tables using sas. *Journal of Statistical Software* **7**, 1–23.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika* **40**, 105–110.

Tahata, K. (2020). Separation of symmetry for square tables with ordinal categorical data. *Japanese Journal of Statistics and Data Science* **3**, 469–484.

Tahata, K. (2022). Advances in quasi-symmetry for square contingency tables. *Symmetry* **14**, 1051.

Tan, T. K. (2017). *Doubly Classified Model with R.* Springer, Singapore.

田畑耕治 (2020). 正方分割表における対称性のモデリング. **数理解析研究所講究録 2157**, 52–55.