

順序カテゴリ正方分割表における 2 変量 t 分布型対称モデルについて

東京理科大学理工学部 生^い亀清貴^き¹
東京理科大学理工学部 富澤貞男

1. はじめに

行と列が順序のある同じ分類からなる $R \times R$ 正方分割表において, (i, j) セル確率を p_{ij} とする ($i = 1, \dots, R; j = 1, \dots, R$). 対称モデル (Bowker, 1948) は次のように定義される:

$$p_{ij} = p_{ji} \quad (i < j).$$

このモデルはセル確率の対称構造を示している. 線形対角パラメータ対称モデル (Agresti, 1983) は次のように定義される:

$$p_{ij} = \theta^{j-i} p_{ji} \quad (i < j).$$

このモデルは対称的なセル確率の比が主対角線からの距離 $j - i$ に依存して指数的に変化するという構造を示している. 特に $\theta = 1$ とおいたこのモデルは対称モデルである.

2 変量正規分布に従う確率変数 U と V を考える, ただし $E(U) = \mu_1$, $E(V) = \mu_2$, $\text{Var}(U) = \text{Var}(V) = \sigma^2$, $\text{Corr}(U, V) = \rho$. このとき結合確率密度関数 $f(u, v)$ は, 次の関係をみたす:

$$\frac{f(u, v)}{f(v, u)} = \delta^{v-u},$$

ただし,

$$\delta = \exp \left[\frac{\mu_2 - \mu_1}{\sigma^2(1 - \rho)} \right].$$

したがって, 本来データが連続量で周辺分散が等しい潜在的な 2 変量正規分布に従うと想定される場合において, いくつかの切断点を設けて正方分割表データを構成するとき, 線形対角パラメータ対称モデルはよく適合すると考えられる. 詳細については Yamamoto et al. (2007) も参考にされたい.

本講演では順序カテゴリ正方分割表の解析において, 潜在分布として周辺分散の等しい 2 変量 t 分布が想定される場合に適切であると考えられるモデルを提案する.

¹〒 278-8510 千葉県野田市山崎 2641
e-mail: iki@is.noda.tus.ac.jp

2. 2変量 t 分布型対称モデル

自由度 m の 2 変量 t 分布に従う確率変数 U と V を考える, ただし $E(U) = \mu_1$, $E(V) = \mu_2$, $\text{Var}(U) = \text{Var}(V) = m\sigma^2/(m-2)$ ($m > 2$), $\text{Corr}(U, V) = \rho$. このとき結合確率密度関数 $h(u, v)$ は,

$$h(u, v) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \left(1 + \frac{Q(u, v)}{m}\right)^{-\frac{m+2}{2}},$$

ただし,

$$Q(u, v) = \frac{1}{\sigma^2(1-\rho^2)} \left[(u - \mu_1)^2 - 2\rho(u - \mu_1)(v - \mu_2) + (v - \mu_2)^2 \right].$$

また $h(u, v)$ は次の関係をみたす:

$$\left(h(u, v)\right)^{-\frac{2}{m+2}} - \left(h(v, u)\right)^{-\frac{2}{m+2}} = \alpha_m(v - u) \quad (u < v),$$

ただし,

$$\alpha_m = \frac{2(\mu_1 - \mu_2)}{m\sigma^2(1-\rho)} \left(2\pi\sigma^2\sqrt{1-\rho^2}\right)^{\frac{2}{m+2}}.$$

$R \times R$ 順序カテゴリ正方分割表に対して次のモデルを提案する. 固定した m ($m > 2$) に対して,

$$p_{ij}^{-\frac{2}{m+2}} - p_{ji}^{-\frac{2}{m+2}} = \beta_m(j - i) \quad (i < j),$$

ただし β_m は未知パラメータとする. このモデルはデータが連続量で周辺分散が等しい潜在的な自由度 m の t 分布に従うと想定されるとき, よく適合すると考えられる. このモデルを t 分布型対称モデルと呼ぶことにする (TS(m)) によって記す. 特に $\beta_m = 0$ とおいたこのモデルは対称モデルである. このモデルは p_{ij} の $-2/(m+2)$ 乗と p_{ji} の $-2/(m+2)$ 乗の差が主対角線からの距離 $j - i$ に比例することを示している.

行変数 X , 列変数 Y の分布関数をそれぞれ $F_i^X = \sum_{k=1}^i \sum_{l=1}^R p_{kl}$, $F_i^Y = \sum_{k=1}^R \sum_{l=1}^i p_{kl}$ とする ($i = 1, \dots, R-1$). 任意の m に対する TS(m) モデルの下で, $\beta_m > 0$ と $p_{ij} < p_{ji}$ ($i < j$) は同値である. すなわち $\beta_m > 0$ と $F_i^X < F_i^Y$ ($i = 1, \dots, R-1$) が同値である. よって, このモデルの下でのパラメータ β_m は, X と Y のどちらが確率的に大きい (小さい) かを推測するのに有効である.

3. 例題

表 1 は, デンマークの既婚女性 (40~59 歳) とその父親を, それぞれ社会階級で分類した順序カテゴリ正方分割表データである (Andersen, 1997). カテゴリは社会階級の高い方から低い方へ並べられている. 表 2 は, S モデル, LDPS モデル, TS(m) モデルに対する尤度比カイ 2 乗値 G^2 を表している.

表 2 より, S モデル, LDPS モデル, TS(m) モデル ($m = 50, 100, 150, 200$) の当てはまりが悪いのに対して, TS(m) モデル ($m = 5, 10, 20, 30$) の当てはまりが良いことがわかる. また, TS(m) モデルの m の値が大きくなるにつれ G^2 値は上昇し, LDPS モデルに対する G^2 値に近づく傾向にある.

TS(5) モデルの下で $\hat{\beta} = -0.471$ であることから, 娘の社会階級よりも父の社会階級の方が高い傾向にある.

表 1: Danish Welfare Study による父親と娘の社会階級

父親の 階級	娘の階級				計
	I-II	III	IV	V	
I-II	12	17	22	3	54
III	8	33	85	95	221
IV	11	26	72	87	196
V	2	18	50	111	181
計	33	94	229	296	652

表 2: 表 1 のデータに適用したモデルに対する尤度比カイ 2 乗値 G^2

モデル	自由度	G^2
S	6	107.96*
LDPS	5	13.00*
TS(5)	5	7.88
TS(10)	5	8.99
TS(20)	5	10.36
TS(30)	5	11.05
TS(50)	5	11.72*
TS(100)	5	12.32*
TS(150)	5	12.53*
TS(200)	5	12.65*

(*印は 5% 有意)

4. シミュレーション

人工データに対して LDPS モデル, TS(m) モデルを当てはめることにより, 尤度比カイ 2 乗値 G^2 がどのような値をとるかを調べたい. 周辺分散が等しい 2 変量 t 分布を想定し, 変数の取り得る範囲に $R - 1$ 個の切断点を設けて観測度数からなる $R \times R$ 正方分割表を作成する. ここでは 2 変量 t 分布の各パラメータを $\mu_2 = \mu_1 + 0.5$, $\sigma^2 = 1$, $\rho = 0.6$ とし, 自由度 $m = 5, 20, 100$ の場合をそれぞれ考える. それぞれの分布から発生する標本数を 5000 とし, 切断点を $\mu_1 - 0.7$, μ_1 , $\mu_1 + 0.7$ とし, 標本からなる 4×4 分割表を作成する. それらの分割表に, 対応する自由度 m の TS(m) モデ

ルと S モデル, LDPS モデルをそれぞれ当てはめ, 尤度比カイ 2 乗値 G^2 を与えたのが表 3 である.

表 3: $m = 5, 20, 100$ で作成したデータに適用したモデルに対する G^2

モデル	$m = 5$		$m = 20$		$m = 100$	
	自由度	G^2	自由度	G^2	自由度	G^2
S	6	851.53*	6	1066.26*	6	983.83*
LDPS	5	21.61*	5	8.74	5	3.34
TS(5)	5	1.55	5	7.68	5	21.62*
TS(20)	5	11.30*	5	3.55	5	5.93
TS(100)	5	19.02*	5	7.12	5	3.56

(*印は 5% 有意)

LDPS モデルは $m = 5$ の場合に対しては当てはまりが悪く, $m = 20, 100$ の場合に対しては当てはまりが良い. 一方, 背後の t 分布に対応する自由度を用いた TS(m) モデルは $m = 5, 20, 100$ のすべての場合に対して当てはまりが良いことがわかる. たとえば自由度 $m = 5$ の t 分布に対しては, TS(5) モデルの適合度は極めて良い.

すなわち, 周辺分散が等しい潜在的 t 分布の自由度が小さい場合は, LDPS モデルよりも対応する TS(m) モデルの方が当てはまりが良いと推測される. また自由度が大きい場合は, LDPS モデルと対応する TS(m) モデルの両方の当てはまりが良く, TS(m) モデルに対する G^2 値は LDPS モデルに対する G^2 値に近づくと推測される.

その他の条件でのシミュレーション解析は, 当日報告する予定である.

参考文献

- Agresti, A. (1983). A simple diagonals-parameter symmetry and quasi-symmetry model. *Statistics and Probability Letters*, **1**, 313-316.
- Andersen, E. B. (1997). *Introduction to the Statistical Analysis of Categorical Data*, Springer, New York.
- Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, **43**, 572-574.
- Yamamoto, H., Iwashita, T. and Tomizawa, S. (2007). Decomposition of symmetry into ordinal quasi-symmetry and marginal equimoment for multi-way tables. *Austrian Journal of Statistics* 36, 291-306.