二変量正規分布を潜在的にもつ分割表に対する連関尺度の提案

浦崎 航[1]
[1]東京理科大学大学院 創域理工学研究科

1 はじめに

カテゴリカル変数は医学, 教育学, 社会科学などの様々な分野で現れ, 解析の対象となることも多い. また, このようなカテゴリカルなデータに対する解析は 1 世紀以上前から現在に至るまで多くの議論や研究の対象となってきた. 特に, r 個のカテゴリと c 個のカテゴリから構成される 2 種類のカテゴリカル変数を考えたとき, それらの組み合わせから得られる r 行 c 列の表は二元分割表と呼ばれている. このような分割表を用いた解析では, 2 つの変数に関係性があるかどうか, つまり, 統計的に独立性が成り立つかの評価を目的に様々な手法が提案されてきた. その手法のなかには, 1 世紀程前より提案された Pearson のカイ二乗検定 (Pearson, 1900) や対数尤度比検定 (Wilks, 1935) を用いた独立性の検定が有名であり, 現在でも R や SAS といった統計解析ソフトウェアにおいて標準的に備わっているほど普遍的であり, 広く用いられている. また, サンプルサイズが小さい場合には, Fisher の正確確率検定という手法も一般的に利用されている. これらの検定手法によって, 二元分割表の変数間に独立性が当てはまらないと判断されたとき, これらの変数間には何らかの関連性があるとした上でのさらなる解析として, 連関性を解析する手法が様々に提案されてきた.

連関性解析の手法については, 「連関モデル」や「連関尺度」という枠組みで様々な提案がされてきた. 連関モデルには, 独立性の検定と同様に, 変数間にどのような関係構造が存在するのかをモデルとして捉え, 検定によって構造の評価を下す手法である. Goodman (1979) が導入した一様連関モデルを皮切りに, 様々なモデルが提案されてきた. 近年では, f-divergence (または ϕ-divergence) に基づいて一般化を行なった連関モデルとして Kateri and Papaioannou (1994), Kateri (2018), Forcina and Kateri (2021) などが提案されている. 連関モデルの詳しい紹介については, Goodman (1985, 1986), Agresti (1983b), Liu and Agresti (2005) などを参照したい. その一方で, 連関尺度については, 独立性からの逸脱度合いを連関性の強さの程度だと捉えて, サンプルサイズに影響を受けずに, 関係性の強さを一定の区間内で定量化する手法である. 連関尺度の例としては, クラメール係数 V^2 (Cramér, 1946) や Theil の不確実係数 U (Theil, 1970) があり, これらを含めても様々な提案がされてきた. 連関尺度についても f-divergence などを用いた一般化を目的とする提案が行われており, Momozaki et al. (2023) や Urasaki et al. (2024) などが提案さ

れている.

これらのような画期的な手法が提案されている一方, 2×2 分割表ではあるものの, 分割表自体の構成方法に関して Pearson (1900) から議論されてきたことがある. それは, 分割表として観測されるカテゴリカルデータは, 本来が連続的な潜在変数 (例えば, 能力, 態度, 傾向など) が閾値で区切られて離散化された結果である, というものである. そのため, 上記で述べてきた手法以外にも, 潜在分布として正規分布や t 分布, 対数正規分布などを仮定した分割表に対する提案が様々にされてきた. 提案例としては, Formann (1993), Goutis (1993) などの, 独立性における検定統計量を修正する, また独立性や連関の不確実性を考慮するなどの研究がある. また, 独立性や連関性とは異なる対称性という枠組みではあるが, Agresti (1983a), Yamamoto and Murakami (2014), Saigusa et al. (2018) などの提案もあり, 興味があるならば読んでおきたい.

本研究の目的は, 二元分割表における f-divergence による連関尺度の提案において Urasaki et al. (2024) で示された, 離散型 divergence による独立性の逸脱と潜在分布に二変量正規分布を仮定したときの相関係数 ρ との関係に基づき, divergence で構成される新たな連関尺度の提案である. また, 潜在分布が仮定される状況下で構成される divergence 型尺度には, どのような特徴が示されるのかを考察することも目的の 1 つである. 本講演では, 新たに提案する連関尺度の性質の紹介と信頼区間の導入を行うとともに, 提案尺度を用いて実データ解析を行なった結果などを示す.

2　先行研究 (f-divergence と潜在分布との関係)

二元分割表における独立性からの逸脱を測る離散 divergence と二変量正規分布を潜在分布に仮定した場合における相関係数 ρ との関係を述べる前に, f-divergence を紹介する.

二元 $r \times c$ 分割表を考え, p_{ij} $(i = 1, \ldots, r;\ j = 1, \ldots, c)$ を i 行 j 列目のセル確率とする. また, $p_{i\cdot}$, $p_{\cdot j}$ をそれぞれ行と列の周辺確率として, $p_{i\cdot} = \sum_{j=1}^{c} p_{ij}$, $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$ とする. このとき, 独立性における $\{p_{ij}\}$, $\{p_{i\cdot}p_{\cdot j}\}$ 間の隔たりを測るための f-divergence は, 以下のように定義される:

$$I_f(\{p_{ij}\}; \{p_{i\cdot}p_{\cdot j}\}) = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{i\cdot}p_{\cdot j} f\left(\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}\right).$$

ここで, $f(x)$ は一回微分可能な狭義凸関数であり, $(0, +\infty)$ 区間で $f(1) = 0$, $\lim_{x\to 0} f(x) = 0$, $0f(0/0) = 0$, $0f(a/0) = a\lim_{x\to\infty}[f(x)/x]$ の条件をもつ (詳しくは, Csiszár and Shields, 2004 を参照).

このとき, 潜在変数として X^*, Y^* を仮定した場合, セル確率 p_{ij} は次のように表される:

$$\begin{aligned} p_{ij} &= P(X = i, Y = j) \\ &= P(x_{i-1} < X^* \le x_i, y_{j-1} < Y^* \le y_j) \\ &= f_{X^*,Y^*}(\tilde{x}_i, \tilde{y}_j)\Delta_{x_i}\Delta_{y_j}, \end{aligned}$$

ただし,

$$x_{i-1} < \tilde{x}_i \le x_i, \quad y_{j-1} < \tilde{y}_j \le y_j$$

であり, $f_{X^*,Y^*}(\tilde{x}_i, \tilde{y}_j)$ は確率変数 X^*, Y^* における同時確率密度関数とする. また, Δ_{x_i} と Δ_{y_j} は, それぞれ $(x_{i-1}, x_i]$ および $(y_{j-1}, y_j]$ の間隔幅である. このように潜在変数を仮定した状況下において, f-divergence $I_f(\{p_{ij}\}; \{p_{i\cdot}p_{\cdot j}\})$ は次のように近似することができる:

$$\begin{aligned} I_f(\{p_{ij}\}; \{p_{i\cdot}p_{\cdot j}\}) &= \sum_{i=1}^{r}\sum_{j=1}^{c} f_{X^*}(\tilde{x}_i) f_{Y^*}(\tilde{y}_j) f\left(\frac{f_{X^*,Y^*}(\tilde{x}_i, \tilde{y}_j)}{f_{X^*}(\tilde{x}_i) f_{Y^*}(\tilde{y}_j)}\right) \Delta_{x_i}\Delta_{y_j} \\ &\xrightarrow[\Delta_{x_i},\Delta_{y_j}\to 0]{} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X^*}(x) f_{Y^*}(y) f\left(\frac{f_{X^*,Y^*}(x,y)}{f_{X^*}(x) f_{Y^*}(y)}\right) dxdy, \end{aligned} \tag{1}$$

また, $f_{X^*}(x)$ および $f_{Y^*}(y)$ は $f_{X^*,Y^*}(x,y)$ についての X^*, Y^* の周辺確率密度関数である.

式 (1) のように f-divergence と潜在分布との関係が近似的に示されたが, 潜在変数 X^* と Y^* が二変量正規分布に従うとする. また, X^* と Y^* の同時確率密度関数を,

$$\begin{aligned} f_{X^*,Y^*}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\Bigg[-\frac{1}{2(1-\rho^2)} \\ \left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}\Bigg] \end{aligned}$$

ただし,

$$-\infty < x, y < +\infty, \quad \sigma_x, \sigma_y > 0$$

とする. ここで, μ_x, σ_x はそれぞれ X^* の平均と標準偏差, μ_y, σ_y はそれぞれ Y^* の平均と標準偏差である. また, X^* および Y^* の相関係数を ρ $(-1 \le \rho \le 1)$ とする. この式において, 2 つの平均 μ_x, μ_y は正値である必要はない. このような同時確率密度関数に対して, 式 (1) の積分が可能である f-divergence の関数として $f(x) = (x^{\lambda+1} - x)/\{\lambda(\lambda+1)\}$ (ただし, $-\infty < \lambda < \infty$) を考える. ここで, $f(x) = (x^{\lambda+1} - x)/\{\lambda(\lambda+1)\}$ を与えたときには, 以下で表される power-divergence (Cressie and Read, 1984; Read and Cressie, 1988) と一致する:

$$I^{(\lambda)}(\{p_{ij}\}; \{p_{i\cdot}p_{\cdot j}\}) = \frac{1}{\lambda(\lambda-1)} \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\left\{\left(\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}\right)^{\lambda} - 1\right\}.$$

そのため, 式 (1) は power-divergence と相関係数 ρ を用いて, 以下で書き換えることができる:

$$I^{(\lambda)}(\{p_{ij}\}; \{p_{i\cdot}p_{\cdot j}\}) \approx \frac{1}{\lambda(\lambda+1)}\left\{(1-\rho^2)^{-\frac{\lambda}{2}}(1-\lambda^2\rho^2)^{-\frac{1}{2}} - 1\right\}, \tag{2}$$

ただし, 積分可能な条件として,

$$|\lambda\rho| < 1,$$

が得られる. ここで, $|\lambda\rho| < 1$ という条件と相関係数 ρ の範囲を考えると, パラメータ λ の範囲は $-1 \le \lambda \le 1$ に限られることが示唆される. 特に, $\lambda = 0, 1$ (ただし, $\lambda = 0$ は $\lambda \to 0$) を与えた場合

には,

$$I^{(0)}(\{p_{ij}\};\{p_{i\cdot}p_{\cdot j}\}) \approx -\frac{1}{2}\log(1-\rho^2), \tag{3}$$

$$I^{(1)}(\{p_{ij}\};\{p_{i\cdot}p_{\cdot j}\}) \approx \frac{1}{2}\left\{(1-\rho^2)^{-1}-1\right\}, \tag{4}$$

という, シンプルな関係性が得られる.

これらの関係性が得られたように, Urasaki et al. (2024) では, f-divergence における独立性評価と潜在分布を近似的に示せるとともに, 特に, 潜在分布に二変量正規分布が仮定され, なおかつ power-divergence の一部のクラスについては式 (2) が成り立つことが示された. 現在判明しているのは power-divergence に限定されるが, 式 (1) の積分が可能な狭義凸関数 $f(x)$ を見つけることでさらなる関係性の発見に繋がると考えられる. また, 式 (3), (4) の関係に基づいて, 二変量正規分布が潜在分布に仮定される二元分割表に対する連関尺度の提案を次章で行う.

3　近似的な divergence と相関係数の関係に基づく連関尺度の提案

近似的な離散 divergence と相関係数の関係として, パラメータ $\lambda = 0, 1$ の 2 つを与えた場合に表された式 (3), (4) を考える. このとき, それぞれのパラメータにおける式 (3), (4) を ρ^2 について解くことで, 潜在分布に二変量正規分布が仮定される分割表に対する連関尺度とする. したがって, 次の 2 つを提案する:

$$\rho^2_{KL} = 1-\exp\left\{-2I_{KL}(\{p_{ij}\};\{p_{i\cdot}p_{\cdot j}\})\right\},$$
$$\rho^2_{P} = 1-\left\{2I_{P}(\{p_{ij}\};\{p_{i\cdot}p_{\cdot j}\})+1\right\}^{-1},$$

ここで, $I_{KL}(\cdot;\cdot)$ および $I_P(\cdot;\cdot)$ のそれぞれは, KL-divergence および Pearson divergence であり, power-divergence に $\lambda = 0, 1$ を適用した場合と一致するため, 以下のように表される:

$$I_{KL}(\{p_{ij}\};\{p_{i\cdot}p_{\cdot j}\}) = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}},$$
$$I_{P}(\{p_{ij}\};\{p_{i\cdot}p_{\cdot j}\}) = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}.$$

これらの提案尺度は, クラメール係数 V^2 や Theil の不確実係数 U などと同様に, divergence に基づいて構成される連関尺度となっている. また, これらの尺度については次の定理を得る.

定理 1. *ρ^2_{KL} は次の性質をもつ:*

1 $0 \le \rho^2_{KL} \le 1-\exp\left\{-2\min\left(-\sum_{i=1}^{r} p_{i\cdot}\log p_{i\cdot}, -\sum_{j=1}^{c} p_{\cdot j}\log p_{\cdot j}\right)\right\}$.
  1-1 $r \ge c$ かつ $p_{i\cdot} = r^{-1}$ のとき, $0 \le \rho^2_{KL} \le 1-\exp\{2\sum_{i=1}^{r} p_{i\cdot}\log p_{i\cdot}\} \le 1-r^{-2}$.
  1-2 $r < c$ かつ $p_{\cdot j} = c^{-1}$ のとき, $0 \le \rho^2_{KL} \le 1-\exp\{2\sum_{j=1}^{c} p_{\cdot j}\log p_{\cdot j}\} \le 1-c^{-2}$.
2 $\rho^2_{KL} = 0$ のとき, 行と列変数には完全な独立構造がある.

3 ρ_{KL}^2 が最大値をとるとき, 行と列変数には完全な連関構造がある.

4 $r, c \to \infty$ のとき, $0 \leq \rho_{KL}^2 < 1$.

定理 2. ρ_P^2 は次の性質をもつ*:*

1 $0 \leq \rho_P^2 \leq 1 - \min(r, c)^{-1}$.

2 $\rho_P^2 = 0$ のとき, 行と列変数には完全な独立構造がある.

3 $\rho_P^2 = 1 - \min(r, c)^{-1}$, 行と列変数には完全な連関構造がある.

4 $r, c \to \infty$ のとき, $0 \leq \rho_P^2 < 1$.

これらの定理から見てとれるように, 提案される尺度は, 既存の divergence に基づいて構成される尺度と異なる性質を備えていることがわかる.

本講演では, 上記で提案された 2 つの尺度についての特徴を深掘りしていくとともに, 信頼区間の構成についてを紹介する. また, 本研究で提案された尺度における応用上の利点についてを, いくつかの数値実験や実データ解析例を通して報告する.

参考文献

Agresti, A. (1983a). A simple diagonals-parameter symmetry and quasi-symmetry model. *Statistics & probability letters*, 1(6):313–316.

Agresti, A. (1983b). A survey of strategies for modeling cross-classifications having ordinal variables. *Journal of the american statistical association*, 78(381):184–198.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton university press.

Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464.

Csiszár, I. and Shields, P. C. (2004). *Information theory and statistics: A tutorial*. Now Publishers Inc.

Forcina, A. and Kateri, M. (2021). A new general class of rc association models: Estimation and main properties. *Journal of Multivariate Analysis*, 184:104741.

Formann, A. K. (1993). Fixed-distance latent class models for the analysis of sets of two-way contingency tables. *Biometrics*, pages 511–521.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, pages 10–69.

Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach

and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review/Revue Internationale de Statistique*, pages 243–270.

Goutis, C. (1993). Bayesian estimation methods for contingency tables. *Journal of the Italian Statistical Society*, 2(1):35–54.

Kateri, M. (2018). ϕ-divergence in contingency table analysis. *Entropy*, 20(5):324.

Kateri, M. and Papaioannou, T. (1994). *f-divergence Association Models*. University of Ioannina.

Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, 14:1–73.

Momozaki, T., Wada, Y., Nakagawa, T., and Tomizawa, S. (2023). Extension of generalized proportional reduction in variation measure for two-way contingency tables. *Behaviormetrika*, 50(1):385–398.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Read, T. R. and Cressie, N. A. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer Science & Business Media.

Saigusa, Y., Goda, S., Yamamoto, K., and Tomizawa, S. (2018). Unrestricted normal distribution type symmetry model for square contingency tables with ordered categories. *J Biom Biostat*, 9(395):2.

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

Urasaki, W., Nakagawa, T., Momozaki, T., and Tomizawa, S. (2024). Generalized cramér' s coefficient via f-divergence for contingency tables. *Advances in Data Analysis and Classification*, 18(4):893–910.

Wilks, S. (1935). The likelihood test of independence in contingency tables. *The Annals of Mathematical Statistics*, 6(4):190–196.

Yamamoto, K. and Murakami, H. (2014). Model based on skew normal distribution for square contingency tables with ordinal categories. *Computational Statistics & Data Analysis*, 78:135–140.

著者連絡先: 〒 278-8510 千葉県野田市山崎 2641
浦崎航 (Tel. 04-7124-1501)
E-mail: 6323701@ed.tus.ac.jp