Conditional covariance formula を用いたポリコリック相関係数

早稲田大学　久保 沙織
早稲田大学　椎名 乾平
早稲田大学　上田 卓司

はじめに

順序カテゴリカルデータから計算された相関係数 r は歪んでいることが広く知られており，とりわけ 2 変数間でカテゴリ数が異なる場合には，その歪みは顕著となる (UCE; unequal category effect)。順序カテゴリカル変数同士で相関を求めたい場合には，ポリコリック相関係数を計算することで，UCE を軽減し，より正確な相関係数が得られる。本発表では，conditional covariance formula (Ross, 2010) と呼ばれる公式を利用することで，ポリコリック相関係数に相当する統計量が簡単に得られることを示す。

方法

2 つのカテゴリカル変数 $X(\in \{1, 2, \ldots, p\})$ と $Y(\in \{1, 2, \ldots q\})$ は，二変量正規分布 $\phi(x', y'|\mu_{X'}, \mu_{Y'}, \sigma^2_{X'}, \sigma^2_{Y'}, \rho)$ に従う n 組のデータペア $(x'_k, y'_k), k = 1, 2, \ldots, n$ が，一方は p 段階の順序付きカテゴリ，もう一方は q 段階の順序付きカテゴリに変換されて生成されたと仮定する。このとき，観測データ $(x_k, y_k), k = 1, 2, \ldots, n$ から，$p \times q$ の分割表 (contingency table) N を作成することができる。ポリコリック相関係数では，分割表 N をもとに，もとの二変量正規分布 $\phi(x', y'|\mu_{X'}, \mu_{Y'}, \sigma^2_{X'}, \sigma^2_{Y'}, \rho)$ のパラメータ ρ を推定することが目的となる。

3 つの変数 A, B, C について，C が所与のときの A と B との共分散を表す conditional covariance formula を，A と B をそれぞれ X' と Y'，そして C を N の各セルの要素と見なして当てはめると，X' と Y' との共分散および分散について以下が導かれる。

$$
\begin{aligned}
Cov(X', Y') &= \left\{ \sum_{i=1}^{p} \sum_{j=1}^{q} \gamma_{ij} Cov(x'_{ij}, y'_{ij}) \right\} + Cov(E[x'_{ij}], E[y'_{ij}]) \\
&= \sum_{i=1}^{p} \sum_{j=1}^{q} \gamma_{ij} \{Cov(x'_{ij}, y'_{ij}) + E[x'_{ij}]E[y'_{ij}]\} - E[X']E[Y'] \\
V(X') &= \sum_{i=1}^{p} \sum_{j=1}^{q} \gamma_{ij} \{V[x'_{ij}] + (E[x'_{ij}])^2\} - (E[X'])^2 \\
V(Y') &= \sum_{i=1}^{p} \sum_{j=1}^{q} \gamma_{ij} \{V[y'_{ij}] + (E[y'_{ij}])^2\} - (E[Y'])^2
\end{aligned}
$$

ここで，x'_{ij} と y'_{ij} は，$p \times q$ 個の矩形領域に区切られた (X', Y') 空間においてそれぞれの領域に含まれる x' と y' を表しており，二重切断二変量正規分布に従う。また，γ_{ij} は各領域に対応する二変量正規分布の体積 (確率) である。$\mu_{X'} = \mu_{Y'} = 0$ および $\sigma^2_{X'} = \sigma^2_{Y'} = 1$ を仮定し，上式の γ_{ij} を n_{ij}/n に置き換えて $Cov(X', Y')/\sqrt{V(X')V(Y')}$ を計算することで，X' と Y' との相関係数の推定値が得られる。

適用例

実データに対して，本研究で導出された X' と Y' との相関係数の推定値と，統計解析環境 R の polycor パッケージを利用した従来のポリコリック相関係数を計算し，その結果を比較検討する。

参考文献

Ross, S. M. (2010). *A first course in probability*, 8th edition, New Jersey: Pearson, p.381.