

集約的シンボリックデータにおける変数間の相関の指標

統計数理研究所 清水 信夫
中央大学 中野 純司
徳島文理大学 山本 由和

1 集約的シンボリックデータ

大量の個体をもつ多変量データが自然なグループに分かれている場合、そのような各々のグループについての推論に興味がある場合が考えられる。このとき、グループを表すためのいくつかの記述統計量の集合をデータと考えたものを集約的シンボリックデータ (Aggregated Symbolic Data, ASD) と呼ぶ。実際のデータにおいては連続変数とカテゴリー変数が両方が多数含まれている場合がよく見られる。われわれはこのような状況において、連続変数とカテゴリー変数いずれの場合に対しても同じ基準でグループ間の非類似度を考え、それを用いてグループのクラスタリングを行う方法を提案してきた。

われわれは特に各グループにおいて、各々の変数、および異なる 2 つずつの変数の組に関して、2 次までのモーメントを用いて求められる記述統計量を考え、それらの集合を ASD と考える。ASD を形成する記述統計量は、連続変数同士に関しては平均や分散共分散行列、カテゴリー変数同士であれば Burt 行列に相当する。また、連続変数とカテゴリー変数に対しても意味のある統計量が得られる。

データ全体および各グループは共通の変数をもつので、それらから求められる各 ASD を比較し表示する際には、各変数ごと、および異なる 2 変数間の各々の関係において、共通の概念で表される指標が必要である。

2 集約的シンボリックデータにおける各変数間の相関

データ全体および各グループにおいて、異なる変数間の関係はそのグループの情報として重要である。2 つの連続変数間の関係は相関係数が有名である。一方、カテゴリー変数についてはカテゴリー値が順序尺度の場合 (順序変数) と名義尺度の場合 (名義変数) が存在するが、順序変数同士の従属性を表す指標としてはポリコリック相関係数、順序変数と連続変数との相関を表す指標としてはポリシリアル相関係数がそれぞれ知られている。ただしこれらは連続変数同士の相関係数とは導出方法が異なっている。

われわれは連続変数もカテゴリー変数も統一的に扱いたい。そのため、例えば連続変数同士の組み合わせにおいては、連続変数を形式的に離散化する。すなわち 2 つの連続変数データがその内部に含まれる十分大きな正方形領域を各次元ごとに極めて細かく等分した分割表を考えると、それぞれのセルおよび周辺度数において正規化された個体が 1 個もしくは 0 個存在するようにできる。そのような巨大かつスパースな分割表に 2 変量標準正規分布を当てはめた場合の相関係数の最尤推定量であるポリコリック相関係数は、連続 2 次元標本に対する相関係数に等しいことが示される。またカテゴリー変数と連続変数の組み合わせの場合の相関も、連続変数の離散化によりカテゴリー変数同士のポリコリック相関係数として考えられる。すなわち変数間の全ての組み合わせに対してポリコリック相関係数を同一の指標として考えることができる。そしてその指標を ASD をより簡略化した記述統計量として考え、実データに適用した例を示す。