連続・離散変換の影響－記述的統計（２）－

統計数理研究所　　馬場康維

１．はじめに

連続型のデータをカテゴリーに変換して解析に用いる場合を想定する。ヒストグラムを描いて、ヒストグラムから平均値や標準偏差を求める場合の誤差の評価は典型的なものである。また、ジニ係数は、累積による数値に基づいているがカテゴリーの切り方の影響を受ける。主成分分析では、５カテゴリーに分類したデータでも、オリジナルの数値のままでも結果はほとんど変わらないことが、発表者により既に報告されている。ここでは、さまざまな統計指標について連続・離散変換が及ぼす影響を考察する。

２．主成分分析とクラスター分析の例

ｎ個体ｐ変数のデータ行列を元にした分析法について考える。離散化について次のような場合いを想定する。長さＬの区間に入っているデータを幅ｃの小区間に分け各区間の中心点をその区間の代表値とする。このような場合に離散化が与える影響について考える。例えば、主成分分析は，座標回転による方法であり，したがって，座標変換によっても個体間の距離は変わらない。主成分を求めるために必要な分散共分散に離散化が与える影響は，級内変動と級間変動の比で見積ることができる。すなわち，分散共分散にあたえる影響は，c/L の２乗のオーダーである。例えば 100 点満点の試験に対して 20 点の幅で離散化したとすればそのオーダーは 1/25 であり主成分分析に与える影響は大きくはない。一方，個体間の個々の距離についてはどうか。１変数について考えると，隣接した二つの級にある二つの個体間の距離ｄは　0＜ｄ＜2ｃ　であり，離散化した場合にはこの範囲にある距離がｃを単位として表現されることになる。したがって、クラスター分析でも上記のような見積もりができる。

３．種々の統計指標・統計値への影響

連続量をカテゴリー化することで影響を受ける例を示す。

例１）ヒストグラムから平均、分散、モーメントを求める

例２）ローレンツ曲線を描き、ジニ係数を求める

例３）クラスター分析では、個体間、クラスター間の距離が影響を受ける

例４）回帰分析の場合、偶然変動に等分散の仮定を置いたとして、直線のあてはめには影響が少ないが、多項式のあてはめの場合には、等分散の仮定が崩れるところにその影響が現れる。

その他、様々な場合を想定して、連続・離散変換の影響について考察する。