研修教材としての擬似データの作成と利用[1]

統計数理研究所　馬場康維
情報・システム研究機構／統計数理研究所　岡本　基
総務省統計研究研修所　野呂竜夫
総務省統計研究研修所　加藤真二

統計学を学ぶ過程で、公的統計の分析実習をおこなうことがしばしばある。分析のための仮説の構築、実証のためのデータの獲得、データ解析まで、一連のステップを学習者一人一人に実行させることにより総合的な分析力の向上を促す学習スタイルは良く利用される方法である。統計の分析能力を身につける上でこの方法は効果的で、それまで学習した多くの手法や理論が、有機的に結びつき統計学全般への理解が深まることを目の当たりにしている。

分析実習では、学習者は自分でデータの探索を行い、手に入る情報を有効に活用して自身の設定した仮説の実証を試みることが多い。その際、個票データが利用できないことから、都道府県あるいは市区町村単位の集計結果を用いて分析を行わざるを得ないのが現実である。集計結果を用いての分析にはそこから導き出せることに限界がある。したがって、個票データの利用が望まれるが、個票データの情報の秘匿の観点から、演習・実習で用いるのは様々なハードルがある。

分析の多くは、変数間の関係を求めることにある。このため、しばしば、多変量解析法が手法として用いられる。馬場の一連の研究では、主成分分析等の線形変換に基づく多変量解析の諸手法では、連続量であるデータをカテゴリーに変換したデータを用いてもオリジナルデータの結果に近い結果が得られることが知られている。そこで、全国消費実態調査を基に、カテゴリー化したデータに乱数を加えることにより、元のデータの情報を保持しつつ、個人情報ではない擬似データを作成することを試みた。この報告では、オリジナルデータと擬似データの結果の比較を行い、擬似データが十分利用に耐えうることを示す。

参考文献

馬場康維（2010）．連続・離散変換による情報の保持と秘匿、日本計算機統計学会第 24 回大会予稿集，pp41-42.

[1] この報告は 2016 年度統計研修所共同研究の成果の一部に基づくものである。