公的統計調査の離散変換による教育用疑似データ作成の試み
―全国消費実態調査を例として―

統計数理研究所　　馬場康維
情報・システム研究機構　　岡本　基
総務省　統計研修所　　山口幸三

１．集計データの限界

官公庁等の公的機関の研修授業や大学の分析実習で公的統計のデータを利用して、統計処理や分析の実習を行うという場面を想定しよう。研修所を大学に置き換えてもほぼ同じ状況であるから、以下では研修所の研修ということにする。こういう研修にはいくつかのテーマがある。

１）データから分析ができるためのスキルを身に付ける（データ分析の基礎の学習）
２）データの記述から発見的に問題を見つける能力を養う（探索の能力養成）
３）仮説を立て、データを用いて仮説を検証する能力を養う（検証能力の養成）
４）モデルを作りデータを用いて予測をする能力を養う（予測の能力の養成）

等である。第１のテーマは、統計手法等の学習であり統計処理の基本的な技術を学習するものであるから、手法等の学習の材料になる適当なデータがあれば事足りる。一方、第２以下のテーマでは、本物のデータが必要である。たとえば、第４のテーマでは、本物でないデータで予測をしても架空の現象の予測になり学習の手ごたえのないものになりかねないが、本物のデータであれば将来の政策立案に役立つ結論が得られる可能性があり、研修の意欲を高めるに十分である。

実際の研修の場では、研修生たちがさまざまな仮説を提起する。しかしそれを実証するために必要な個票からの集計ができないために、仕方がなく、都道府県単位の集計データから項目間の相関構造を探るというような代替的な方法に頼らざるを得ないのが現状である。研修生にも教育者の側にも隔靴掻痒の感が残り不満足な研修になる恐れがある。

２．教育用擬似データの生成

個人情報の秘匿の観点から、個票が使えないのであれば、結果的に個票を用いたのと同じような結論が得られるデータを生成すればよい。擬似データの条件は、単純である。

１）単純集計の結果が個票から求めたものとほぼ同じになる
２）変数間の相関構造が個票から求めたものとほぼ同じになる

という条件を満たせば良い。連続量であれば先ずカテゴリーに変換し、本物らしく見せるために再連続化を行うという手順を踏むことによって利用価値のある擬似データが比較的簡単に得られることになる。

ここでは、全国消費実態調査を例として、教育用擬似データの作成の試みについて報告する。

参考資料

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.