全国消費実態調査の匿名データを用いた新疑似ミクロデータの作成

BioStat 研究所（株）　高橋 行雄
(公財)統計情報研究開発センター　周防 節雄
(独)統計センター 統計情報・技術部　宮内 亨

はじめに　(独)統計センターで 2012 年から提供された教育用擬似ミクロデータを用いて SAS ユーザー総会で「Let's データ分析コンテスト」を 2013 年から 4 回開催してきたが, 2016 年度末で教育用擬似ミクロデータの提供が打切られた. そこで, 2004 年全国消費実態調査(全消)の匿名データを用いて, SAS ユーザー会世話人有志が新擬似ミクロデータを新たに作成し, 今年のコンテスト用に供した. 新擬似ミクロデータ作成について, 公表された統計表だけを利用して作成することが統計センターから課せられた必須条件であった. このため, 匿名データから, ①世帯に関する情報 14 項目と集計乗率, ②14 次元クロス表のセル毎に収支に関する 203 項目の対数変換した平均値と標準偏差, ③年間収入 3 階級別の主要 21 項目間の相関行列の表を作成しウェブ上に公表. その公表した情報から新擬似ミクロデータを作成, SAS データセットと CSV 形式の両方でウェブ上に公開した.

収支項目　多彩な分析ができるよう, これまで提供されてきた教育用擬似ミクロデータの収支項目 183 個の他に 20 項目を新たに加え 203 項目としたほか, 質的項目に「世帯分類」情報も加えた.

多次元クロス表　2004 年全消の匿名データは, 二人以上の世帯及び単身世帯別のファイルで, それぞれ 1780 項目, 全 47,797 レコードから成る. 世帯属性に関連する質的項目で多次元クロス表を作成すると, 値が匿名データそのものとなる度数 1 のセルが多発する. このため項目を絞り込み 14 項目(延べ 53 カテゴリー)とした. 14 次元クロス表(14,246 セル)において, 度数1(8,406 セル)及び度数 2(2,261 セル)となるセルについては, 度数がそれぞれ 3 と 4 になるよう誤差を与えたレコードを追加した.

追加レコードに与える誤差　度数 1 及び度数 2 のセルに該当する 21,334 レコードの 203 項目に対して常用対数変換を行い, それぞれに平均=0, SD=0.02 の正規乱数を加えた. こうすれば, 元の円単位データに一律 4.7%の誤差変動を与えたことになる. 匿名データにこれらを追加し 69,131 レコードとし, 新擬似ミクロデータ作成に使用する各種統計表の作成のための元ファイルを作成した.

14 次元番号付き統計量　14 次元クロス表(サイズ:14,242×14)のすべてのセルに付与した一連番号を 69,131 レコードに与え, このセル番号別に対数変換した 203 項目に対し対数平均と SD 求めた. また, 新擬似ミクロデータ作成の際に必要となる 0 円となっているレコード数も項目として加えた統計表(サイズ:14,242×203×3)を作成した.

相関行列　対数変換した主要 21 項目に対して, 年間収入 3 階級別の対数相関行列表を作成し, 14 次元クロス表, 14 次元番号付き統計表と一緒に Excel 形式でウェブ上に公開した(高橋, 周防).

新擬似ミクロデータの作成　ウェブ上に公表された情報だけを用いて新擬似ミクロデータを作成した(高橋, 周防, 宮内). 作成時の留意点は以下の通り. ①年間収入 3 階級別の 69,131 世帯分の 21 次元正規乱数の作成. ②主要 21 項目の 14,246 個のセルの対数平均値と SD に対して 21 次元正規乱数を適用し, 69,131 世帯データの作成. ③主要 21 項目以外の収支 182 項目について 14,246 個のセルの対数平均値と SD に正規乱数の適用し 69,131 世帯データを作成. ④収支データが 0 円の割合を保つために一様乱数の適用. ⑤下位の収支項目の合計がその上位の項目の収支金額となる様に「足し上げ」を実施. ⑥正規乱数の適用の際に, 過剰な発散を防ぐための制約条件を設定. ⑦作成された新擬似ミクロデータに対する足し上げ計算の確認. ⑧作成された新擬似ミクロデータと匿名データの各種の統計量を比較し, 元の変数の分布状況が保持されていることを確認. ⑨何時でも誰でも自由にダウンロードできるように, 以下の URL において新擬似ミクロデータを公開.

SAS データセット　http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroData.zip
CSV 形式　http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroDataCSV.zip

今後の展望　2004 年に加え, 1989 年, 1994 年, 1999 年の全国消費実態調査の匿名データについても擬似ミクロデータ化し, 来年以降の「Let's データ分析コンテスト」に供したい.