全国消費実態調査 4 回分の匿名データから擬似ミクロデータの作成

BioStat 研究所（株）　　高橋 行雄

はじめに　全国消費実態調査 2004 年度の匿名データを用いた擬似ミクロデータを作成し，2017 年および 2018 年の SAS ユーザー総会の「Let's データ分析コンテスト」に供した．さらに 1989，1994，および 1999 年度の匿名データについても 2004 年度に準じて擬似ミクロデータを作成した．調査項目および符号内容が年度毎に異なり，別々の擬似ミクロデータとすることも検討したが，経年変化を容易に分析できるように 4 年度分を一括した擬似ミクロデータとした．

メタデータを活用した連結　統計センターの WEB に公開されている各年度の「データレイアウト・符号表」を用いて，各年度別にメタデータを作成した．各年度の変数名は，接頭語を（S，T，U，V）とし 4 桁の通し番号を付けた．各年度の「年間収入」の変数名は，（S0740，T0502，U0405，V0399）であり，2004 年度の変数名 V0399 に他の年度の変数名を変更することにした．このように疑似ミクロ化の対象の全ての変数についてファイル操作で変数名を変更しメタデータに加えた．共通化した変数名を持つ各年度の匿名データについて，変数名でのマッチ機能を活用しファイルの連結を行い 4 年度分で 192,599 レコードのファイルを生成した．なお，統計センターから提供を受けた各年度の匿名データは，単身世帯と二人以上の世帯に分かれていたが，それらも全て連結した．

14 次元クロス表　世帯に関する項目についての多次元クロスを作成すると，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．このため項目と符号を絞り込み 14 項目（延べ 53 カテゴリー）とした．年度別 14 次元クロス表（53,772 セル）を作成すると，度数1（30,954 セル）および度数 2（8,331 セル）が発生する．度数がそれぞれ 3 と 4 になるように 30,954×2＋8,331×2=78,570 レコードを匿名データ 192,599 レコードに追加し，年度別 14 次元クロス表を改めて作成し，公表のための統計表とした．

収支項目に与える誤差　追加後のファイルは 271,169 レコードとなり，14 次元クロス表で度数 1 の場合は全て度数 3 となるが，収支項目の平均は匿名データそのものが再現されてしまう．そこで，収支 203 項目に対して常用対数変換を行い，追加した 78,570 レコードに平均=0，SD=0.02 の正規乱数を加えた．この操作で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．

14 次元番号付き統計量　14 次元クロス表の通し番号を 271,197 レコードに与え，この番号別に対数変換した収支 203 項目に対し対数平均と SD 求めた．また，擬似ミクロデータ作成の際に必要となる 0 円（対数では欠測値）を除いたレコード数も項目として加え，収支項目に関する統計表（サイズ：53,772×203×3）を作成した．

相関行列　対数変換した主要な収支 21 項目に対して，年度別年間収入 3 階級別の対数相関行列表を作成し，14 次元クロス表，14 次元番号付き統計量と一緒に Excel 形式でウェブ上に公開する予定である．

擬似ミクロデータの作成　すでに作成した 2004 年度分の擬似ミクロデータ作成の手順に準じ，公表予定の Excel 形式の統計表のみを用い，4 年度分の擬似ミクロデータを一括して作成した．擬似ミクロデータの公表形式は，分析コンテスト用であれば，変数名付の SAS データセット形式が望ましいが，一般的には，項目名を変数ラベルとして付与することが親切であり，世帯に関する変数については，符号のみならず符号内容を含めることも親切と思われる．公表形式については，今後の課題である．

文献　高橋行雄，周防節雄，宮内亨（2017），全国消費実態調査(2004 年)の匿名データから JMP による新擬似ミクロデータの作成，官民オープンデータ利用活用の動向及び人材育成の取組（平成 29 年度）報告要旨集：1-40．http://www.nstac.go.jp/services/pdf/171117_1-2.pdf　http://www.nstac.go.jp/services/pdf/171117_1-1.pdf