質的変数に関わる擬似ミクロデータ作成の試み

滝澤有美（統計センター）

1　はじめに

平成 21 年の新統計法(平成 19 年法律第 53 号)全面施行に伴い、総務省統計局は公的統計の二次的利用拡大に向けた匿名データの提供及びオーダーメード集計を開始した。その後、研究者、教育関係者等から「二次的利用推進のためには、統計教育・訓練用データが必要」との意見が高まり、統計委員会においても統計教育・訓練用データの必要性に関する指摘を受けた。この流れに基づき、独立行政法人統計センターにて、教育機関における授業や演習での利用を想定した「擬似ミクロデータ」に関する研究を行うこととなった。

擬似ミクロデータは、調査票情報から集計した高次元のクロス集計表の確率分布に基づき、乱数を発生させて作成したミクロデータ形式のデータである。調査票情報とは異なるが、公表結果表に近い統計量の再現を目的としたものである。

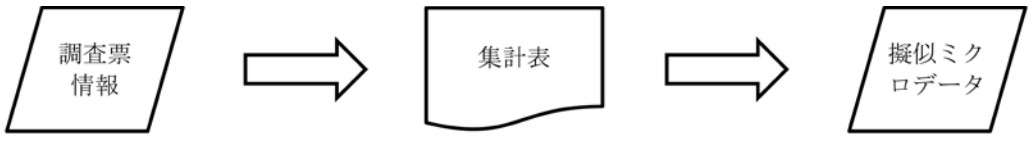
当センターでは、公的統計の二次的利用を推進するための研究活動の一環として、平成 16 年全国消費実態調査データを用いて量的変数を対象とした擬似ミクロデータの作成方法に関する研究を進めてきた。この研究では、支出・収入の統計量（平均、分散及び共分散）に基づく多変量正規乱数を発生させ擬似的な値を生成している。

作成した擬似ミクロデータは､平成23 年８月より当センターHPにて試行提供中である。(http://www.nstac.go.jp/services/giji-microdata.html)

当センターでは、新たに質的変数を対象とした擬似ミクロデータ作成方法の研究を行ったので、その研究結果について報告する。

2　作成方法

擬似ミクロデータの基本的な作成の流れは、以下のとおりである。



一連番号	性別	年齢	年収(万円)
0001	1	20	0
0002	1	21	20
0003	1	24	100
0004	1	25	80
0005	1	27	120
0006	1	29	100
:	:	:	:

→

性別	年齢階級	度数	年収平均(万円)	年収標準偏差
1	20～24	10	40	52.9
1	25～29	20	100	20.0
:	:	:	:	

→

性別	年齢階級	年収(万円)
1	20～24	26
1	20～24	61
1	20～24	52
1	25～29	94
1	25～29	104
1	25～29	108
:	:	:

ただし、質的変数のデータは量的変数と異なり、平均値、相関係数、正規分布等に基づく乱数から作成することが困難であるため、乱数発生方法に新たな工夫が必要となる。

本研究では、平成 14 年就業構造基本調査の「有業者」のデータを用いて、以下の方法で擬似ミクロデータを作成した。

(1) 公表結果表での使用頻度等を考慮し、公表結果表で多く用いられる基本事項の変数（以下「基本変数」）と、これ以外の「加工対象変数」とに分けて、使用変数の選定を行う。

(2) (1)で選定した全変数を用いてクロス集計表を作成し、度数 3 未満のセルについては「加工対象変数」を加工して度数 3 以上にする処理を行う。

(3) (2)のクロス集計表について「基本変数」でのグルーピングを行う。

(4) (3)のグループごとに「加工対象変数」で集計し、各セルの集計用乗率を求める。

(5) 一様乱数を発生させ、これに基づき(4)の「加工対象変数」のセルを抽出する。

この作成方法は、公表結果表で多く用いられる「基本変数」の分布及びデータ内の変数の関連性をよく再現し、統計教育・訓練用のデータに適ったものと考えている。

今回は、作成したデータを公表結果表の数値と比較検証した結果について「作成方法の試み」として報告を行う。

主要参考文献

秋山裕美ほか(2012) 「教育用擬似ミクロデータの開発とその利用～平成16年全国消費実態調査を例として～」（『製表技術参考資料』No.16）