全国消費実態調査の匿名データから SASによる新擬似ミクロデータの作成

周防節雄　(公財) 統計情報研究開発センター
高橋行雄　BioStat研究所 (株)
宮内亨　　(独) 統計センター

2017/11/17
平成29年度共同研究集会
「官民オープンデータ利活用の動向及び人材育成の取組」(29-共研-5010)
@ 統計数理研究所

はじめに

目的
全消(2004年)の匿名データ → ＳＡＳ擬似ミクロデータ

方法
匿名データ → 基本統計表 → 擬似ミクロデータ

結果
作成したデータの品質？

全消　2004年
匿名データ

14個の
世帯属性変数

203個の
収支項目変数

OBS	_STAT_	cellNo	label	Y1	Y2	Y3	～Y203
1	N	1	件数	3	3	3	
2	MIN	1	最小値	215	392725	198800	
3	MAX	1	最大値	242	458785	220203	
4	lg10MEAN	1	log10平均	2.36	5.63	5.32	
5	lg10STD	1	log10標準偏差	1.06	4.43	3.96	
6	wgtMEAN	1	平均乗率	187.64	187.64	187.64	
7	N	2	件数	3	3	3	
8	MIN	2	最小値	270	366231	231725	
9	MAX	2	最大値	384	734995	362068	
10	lg10MEAN	2	log10平均	2.56	5.75	5.48	
11	lg10STD	2	log10標準偏差	1.52	5.26	4.79	
12	wgtMEAN	2	平均乗率	2253.21	2253.21	2253.21	

表1 14次元重み付き基本統計量

表1
14重クロス表

1～16,668
「セル番号」

＋　符号表

擬似ミクロデータ

秘匿性のために

出現頻度1のセル
コピーを2つ追加　→　出現頻度3にする

出現頻度2のセル
コピーを一つずつ追加→出現頻度4にする

コピーしたデータには
乱数を使って若干の誤差変動を与えた

2．擬似ミクロデータの構造

変数リスト（付録1）

変数名	項目名
○世帯事項	
Year	調査年
No	レコード連番号
X01	大都市圏の別
X02	世帯区分
X03	世帯人員
X04	就業人員
X05	住居の構造
X06	住居の建て方
X07	住居の所有関係
X08	世帯主の性別
X09	世帯主の年齢
X10	企業区分･従業者規模
X11	家族分類
X12	未就学児の有無
X13	学校に通う世帯員の有無
X14	65歳以上の世帯員数
Weight	集計用乗率
○収支項目(単位千円)	
Y1	年間収入

世帯属性

Y38	支出総額	Y94	家庭用耐久財	Y150	(特掲)イ
Y39	実支出	Y95	家事用耐久財	Y151	その他の
Y40	消費支出	Y96	冷暖房用器具	Y152	諸雑費
Y41	食料	Y97	一般家具	Y153	理美容
Y42	穀類	Y98	室内装備･装飾品	Y154	理美容
Y43	米	Y99	寝具類	Y155	身の回
Y44	パン	Y100	家事雑貨	Y156	たばこ
Y45	めん類	Y101	家事用消耗品	Y157	その他の
Y46	他の穀類	Y102	家事サービス	Y158	こづかい
Y47	魚介類	Y103	被服及び履物	Y159	交際費
Y48	生鮮魚介	Y104	和服	Y160	交際費
Y49	塩干魚介	Y105	洋服	Y161	交際費
Y50	魚肉練製品	Y106	男子用洋服	Y162	交際費
Y51	他の魚介加工品	Y107	婦人用洋服	Y163	交際費
Y52	肉類	Y108	子供用洋服	Y164	他の物品
Y53	生鮮肉	Y109	シャツ･セーター類	Y165	贈与金
Y54	加工肉	Y110	男子用シャツ･セーター類	Y166	他の交際
Y55	乳卵類	Y111	婦人用シャツ･セーター類	Y167	仕送り金
Y56	牛乳	Y112	子供用シャツ･セーター類	Y168	(再掲)教
Y57	乳製品	Y113	下着類	Y169	(再掲)教

収支項目

収支項目の203変数の関係

親子関係の変数　（１８６個）
足し上げ構造
親の変数＝子の全変数の合計

付録3

再掲･特掲関係の変数（１６個）

再掲･特掲の変数
大小関係
それにぶら下がる変数

無関係の変数
1個
「年間収入」

表1（14重クロスの基本統計量）
＋
変数間の密接な関係

↓

擬似ミクロデータ

擬似ミクロデータ作成の作業手順

（付録4を参照）

作成された擬似ミクロデータの問題点

JMPによる作成作業で行った
「ゼロ円収支項目の出現割合調整」
「年間収入3階区分別主要21項目間の相関係数行列」？
組み込まなかった

足し上げ関係は全て補正した
再掲･特掲関係の一部の変数に再調整が必要

検証と改善案

匿名データ (47,797 observations)
↓
我々の擬似ミクロデータ (73,519 observations)

最初の段階で
出現頻度1と2のデータをコピーした結果

秘匿性のために
必要なのか？

検証と改善案

擬似ミクロデータと匿名データについて
203変数
①平均値、②標準偏差、③最小値、④最大値
（集計用乗率：不使用／使用）　付録5

平均値
集計用乗率なし：匿名データの平均値にかなり近い
集計用乗率あり：両者の差が広がっている

箱ひげ図　付録6
擬似ミクロデータに外れ値多い

検証と改善案

外れ値の値と出現割合を匿名データにほぼ見合うようにして、新擬似ミクロデータの変数を発生させれば散らばり具合をかなりコントロールできる

↓

提案
公開する「14次元重み付き基本統計量」(表1)の他に、
公表の許容範囲内で、セル毎の外れ値に関する情報も表形式にまとめて公表し、
擬似ミクロデータ作成時に利用できるようにする。

？

まとめ

他の年次の全消匿名データも使って、
複数回次の擬似ミクロデータの作成。

利用者にとって、どのような家族分類が最適？

最終的に擬似ミクロデータの作成方法が確定した後、
作成マニュアルを整備･公開

謝辞

本研究の推進に全国消費実態調査の匿名データを使用したが、その際、㈱ＳＡＳジャパンインスティテュートから使用料の援助を受けた。また、匿名データの利用に際しては、独立行政法人統計センターから便宜を図っていただいた。ここに記して謝意を述べる。

ご静聴、
ありがとうございました。