統計数理(2024)
第 72 巻 第 2 号 261–271

特集「公的統計―社会の重要な情報基盤―」
［統計ソフトウェア］

公的統計を利用した教育の実践例

佐藤 正昭†

（受付 2024 年 2 月 2 日；改訂 3 月 22 日；採択 3 月 27 日）

要　　旨

公的統計を利用した統計教育の一端を紹介するとともに，その際必要となる R survey パッケージの分析利用の方法を解説する．

キーワード：ミクロデータ，公的統計調査匿名データ，統計教育，survey パッケージ．

1. はじめに

公的統計は，社会経済について知ることができる国民の情報基盤であることから，公的統計の教育利用が有意義なものとなることは論を待たないところであり，様々な形で教育利用が行われている．大別すると，e-Stat から入手できる既製品とも言える集計表の利用，ミクロデータのひとつである公的統計調査匿名データの利用の 2 つの方法がある．それらの実践例について，筆者の経験の一端を紹介し，さらなる教育利用の広がりを期待するものである．ここで，e-Stat と公的統計調査匿名データという用語について説明しておく．日本の統計が閲覧できる政府統計ポータルサイトとして，「e-Stat」と名付けられた政府統計の総合窓口 (2024)がインターネット上で運営されている. 統計法第 2 条に定義されているように，公的統計(同条第 3 項)を作成するために実施される統計調査(同条第 5 項)の調査票情報(同条第 11 項)から，匿名データ(同条第 12 項)が作成され提供されている．この匿名データについて，「公的統計調査匿名データ」と言うこととする．

2. e-Stat により提供されるデータを利用した統計教育

図 1 で示す e-Stat を通じて，各府省が作成している様々な分野の統計データ(集計表)や関連する情報を，分野，府省名，及びキーワードにより検索して，ダウンロードできるようになっている. さらには，地方公共団体などが提供している統計データの所在源情報も「統計関係リンク集」として掲載されている. e-Stat により提供されている様々な分野の統計データから主要なものを教育用標準データセットとしてパッケージ化した SSDSE というものが用意されており，また，上記で示したように様々な分野の統計調査や加工統計の集計表が利用可能で，それらを利用した統計データ分析教育の実践例が積み重ねられているので，筆者の実践例は省略するが，データを入手する手法として，web スクレイピングの技術も教育の中で取り入れることが重要ではないかと考えている．例えば，メッシュ統計は，第一次メッシュ区画別に境界データと統計データが e-Stat 上で提供されているので，広範囲にわたるデータを入手しようとすると，web スクレイピングによる効率化が求められる．web スクレイピングを実現す

† 滋賀大学 データサイエンス学部：〒 552–8522 滋賀県彦根市馬場 1–1–1

図 1．e-Stat.

る手法としては，様々なものがあるが，筆者は，Windows Subsystem for Linux（WSL），Docker Google-chrome, RSelenium パッケージの組合せで，Windows PC 上で実践している．メッシュ統計だけでなく，他のデータの web スクレイピングにも応用可能であろう．

3. 大学院における公的統計調査個票データを利用した研究指導

教育の実践例からは外れるが，関連することとして紹介する．滋賀大学においては，地方公共団体の職員が，社会人派遣枠で，このところ数名，大学院博士前期過程に入学している．筆者もこれまで 2 人の研究指導を担当した．そのうちの 1 人は，統計法第 33 条第 1 項の規定に基づき，調査票情報のオンサイト利用を活用して，研究を行い論文を執筆することができた．統計法第 33 条第 1 項の規定に基づいた調査票情報の利用は，大学院生が自治体職員であったため可能となった特殊な例ではあるが，今後は，政府が推進している，根拠に基づいた政策立案と訳される Evidence-Based Policy Making（以下，EBPM という. 行政改革推進本部, 2024 も参照）に関連する分析を含めてデータ分析を実践できる公務分野の職員の養成がさらに求められると予想されるところであり，このような形態の研究指導の需要も増加するのではないかと考えている．e-Stat からもたどることができる miripo と名付けられたミクロデータ利用ポータルサイト（2024）の情報や関連する研究会の情報に接していると，調査票情報の研究利用が推進されていると理解できるので，そのような方針とも合致すると考えている．（ただし，筆者の理解では，修士論文を執筆するためという理由だけでは，利用は承認されないので注意を.）

4. 公的統計調査匿名データを利用した統計教育，今後の課題

学部での演習といった，公的統計調査匿名データを利用した教育の意義やメリット，並びに演習環境の例については，佐藤（2023）に記したので参考にしていただきたい．また，全国でどのような教育実践例があるかは，上記で紹介した miripo（ミクロデータ利用ポータルサイト, 2024）において検索可能である．教育利用においては，筆者が利用承認を受けた経験では，総務省令に基づいた利用環境の整備が求められる．具体的には，分析に利用する PC がインターネットに接続されていないことが必要，データを USB に保存して教員が管理する必要，さらには，教員が管理する演習室や研究室を用意することが必要である．したがって，学生が自習しようとしても，できないのが現状である．教員が，分析例を示して，学生が再現実験するだけならばよいが，探究的に学生が試行錯誤しながら分析をすすめることをしようとすると，自習環境の整備も必要と考えている．具体的には，学内 LAN 上に，Linux サーバーを設置し，学生は閉じた演習室の環境において，USB を利用したシンクライアントを立ち上げ，バーチャル

デスクトップ環境で分析する自習環境が構築できないかと考えており，最近テストを開始したところである．

5. R survey パッケージを利用した匿名データの分析方法

5.1 R survey package 利用の意義

公的統計調査匿名データを集計するためには，一般には，抽出率を考慮した集計が必要となる．大学学部レベルの演習において，そのような集計を簡単に行うことができるソフトウエアはいろいろと考えられるが，経費を抑えるという観点からは，例えば，R を使うことも一案であり，R のパッケージ survey を利用することができる．もともと，survey パッケージは，標本抽出やそのためのシミュレーションなどを行うためのパッケージであることから，標本抽出における抽出率を考慮した分析機能（クロス集計や多変量解析，検定など）も実装されている．詳細は Lumley et al.（2024）を参照．ところが，既存の検索システムや，国立情報学研究所（CiNii research）において，「R survey package パッケージ」と検索してもたところ，survey パッケージを活用した分析方法一般を解説した記事は見当たらなかったので，ここに当該パッケージの分析機能の利用方法の概要を解説し，参考にしてもらうこととしたい．参考に，他のソフトウエアの例を記しておくと，SAS ソフトウエアにおいては，weight statement を指定することが可能なプロシージャーでは，抽出率を考慮した集計が可能である．

5.2 データの準備

申請手続きをすることなく入手できる「一般用ミクロデータ」を用いて解説することとする．そのために，独立行政法人統計センター（2024）から，利用規約に同意し，簡単なアンケートに答えて，平成 21 年全国消費実態調査（詳細品目）ippan_2009zensho_s.zip をダウンロードし，zip ファイルを解凍する．その中には，コードブックとも呼ばれる符号表（ippan_2009zensho_s.xls）も含まれており，符号表は 2 つのシートから構成されている．全国消費実態調査についての情報は，総務省統計局（2024）を参照されたい．以下，ippan_2009zensho_s_dataset.csv を用いて，survey パッケージの利用方法を解説していく．

5.3 データの読み込み

```
#R version 4.3.2 (2023-10-31 ucrt) -- "Eye Holes"にて動作確認
#RStudio 2023.12.0 Build 369
#survey_4.2-1.zip
setwd("c:/data/pubstat") #各自の環境に応じてフォルダ(ディレクトリ)を指定
library(tidyverse)
library(survey)

#符号表を参照して，明示的に，f 指定，d 指定を，430 の変数にそれぞれ割り当てる
num_of_factor <- "f"
for( i in seq(6)){
  num_of_factor<-paste(num_of_factor,"f" ,sep = "")
}
str_count(num_of_factor,"f")

num_of_numeric <- "d"
for( i in seq(422)){
```

```
  num_of_numeric<-
    paste(num_of_numeric,"d" ,sep = "")
}
str_count(num_of_numeric,"d")

col_format<-
  paste(num_of_factor,num_of_numeric,sep= "")
col_format
nchar(col_format)
```

出力結果例 1

```
> nchar(col_format)
[1] 430
```

430 の変数に対応して，フォーマット指定の文字列の文字数が 430 となっていることが確認できる．

```
#emEditor などを使って，一般ミクロデータのヘッダーを確認の上，読み込み
zensho_s2009 <-
  read_csv("ippan_2009zensho_s_dataset.csv",
           col_types =col_format,
           skip=8
  )
dim(zensho_s2009)
str(zensho_s2009)
#R studio おいて，430 個の変数について，意図どおり型が指定されているか確認
View(zensho_s2009)
#出力結果は，図 2 のとおりである．
```

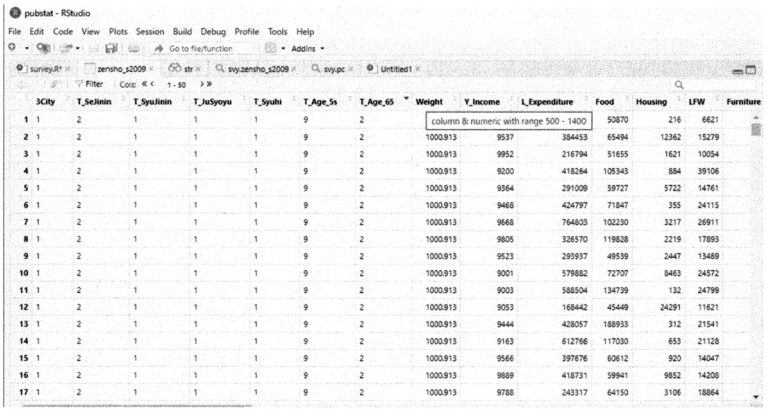
出力結果例 2

```
> dim(zensho_s2009)
[1] 45811   430
```

上記で示した符号表の 1 枚目のシート「符号表」に変数の数が 430 と記してあり，2 枚目のシート「基本数」というシートに，オブザベーション数が 45,811 と記してあるので，R の出力結果と照合し確認するとよい．なお，変数の名称は，csv ファイルのヘッダーにも記されており，符号表にも記されているので，R の実行例では，その名称をそのまま用いている．

5.4 surveydesign オブジェクトの作成

survey パッケージを利用して，分析機能を利用するためには，まず，surveydesign 型のオブジェクトを作成する必要がある．surveydesign() 関数を用いてオブジェクトを作成することになり，その際の引数としては，標本抽出を行うわけではないので，層が一つしかないと考えて ids=~1 と指定し，加えて抽出率の逆数(重み，weight)が格納されている変数の名称を指定する



	3City	T_SeJinin	T_SyuJinin	T_JuSyoyu	T_Syuhi	T_Age_5s	T_Age_65	Weight	Y_Income	L_Expenditure	Food	Housing	LFW	Furniture
1	1	2	1	1	1	9	2	1000.913	9537	384453	65494	12362	15279	
2	1	2	1	1	1	9	2	1000.913	9952	216794	51655	1621	10054	
3	1	2	1	1	1	9	2	1000.913	9200	418264	105343	884	39106	
4	1	2	1	1	1	9	2	1000.913	9364	291009	59727	5722	14761	
5	1	2	1	1	1	9	2	1000.913	9468	424797	71847	355	24115	
6	1	2	1	1	1	9	2	1000.913	9668	764803	102230	3217	26911	
7	1	2	1	1	1	9	2	1000.913	9805	326570	119828	2219	17893	
8	1	2	1	1	1	9	2	1000.913	9523	293937	49539	2447	13489	
9	1	2	1	1	1	9	2	1000.913	9001	579882	72707	8463	24572	
10	1	2	1	1	1	9	2	1000.913	9003	586504	134739	132	24799	
11	1	2	1	1	1	9	2	1000.913	9053	168442	45449	24291	11621	
12	1	2	1	1	1	9	2	1000.913	9444	428057	188933	312	21541	
13	1	2	1	1	1	9	2	1000.913	9163	612766	117030	653	21128	
14	1	2	1	1	1	9	2	1000.913	9566	397676	60612	920	14047	
15	1	2	1	1	1	9	2	1000.913	9889	418731	59941	9852	14208	
16	1	2	1	1	1	9	2	1000.913	9788	243317	64150	3106	18864	
17	1	2	1	1	1	9	2	1000.913						

図 2. 出力結果例 3. Rstudio において，ポップアップ機能で変数の型を確認.

ことがポイントとなる．下記の例では，svy.zensho_s2009 が surveydesign 型オブジェクトであり，このオブジェクトを survey パッケージに含まれる分析用の各種関数に渡して分析を実行していくこととなる．各関数を利用する場合は，重みとなる変数を明示的に指定する必要がなくなるメリットがある．

```
svy.zensho_s2009 <-
  svydesign(ids=~1,data=zensho_s2009,weights=zensho_s2009$Weight)
class(svy.zensho_s2009)#型を確認
```

出力結果例 4

```
> class(svy.zensho_s2009)
[1] "survey.design2" "survey.design"
```

surveydesign 型オブジェクトであることが示される．

5.5 属性別の分布確認，基本数の確認

符号表ファイルの基本数シートの数値と一致するか確認するために svytable() 関数を利用する．以下の実行例では，引数として，年齢 5 歳階級を示す属性変数「T_Age_5s」の名称が「~」（チルダー）をつけて渡され，次に上記で作成した surveydesign 型のオブジェクト名が渡されている．svytable() 関数では，属性変数を複数指定して，クロス集計を行うこともできる．ここでは，基本数シートに掲載されている重み考慮ありの総世帯数 31,761,998 に一致することが確認できる．

```
svytable(~T_Age_5s,design=svy.zensho_s2009)
sum(svytable(~T_Age_5s,design=svy.zensho_s2009))
#クロス集計の例
svytable(~T_Age_5s +T_Syuhi,design=svy.zensho_s2009)
```

出力結果例 5　2 次元クロス集計表

```
> svytable(~T_Age_5s +T_Syuhi,design=svy.zensho_s2009)
        T_Syuhi
T_Age_5s       1       2
       1  758995       0
       2 1859482       0
       3 2779608       0
       4 2923673       0
       5 2878926       0
       6 3039765       0
       7 3424996       0
       8 2936320       0
       9 3101564       0
       0       0 8058669
```

基本数シートに掲載されている重み(ウエイト)考慮ありの場合の属性別世帯数に一致していることが確認できる．

クロス集計を行う場合，上記の例では，集計値として世帯数をカウントしていたが，集計値を平均値としたい場合もある．その場合は，svyby() 関数を利用するのも一案である．集計値を平均値としたい場合は，下記の例のように引数として「svymean」を指定する．引数として，集計対象となる変数名や属性変数名を渡す場合の記述方法は，これまでの実行例から類推できる．下記の実行例で出力されたクロス集計表の値は，基本数のシートにおいて示されている数値と一致する．

```
#年齢 5 歳階級別収入平均値
svyby(~Y_Income,
      ~T_Age_5s,
      svy.zensho_s2009,
      svymean
)
#年齢 5 歳階級別(収入平均値，消費支出平均値)
svyby(~Y_Income+L_Expenditure,
      ~T_Age_5s,
      svy.zensho_s2009,
      svymean
)
#年齢 5 歳階級別就業状態別(収入平均値，消費支出平均値)
svyby(~Y_Income+L_Expenditure,
      ~T_Age_5s+T_Syuhi,
      svy.zensho_s2009,
      svymean
)
```

5.6 ヒストグラムの作成

最初に，重みの分布を通常の hist() を使ってみて以下の実行例でみてみると，ばらつきがあることがわかる．

```
hist(zensho_s2009$Weight,plot=FALSE)
```

次に，年間収入を表している数値変数「Y_Income」についてヒストグラムを図示する実行例は以下のとおりとなる．この関数でも，変数名の前に「~」（チルダー）をつけることがポイントとなる．Lumley et al.（2024）を参照するとわかるとおり，残念ながら，「plot=FALSE」オプションはサポートしていない．

```
#ヒストグラムを図示する場合の例
svyhist(~Y_Income, svy.zensho_s2009, main="Survey weighted",
freq=TRUE,col="purple",xlim=c(0,5e4),ylim=c(0,3.0e7))#明示的に freq 指定が必要
```

5.7 箱ひげ図の作成

箱ひげ図を描く関数は svyboxplot() であるが，数値変数を指定した上で，属性別に一挙に箱ひげ図を描くことができる仕様となっているので，数値変数「Y_Income」に「~」（チルダー）をつけて属性変数を指定する．属性別の作成が必要でない場合は，「1」を指定すればよい．

```
#箱ひげ図
svyboxplot(Y_Income~1,svy.zensho_s2009,all.outliers=TRUE)
#属性別箱ひげ図
svyboxplot(Y_Income~T_Age_5s,svy.zensho_s2009,all.outliers=TRUE)
```

5.8 四分位数の算出

四分位数の算出は，以下の関数を利用する．引数の渡し方は，今までの例から簡単に類推できるかと思う．

```
#四分位階級
svyquantile(~Y_Income,svy.zensho_s2009, c(.25,.5,.75))
```

出力結果例 6

```
> #四分位階級
> svyquantile(~Y_Income,svy.zensho_s2009, c(.25,.5,.75))
$Y_Income
     quantile ci.2.5 ci.97.5      se
0.25     3800   3772    3829 14.5407
0.5      5508   5472    5544 18.3672
0.75     8003   7946    8071 31.8875

> income_cdf<-svycdf(~Y_Income,svy.zensho_s2009)
#5508 の位置を確認すると中位数であることがわかる
> income_cdf[[1]](5508)
[1] 0.5000038
```

この関数は面白い使い方ができる．例えば，年間収入が 6000 千円の場合，何パーセント点にあたるのか，教えてくれる．実行例は以下のとおりである．

```
income_cdf<-svycdf(~Y_Income,svy.zensho_s2009)
income_cdf
income_cdf[[1]](6000)
```

5.9 回帰分析

回帰分析を行うためには，svyglm() 関数を利用するが，通常の glm() 関数と同様に引数を渡せばよく，引数「family=""」も有効である．疑似的にオブザベーションが作成されている一般ミクロデータを利用しているため，下記の例では，回帰分析の結果はあくまで実行例である．

```
reg_income<-
  svyglm(Y_Income~Food, design=svy.zensho_s2009)
summary(reg_income)
library(performance)
performance(reg_income)#評価指標を出力
```

5.10 主成分分析

主成分分析を行うためには，svyprcomp() 関数を利用するが，通常の prcomp() 関数と同様に引数を渡せばよい．関数より作成されるオブジェクトも通常の prcomp() の場合と同様に扱えばよい．class() 関数を使えば，そのことがわかる．なお，疑似的にオブザベーションが作成されている一般ミクロデータを利用しているため，下記の例では，統計的な意味はあまりなく，結果はあくまで実行例である．

```
svy.pc<- svyprcomp(
  ~Y_Income+L_Expenditure+Food+Housing+LFW+Furniture+Clothes+Health
  +Transport+Education+Recreation+OL_Expenditure,
  design=svy.zensho_s2009,scale=TRUE,scores=TRUE)
class(svy.pc)
svy.pc$rotation#主成分負荷量を表示
svy.pc$x#主成分得点を表示
dim(svy.pc$x)
svy.pc$rotation[,1]#PC1 主成分負荷量を表示
round(sum(svy.pc$rotation[,1]*svy.pc$rotation[,2]),1)#内積により
直交していることの確認

#累積寄与率計算など
vars <- svy.pc$sdev^2
pov <-vars/sum(vars)
cpov <- cumsum(pov)
cpov#累積寄与率の表示
screeplot(x=svy.pc,type="lines",main="Scree Plot",npcs=12)
#npcs を明示しないと 10 成分分しか表示されない
```

出力結果例 7
```
#主成分分析の結果の一部
> dim(svy.pc$x)
[1] 45811    12

> round(sum(svy.pc$rotation[,1]*svy.pc$rotation[,2]),1)
#内積により直交していることの確認
[1] 0

> #累積寄与率計算など
> vars <- svy.pc$sdev^2
> pov <-vars/sum(vars)
> cpov <- cumsum(pov)
> cpov
 [1] 0.2792098 0.3799106 0.4724950 0.5539764 0.6318810 0.7077415
 0.7811314  0.8526595  0.9166016 0.9625722 1.0000000 1.0000000
#当然に 1 になる.
```

```
screeplot(x=svy.pc,type="lines",main="Scree Plot",npcs=12)
#出力結果は，図 3 のとおりである.
```

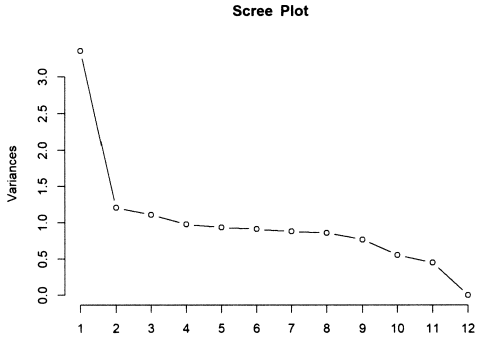


図 3. 出力結果例 8. Rstudio において Screeplot 図を確認.

参 考 文 献

独立行政法人統計センター (2024). 一般用ミクロデータの利用, https://www.nstac.go.jp/use/archives/ippan-microdata/（最終アクセス日 2024 年 4 月 30 日）.

行政改革推進本部 (2024). 政府の行政改革／ EBPM の推進, https://www.gyoukaku.go.jp/ebpm/index.html（最終アクセス日 2024 年 4 月 30 日）.

Lumley, T., Gao, P. and Schneider, B. (2024). Package 'survey' Analysis of Complex Survey Samples, https://cran.r-project.org/web/packages/survey/survey.pdf（最終アクセス日 2024 年 4 月 30 日）.

ミクロデータ利用ポータルサイト (2024). miripo, https://www.e-stat.go.jp/microdata/（最終アクセス日 2024 年 4 月 30 日）.

佐藤正昭 (2023). 公的統計調査匿名データを活用した教育活動の可能性について, 月刊「統計」, 2023 年 10 月号, 54–55.

政府統計の総合窓口 (2024). e-Stat, https://www.e-stat.go.jp/（最終アクセス日 2024 年 4 月 30 日）.

総務省統計局 (2024). 平成 21 年(2019 年)全国消費実態調査, https://www.stat.go.jp/data/zensho/2009/index.html（最終アクセス日 2024 年 4 月 30 日）.

Practical Examples of Education Using Official Statistics

Masaaki Sato

Faculty of Data Science, Shiga University

This paper introduces a component of statistical education using official statistics and explains the method of analysis using the R survey package, which is necessary for that purpose.

Key words: Microdata, anonymous data from official statistical surveys, statistical education, survey package.