DOI:10.13546/j.cnki.tjyjc.2017.22.008

基于R软件的分层随机抽样方法

卢玉桂,黄基廷

(河池学院 数学与统计学院,广西 宜州 546300)

摘 要:文章以分层随机抽样为例,介绍了在完整抽样框下,运用R软件完成样本的抽取与总体参数的估计的方法,以及在仅样本数据的情况下(非完整抽样框),如何运用R软件完成总体参数的估计。

关键词:R软件;分层随机抽样;完整抽样框;非完整抽样框

中图分类号:O212.4 **文献标识码**:A **文章编号**:1002-6487(2017)22-0036-04

0 引言

一般而言,抽样调查涉及两个重要的过程,即样本的选取与总体参数的估计。R软件提供了进行抽样的sampling包和对抽样结果进行估计的survey包,运用R软件可进行样本的选取与总体参数的估计,但目前关于如何运用R软件进行抽样和分析的介绍很少,特别是对于非完整抽样框下如何进行总体参数的估计更是无人提及,这不利于学生理解与掌握有关理论方法。因此,本文以分层随机抽样为例,介绍在完整抽样框下如何运用R软件实现样本选取与总体参数的估计,以及非完整抽样框下如何运用R软件实现总体参数的估计,使学生更好地理解与掌握分层随机抽样的有关理论及R软件实现。

1 分层随机抽样的相关理论

所谓分层随机抽样[1],就是将总体的 N 个单元按某个变量划分为"不重不漏"的 L 个子总体(层),并在每一层中独立地按简单随机抽样方法抽取样本,总的样本量 n 由各层样本组成,总体参数的估计值由各层样本参数加权汇总得到的抽样方法。

假设在分层随机抽样中,L 个子总体的单元总数依次记为 $N_1, N_2, \cdots, N_L$,从每层独立抽取的样本量依次记为 $n_1, n_2, \cdots, n_L$,第 h 层的 N_h 个总体指标值记为 $Y_{h1}, Y_{h2}, \cdots, Y_{hN_h}$,第 h 层的 n_h 个的样本指标值记为 $y_{h1}, y_{h2}, \cdots, y_{hn_h}$,则有第 h 层的层权 $W_h=\frac{N_h}{N}$,第 h 层的抽样比 $f_h=\frac{n_h}{N_h}$,第 h 层的总体均值 $\bar{Y}_h=\frac{1}{N_h}\sum_{i=1}^{N_h}Y_{hi}$,第 h 层的样本均值 $\bar{y}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi}$,第 h 层的总体方差和样本方差分别为 $S_h^2=\frac{1}{N_h-1}\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y}_h)^2$ 和 $s_h^2=\frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2$。

1.1 简单估计

由于分层随机抽样是各层都独立的按简单随机抽样抽取样本,所以第 h 层的样本均值 $\bar{y}_h$ 是其总体均值 $\bar{Y}_h$ 的无偏估计量。容易验证,分层随机抽样的简单估计 $\bar{y}_{st}=\sum_{h=1}^{L}W_h\bar{y}_h$ 是总体均值 $\bar{Y}$ 的无偏估计量。无偏性是衡量估计量优劣的一个标准,但不是唯一标准,一般还需要考虑进度的高低,精度通常用估计量的标准差来衡量。估计量 $\bar{y}_{st}$ 方差 $V(\bar{y}_{st})$ 的无偏估计为:

$$v(\bar{y}_{st})=\sum_{h=1}^{L}W_h\frac{1-f_h}{n_h}s_h^2 \quad (1)$$

估计量 $\bar{y}_{st}$ 的标准差为:

$$s(\bar{y}_{st})=\sqrt{v(\bar{y}_{st})}=\sqrt{\sum_{h=1}^{L}W_h\frac{1-f_h}{n_h}s_h^2} \quad (2)$$

由式(2)可知,分层随机抽样的精度与层内方差 s_h^2 的大小有关,因而在选择分层变量时应选取层内差异小,层间差异大的变量,从而提高抽样估计的精度。

在进行分层随机抽样调查时,若存在与调查的主要变量 Y 高度相关的辅助变量 X,利用辅助变量 X 的信息将有利于提高抽样估计的精度。借助辅助变量 X 进行参数估计时的方法有比估计和回归估计两种方式。

1.2 比估计

分层随机抽样的比估计包括分别比估计和联合比估计两种,其中,分别比估计是先计算各层的比估计量 R_h,然后再进行加权平均;而联合比估计则是先对比率的分子、分母进行加权平均,然后构造比估计。一般情况下,在各层样本量 n_h 都比较大时,常用分别比估计进行估计。反之,在总样本量 n 比较大时,则用联合比估计进行估计。

假设 X 为与主要变量 Y 高度相关的辅助变量,$X_1, X_2, \cdots, X_L$ 为辅助变量 X 的 L 个总体总值,$x_{h1}, x_{h2}, \cdots, x_{hn_h}$ 为

基金项目:广西高校中青年教师基础能力提升项目(KY2016LX279);河池学院课程教学模式改革项目(2015KTJY11)

作者简介:卢玉桂(1988—),女,广西崇左人,硕士,讲师,研究方向:抽样调查与数据分析。

(通讯作者)黄基廷(1964—),男,广西天等人,副教授,研究方向:应用数学。

第 h 层辅助变量 X_h 的 n_h 个样本指标值，则第 h 层辅助变量的总体均值 $\bar{X}_h=\frac{X_h}{N_h}$，第 h 层辅助变量的样本均值 $\bar{x}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}x_{hi}$，第 h 层辅助变量的样本方差为 $s_{xh}^2=\frac{1}{n_h-1}\sum_{i=1}^{n_h}(x_{hi}-\bar{x}_h)^2$，第 h 层的样本协方差为 $s_{xyh}=\frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y})(x_{hi}-\bar{x})$，以及均值 $\bar{X}$ 的分层简单估计为 $\bar{x}_{st}=\sum_{h=1}^{L}W_h\bar{x}_h$。

在各层样本量 n_h 都比较大时，第 h 层比率 R_h 的估计量 $\hat{R}_h=\bar{y}_h/\bar{x}_h$ 近似无偏，则总体均值 $\bar{Y}$ 的分别比估计 $\bar{y}_{RS}$ 近似无偏，并有：

$$\bar{y}_{RS}=\sum_{h=1}^{L}W_h\hat{R}_h\bar{X}_h=\sum_{h=1}^{L}W_h\frac{\bar{y}_h}{\bar{x}_h}\bar{X}_h$$

方差 $V(\bar{y}_{RS})$ 的估计为：

$$v(\bar{y}_{RS})\approx\sum_{h=1}^{L}\frac{W_h^2(1-f_h)}{n_h}(s_h^2+\hat{R}_h^2s_{xh}^2-2\hat{R}_hs_{yxh}) \quad (3)$$

在总样本量 n 比较大时，比率 R 的估计量 $\hat{R}_C=\bar{y}_{st}/\bar{x}_{st}$ 近似无偏，则总体均值 $\bar{Y}$ 的联合比估计 $\bar{y}_{RC}$ 近似无偏，并有：

$$\bar{y}_{RC}=\hat{R}_C\bar{X}=\frac{\bar{y}_{st}}{\bar{x}_{st}}\bar{X}$$

方差 $V(\bar{y}_{RC})$ 的估计为：

$$v(\bar{y}_{RC})\approx\sum_{h=1}^{L}\frac{W_h^2(1-f_h)}{n_h}(s_h^2+\hat{R}_C^2s_{xh}^2-2\hat{R}_Cs_{yxh}) \quad (4)$$

2 完整抽样框下样本选取与估计

在实施调查前，应先编制抽样总体的抽样框，然后从该抽样框中按某种抽样方法抽取样本单元，完成样本单元的抽取之后，才能开展进一步的抽样调查工作。对于完整抽样框下的抽样调查（即抽样总体的所有指标值已经得到），R 软件提供了进行样本选取与总体参数估计的软件包。本文以《中国统计年鉴 2014》中分地区的城镇居民人均消费支出和人均可支配收入数据为例，介绍如何运用 R 软件实现完整抽样框下分层随机抽样的样本选取与估计。

例 1：为调查 2013 年我国城镇居民人均消费支出，以全国 31 个省（市、自治区）为抽样单元，并根据《中国统计年鉴 2014》将其划分为东部、中部、西部和东北地区 4 层，数据集中包含编号、地区、层、人均消费支出（y）、人均可支配收入（x）等变量，具体调查数据见表 1。现用 R 软件从 4 个层中独立的按简单随机抽样抽取各抽取 2 个样本单元，并求各地区人均消费支出均值的估计。即有 $N=31, n=8, h=4$。

由于 R 软件不能直接读取 Excel 表，故要先将其转换为其他格式的文件，如转换为“CSV（逗号分隔）”文件。具体方法为：打开 example1.xls，并点击文件 → 另存为，保存类型选择 CSV（逗号分隔），文件名为 example1，完成数据类型转换，并得到一个新的文件 example1.csv。

表 1　　2013 年我国分地区城镇居民人均消费支出调查数据

编号	地区	层	y	x	编号	地区	层	y	x
1	北京	1	26274.89	40321	17	内蒙古	3	19249.06	25496.67
2	天津	1	21711.86	32293.57	18	广西	3	15417.62	23305.38
3	河北	1	13640.58	22580.35	19	重庆	3	17813.86	25216.13
4	上海	1	28155	43851.36	20	四川	3	16343.45	22367.63
5	江苏	1	20371.48	32537.53	21	贵州	3	13702.87	20667.07
6	浙江	1	23257.19	37850.84	22	云南	3	15156.15	23235.53
7	福建	1	20092.72	30816.37	23	西藏	3	12231.86	20023.35
8	山东	1	17112.24	28264.1	24	陕西	3	16679.69	22858.37
9	广东	1	24133.26	33090.05	25	甘肃	3	14020.72	18964.78
10	海南	1	15593.04	22928.9	26	青海	3	13539.5	19498.54
11	山西	2	13166.19	22455.63	27	宁夏	3	15321.1	21833.33
12	安徽	2	16285.17	23114.22	28	新疆	3	15206.16	19873.77
13	江西	2	13850.51	21872.68	29	辽宁	4	18029.65	25578.17
14	河南	2	14821.98	22398.03	30	吉林	4	15932.31	22274.6
15	湖北	2	15749.5	22906.42	31	黑龙江	4	14161.71	19596.96
16	湖南	2	15887.11	23413.99					

打开 R 软件，点击文件 → 改变工作目录，找到存放 example1.csv 文件的工作目录，点击确定即可。完成改变工作目录的工作之后，点击文件 → 新建程序脚本 → 保存，打开一个新的 R 程序编辑器，并将其保存为 example1.R 文件。完成这一系列工作之后，开始编写程序。具体步骤如下：

（1）准备工作。将 sampling 包和 survey 包下载并加载到 R 软件中，其中，sampling 包用于抽取样本，survey 包用于估计（注意，运行 install.packages()时，会弹出一个对话框，此处选择 China(Beijing4)[https]映像进行下载）。除了此之外，还需加载一个基础包 grid。R 程序如下：

```
>install.packages('sampling')
>library(sampling)
>install.packages('survey')
> library(survey)
>library(grid)
```

（2）抽样。将总体数据文件 example1.csv 导入 R 软件，并完成变量总体单元数 N、各层单元总数 Nh、各层的层权 Wh、层数 L 和各层抽取的样本量 nh 的定义。然后，调用分层抽样函数“strata”抽取样本（注意，strata 函数包含四个参数，其含义依次为总体的数据集、分层变量、各层样本量和各层抽样所用的抽样方法）。最后，调用 getdata 从总体数据集 example 中提取样本单元，完成样本单元的抽取。R 程序如下：

```
>example=read.csv("example1.csv")
>N=nrow(data)
>Nh=table(example$层);
>Wh=Nh/N
>L=4
>nh=rep(2,L)
>st=strata(example,"层",nh,"srswor");
>example.st=getdata(example,st)
```

（3）抽样设计。先定义每个样本单元的权重变量 pw

和fpc变量，其中，pw变量为各样本单元进入样本概率的倒数，fpc变量为每个单元所在层的单元总数Nh。然后，将变量pw和fpc加入样本数据集example.st中。此时可调用svydesign函数定义抽样设计dst，其中，参数id用“~0”表示（不包含群变量）；参数strata定义分层变量；weights定义为权重变量；参数data定义样本数据集；参数fpc定义fpc变量。R程序如下：

```
>pw=1/st$Prob
>fpc=as.numeric(table(example$层)[example.st$层])
>agst=as.data.frame(cbind(example.st,pw,fpc))
>dst<-svydesign(id=~1,strata=~层,weights=~pw,data=agst,fpc=~fpc)
>summary(dst)
```

（4）简单估计。完成以上抽样工作后，可调用svymean计算目标变量y均值的简单估计和标准误。R程序如下：

```
>svymean(~y,dst,deff=TRUE)
```

（5）分别比估计。先调用svyratio函数计算各层的比率估计rsh，其中，四个参数的依次为目标变量y、辅助变量x、抽样设计dst和separate=TRUE（取值TRUE表示分别比估计），然后计算辅助变量x的各层总体均值，最后调用predict函数对目标变量y的均值进行估计，其中，第一个参数为各层的比率估计rsh，第二个参数为辅助变量x的各层总体均值乘以各层的层权。R程序如下：

```
>rsh<-svyratio(~y,~x,dst,separate=TRUE)
>ymean<-data.frame(x=tapply(example$x,INDEX=example$层,FUN=mean))
>predict(rsh,ymean$x*Wh)
```

（6）联合比估计。先调用svyratio函数计算总比率rc，其中，前三个参数与分别比率估计的定义一致，区别在于separate为缺省值FALSE（（取值FALSE表示联合比估计））。然后计算辅助变量x的总体均值。最后，调用predict函数进行估计，其两个参数依次为总比率rc和辅助变量x的总体均值。R程序如下：

```
>rc<-svyratio(~y,~x,dst)
>ymean<-data.frame(x=mean(example$x))
>predict(rc,ymean$x)
```

运行以上R程序代码，即可完成完整抽样框下分层随机抽样的样本抽取和总体参数的估计。运行一次所抽得的8个样本单元分别为层1的福建和海南、层2的山西和江西、层3的内蒙古和甘肃、层4的辽宁和吉林。与之对应的我国城镇居民人均消费支出的简单估计、分别比率估计和联合比率估计结果见表2。真值为17190.595。

表2　　估计结果

估计方法	估计值	标准差的估计
简单估计	16452.9	1131.8
分别比估计	27169.68	255.4094
联合比估计	17591.47	219.4147

由表2可知，联合比估计的估计值与真值17190.6最为接近，其估计精度也是最高的，这主要是因为总体样本量8相对于总体规模31来说足够大；简单估计的估计精度最差，这主要是因为简单估计没有利用辅助变量的信息；而分别比估计虽然借助辅助变量进行估计，但是在各层样本量较小时（例1中各层样本量仅为2），其偏倚较大。

3　非完整抽样框下目标变量的估计

在一般的抽样调查工作中，调查所得到的数据仅为样本数据，不是完整的总体数据（全面抽样、普查除外），即在实施抽样调查时，通常的做法是编制总体的抽样框后，运用R软件抽取样本单元，然后对被抽中的总体单元进行调查，从而获取样本单元的指标值。此时，不能用例1的方法对目标变量的总体参数进行估计，为说明如何运用R软件对这类数据完成参数估计，将引用文献[1]的例子进行简单介绍。

例2：为调查某公司职员的平均工资$\bar{Y}$，将该公司的职员分为一般职员和高级管理人员两层，其中，一般职员总数$N_1=390$，高级管理人员总数$N_2=84$。现已知一般职员和高级管理人员刚入职时的工资总额分别为$X_1=5523965$和$X_2=2541660$。运用分层随机抽样分别从一般职员层和高级管理人员层分别抽取$n_1=15$和$n_2=10$名职员进行调查。具体数据见表3。

表3　　某公司工资收入调查数据

Id	层	x	y	Id	层	x	y	Id	层	x	y
1	1	8000	19200	10	1	10200	27450	19	2	29400	80250
2	1	16500	34800	11	1	15000	35750	20	2	37500	87500
3	1	10500	22950	12	1	15750	36750	21	2	23250	49625
4	1	15550	34400	13	1	13750	33250	22	2	31500	79500
5	1	18000	38350	14	1	21240	44875	23	2	34200	83500
6	1	16500	37800	15	1	10750	21300	24	2	33000	82250
7	1	14250	31650	16	2	26250	64750	25	2	35540	85500
8	1	13500	32550	17	2	22750	46000				
9	1	15000	36050	18	2	20000	41500				

类似于例1，将表3的数据保存在example2.csv中，并利用R软件完成该公司职员的平均工资的估计（包括简单估计、分别比估计和联合比估计）。具体如下：

（1）准备工作。由于此时不需要抽取样本，因而只需下载并加载survey包和加载grid基础包，然后将样本数据文件example2.csv导入R软件，并完成总体单元数N、各层单元总数Nh、层权Wh、辅助变量X各层总体总量Xh、辅助变量X各层总体均值Xhmean的定义工作。R程序如下：

```
>install.packages('survey')
>library(survey)
>library(grid)
>example=read.csv("example2.csv")
>N=474
>Nh=c(390,84)
>Wh=Nh/N
>Xh=c(5523965,2541660)
>Xhmean=Xh/Nh
```

（2）抽样设计。定义样本权重变量pw和fpc变量，并将其加入样本数据集example，得到新的数据集agst。调用svydesign函数完成分层抽样设计dst的定义。（注意，尽管不需要抽取样本，但是，依然需要定义抽样设计，否则无法完成后面的估计）。R程序如下：

```
>pw=c(rep(390/15,15),rep(84/10,10))
>fpc=c(rep(390,15),rep(84,10))
>agst=as.data.frame(cbind(example,pw,fpc))
>dst<-svydesign(id=~1,strata=~层,weights=~pw,data=agst,fpc=~fpc)
```

（3）简单估计。调用svymean计算目标变量y均值的简单估计和标准误。R程序如下：

```
>svymean(~y,dst,deff=TRUE)
```

（4）分别比估计。先调用svyratio函数计算各层的比率估计rsh，然后定义辅助变量x的各层总体均值，最后调用predict函数对目标变量y的均值进行估计。各参数的含义与例1类似。R程序如下：

```
>rsh<-svyratio(~y,~x,dst,separate=TRUE)
>ymean<-data.frame(x=Xhmean)
>predict(rsh,ymean$x*Wh)
```

（5）联合比估计。先调用svyratio函数计算总比率rc，然后定义辅助变量x的总体均值。最后调用predict函数对目标变量y的均值进行估计。各参数的含义与例1类似。R程序如下：

```
>rc<-svyratio(~y,~x,dst)
>ymean<-data.frame(x=sum(Xh)/N)
>predict(rc,ymean$x)
```

运行以上R程序代码后，完成非完整抽样框下分层随机抽样总体参数的估计，获得公司职员平均工资的简单估计、分别比估计和联合比估计结果见表4。

表4　　估计结果

估计方法	估计值	标准差的估计
简单估计	39131.6	1737.6
分别比估计	39267.5	558.2199
联合比估计	39250.43	576.0907

由表4可知，分别比估计与联合比估计的效果基本一致，简单估计的效果最差。

4　结束语

尽管SPSS软件有专门用于抽样的模块，实施起来也十分方便，但只能满足基本抽样方法，此外，SPSS软件还涉及软件收费和版权问题，这限制了其使用范围。R软件作为一款完全免费和公开的统计分析软件，它提供了大量统计分析模块，其中就有用于抽样和估计的sampling包和survey包，我们只需要在现有抽样模块的基础上编写新的抽样模块，就可以快速地完成随机样本的抽取和总体参数的估计。因此，在抽样调查过程中，应注意R软件在实际案例中的灵活运用，做到理论与实际相结合。

参考文献：

[1]金勇进等.抽样技术(第四版)[M].北京:中国人民大学出版社,2015.

[2]薛毅,陈立萍.统计建模与R软件[M].北京:清华大学出版社,2006.

[3]汤银才.R语言与统计分析[M].北京: 高等教育出版社,2005.

[4]朱春华. 如何运用SPSS软件进行简单随机抽样[J]. 统计与决策,2014,(9).

（责任编辑/亦　民）

Study on Stratified Random Sampling Based on R Software

Lu Yugui, Huang Jiting

(Department of Mathematics and Statistics, Hechi University, Yizhou Guangxi 546300, China)

Abstract: This paper takes the stratified random sampling as an example, and proposes a method for sample extraction and population parameter estimation by using R software under the frame of complete sampling. And the paper also introduces how to use R software to complete the population parameter estimation in the case of only sample data (non-complete sampling frame).

Key words: R software; stratified random sampling; complete sampling frame; non-complete sampling frame