Simple Random Sampling, Proportional Stratified Sampling and One Stage Cluster Sampling

Jose C. Torres Hernandez

December 7, 2018

1 Introduction of the data

The data set for the study is the Academic Performance Index for California schools. This index is based on standardized tests. More specifically the Standardized Testing and Reporting (STAR) Program, the California High School Exit Examination (CAHSEE) and the California Alternate Performance Assessment (CAPA). Also the index is computed for schools with at least 100 students. The entire data is composed of 6194 observation with 37 variables. The variables of interest is meals which is the percentage of students eligible for subsidized **meals** in a given school.

1.1 Motivation(Purpose)

Three sampling techniques will be used to compute the mean percentage of subsidized meals along with its standard error for California schools. The results will help determine how well the apipop survey reflects california schools

My real motivation for choosing this project was to learn new statistic topics.

1.2 Background

Three types of sampling methods will be used to estimate the mean percentage of subsidized meals in California. The population size is 6184 California schools. These cover 757 schools districts. We will be selecting samples from this population using Simple Random Sampling, Proportional Stratified Sampling and One Stage Cluster Sampling. These three sampling techniques are based on randomization. That means every element in the population has an equal probability of getting picked.

Simple random sampling is used if there is no prior knowledge about the population. This definitely the case here. In Stratified sampling we divide the population into three strata/groups Elementary, Middle and High Schools. Then we can do randomly sample these three strata/groups. In cluster sampling the groups are based on School Districts. There are 757 total districts in the population (entire data set). When we randomly sample 15 districts will be chosen.

2 Data Preparation

2.1 for Simple Random Sampling

No data preparation was needed to be done with the *apipop* data frame.

2.2 for Proportional Stratified Sampling

First we create a new variable in the *apipop* called N for the stratum size. The three strata are Elementary, Middle and High School with sizes 4421, 1018 and 755 respectively.

```
apipop$N <- NA

apipop$N[apipop$stype=='E'] <- length(apipop$sname[apipop$stype=='E'])
apipop$N[apipop$stype=='M'] <- length(apipop$sname[apipop$stype=='M'])
apipop$N[apipop$stype=='H'] <- length(apipop$sname[apipop$stype=='H'])
```

Since we are doing proportional stratified sampling we need to find what proportion of schools are Elementary, Middle and High Schools.

```
props= table(apipop$stype)/length(apipop$stype); props
round(props*100)
```

These values were multiplied by 100 and rounded to a whole numbers 71, 16 and 12 respectively. Each stratified sample done will include resulting in 71 elementary schools, 16 middle schools and 12 high schools.

```
round(props*100)
```

The *apipop* was reorder to make it easier to sample each strata correctly.

```

elem_data = apipop[apipop$stype=='E',]
midd_data = apipop[apipop$stype=='M',]
high_data = apipop[apipop$stype=='H',]
```

The new data frame is called *apipop_reorder* and rows 1 to 4421 are elementary schools, 4422 to 5439 are Middle Schools and 5440 to 6194 are High Schools.

```
apipop_reorder = rbind(elem_data, midd_data,high_data)
```

2.3 for One Stage Cluster Sampling

A new column is added to *apipop* called fpc which is the number of clusters in the population which is 757 clusters. This column is then given to the one-stage cluster sampling design function svydesign(). It is a survey design object that includes information about the design and the data. [3]

```
apipop$fpc <- length(unique(apipop$dname))
```

3 Estimations of Sample mean and its estimated Standard Error

3.1 Theoretical

The following equations are used to compute the theoretical values for mean and SE given a sampling technique.

3.1.1 Proportional Stratified Sampling

First of all the number 3 above the summation says we will have 3 strata (Elementary, Middle and High School). The means for for each strata are computed

```
mus=tapply(apipop$meals,apipop$stype ,mean)
```

along with the size N_i of each strata 4421 (Elemtary Schools), 1018 (Middle Schools) and 755 (High schools) which total to N= 6194

```
len_Es=length(apipop$sname[apipop$stype=='E'])
len_Ms=length(apipop$sname[apipop$stype=='M'])
len_Hs=length(apipop$sname[apipop$stype=='H'])
```

Now we can compute $\bar{y}_{st}$

```
ybar_str = (1/nrow(apipop))*( len_Es*mus[[1]] + len_Ms*mus[[3]]+ len_Hs*mus[[2]] )
```

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{3} N_i \bar{y}_i$$

In order to compute the variance of $\bar{y}_{st}$ the standard deviation s_i for each strata are needed and the n_is are the sample sizes 71, 16 and 12 as mentioned in section 2.2 .

```
SDs=tapply(apipop$meals,apipop$stype  ,sd)
```

Now the $\hat{V}(\bar{y}_{st})$ can be computed by

```
var_ybar_str = (1/nrow(apipop)^2)* ( ((len_Es^2)*(1- (71/len_Es))*((SDs[[1]]^2)/71))  + ((len_Ms^2)*(1- (16/len_Ms))*((SDs[[3]]^2)/16)) +  ((len_Hs^2)*(1- (12/len_Hs))*( (SDs[[2]]^2)/12)) )
```

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{3} N_i^2 \left(1 - \frac{n_i}{N_i}\right)\frac{s_i^2}{n_i}.$$

The theoretical values for $\bar{y}_{st}$ and $SD(\bar{y}_{st})$ are 48.03568 and 2.962979 .

3.1.2 One Stage Cluster Sampling

In order to calculate $\bar{\bar{y}}_c$ n is 757 which is the number of cluster (schools districts) in *apipop*. The $\bar{y}_i$ is the mean for each cluster/school district. Summing these and dividing by n=757 gives $\bar{\bar{y}}_c$.

```
ybarbar=sum(tapply(apipop$meals, apipop$dname, mean))/unique(apipop$fpc)
```

$$\bar{\bar{y}}_c = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

On to computing the variance of $\bar{\bar{y}}_c$ where N=6194 (population size), n=15 (cluster/districts sampled) and the rest of the terms $\bar{y}_i$, $\bar{\bar{y}}_c$ are already computed above.

```
S_bSquared=(1/( unique(apipop$fpc)-1))*sum( (tapply(apipop$meals, apipop$dname, mean) - ybarbar)^2 )

var_ybarbar=((nrow(apipop) - 15)/ (nrow(apipop)*15 )) * (S_bSquared)
sd_ybarbar = sqrt(var_ybarbar)
sd_ybarbar
```

$$\hat{V}(\bar{\bar{y}}_c) = \frac{N-n}{Nn}S_b^2 \quad where \quad S_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}}_c)^2$$

The resulting theoretical values are 42.1954 and 6.758054 for $\bar{\bar{y}}$ and $\hat{SD}(\bar{\bar{y}}_c)$ respectively.

3.2 Simulation

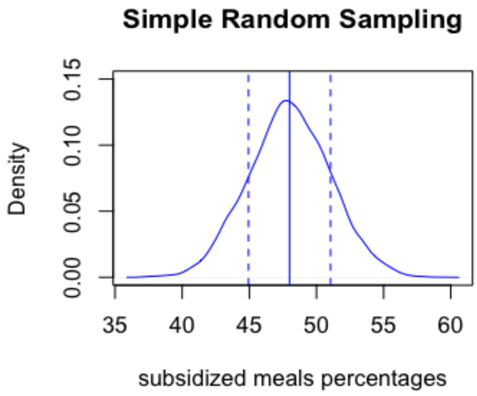
In every technique 10,000 sample simulations will be computed. The sampling will be done according to the sampling technique used.

3.2.1 Simple Random Sampling as comparison to above two

Simple random sampling with size of 99 (the props argument) is done for meals percentages. At every iteration a mean is computed and appended to an empty vector. Then a mean of those means is output along with a sd of all those means.

```
srsAPI= pop_srs(apipop, 10000, props)
```

The simulated values are $\bar{y}$47.99395 and SE for $\bar{y}$ is 3.053388 and both the values are comparable to the calculated theory above. In the below plot The solid line is the estimated μ and the dash lines are one SE from the mean.

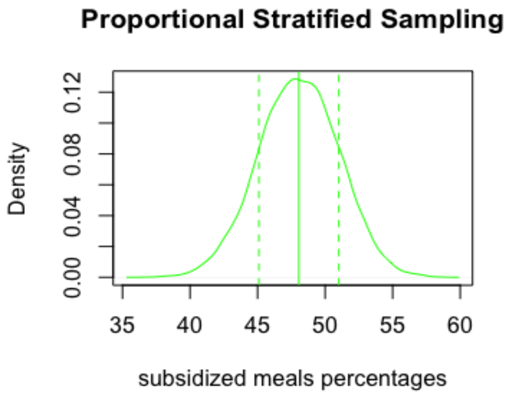


3.2.2 Proportional Stratified Sampling

In proportional stratified sampling the three strata are sampled according to their proportion in the total population. Then a survey design object that includes information about the design and the sample at each of the 10,000 simulations. After each sample a function in R is used to compute the estimated μ and the SE for Stratified sampling.

```
stratAPI= strat_sampl_pop(data=apipop_reorder , Nsim=10000, prop=props)
```

The simulated estimated μ and SE are 48.05325 and 2.959094. These values are similar to the computed theory values.

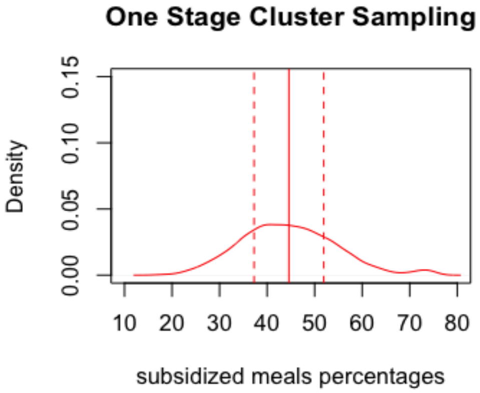


3.2.3 One Stage Cluster Sampling

In each simulation random 15 from 757 districts are sampled each time. The size of the sampling varies by cluster.

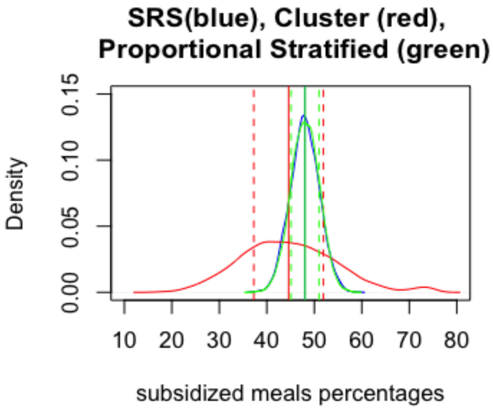
```
clust_samp_meals=clust_sampling(data= apipop , Nsim = 10000, clust_var ="dname", clust_size
    =15)
```

The simulated for μ and SE are 44.56698 and 7.326647 which relatively close to the theoretical values. The cluster sampling with 15 cluster at a time give a larger spread compares to stratified and SRS.



4 Conclusion/Further Work

Comparing all three sampling methods the estimated μ are close to each other. In the other hand their SE's are close for SRS and stratified bu cluster sampling has a larger SE.



An interesting aspect to further study is the cluster sampling. The below plot shows the cluster sampling density plot in red but with 100 districts instead of 15. The estimated mean is closer to the estimated means of the SRS and stratified sampling and the SE of the the distribution is smaller. Using the 100 districts gives simulated mean of 46.7843 and SE of 4.120074 . As a result to sampling 100 districts a new distribution appeared on the right tail. This could be due to a sample bias towards elementary schools which makes up 71% of the population.

References

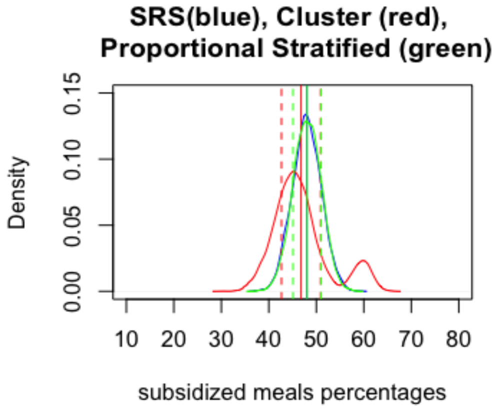
[1] Simple Random Sampling in R, Timothy R. Johnson , $https : //rpubs.com/trjohns/survey - srs$

[2] Stratified Random Sampling Analysis in R, Timothy R. Johnson , $https : //rpubs.com/trjohns/survey- stratified$

[3] Cluster Sampling Analysis with R, Timothy R. Johnson , $https : //rpubs.com/trjohns/survey - cluster$

[4] Sampling Theory Chapter 9 Cluster Sampling, Shalabh IIT Kanpur, $http : //home.iitk.ac.in/ shalab /sampling/chapter9 - sampling - cluster - sampling.pdf$ pg 13-

[5] Package "Sampling" , Yves Tille & Alina Matei ,Version 2.8 December 22, 2016 pg. 14. $https : //cran.r - project.org/web/packages/sampling/sampling.pdf$


SRS(blue), Cluster (red),
Proportional Stratified (green)
Density
0.00
0.05
0.10
0.15
10
20
30
40
50
60
70
80
subsidized meals percentages