260677676_Assignment_3_MATH525

Dan Yunheum Seol

4/1/2019

5.4

Survey evidence is often introduced in court cases involving trademark violation and employment discrimination. There has been controversy, however, about whether nonprobability samples are acceptable as evidence in litigation. Jacoby and Handlin (1991) selected 26 from a list of 1285 scholarly journals in the social and behavioral sciences. They examined all articles published during 1988 for the selected journals, and recorded (1) the number of articles in the journal that described empirical research from a survey (they excluded articles in which the authors analyzed survey data which had been collected by someone else) and (2) the total number of articles for each journal which used probability sampling, nonprobability sampling, or for which the sampling method could not be determined. The data are in file `journal.dat`.

(a)

Explain why this is a cluster sample.

Let us read in the dataset first:

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
journal <- read_csv("/cloud/project/Assignments/260677676_MATH525_Assignment3_files/journal.csv")
```

```
## Parsed with column specification:
## cols(
##   numemp = col_double(),
##   prob = col_double(),
##   nonprob = col_double()
## )
```

```r
head(journal)
```

```
## # A tibble: 6 x 3
##   numemp  prob nonprob
##    <dbl> <dbl>   <dbl>
## 1     17     0      17
## 2      1     0       1
## 3      3     0       3
## 4      3     0       2
## 5     23     0      19
## 6       3     0       3
```

```
#Add a column
journal$indeterm = journal$numemp - (journal$prob + journal$nonprob)
head(journal)
```

```
## # A tibble: 6 x 4
##   numemp  prob nonprob indeterm
##    <dbl> <dbl>   <dbl>    <dbl>
## 1     17     0      17        0
## 2      1     0       1        0
## 3      3     0       3        0
## 4      3     0       2        1
## 5     23     0      19        4
## 6      3     0       3        0
```

Jacoby and Handlin drew a cluster sample since only a selected number of clusters(journals) have been drawn. If they wanted to draw a stratified sample, they must have picked a random sample from each journal. In fact, since all SSU's (articles) from the selected PSU's (journals) have been taken, this is a one-stage cluster sample.

(b)

Estimate the proportion of articles in the 1285 journals that use nonprobability sampling, and give the standard error of your estimate.

Let us denote $\Omega :=$ indices 26 journals selected

$\Omega_i :=$ articles published in one of the 26 journals selected

$y_{ij} := $ the j-th article in i-th journal that use nonprobability sampling; here estimating the proportion is equivalent to estimating mean

$m_i = M_i$ the number of articles chosen

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\Omega}\sum_{j\in\Omega_i} y_{ij}}{\sum_{i\in\Omega} M_i} = 0.9256757$$

```
y = journal$nonprob
M = journal$numemp

ybar = sum(y)/sum(M)
ybar
```

```
## [1] 0.9256757
```

Now let us find the standard error

$$SE(\hat{\bar{y}}_r) = \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i\in\Omega}(t_i - \hat{\bar{y}}_r M_i)^2}{n-1}} = 0.03398672$$

```
summ.omega <- t(journal$nonprob - ybar*journal$numemp) %*% (journal$nonprob - ybar*journal$numemp)
ss.omega <- summ.omega[1][1]
Mbar <- mean(journal$numemp)
N <-1285
n <- 26
```

```
sigma.2 <- (1-n/N)*(1/(n*Mbar^2))*ss.omega/(n-1)
sigma <- sqrt(sigma.2)
sigma
```

```
## [1] 0.03398672
```

(c)

The authors conclude that, because “an overwhelming proportion of . . . recognized scholarly and practitioner experts rely on non-probability sampling designs,” courts “should have no problem admitting otherwise well-conducted non-probability surveys and according them due weight” (p. 175). Comment on this statement.

The premise that authors have brough is valid, but the conclusion they have made seems not as logical. First, their claim that an overwhelming proportion of recognized scholarly and practitioner experts rely on non-probability sampling designs is true. If you take 95% confidence interval, this becomes evident:

```
c(ybar-1.96*sigma, ybar+1.96*sigma)
```

```
## [1] 0.8590617 0.9922897
```

We have sufficient evident that at least 86% of the experts have been using non-probability sampling designs.

However, we should remark two points: (1) Just because some method has been mainstream among the researchers does not prove that it is indeed the right method to use; (2) We cannot apply the same logical steps that researchers adopt to court cases. The authors have committed a fallacy.

5.11

An accounting firm is interested in estimating the error rate in a compliance audit it is conducting. The population contains 828 claims, and the firm audits an SRS of 85 of those claims. In each of the 85 sampled claims, 215 fields are checked for errors. One claim has errors in 4 of the 215 fields, 1 claim has 3 errors, 4 claims have 2 errors, 22 claims have 1 error, and the remaining 57 claims have no errors. (Data courtesy of Fritz Scheuren.)

(a)

Treating the claims as psus and the observations for each field as ssus, estimate the error rate, defined to be the average number of errors per field, along with the standard error for your estimate.

We estimate the total number of errors first and then estimate the average number of errors per field.

We use formulae

$$s_t^2 = \frac{1}{n-1}\sum_{i\in\Omega}(t_i - \frac{\hat{t}}{N})^2$$

$$\hat{t} = \sum_{i\in\Omega}\frac{N}{n}t_i$$

$$\widehat{V}(\hat{t}) = N^2(1-\frac{n}{N})\frac{s_t^2}{n}$$

$$SE(\hat{t}) = \sqrt{\widehat{V}(\hat{t})}$$

$$\hat{\bar{y}} = \frac{\hat{t}}{NM}$$

$$\widehat{V}(\hat{\bar{y}}) = (1 - \frac{n}{N})\frac{s_t^2}{M^2 n}$$

$$SE(\hat{\bar{y}}) = \sqrt{\widehat{V}(\hat{\bar{y}})}$$

```
N = 828
n = 85
Omega <- c(0, 1, 2, 3, 4)
Y <- c(57, 22, 4, 1, 1)

M = 215

#We estimate the total

t.hat <- (N/n)*sum(Omega*Y)

# we estimate the average
y.bar  <- t.hat/(M*N)
#Omega*Y

st2.q11 <- (1/(n-1))*t((Omega*Y)-t.hat/N%*%((Omega*Y)-t.hat/N))
st2.q11
```

```
##            [,1]
## [1,] 0.01190476
## [2,] 0.26166446
## [3,] 0.09455306
## [4,] 0.03369375
## [5,] 0.04616533
```

```
# we estimate SE(ybar)
sigma.2.q11.bar <- (1-n/N)*(st2.q11[1][1]/(n*M^2))
sigma.q11.bar <- sqrt(sigma.2.q11.bar)

#we estimate SE(total)
sigma.2.q11.tot <- N^2*(1-n/N)*(st2.q11[1][1]/n)
sigma.q11.tot <- sqrt(sigma.2.q11.tot)

print("Our y bar is:")
```

```
## [1] "Our y bar is:"
```

```
y.bar
```

```
## [1] 0.002024624
```

```
print("with standard error")
```

```
## [1] "with standard error"
```

```
sigma.q11.bar
```

```
## [1] 5.214248e-05
```

(b)

Estimate (with standard error) the total number of errors in the 828 claims.

```
print("Estimated total:")
```

```
## [1] "Estimated total:"
```

```
t.hat
```

```
## [1] 360.4235
```

```
print("with standard error")
```

```
## [1] "with standard error"
```

```
sigma.q11.tot
```

```
## [1] 9.282404
```

(c)

Suppose that instead of taking a cluster sample, the firm had taken an SRS of $85 * 215 = 18,275$ fields from the $178,020$ fields in the population. If the estimated error rate from the SRS had been the same as in (a), what would the estimated variance $\widehat{V}(\widehat{p}_{SRS})$ be? How does this compare with the estimated variance from (a)?

This implies

$$n_{SRS} = 18275$$

$$N_{SRS} = 178020$$

We will have SRS variance

$$(1 - \frac{n_{SRS}}{N_{SRS}})\frac{\hat{\bar{y}}(1 - \hat{\bar{y}})}{n_{SRS}}$$

We will have

```
n.SRS = 18275
N.SRS = 178020

print("SRS variance ")
```

```
## [1] "SRS variance "
```

```
SRS.var = (1-n.SRS/N.SRS) *((y.bar*(1-y.bar))/n.SRS)
SRS.var
```

```
## [1] 9.921224e-08
```

```
print("cluster variance")
```

```
## [1] "cluster variance"
```

```
sigma.2.q11.bar
```

```
## [1] 2.718838e-09
```

```
print("Ratio")
```

```
## [1] "Ratio"
```

```
sigma.2.q11.bar/SRS.var
```

```
## [1] 0.02740426
```

Cluster sampling clearly performs worse than SRS. # 5.24

Suppose in a two-stage cluster sample that all population cluster sizes are equal ($M_i = M \ \forall i$), and that all sample sizes for the clusters are equal ($m_i = m \ \forall i$).

(a)

Show (5.30).

If $\forall i \ M_i = M$ and $m_i = m$, then

$$V(\hat{\bar{y}}_{unb}) = (1 - \frac{n}{N})\frac{MSB}{nM} + (1 - \frac{m}{M})\frac{MSW}{nm}$$

We start from

$$V(\hat{\bar{y}}) = \frac{1}{(NM)^2}\left[N^2\left(1 - \frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\left(1 - \frac{m}{M}\right)M^2\frac{S_i^2}{m}\right](*)$$

and

$$S_t^2 = M(MSB)$$

$$\sum_{i=1}^{N} S_i^2 = \frac{1}{M-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(\overline{y}_{ij} - \overline{y}_{iU}) = \frac{SSW}{M-1} = N(MSW)$$

(**)

Remark that

$$(*) = \frac{1}{M^2}\left(1 - \frac{n}{N}\right)\frac{S_t^2}{n} + \frac{1}{nN}\sum_{i=1}^{N}\left(1 - \frac{m}{M}\right)\frac{S_i^2}{m}$$

Due to (**)

$$\frac{1}{M^2}\left(1 - \frac{n}{N}\right)\frac{M(MSB)}{n} + \frac{1}{nN}\left(1 - \frac{m}{M}\right)\frac{N(MSW)}{m} =$$
$$\left(1 - \frac{n}{N}\right)\frac{(MSB)}{nM} + \left(1 - \frac{m}{M}\right)\frac{(MSW)}{nm}$$

as wanted

(b)

Show that

$$MSW = S^2(1 - R_a^2)$$

and

$$MSB = S^2[\frac{N(M-1)R_a^2}{N-1} + 1]$$

We know $SSTO = SSW + SSB$ on top of $R_a^2 = 1 - \frac{MSW}{S^2}$.

We have

$$R_a^2 = 1 - \frac{MSW}{S^2} \implies MSW = S^2(1 - R_a^2)$$

$$MSB = \frac{1}{N-1}[SST - N(M-1)MSW] = \frac{S^2}{N-1}[(NM-1) - N(M-1)(1-R_a^2)]$$

$$= \frac{S^2}{N-1}[NM - 1 - (NM - N)(1 - R_a^2)] =$$

$$\frac{S^2}{N-1}[NM - 1 - NM + N + (NM - N)R_a^2] =$$

$$\frac{S^2}{N-1}[N - 1 + N(M-1)R_a^2] =$$

$$S^2[1 + \frac{N(M-1)}{N-1}R_a^2]$$

(c)

Using (a) and (b) express $V(\hat{\bar{y}})$ as a function of n, m, N, M and R_a^2

$$\left(1 - \frac{n}{N}\right)\frac{(MSB)}{nM} + \left(1 - \frac{m}{M}\right)\frac{(MSW)}{nm}$$

$$\left(1 - \frac{n}{N}\right)\frac{S^2[1 + \frac{N(M-1)}{N-1}R_a^2]}{nM} + \left(1 - \frac{m}{M}\right)\frac{S^2(1 - R_a^2)}{nm}$$

(d)

Show that provided that S^2, population sizes, and sample sizes fixed and if $\frac{m-1}{m} > \frac{n}{N}$ then $V(\hat{\bar{y}})$ is an increasing function with respect to R_a^2.

From part (c), we can arrive to the conclusion that the coefficients for $S^2 R_a^2$ term

$$\left(1 - \frac{n}{N}\right)\frac{N(M-1)}{nM(N-1)} - \left(1 - \frac{m}{M}\right)\frac{1}{nm} =$$

then

$$\frac{m(N-n)(M-1)}{nmM(N-1)} - \frac{(M-m)(N-1)}{nmM(N-1)} =$$

$$\frac{m(N-n)(M-1) - (M-m)(N-1)}{nmM(N-1)} =$$

$$\frac{(N-n)(Mm-m) - MN + mN + M - m}{nmM(N-1)} =$$

$$\frac{NMm - Mnm - mN + nm - MN + mN + M - m}{nmM(N-1)} =$$

$$\frac{NMm - Mnm + nm - MN + M - m}{nmM(N-1)} =$$

$$\frac{M(Nm - nm - N + 1) + m(n-1)}{nmM(N-1)} =$$

$$\frac{M(N(m-1) - nm + 1) + m(n-1)}{nmM(N-1)}$$

(***)

if $N(m-1) > nm$ then (***) is strictly positive. Thus $V(\hat{\bar{y}})$ has a positive coefficient for R_a^2 and is thus increasing with it.

5.25

(a)

Show that $\hat{t}_{unb} = \hat{t_r}$, and, hence, that

$$\hat{\bar{y}}_{unb} = \hat{\bar{y}}_r$$

We define

$$\hat{t_r} := \hat{\bar{y}}_r(\sum_{i \in \Omega} M_i)$$

$$\hat{t_{unb}} = \sum_{i \in \Omega} t_i = \sum_{i \in \Omega} \sum_{j \in \Omega_i} y_{ij} = \frac{\sum_{i \in \Omega} \sum_{j \in \Omega_i} y_{ij}}{\sum_{i \in \Omega} M_i}(\sum_{i \in \Omega} M_i) = \frac{\sum_{j \in \Omega_i} y_{ij}}{nM}(nM) = \hat{\bar{y}}_r(\sum_{i \in \Omega} M_i) = \hat{t_r}$$

(b)

$$SSB = \sum_{i=1}^{n} \sum_{j=1}^{m} (\bar{y}_i - \hat{\bar{y}})^2$$

$$SSW = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2$$

$$SSTO = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \hat{\bar{y}})^2$$

(c)

We should define a sampling indicator RV:

$$Z_i = \begin{cases} 1 & i \in \Omega \\ 0 & \text{else} \end{cases}$$

We have

$$\widehat{SSW} = \sum_{i \in \Omega} \sum_{\Omega_i} (y_{ij} - \bar{y}_i)^2$$

We obtain that

$$E[\widehat{SSW}] = E[\sum_{i=1}^{N} Z_i \sum_{j\in\Omega_i} (y_{ij} - \overline{y}_i)^2] = E[\sum_{i=1}^{N} Z_i E[\sum_{j\in\Omega_i} (y_{ij} - \overline{y}_i)^2 | Z_i \ ]] =$$

$$E[\sum_{i=1}^{N} Z_i (m-1) E[\sum_{j\in\Omega_i} S_i^2 | Z_i = 1 \ ]] =$$

$$(m-1)\frac{n}{N}\sum_{i=1}^{N} S_i^2$$

Furthermore,

$$E[msw] = \frac{1}{N}\sum_{i=1}^{N} S_i^2 = MSW$$

where $M_i = M$ and $m_i = m$, and

$$\hat{\overline{y}}_{unb} = \frac{1}{n}\sum_{i\in\Omega}\overline{y}_i = \frac{1}{n}\sum_{i=1}^{N} Z_i \overline{y}_i$$

for the sample estimator of SSB,

$$\widehat{SSB} = \sum_{i\in\Omega}\sum_{j\in\Omega_i}(\overline{y}_i - \hat{\overline{y}}_{unb})^2$$

It follows that

$$E[\widehat{SSB}] = mE[\sum_{i\in\Omega}(\overline{y}_i^2 - 2\overline{y}_i\hat{\overline{y}}_{unb} + \hat{\overline{y}}_{unb}^2)] =$$

$$mE[\sum_{i\in\Omega}(\overline{y}_i^2 - n\hat{\overline{y}}_{unb}^2)] = mE[\sum_{i=1}^{N}(Z_i\overline{y}_i^2|Z_i) - mnE[\hat{\overline{y}}_{unb}^2]] =$$

$$mE[\sum_{i=1}^{N} Z_i\{V(\overline{y}_i|Z_i) + \overline{y}_{iU}^2\}] - mn\{V[\hat{\overline{y}}_{unb} + \overline{y}_U^2]\} =$$

$$\frac{mn}{N}\sum_{i=1}^{N}[(1-\frac{m}{M})\frac{S_i^2}{m} + \overline{y}_{iU}^2] - \frac{m}{M^2}(1-\frac{n}{N})S_t^2 - \frac{m}{n}\sum_{i=1}^{N}(1-\frac{m}{M})\frac{S_i^2}{m} - mn\overline{y}_U^2 =$$

$$\frac{m(n-1)}{N}(1-\frac{m}{M})\sum_{i=1}^{N}\frac{S_i^2}{m} + mn[\frac{1}{N}\sum_{i=1}^{N}\overline{y}_{iU}^2 - \overline{y}_U^2] - \frac{m}{M}(1-\frac{n}{N})\frac{SSB}{N-1} =$$

$$(n-1)(1-\frac{m}{M})MSW + (n-1)\frac{m}{M}MSB$$

It follows that

$$E[msb] = (1-\frac{m}{M})MSW + \frac{m}{M}MSB$$

(d)

$$E[\widehat{MSB}] = \frac{M}{m}E[msb] - (\frac{M}{m} - 1)E[msw] =$$

$$MSB + (\frac{M-m}{m} - \frac{M-m}{m})(MSW) = MSB$$

Returning an unbiased estimator.

(e)

We know

$$\hat{V}(\hat{t}_{unb}) = N^2(1-\frac{n}{N})\frac{S_t^2}{n} + \frac{N}{n}\sum_{i\in\Omega}(1-\frac{m_i}{M_i})M_i^2\frac{s_i^2}{m_i}(*)$$

and

$$s_t^2 = \frac{M^2}{m}msb(**)$$

finally

$$\sum_{i\in\Omega} s_i^2 = n*msw(***)$$

Using all 3

$$\widehat{V}(\hat{\bar{y}}) = \frac{1}{(NM)^2}[N^2(1-\frac{n}{N})\frac{S_t^2}{n} + \frac{N}{n}\sum_{i\in\Omega}(1-\frac{m}{M})\frac{M^2}{m}s_i^2] =$$

$$(1-\frac{n}{N})\frac{msb}{mn} + \frac{1}{N}(1-\frac{m_i}{M_i})\frac{msw}{m}$$

5.38

IPUMS exercises.

```
set.seed(329)
ipums <- read.csv("/cloud/project/Assignments/260677676_MATH525_Assignment3_files/ipums.csv")

#Counts of PSUs would be

PSUs = ipums %>% summarise(Num_Clusters = n_distinct(Psu))
PSUs
```

```
##   Num_Clusters
## 1           90
```

90 PSU's are there for us. Let us construct a frame of SSU's for a given PSU

```
set.seed(329)
Mi.tbl = ipums %>% group_by(Psu) %>% summarise(Mi=n())
Mi.tbl
```

```
## # A tibble: 90 x 2
##      Psu    Mi
##    <int> <int>
##  1     1   904
##  2     2  1082
##  3     3  1286
##  4     4  1094
##  5     5  1077
##  6     6  1020
##  7     7   951
##  8     8   928
##  9     9   985
## 10    10   974
## # ... with 80 more rows
```

(b)

In Assignment 2 I have obtained 661 as the ideal number thus the cost

```
661*50
```

```
## [1] 33050
```

The total cost here will be $100 * 10 + 20 * interview$ and set it too 33050

```
interview <- (33050-1000)/20
interview
```

```
## [1] 1602.5
```

Need to take 1603 interviews.

Let us take a sample of size 10 on PSU's

```
set.seed(329)
#We take a cluster sample of size 10
os.cluster.sample <- ipums %>% filter(Psu %in% sample(unique(Psu), size=10))

pnum <- os.cluster.sample %>% summarise(Num_Clusters =n_distinct(Psu))
pnum # check whether it's 10
```

```
##   Num_Clusters
## 1           10
```

Let us list the counts of SSU's for obtained PSU's

```
Mi.tblB <- os.cluster.sample %>% group_by(Psu) %>% summarise(Mi=n())
#SSU tables for the sample
Mi.tblB
```

```
## # A tibble: 10 x 2
##      Psu    Mi
##    <int> <int>
##  1     2  1082
##  2     8   928
##  3    12   418
##  4    22   384
##  5    45   288
##  6    48   266
##  7    55   548
##  8    78   437
##  9    84   599
## 10    89   522
```

```
#total count of SSUs
Mi.tblB %>% ungroup() %>% summarise(numSSU=sum(Mi))
```

```
## # A tibble: 1 x 1
##   numSSU
##    <int>
## 1   5472
```

Now we find finite population correction at the PSU level

```
os.cluster.sample <- os.cluster.sample %>% mutate(Psu.fpc = PSUs %>% pull(Num_Clusters))
os.cluster.sample %>% select(Psu, Psu.fpc) %>% slice(60:70)
```

```
##    Psu Psu.fpc
## 1    2      90
## 2    2      90
## 3    2      90
## 4    2      90
## 5    2      90
## 6    2      90
## 7    2      90
## 8    2      90
## 9    2      90
## 10   2      90
## 11   2      90
```

Do similarly at SSU level

```
os.cluster.sample <- os.cluster.sample %>% inner_join(Mi.tblB, by="Psu") %>% rename(Ssu.fpc=Mi)
os.cluster.sample %>% select(Psu, Psu.fpc, Ssu.fpc) %>% slice(60:70)
```

```
##    Psu Psu.fpc Ssu.fpc
## 1    2      90    1082
## 2    2      90    1082
## 3    2      90    1082
## 4    2      90    1082
## 5    2      90    1082
## 6    2      90    1082
## 7    2      90    1082
## 8    2      90    1082
## 9    2      90    1082
## 10   2      90    1082
## 11   2      90    1082
```

```
counts <- os.cluster.sample %>% count(Psu) %>% mutate(prop=n/sum(n), nsampled=round(prop*interview))
counts
```

```
## # A tibble: 10 x 4
##      Psu     n   prop nsampled
##    <int> <int>  <dbl>    <dbl>
##  1     2  1082 0.198       317
##  2     8   928 0.170       272
##  3    12   418 0.0764      122
##  4    22   384 0.0702      112
##  5    45   288 0.0526       84
##  6    48   266 0.0486       78
##  7    55   548 0.100       160
##  8    78   437 0.0799      128
##  9    84   599 0.109       175
## 10    89   522 0.0954      153
```

Let us give a sanity-check on the sample size

```
sum(counts$nsampled)
```

```
## [1] 1601
```

```
#It is off by 2
```

```
os.cluster.sample <- os.cluster.sample %>% inner_join(counts, by="Psu") %>% rename(mi = nsampled)
os.cluster.sample %>% select(Psu, Psu.fpc, Ssu.fpc, mi) %>% slice(60:70)
```

```
##    Psu Psu.fpc Ssu.fpc  mi
## 1    2      90    1082 317
## 2    2      90    1082 317
## 3    2      90    1082 317
## 4    2      90    1082 317
## 5    2      90    1082 317
## 6    2      90    1082 317
## 7    2      90    1082 317
## 8    2      90    1082 317
## 9    2      90    1082 317
## 10   2      90    1082 317
## 11   2      90    1082 317
```

Let us construct two-stage sampling by

```
ts.cluster.sample <- os.cluster.sample %>% group_by(Psu) %>% filter(Ssu %in% sample(unique(Ssu), size =
ts.cluster.sample
```

```
## # A tibble: 1,601 x 22
## # Groups:   Psu [10]
##    Stratum   Psu   Ssu Inctot   Age   Sex  Race Hispanic Marstat Ownershg
##      <int> <int> <int>  <int> <int> <int> <int>    <int>   <int>    <int>
##  1       1     2     6    505    18     1     1        0       1        1
##  2       1     2     9   4255    22     1     2        1       1        1
##  3       1     2    15   6245    23     1     1        0       5        1
##  4       1     2    17   7005    21     1     1        0       3        2
##  5       1     2    21  13005    16     1     1        0       5        1
##  6       1     2    25      0    16     1     1        0       5        1
##  7       1     2    26   4305    20     1     1        0       5        2
##  8       1     2    33  11005    21     1     1        0       1        1
##  9       1     2    39      0    19     1     1        0       5        1
## 10       1     2    49      0    16     1     2        0       5        2
## # ... with 1,591 more rows, and 12 more variables: Yrsusa <int>,
## #   School <int>, Educrec <int>, Labforce <int>, Occ <int>, Classwk <int>,
## #   VetStat <int>, Psu.fpc <int>, Ssu.fpc <int>, n <int>, prop <dbl>,
## #   mi <dbl>
```

(c)

```
library(survey)
```

```
## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':
##
##     expand

## Loading required package: survival

##
## Attaching package: 'survey'
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
library(srvyr)
```

```
##
## Attaching package: 'srvyr'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
ts.design <- svydesign(id = ~Psu+Ssu, fpc = ~Psu.fpc+Ssu.fpc, data=ts.cluster.sample)
svymean(~Inctot, ts.design)
```

```
##          mean    SE
## Inctot 8725.1 374.9
```

Therefore the estimated mean is 8568.2 with standard error 509.6

```
svytotal(~Inctot, ts.design)
```

```
##            total       SE
## Inctot 429695774 54906360
```

Estimated total being 421967270 with standard error 55057973

We can compare its performance against SRS in assignment 1 that returned the mean of 8884.9 with SE 401.7 and total of 475000000 with SE 21475362

```
SRS <- c(8884.9, 401.7, 475000000, 21475362)
TSCLUS <- c(8568.2, 509.6, 421967270, 55057973)

label <- c("Estimated Mean", "Mean SE", "Estimated Total", "Total SE")
tbl <- tibble(label, SRS, TSCLUS)
tbl <- column_to_rownames(tbl, var = "label")
tbl
```

```
##                         SRS      TSCLUS
## Estimated Mean       8884.9      8568.2
## Mean SE               401.7       509.6
## Estimated Total 475000000.0 421967270.0
## Total SE         21475362.0  55057973.0
```

We can remark that the cluster sample keeps underestimating the mean and the total. Furthermore, it has a wider variance for both cases.