260677676 MATH525 Assignment 1 ver4

Dan Yunheum Seol

##2.24 For some $c_0, c_1, k \in \mathbb{R}$

$$L(n) = k * Var(\bar{y}_s) = k(1 - \frac{n}{N})\frac{S^2}{n}$$

$$C(n) = c_0 + c_1 n$$

find

$$n = argmin_{n \in \mathbb{N}} L(n) + C(n)$$

We differentiate $L(n) + C(n)$ with respect to n and set it to zero

$$\frac{\partial L(n) + C(n)}{\partial n} = \frac{\partial}{\partial n}\{kS^2(\frac{1}{n} - \frac{1}{N}) + c_0 + c_1 n\}$$

$$= \frac{-kS^2}{n^2} + c_1 = 0$$

$$\implies c_1 = \frac{kS^2}{n^2} \implies n^2 = \frac{kS^2}{c_1}$$

$$\implies n = S\sqrt{\frac{k}{c_1}}$$

Let us differentiate $L(n) + C(n)$ once more to verify that it is a minimum.

$$\frac{\partial}{\partial n}\{\frac{-kS^2}{n^2}\} = \frac{2kS^2}{n^3} \geq 0$$

So indeed it is a minimum as long as $k \geq 0$, but by the nature of cost function it should be nonnegative. So choose $\lceil n \rceil$

##2.26 Show that for systematic sampling $\forall i = 1, ..., N$

$\pi_i = P(i \in S) = \frac{n}{N}$, but not necessarily SRS, i.e.

$$P(S) \neq \frac{1}{\binom{N}{n}}$$

Let $K = \frac{N}{n}$ be assumed to be a positive integer (thus assumed $n|N$. and let $R = \{1, 2, ..., K\}$ then we know the set of population U can be partitioned into

$$\{1 + K, 1 + 2K, ...1 + (n-1)K\} = [1]_K$$

$$\{2, 2 + K, ..., 2 + (n-1)K\} = [2]_K$$

. . .

$$\{K, 2K, ..., nK\} = [0]_K = [K]_K$$

So each element $i \in U$ $\pi_i = \frac{1}{K} = \frac{n}{N}$, but we only have K possible samples. Therefore for all possible sample of size n in this design

$$P(S) = \frac{1}{K} = \frac{n}{N}$$

##2.28

SRS with replacement:

Let $y_i \in U$. y_i can appear k many times in the sample of size n where $k \in \{0, ..., n\}$ \ Define

$$Q_i = \{\text{Number of times } y_i \text{ appear in a sample }\}$$

Define our estimator of the population total

$$\widehat{t} = \frac{N}{n}\sum_{i=1}^{N} Q_i y_i$$

$$t = \sum_{i=1}^{N} y_i$$

##(a)

Argue that

$$Q = (Q_1, Q_2, ..., Q_N) \sim multinomial(\{n, \frac{1}{N}, ..., \frac{1}{N}\})$$

Since this is a simple random sample, every population unit will have a chance of

$$\frac{1}{N}$$

of being chosen. If we do this n times

$P(Q_1 = q_1, Q_2 = q_2, ..., Q_N = q_n) = p(Q, n) * \frac{1}{N}^n$ for some $p(Q, n)$. Since it shows the number of partitioning n into $q1, q2, ..., q_N$, $p(Q, n)$ will be the multinomial coefficient. i.e.

$$p(Q, n) = \binom{n}{q_1, q_2, ..., q_N}$$

$$P(Q_1 = q_1, Q_2 = q_2, ..., Q_N = q_n) = \binom{n}{q_1, q_2, ..., q_N} * \frac{1}{N}^n =$$

$$\binom{n}{q_1, q_2, ..., q_N} \frac{1}{N}^{q_1} \frac{1}{N}^{q_2} \cdots \frac{1}{N}^{q_N}$$

which is a case of multinomial pmf.

$$P(Q_1 = q_1, Q_2 = q_2, ..., Q_N = q_N) = \binom{n}{q_1, q_2, ..., q_N} \prod_{i=1}^{N} p_i^{q_i}$$

##(b)

For a multinomial random vector

$$Y = (Y_1, ..., Y_N) \sim multinomial(n, \vec{p} = (p_1, ..., p_N))$$

, we have its expected value

$$E[Y] = n\vec{p} = (np_1, np_2, ..., np_N)$$

For our case, we will have

$$E[Q] = n\vec{p}_Q = (\frac{n}{N}, \frac{n}{N}, ..., \frac{n}{N})$$

There $\forall i$ we have

$$E[Q_i] = \frac{n}{N}$$

$$Var[Q_i] = n(\frac{1}{N})(1 - \frac{1}{N})$$

$$Cov(Q_i, Q_j) = -\frac{n}{N^2} \ i \neq j$$

Now, for $E[\widehat{t}] = E[\frac{N}{n}\sum_{i=1}^{N} Q_i y_i]$

$$E[\frac{N}{n}\sum_{i=1}^{N} Q_i y_i] = \frac{N}{n}\sum_{i=1}^{N} y_i E[Q_i] = \frac{N}{n}\sum_{i=1}^{N} y_i \frac{n}{N} = \frac{N}{n}\frac{n}{N}\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} y_i$$

Proving that it is an unbiased estimator of t.

##(c)

$$Var(\widehat{t}) = Var(\frac{N}{n}\sum_{i=1}^{N} Q_i y_i) = \frac{N^2}{n^2} Var(\sum_{i=1}^{N} Q_i y_i) =$$

$$= \frac{N^2}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j Cov(Q_i, Q_j) =$$

$$\frac{N^2}{n^2}[\sum_{i=1}^{N} y_i^2 Var(Q_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} Cov(Q_i, Q_j)] =$$

$$\frac{N^2}{n^2}[\frac{n}{N}\frac{N-1}{N}\sum_{i=1}^{N} y_i^2 - \frac{n}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j] =$$

$$\frac{N^2}{n^2}[\frac{n}{N}\sum_{i=1}^{N} y_i^2 - \frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j] =$$

$$\frac{N}{n}\sum_{i=1}^{N} y_i^2 - \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j = \frac{N}{n}[N\sum_{i=1}^{N} y_i^2 - \frac{1}{N}(\sum_{i=1}^{N} y_i)^2] =$$

$$\frac{N}{n}(N-1)S^2$$

##2.29 We can interpret this with recursive definition of

$$S_0 = \{1, 2, ..n\}, u_k \sim U(0,1)$$

given S_{k-1}

$$S_k | S_{k-1} =$$

$$\begin{cases} S_{k-1} & u_k > \frac{n}{n+k} \\ S_{k-1} \backslash \{j\} \cup \{n+k\} & u_k \leq \frac{n}{n+k} \end{cases}$$

where $j \sim$ Discrete Uniform(S_{k-1}). Remark that

$$P(S_k = S_{k-1} | S_{k-1}) = P(u_k \leq \frac{n}{n+k}) = \frac{n}{n+k}$$

and

$$P(S_k \neq S_{k-1} | S_{k-1}, j) = P(u_k > \frac{n}{n+k}) = 1 - \frac{n}{n+k} = \frac{k}{n+k}$$

We show this by induction on N.

###Base Case

$N=0$: You can only pick the empty sample (sample size of zero), so it trivially holds. \ $N=1$

Likewise, you can only pick one sample. Let $U=\{1\}$ be a population of 1. Then the only possible sample of size n (nonzero) is to have sample $S=U$. then

$$P(S)=P(U)=1=\frac{1}{\binom{1}{1}}$$

Thus it is an SRS.

###Inductive step.

Suppose S_{N-1-n} is an SRS of size n for some $N\in\mathbb{N}$ and all possible n. Then for S_{N-n}, given S_{N-1-n}

$$S_{N-n}|S_{N-1-n}=$$

$$\begin{cases} S_{N-1-n} & P(S_{N-n}=S_{N-1-n}|S_{N-1-n})=P(u_{N-n}>\frac{n}{n+(N-n)})=\frac{N-n}{N} \\ S_{N-1-n}\backslash\{j\}\cup\{N-n\} & P(S_{N-n}\neq S_{N-1-n}|S_{N-1-n},j)=P(u_{N-n}\leq\frac{n}{n+(N-n)})=\frac{n}{N} \end{cases}$$

Due to our inductive hypothesis $P(S_{N-1-n})=\frac{1}{\binom{N-1}{n}}=\frac{n!(N-1-n)!}{(N-1)!}$ so eventually we have

$$P(S^*_{N-n}=S^*_{N-1-n})=P(S_{N-n}=S_{N-1-n}|S_{N-1-n})P(S_{N-1-n})=\frac{n!(N-1-n)!}{(N-1)!}*\frac{N-n}{N}=$$

$$\frac{n!(N-n)!}{N!}=\frac{1}{\binom{N}{n}}$$

For the other case, let S^*_{N-1-n}, S^*_{N-n} be any particular sets constructed by following scheme.

$$P(N\in S^*_{N-n})=\frac{n}{N}=\frac{\binom{N-1}{n-1}}{\binom{N}{n}}$$

but if $s\neq N$ s is from the S_{N-1-n}, which is a random sample. You can understand it as taking a sample size of n from S_{N-1-n}, and choose which one to replace with N (or equivalently choosing which n-1 elements to preserve). Especially since any element in S_{N-1-n} has an equal chance of being chosen, and due to our inductive hypothesis, marginally the element chosen to be replaced can be any one of N-n values not chosen once having been replaced.

$$P(j\notin S^*_{N-n},j\in S^*_{N-1-n},S^*_{N-1-n})=$$
$$P(j\notin S^*_{N-n},j\in S^*_{N-1-n},|S^*_{N-1-n})P(S^*_{N-1-n})$$
$$=\frac{1}{n}\frac{1}{\binom{N-1}{n}}=\frac{1}{n}\frac{(n-1)!(N-1-n)!}{(N-1)!}=\frac{1}{N-n}\frac{1}{\binom{N-1}{n-1}}$$

Now, there can be $N-n$ for i, the values replaced, so we wil; get the

$$P(j\notin S^*_{N-n},j\in S^*_{N-1-n})=\frac{1}{N-n}$$

$$P(S^*_{N-n}|j\notin S^*_{N-n},j\in S^*_{N-1-n})=\frac{P(S^*_{N-n},j\notin S^*_{N-n},j\in S^*_{N-1-n})}{P(j\notin S^*_{N-n},j\in S^*_{N-1-n})}$$

$$\frac{1}{N-n}\frac{1}{\binom{N-1}{n-1}}*(N-n)=\frac{1}{\binom{N-1}{n-1}}$$

but by our construction,

$$P(S^*_{N-n}|j\notin S^*_{N-n},j\in S^*_{N-1-n})=P(i\in(S_{N-n}\cap S_{N-1-n})\ |N\in S_{N-n})=\frac{1}{\binom{N-1}{n-1}}$$

and since

$$P(S^*_{N-n}|j \notin S^*_{N-n}, j \in S^*_{N-1-n}) = P(S^*_{N-n}|N \in S^*_{N-n}) \implies$$

We have

$$P(S^*_{N-n}) = P(S^*_{N-n}|N \in S^*_{N-n})P(N \in S^*_{N-n}) = \frac{1}{\binom{N-1}{n-1}}\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{1}{\binom{N}{n}}$$

Proving our claim. ## 2.37

```
## -- 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()

## 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

##
## Attaching package: 'srvyr'

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

```
ipums <- read_csv('ipums.csv')
```

```
## Parsed with column specification:
## cols(
##   Stratum = col_double(),
##   Psu = col_double(),
##   Inctot = col_double(),
##   Age = col_double(),
##   Sex = col_double(),
##   Race = col_double(),
##   Hispanic = col_double(),
##   Marstat = col_double(),
##   Ownershg = col_double(),
```

```
##   Yrsusa = col_double(),
##   School = col_double(),
##   Educrec = col_double(),
##   Labforce = col_double(),
##   Occ = col_double(),
##   Classwk = col_double(),
##   VetStat = col_double()
## )
```

```r
head(ipums)
```

```
## # A tibble: 6 x 16
##   Stratum   Psu Inctot   Age   Sex  Race Hispanic Marstat Ownershg Yrsusa
##     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>    <dbl>  <dbl>
## 1       1     1   4105    18     1     2        0       5        0      0
## 2       1     1   7795    20     1     1        0       5        2      0
## 3       1     1  16985    24     1     1        0       1        1      0
## 4       1     1   7045    21     1     1        0       1        2      0
## 5       1     1   2955    23     1     1        0       5        2      0
## 6       1     1      0    17     1     1        0       5        1      0
## # ... with 6 more variables: School <dbl>, Educrec <dbl>, Labforce <dbl>,
## #   Occ <dbl>, Classwk <dbl>, VetStat <dbl>
```

```r
dim(ipums)
```

```
## [1] 53461    16
```

(a)

The process of topcoding would shadow the true distribution of total personal income if one takes a model-based approach. It would also make it harder to perform a more accurate inference.

(b)

```r
ipums_complete = ipums %>% filter(!is.na(Inctot))
#Verify that none of the data is missing
dim(ipums_complete)
```

```
## [1] 53461    16
```

```r
set.seed(329)

srs_pilot = ipums_complete %>% slice(sample(1:nrow(ipums_complete),
                                            size=50, replace=F))
dim(srs_pilot)
```

```
## [1] 50 16
```

```r
head(srs_pilot)
```

```
## # A tibble: 6 x 16
##   Stratum   Psu Inctot   Age   Sex  Race Hispanic Marstat Ownershg Yrsusa
##     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>    <dbl>  <dbl>
## 1       7    61   8510    43     1     1        0       1        1      0
## 2       6    54      0    57     2     1        0       1        1      0
```

```
## 3       6    55   -805    33      2      1         0        3         2       0
## 4       6    52      0    46      2      1         0        3         1       0
## 5       9    81  10535    65      1      1         0        1         1       0
## 6       2    20   8300    30      1      1         1        1         2       0
## # ... with 6 more variables: School <dbl>, Educrec <dbl>, Labforce <dbl>,
## #   Occ <dbl>, Classwk <dbl>, VetStat <dbl>
```

```
srs_pilot %>%
  summarise(SampleMean=mean(Inctot),
                        SampleVar = var(Inctot),
                        SampleSD = sd(Inctot))  %>%
  gather(stat,val) %>%
  kable(.,format="latex",digits=0) %>%
  kable_styling(.)
```

stat	val
SampleMean	8321
SampleVar	85300547
SampleSD	9236

From lectures we have learned that

$$n = \frac{S^2 Z^2_{0.025}}{e^2 + \frac{S^2 Z^2_{0.025}}{N}} = \frac{n_0}{1 + \frac{n_0}{N}}$$

where $N = 53461$, $e = 700$ and $n_0 = \frac{S^2 Z^2_{0.025}}{e^2}$

Let me use the formula using the sample variance $\widehat{S^2}$

```
#Number of population unit
N = 53461
#Margin of error
e = 700

pilot_summary = srs_pilot %>%
  summarise(SampleMean=mean(Inctot),SampleVar = var(Inctot), SampleSD = sd(Inctot))

#Extracting the sample variance
S.2 = pilot_summary$SampleVar
#Normal quantile (1.96)
Z.a = qnorm(0.975)
#First candidate for the sample size
n_a =  (S.2*Z.a^2)/(e^2 + (S.2 * Z.a^2)/N)

#Now find the second candidate for the sample size using another formula and check whether the two cand

n_0 = (S.2 * Z.a^2)/(e^2)

n_b = (n_0)/(1+ n_0/N)

n_a
```

```
## [1] 660.47
```

```r
n_b
```

```
## [1] 660.47
```

Our ideal sample size is $ceiling(660.47) = 661$ population units.

###(c)

```r
set.seed(329)
srs_661 = ipums_complete %>% slice(sample(1:nrow(ipums_complete),
                                         size=661, replace=F))
dim(srs_661)
```

```
## [1] 661  16
```

```r
head(srs_661)
```

```
## # A tibble: 6 x 16
##   Stratum   Psu Inctot   Age   Sex  Race Hispanic Marstat Ownershg Yrsusa
##     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>    <dbl>  <dbl>
## 1       7    61   8510    43     1     1        0       1        1      0
## 2       6    54      0    57     2     1        0       1        1      0
## 3       6    55   -805    33     2     1        0       3        2      0
## 4       6    52      0    46     2     1        0       3        1      0
## 5       9    81  10535    65     1     1        0       1        1      0
## 6       2    20   8300    30     1     1        1       1        2      0
## # ... with 6 more variables: School <dbl>, Educrec <dbl>, Labforce <dbl>,
## #   Occ <dbl>, Classwk <dbl>, VetStat <dbl>
```

```r
#Estimated total income for the population
srs_661 %>%
  summarise(SampleMean=mean(Inctot),
                        SampleVar = var(Inctot),
                        SampleSD = sd(Inctot),
                        t_hat = mean(Inctot)*N,
                        se_t_hat = N*sqrt((1 - 661/N))*sd(Inctot)/sqrt(661))  %>%
  gather(stat,val) %>%
  kable(.,format="latex",digits=0) %>%
  kable_styling(.)
```

stat	val
SampleMean	9309
SampleVar	115375475
SampleSD	10741
t_hat	497663677
se_t_hat	22196859

```r
srs_design = survey::svydesign(id=~1,data=srs_661, fpc=rep(N,661))

svytotal(~Inctot,srs_design)
```

```
##            total       SE
## Inctot 497663677 22196859
```

```r
confint(svytotal(~Inctot, srs_design))
```

```
##            2.5 %    97.5 %
```

```
## Inctot 454158633 541168721
```

```r
ggplot(srs_661,aes(x=Inctot)) + geom_histogram(fill="lightblue",col="black")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```