3430 Assignment 2

Ravish Kamath: 213893664

12 February, 2022

Question 1

In the without replacement sampling example of Table 3.2, demonstrate that an unbiased estimate of the population size, N, is provided by

$$\sum_{i=1}^{n} w_i$$

Solution

This is re-creating our samples, denoted as S, from a population of $N = 4$, and having each S_i contain $n = 2$ elements **without** replacement

```
u = c(1,2,3,4)
N = 4
n = 2
sample = t((combn(1:N, n)))
sample
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    1    4
## [4,]    2    3
## [5,]    2    4
## [6,]    3    4
```

We now calculate the probabilities, denoted as $P(S)$, for each of the elemnts of S. Take note that our population elements have unequal probability of selection, represented by δ_i, as shown below. So in order to accommodate the δ_i, we take the inclusion probability, denoted by π_i, for each element of our population.

```
delta = c(0.1, 0.1, 0.4, 0.4)
prob = round(delta[sample[,1]]*delta[sample[,2]]/(1-delta[sample[,1]])+
      delta[sample[,2]]*delta[sample[,1]]/(1-delta[sample[,2]]),4)
pi = rep(0, length(u))
for(i in 1:length(u)){
  pi[i] = sum(prob[sample[,1]== i]) + sum(prob[sample[,2] == i])
}
```

```
data.frame(sample,prob)
```

```
##   X1 X2   prob
## 1  1  2 0.0222
## 2  1  3 0.1111
## 3  1  4 0.1111
## 4  2  3 0.1111
## 5  2  4 0.1111
## 6  3  4 0.5333
```

In order to get the weights, w_i, we must reciprocate our π_i. Next, we sum up the w_i for each element per sample.

```
wi = 1/pi
wi
```

```
## [1] 4.091653 4.091653 1.323627 1.323627
```

```
sumWeights = rep(0, length(sample[,1]))
sumWeights = wi[sample[,1]] + wi[sample[,2]]
data.frame(sample,prob,sumWeights)
```

```
##   X1 X2   prob sumWeights
## 1  1  2 0.0222   8.183306
## 2  1  3 0.1111   5.415280
## 3  1  4 0.1111   5.415280
## 4  2  3 0.1111   5.415280
## 5  2  4 0.1111   5.415280
## 6  3  4 0.5333   2.647253
```

Finally, to demonstrate that this is an unbiased estimate of the population size N, which in our case $N = 4$, we take the expectation of the $\sum_{i=1}^{n} w_i$ for each sample. As shown below the $E(\sum_{i=1}^{n} w_i)$ gives us our true population size, N.

```
sum(sumWeights*prob)
```

```
## [1] 4
```

Question 2

Following textbook exercise 3.17, again $u = 1, 2, 3, 4$ $\delta_i = (.1, .1, .4, .4)$ but this time, take a sample of size $n = 3$ without replacement. Use R to show that an unbiased estimate of the population size $N = 4$ is provided by $\sum_{i=1}^{3} w_i$

Solution

A similar approach to Question 1, we will now take a sample of $n = 3$ from a population of $N = 4$ elements.

```
u = c(1,2,3,4)
N = 4
n = 3
sample = t((combn(1:N, n)))
sample
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    1    2    4
## [3,]    1    3    4
## [4,]    2    3    4
```

We now calculate the probabilities for each of the sample. Take note that our N elements have unequal probability weights, represented by δ_i, as shown below. So in order to accommodate the δ_i, we take the inclusion probability, denoted by π_i, for each element of our population.

```
delta = c(0.1, 0.1, 0.4, 0.4)
prob = round(delta[sample[,1]]*(delta[sample[,2]]/(1-delta[sample[,1]]))*
       (delta[sample[,3]]/(1-delta[sample[,1]]-delta[sample[,2]]))+
       delta[sample[,1]]*(delta[sample[,3]]/(1-delta[sample[,1]]))*
       (delta[sample[,2]]/(1-delta[sample[,1]]-delta[sample[,3]]))+
       delta[sample[,2]]*(delta[sample[,3]]/(1-delta[sample[,2]]))*
       (delta[sample[,1]]/(1-delta[sample[,2]]-delta[sample[,3]]))+
       delta[sample[,2]]*(delta[sample[,1]]/(1-delta[sample[,2]]))*
       (delta[sample[,3]]/(1-delta[sample[,2]]-delta[sample[,1]]))+
       delta[sample[,3]]*(delta[sample[,1]]/(1-delta[sample[,3]]))*
       (delta[sample[,2]]/(1-delta[sample[,3]]-delta[sample[,1]]))+
       delta[sample[,3]]*(delta[sample[,2]]/(1-delta[sample[,3]]))*
       (delta[sample[,1]]/(1-delta[sample[,3]]-delta[sample[,2]])),4)

pi = rep(0, length(u))
for(i in 1:length(u)){
  pi[i] = sum(prob[sample[,1]== i]) + sum(prob[sample[,2] == i]) + sum(prob[sample[,3]== i])
}
data.frame(sample,prob)
```

```
##   X1 X2 X3   prob
## 1  1  2  3 0.0556
## 2  1  2  4 0.0556
## 3  1  3  4 0.4444
## 4  2  3  4 0.4444
```

In order to get the weights, represented by w_i we must now reciprocate π_i. Then, we sum up the w_i for each element per sample.

```
wi = 1/pi
wi
```

```
## [1] 1.799856 1.799856 1.058873 1.058873
```

```
sumWeights = rep(0, length(sample[,1]))
sumWeights = wi[sample[,1]] + wi[sample[,2]] + wi[sample[,3]]
data.frame(sample,prob,sumWeights)
```

```
##   X1 X2 X3   prob sumWeights
## 1  1  2  3 0.0556   4.658585
## 2  1  2  4 0.0556   4.658585
## 3  1  3  4 0.4444   3.917603
## 4  2  3  4 0.4444   3.917603
```

Finally,to demonstrate that this is an unbiased estimate of the population size N, which in our case $N = 4$, we take the expectation of our weights from each sample. As shown below the $E(\sum_{i=1}^{n} w_i)$ gives us our true population size, N.

```
sum(sumWeights*prob)
```

```
## [1] 4
```

Question 3

Expanding on Exercise 3.18, data on K–12 education variables and populations for all 50 states are available via links in Electronic Section 3.0. Using a sample size of $n = 5$, select repeated random samples without replacement from this population of states and calculate the mean number of teachers per state for each sample. Plot the sample means, thereby generating a simulated sampling distribution for the sample mean for samples of size 5.

a. Describe the shape of the simulated sampling distribution. Does it look normal? Why or why not?
b. Calculate the standard deviation for the set of generated sample means. Is it close to the theoretical value of

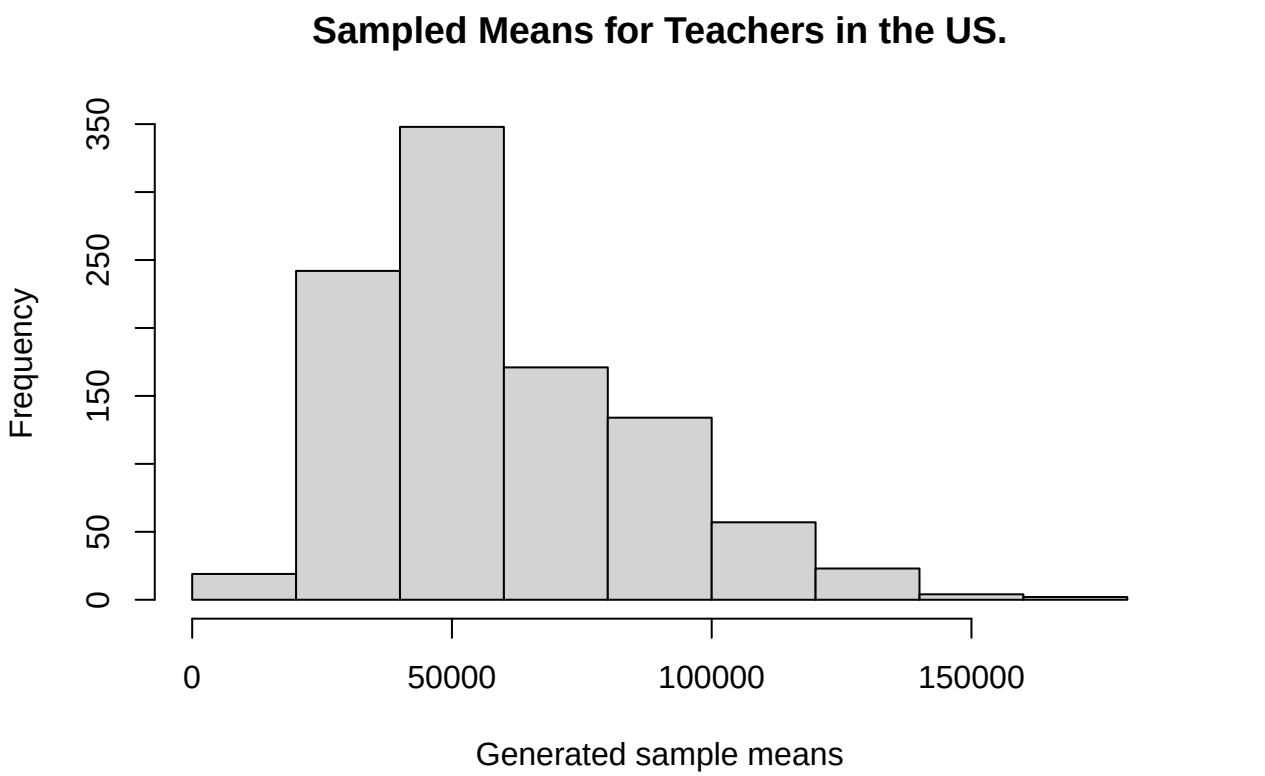
$$\frac{\sigma}{\sqrt{n}} = \frac{63,650}{\sqrt{5}} = 28,465$$

Solution

Part A

In this question, we are trying to calculate the sample mean, denoted by $\overline{y}$, of the number of teachers per state in a sample of $\mathbf{n = 5}$ from $\mathbf{N = 50}$ states. We will repeat this process a **1000** times.

```
ybar = rep(0,1000)
for(i in 1:1000){
  n = sample(1:50, 5, replace = FALSE)
  teachers = schools[n,]$Teachers
  ybar[i] = mean(teachers)
}
hist(ybar, xlab = 'Generated sample means',
     main = 'Sampled Means for Teachers in the US.')
```



After calculating the $\overline{y}$ a 1000 times, and plotting the $\overline{y}$ on a histogram, we see that though it is skewed to right. Hence we do not have a normal distribution of the mean population of the teachers in the United States. This could be due to the fact that our sample size of $n = 5$ is fairly small. If we were to increase our sample size, we may have a more normal distribution.

Part B

When calculating the standard deviation,denoted by s for our generated sample means, it does come quite close to the theoretical $\sigma = 28,465$.

```
sd(ybar)
```

```
## [1] 26954.27
```

Question 4

A study to assess the attitudes of accountants toward advertising their services involved sending questionnaires to 200 accountants selected from a list of 1400 names. A total of 82 usable questionnaires were returned. The data summary for one question is shown in the accompanying table.

a. Estimate the population proportion virtually certain to advertise in the future.
b. Estimate the population proportion having at least a 50–50 chance of advertising in the future.
c. Among those who advertised in the past, estimate the population proportion some- what unlikely to advertise again.
d. Among those who advertised in the past, estimate the population proportion having at least a 50–50 chance of advertising again.

Place bounds on the errors of estimation in all cases. Do parts (c) and (d) require further assumptions over those made for parts (a) and (b)?

Solution

Let us first re-create the data from the questionnaires provided. Let it be known that our population is 1400 names, denoted by N, and our sample, S, for all respondents is $n = 82$, and those who have advertised in the past $n' = 46$.

```
response = c('Virtual certainty', 'Very likely', 'Somewhat likely', 'About 50-50',
             'Somewhat unlikely', 'Very unlikely','Absolutely not', 'No response')
allRespondents = c(.22,.04,.19,.18,.06,.12,.15,.04)
advertisedPast = c(.35,.05,.35,.15,.10,0,0,0)
df = data.frame(response, allRespondents, advertisedPast)
df
```

```
##            response allRespondents advertisedPast
## 1 Virtual certainty           0.22           0.35
## 2       Very likely           0.04           0.05
## 3   Somewhat likely           0.19           0.35
## 4       About 50-50           0.18           0.15
## 5 Somewhat unlikely           0.06           0.10
## 6     Very unlikely           0.12           0.00
## 7    Absolutely not           0.15           0.00
## 8       No response           0.04           0.00
```

```
N = 1400
n = 82
nPrime = 46
```

Part A

Based on the calculations below, we can see that the estimated population proportion, denoted as $\hat{p}$, for respondents that are virtually certain to advertise in the future is **0.22** with a bound on the errors of **0.0893**.

```
p_hat = df[1,2]
q_hat = 1-p_hat
p_hat
```

```
## [1] 0.22
```

```
var_hat = (1-(n/N))*((p_hat*q_hat)/(n-1))
B = 2*(sqrt(var_hat))
round(B,4)
```

```
## [1] 0.0893
```

Part B

Based on the calculations below, we can see that the $\hat{p}$ for respondents having *at least* a 50-50 chance of advertising in the future is **0.63** with a bound on the errors of **0.1041**.

```
p_hat = sum(df[1:4,2])
q_hat = 1-p_hat
p_hat
```

```
## [1] 0.63
```

```
var_hat = (1-(n/N))*((p_hat*q_hat)/(n-1))
B = 2*(sqrt(var_hat))
round(B,4)
```

```
## [1] 0.1041
```

Part C

Based on the calculations below, we can see that the $\hat{p}$ for respondents who have advertised in the past and are unlikely to advertise again is **0.1** with a bound on the errors of **0.088**.

```
p_hat = df[5,3]
p_hat
```

```
## [1] 0.1
```

```
q_hat = 1-p_hat
var_hat = (1-(nPrime/N))*((p_hat*q_hat)/(nPrime-1))
B = 2*(sqrt(var_hat))
round(B,4)
```

```
## [1] 0.088
```

Part D

Based on the calculations below, we can see that the $\hat{p}$ for respondents who have advertised in the past and having *at least* a 50-50 chance of advertise again is **0.9** with a bound on the errors of **0.088**.

```
p_hat = sum(df[1:4,3])
p_hat
```

```
## [1] 0.9
```

```
q_hat = 1-p_hat
var_hat = (1-(nPrime/N))*((p_hat*q_hat)/(nPrime-1))
B = 2*(sqrt(var_hat))
round(B,4)
```

```
## [1] 0.088
```

In order to calculate the correct bounds for part (c) and (d), we need to have the assumption that it is independent. In part (a) and (b), our sample contains all the accountants that have been sampled. However in part (c) and (d), the sample of 42 accountants is taken from the full sample, due to the fact that we have more information that they have advertised previously. Hence we do need to make the assumption that it is independent, in order to calculate the correct bounds.

Quesiton 5

Redo Exercise 3.18 on the complete data set with 50 states.

The table below provides data for the 2001 school year on some K–12 education variables as well as populations for the New England states. For samples of size $n = 2$ taken with probabilities proportional to the populations of the states, find all possible estimates of the total number of teachers in the New England states and demonstrate that the estimator is unbiased. Do this for

a. Sampling with Replacement
b. Sampling without Replacement

Solution

Part A

The schools data frame provides multiple data variables for the 2001 school year of all 50 states. To estimate the total number of teachers in all, $N = 50$ states, we must focus on the population and teacher variables in our data. Furthermore we will sample, the total number of teacher, denoted by S, with a size $n = 2$ **with** replacement. The total number of S in this case will be **2500**.

```
teachers = schools$Teachers
states = schools$State
popu = schools$Pop
N = dim(schools)[1]
n = 2
sample=expand.grid(1:N,1:N)
```

To calculate the probabilities, $P(S)$ for each sample, it is given in the question that the $P(S)$ are proportional to the population of the states. Hence our $P(S)$ for each sample will be dependent on the element's corresponding δ_i value. Finally we can calculate our estimated total number of teachers for each sample, denoted by $\hat{\tau}$.

```
delta =schools$Pop/sum(schools$Pop)
prob = rep(0,N)
prob = delta[sample[,1]]*delta[sample[,2]]
tau_hat = 1/n*(teachers[sample[,1]]/delta[sample[,1]]
               +teachers[sample[,2]]/delta[sample[,2]])
```

To demonstrate that our $\hat{\tau}$ is unbiased, we take the expectation of our estimated total population.

```
prob%*%tau_hat
```

```
##          [,1]
## [1,] 2992790
```

```
sum(schools$Teachers)
```

```
## [1] 2992790
```

As proved above, we come to the same total number of teachers, which proves that the $E(\hat{\tau})$ is the theoretical population total, τ.

Part B

Similar to Part A, we continue the same procedure, however there will be **without replacement**, which means our total number of S will be **1225**.

```
sample = t(combn(1:N,n))
```

To calculate the $P(S)$ for each S, we use a similar approach to Part A, except **without** replacement does change how we multiply our individual probability for each element of the S_i, as shown below.

```
prob = rep(0,N)
prob = delta[sample[,1]]*delta[sample[,2]]/(1-delta[sample[,1]]) +
       delta[sample[,2]]*delta[sample[,1]]/(1-delta[sample[,2]])

pi=rep(0,length(schools$Teachers))
for(i in 1:length(schools$Teachers)){
  pi[i]=sum(prob[sample[,1]==i])+sum(prob[sample[,2]==i])
}
```

Finally we can calculate our $\hat{\tau}$ for each S as represented by the variable .

```
tau_hat = rep(0,dim(sample)[1])
tau_hat = teachers[sample[,1]]/pi[sample[,1]]+teachers[sample[,2]]/pi[sample[,2]]
```

To demonstrate that our $\hat{\tau}$ is unbiased, we take the expectation of our estimated $\hat{\tau}$.

```
prob%*%tau_hat
```

```
##          [,1]
## [1,] 2992790
```

```
sum(schools$Teachers)
```

```
## [1] 2992790
```

As proved above, we come to the same total number of teachers, which proves that the $E(\hat{\tau})$ total is the theoretical population total, τ.