STA304 A1

Fariha

2020/02/05

1. (a) Finding μ and σ^2 squared of Y using the formulas $\sum \frac{Yi}{N}$ and $\sum \frac{(Yi-\mu)^2}{N}$ respectively

We're going to call this set Y = {y1,y2,y3,y4}

The set of Samples of size 2 taken from Y is S = {s1,s2,s3,s4,s5,s6}

```
y <-c(3,1,0,5)
```

```
mean(y)
```

```
## [1] 2.25
```

```
var(y)
```

```
## [1] 4.916667
```

(b) **WITHOUT REPLACEMENT**

Plan 1: Consider a simple random sample without replacement (SRS) design with sample size n = 2.

(i) Find the number of possible SRSs of size n = 2.

There are 6 possible SRSs of size n = 2

Assigning each Yi 1 <= i <= 4 to it's value

```
y1<-c(3)
y2<-c(1)
y3<-c(0)
y4<-c(5)
```

List every possible sample

Assigning each Si| 1 <= i <= 6 to its set of 2 Yi's

```
s1 <-c(y1,y2)
s2<-c(y1,y3)
s3 <-c(y1,y4)
s4<-c(y2,y3)
s5<-c(y2,y4)
s6<-c(y3,y4)
```

For each sample, what is the probability that it is the one selected? Because there are 6 different samples, the probablity of each individual samples being selected is $\frac{1}{6}$ or 0.16

For each sample, determine $\overline{y}$, s^2. Now, for each sample Si, we find y-bar using the formula;$\sum \frac{Yi}{n-1}$

```
mean(s1)
```

```
## [1] 2
```

```
mean(s2)
```

```
## [1] 1.5
```

```
mean(s3)
```

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

```
mean(s4)
```

```
## [1] 0.5
```

```
mean(s5)
```

```
## [1] 3
```

```
mean(s6)
```

```
## [1] 2.5
```

and for each sample Si, we find s^2 using the formula;$\sum \frac{(yi-\overline{y})^2}{n-1}$

```
var(s1)
```

```
## [1] 2
```

```
var(s2)
```

```
## [1] 4.5
```

```
var(s3)
```

```
## [1] 2
```

```
var(s4)
```

```
## [1] 0.5
```

```
var(s5)
```

```
## [1] 8
```

```
var(s6)
```

```
## [1] 12.5
```

(iii) Find E($\overline{y}$), V($\overline{y}$), Bias($\overline{y}$) and MSE($\overline{y}$).

```
y_bar<-c(2,1.5,4,0.5,3,2.5)
sigma_squared<- c(2,4.5,2,0.5,8,12.5)
```

```
mean(y_bar)
```

```
## [1] 2.25
```

```
var(y_bar)
```

```
## [1] 1.475
```

```
bias<- mean(y_bar)-mean(y)
bias
```

```
## [1] 0
```

```
MSE <- var(y_bar) + (bias)^2
MSE
```

```
## [1] 1.475
```

(iv) Find E(s^2), V(s^2)

```
mean(sigma_squared)
```

```
## [1] 4.916667
```

```
var(sigma_squared)
```

```
## [1] 20.74167
```

(v) Are $\overline{y}$ and s^2 unbiased estimators for μ and σ^2?

Yes

(c) **WITH REPLACEMENT**

Plan 2: Consider a simple random sample with replacement (SRSWR) design with sample size n = 2.

We're going to call this set Y = {y1,y2,y3,y4}

The set of Samples of size 2 taken from Y is R = {r1,r2,r3,r4,r5,r6,r7,r9,r10}

```
y <-c(3,1,0,5)
```

(i) Find the number of possible SRSs of size n = 2.

There are 10 possible SRSs of size n = 2

Finding μ and σ^2 squared of Y using the formulas $\sum \frac{Yi}{N}$ and $\sum \frac{(Yi-\mu)^2}{N}$ respectively

```
mean(y)
```

```
## [1] 2.25
```

```
var(y)
```

```
## [1] 4.916667
```

Assigning each Yi 1 <= i <= 4 to it's value

```
y1<-c(3)
y2<-c(1)
y3<-c(0)
y4<-c(5)
```

List every possible sample Assigning each Ri| 1 <= i <= 10 to its set of 2 Yi's

```
r1 <-c(y1,y1)
r2 <-c(y1,y2)
r3<-c(y1,y3)
r4 <-c(y1,y4)
r5<-c(y2,y2)
r6<-c(y2,y3)
r7<-c(y2,y4)
r8<-c(y3,y3)
r9<-c(y3,y4)
r10<-c(y4,y4)
```

For each sample, what is the probability that it is the one selected?

Because there are 10 different samples, the probablity of each individual samples being selected is $\frac{1}{10}$ or 0.1 For each sample, determine $\overline{y}$, sˆ2. Now, for each sample Ri, we find y-bar using the formula;$\sum \frac{Yi}{n-1}$

```
mean(r1)
```

```
## [1] 3
```

```
mean(r2)
```

```
## [1] 2
```

```r
mean(r3)
```

```
## [1] 1.5
```

```r
mean(r4)
```

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

```r
mean(r5)
```

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

```r
mean(r6)
```

```
## [1] 0.5
```

```r
mean(r7)
```

```
## [1] 3
```

```r
mean(r8)
```

```
## [1] 0
```

```r
mean(r9)
```

```
## [1] 2.5
```

```r
mean(r10)
```

```
## [1] 5
```

and for each sample Ri, we find sˆ2 using the formula; $\sum \frac{(yi-\overline{y})^2}{n-1}$

```r
var(r1)
```

```
## [1] 0
```

```r
var(r2)
```

```
## [1] 2
```

```r
var(r3)
```

```
## [1] 4.5
```

```
var(r4)
```

```
## [1] 2
```

```
var(r5)
```

```
## [1] 0
```

```
var(r6)
```

```
## [1] 0.5
```

```
var(r7)
```

```
## [1] 8
```

```
var(r8)
```

```
## [1] 0
```

```
var(r9)
```

```
## [1] 12.5
```

```
var(r10)
```

```
## [1] 0
```

(iii) Find $\mathrm{E}(\overline{y})$, $\mathrm{V}(\overline{y})$, $\mathrm{Bias}(\overline{y})$ and $\mathrm{MSE}(\overline{y})$.

```
y_bar<-c(3,2,1.5,4,1,0.5,3,0,2.5,5)
sigma_squared<- c(0,2,4.5,2,0,0.5,8,0,12.5,0)
```

```
mean(y_bar)
```

```
## [1] 2.25
```

```
var(y_bar)
```

```
## [1] 2.458333
```

```
bias<- mean(y_bar)-mean(y)
bias
```

```
## [1] 0
```

```r
MSE <- var(y_bar) + (bias)^2
MSE
```

```
## [1] 2.458333
```

and then the expectation and variance of y-bar

(iv) Find E(sˆ2), V(sˆ2)

```r
mean(sigma_squared)
```

```
## [1] 2.95
```

```r
var(sigma_squared)
```

```
## [1] 17.96944
```

(v) Are $\overline{y}$ and sˆ2 unbiased estimators for μ and σ^2?

Yes

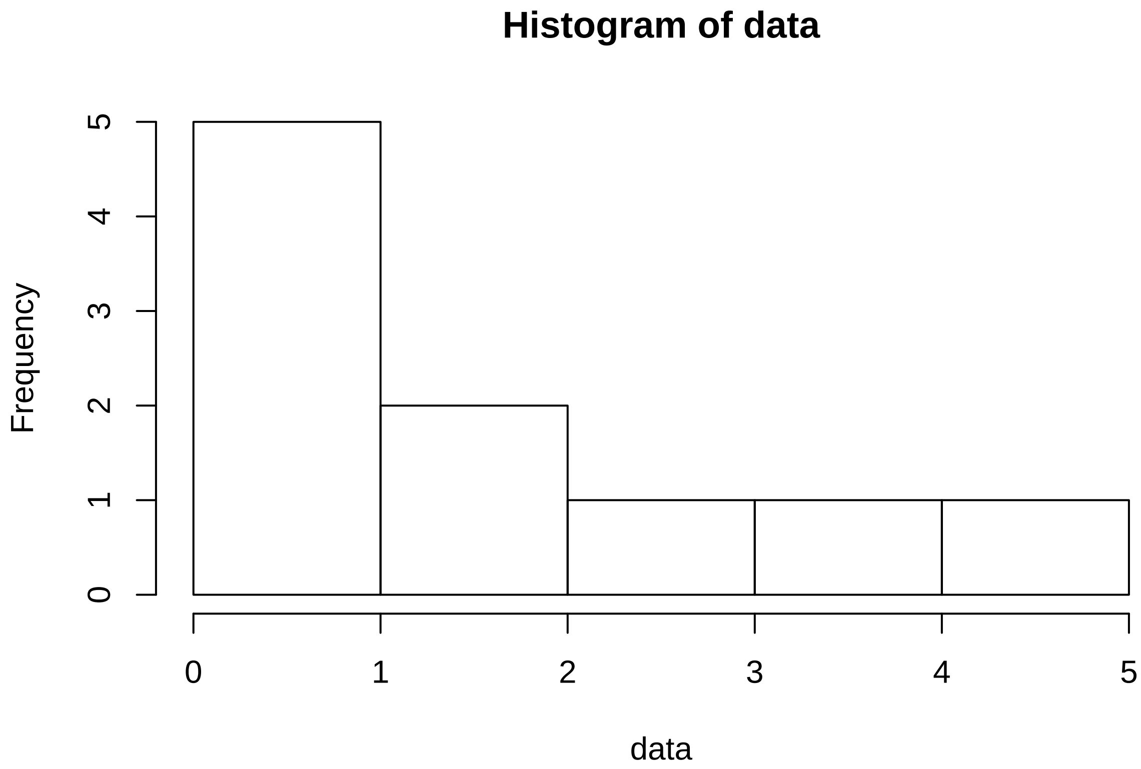
(d) Which of the two sampling plans in (b) and (c) do you think is better for estimating μ? Why?

I think sampling without replacemnt is better to estimate μ because it gives each unit has equal probablity of being chosen

2. Consider all the digits of your U of T student number as a population.

(a) Use a Histogram to visually describe the population digits.

```r
data<-c(1,0,0,2,2,4,0,5,3,1)
hist(data)
```



(b) Find the proportion of 0's and 1's.

```
mean(data=="0")
```

```
## [1] 0.3
```

```
mean(data=='1')
```

```
## [1] 0.2
```

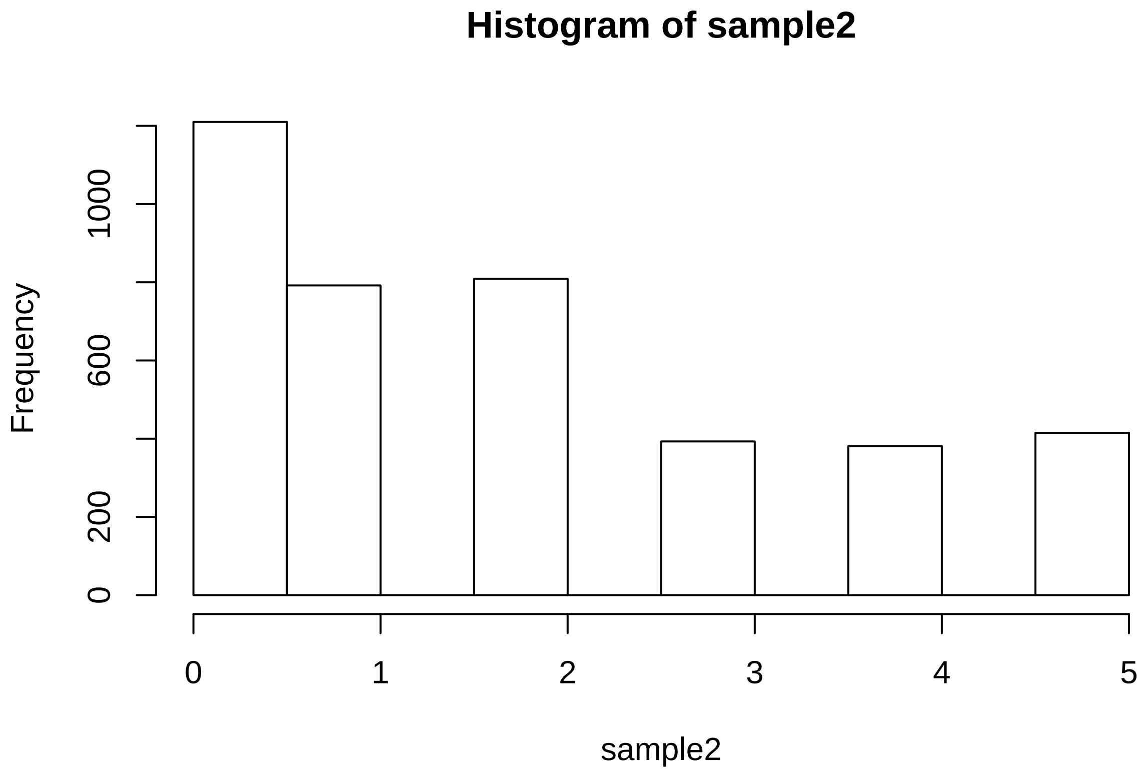
(c) Draw a random sample of 4 digits with replacement and estimate the proportion of 1's.

```
sample1<-sample(data, size=4, replace=TRUE)
mean(sample1=="1")
```

```
## [1] 0.25
```

(d) Generate 1000 random samples of 4 digits with replacement from the population of digits. Plot the distribution of the sample proportion of 1's using a Histogram.

```
sample2<-replicate(1000, sample(data, size=4, rep=TRUE))
hist(sample2)
```



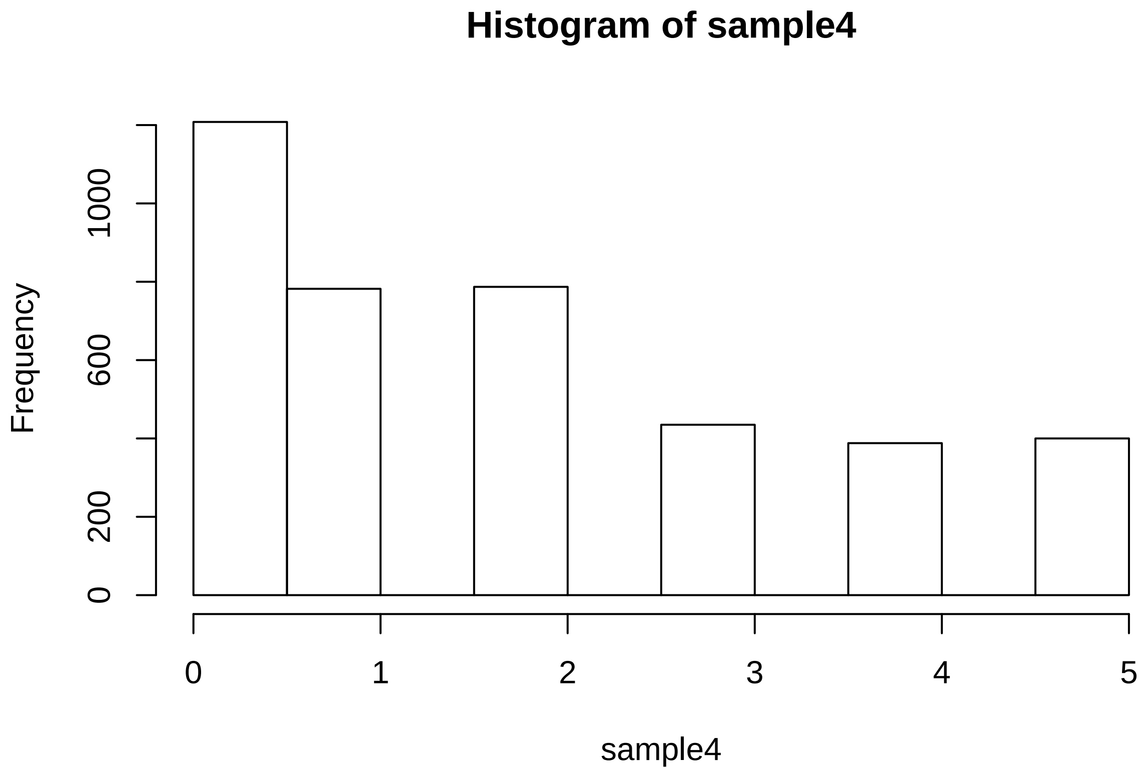
(e) Draw a random sample of 4 digits without replacement and redo (c)

```
sample3<-sample(data, size=4, replace=FALSE)
mean(sample3=="1")
```

```
## [1] 0
```

(f) Generate 1000 random samples of 4 digits without replacement and redo (d)

```
sample4<-replicate(1000, sample(data, size=4, rep=FALSE))
hist(sample4)
```



3. Find a recent survey report in a newspaper, academic journal, or magazine. Include the reference in your assignment and a copy of the full survey in your appendix.

(a) Describe the survey and its question(s) of interest.

In this study, the association between study habits and academic performance of students is examined.

(b) What are the target population and sampled population, sampling frame, sampling unit and observation unit?

A simple random sample of 270 students was taken from two colleges Govt. Allama Iqbal College for Women, Sialkot and Govt. Technical College for boys, Sialkot. The target population for this survey are college students. Sampled population are students of two colleges in Pakistan, Govt. Allama Iqbal College for Women, Sialkot and Govt. Technical College for boys, Sialkot, Sampling frame is list of names of students from the two colleges collected from the school administration. The sampling unit in this study are students, and observation units are human beings.

(c) What conclusions were drawn about the survey in the article?

Conclusions that were drawn from this study are; the relationship between Gender and Stress of Exam is significant, the relationship between Area and effect of family background on result of student is significant, the relationship between Gender and thinking that late night study is harmful is insignificant, the relationship between Gender and coming class late is significant, the relationship between Area and using of time between class for study is insignificant, the relationship between Area and Library is favorite place for study is significant.

(d) What biases do you think might occur in the results? Discuss two possible sources of error in the survey and propose possible ways to alleviate the error.

Biases include; the fact that they only sampled out of convenience of two schools in Pakistan, not representative of all students in Pakistan, let alone all students. To avoid such a bias, they could have expanded their horizons and sampled outside the city as well. Another potential bias may have been; the women's school may have better instructors because students are paying more and hence, their academic performance may be higher compared to the boys school. To avoid this bias, more schools should be sampled from, including a co-ed school as well.

4. The best way to gain understanding of a sampling and estimation method is to carry it out on some real population of interest to you. Read section 2.6 in the required textbook.

Choose a human target population and a population parameter of interest (for example, a mean, a proportion or a total). Suppose you are interested in planning a sample survey and you establish a questionnaire regarding your variable of interest.

Answer the following.

(a) Clearly state your objective. Define your target population. Find a suitable sampling frame. Describe your sample design.

The objective of this survey is to determine if there is any relationship between GPAs and post secondary school attended. The parameter of interest is the proportion of 4.0 GPAs in one school compared to another. We are trying to compare the difficulty levels of schools across Ontario. My target population are students of the major universities in Ontario. A suitable sampling frame would be a list of all students attending the 11 major universities in Ontario from which a sample will be taken. The survey will be carried out using stratified sampling. The entire target population will be divided into 11 stratas; UofT, Ryerson, York, McMaster, Western, Waterloo, Brock, Laurier, Queens, Guelph, and Ottawa, and then a random sample will be collected from each strata. An anonymous online survey consisting of GPA of student, name of institution, program, and year of study will be sent out to each student in these schools and will be highly recommended to complete. The data will be organized by school but also by program and year of study to account for some extenuating factors.

(b) Describe three potential sources of error in questionary design that you would adress in order to minimize the bias.

Double error- each student will be sent a pin that can only be used once to log in into the questionnaire, No answer error- this error can be combatted through an incentive offered Wrong answers- a photo of complete academic history must be provided as well will the rest of the information to avoid GPA estimates, or embarrassed students not wanting to state their GPA.