Math 255 - Homework 1

Colin Pi

Due in class, Friday April 5

Problem 1

Lohr textbook ch. 1 exercise 6 (see prompt on page 19)

- Target population: Female readers of the magazine
- Sampling frame: The list of readers who see the survey in the magazine
- Sampling unit: Individual reader participating in the survey
- Observation unit: Indivisual reader participating in the survey
- Possible source of selection bias: readers who did not see the survey do not have a chance to participate, leading to undercoverage issue. Also, not everyone who sees the prompt may participate in the survey, creating a non-response bias.
- Inaccuracy of responses: Some questions may be constructed in a way that motivate respondents to exaggerate the healthiness of their lifestyles, which is perceived as more socially desirable (leading/loading questions).

Problem 2

Lohr textbook ch. 1 exercise 7

- Target population: Cows in a region
- Sampling frame: The list of local farms
- Sampling unit: Farm
- Observation unit: Cow in a farm
- Possible source of selection bias: If some of the sampled farms decided not to participate in survey and do not provide information (for whatever reason) may result in nonresponse bias.
- Inaccuracy of responses: If the weights of certain farms are not properly calibrated (or malfunctioning), it may cause inaccuracy in responses.

Problem 3

Lohr textbook ch. 1 exercise 13

- Target population: The attendees of 2005 Joint Statistical Meetings
- Sampling frame: The list of conference registrants of 2005 Joint Statistical Meetings
- Sampling unit: Individual attended 2005 Joint Statistical Meetings
- Observation unit: Individual attended 2005 Joint Statistical Meetings
- Possible source of selection bias: The survey is voluntary; therefore, it is not completely free from nonresponse bias.
- Inaccuracy of responses: Reponse errors from misleading wordings or loading questions may create inaccuracy of responses.

Problem 4

Lohr textbook ch. 1 exercise 22
The statement "that's probably just as well" motivates respondents to be more supportive of the idea that female president is socially not desirable, an example of leading question. Also the question is double-barreled.

Two different statements are asked in the same question (1. There won't be a woman president of the U.S. for a long time; 2. That is probably just as well). Even if the respondent agree with one statement but not necessarily agree with the other, she has to respond "Yes" to the question. It gives no choice for the respondents agree on one statement but disagree on the other. Moreover, people tend to agree with a statement when it is given as agree/disagree question than when that statement is presented as one in a list of the options (forced-choice).

Problem 5

Consider the 1936 election polls that were described in the Big Data reading (Financial Times article).

(a) Describe the **target population** for both the Gallup and Literary Digest polls, then describe the **sampled population** for the Literary Digest poll.

Gallup

- Target population: voting population (electorates of 1936 Presidential Election) who is willing to vote.

Literary Digest polls

- Target population: voting population (electorates of 1936 Presidential Election) who is willing to vote.
- Sampled population: 2.4 million people who responded to the postal opinion poll (of 10 million eligible voters who are listed in the telephone directories and automobile registrations).

(b) For the Literary Digest poll, describe one source of undercoverage bias and one source of non-response bias.

- Undercoverage: People who are not in the list of automobile registrations or telephone directories did not have a chance to be selected as a sample, even if they are eligible to vote. People who are in these lists are way more prosperous relative to people not in these lists. Therefore, the poll is disproportionally representative to the consensus of a higher income class.
- Non-response bias: Out of 10 million polls sents, only 2.4 million of them are returned, suggesting that around 80% of the recipients decided not to participate in the poll. Moreoever, Landon supporters turned out to be more responsive to the poll than FDR supporters, implying that the poll does not truly represent the opinions of the voting population.

Problem 6

Probelm 6 has most, but not all, R commands provided in the Homework1.Rmd file linked to on the class moodle page. Upload this .Rmd file to your Math 255 folder on the Mirage server (create this folder in Mirage). Then open the .Rmd in RStudio and compile it into a pdf document using the Knit PDF button. See the R/Rstudio resource link at the top of the course Moodle page for more info on R Markdown and come see me if you've never use R Markdown before.

Work through this assignment and answer all questions in the R Markdown (.Rmd) file. Compile your completed assignment into a pdf (or html, either is fine). You can answer Problems 1-5 in this .Rmd file too, or write these up by hand.

Idea

In class we compared sampling designs for a very simple problem (sampling 2 elements from a population of 3 elements) using (probability) theory to study properties of the two sampling design . Another method of comparing sampling designs is to use a computer simulation. You will use a simulation study to compare two methods of random sampling from a finite population: sampling with replacement vs. sampling without replacement (commonly known as a Simple Random Sample - SRS). You will make comparisons of the behavior of the sample mean under these two different sampling designs. In particular, you will look at the *bias* and *standard error* of the sample mean estimator.

The population

The file StatsClassSurvey.csv is a comma separated values file that contains responses from a large group of statistics students. This will be your population. Read the data into R and explore the contents of this file.

```
> pop <- read.csv("http://math.carleton.edu/kstclair/data/StatsClassSurvey.csv")
> names(pop)
[1] "Sex"      "Height"   "Exercise" "Class"    "Driving"
> head(pop)
     Sex Height Exercise   Class Driving
1 Female   63.0      180 Math115 Average
2 Female   65.0      420 Math115 Average
3 Female   66.0      300 Math115 Average
4 Female   67.0      360 Math115 Average
5 Female   66.0      120 Math115 Average
6 Female   61.5      700 Math115 Average
```

Note that `pop` is an **R object** called a **data frame**. Click on the file name in the Data section of your Rstudio **Environment** pane to see the data.

(a) Questions: How many variables are in this file? Which are categorical and which are quantitative?

There are five variables. Sex, Class, and Driving are categorical variables. Height and Exercise are quantitative variables.

The population parameter

We will look at the height variable (measured in inches) and investigate how to estimate the average height in the population.

```
> pop$Height  # all heights in the population
  [1] 63.00 65.00 66.00 67.00 66.00 61.50 62.00 70.00 71.00 65.00 62.00
 [12] 64.00 64.00 70.00 72.00 70.00 74.00 72.00 66.00 66.00 72.00 74.00
 [23] 75.00 60.00 70.00 74.00 69.00 64.00 65.00 62.00 66.00 65.00 66.50
 [34] 61.00 62.00 63.00 64.00 65.00 65.00 65.00 66.00 65.00 75.00 70.00
 [45] 60.00 72.00 73.00 77.00 66.50 70.00 72.00 75.00 66.00 65.00 65.00
 [56] 65.00 65.00 64.00 64.00 70.00 69.00 70.00 71.00 68.00 70.08 67.00
 [67] 76.00 75.00 64.00 66.00 64.50 67.00 75.00 66.00 72.00 70.00 75.00
 [78] 64.00 75.00 69.00 62.00 65.00 65.00 67.00 62.00 66.00 68.00 70.00
 [89] 70.00 71.00 67.00 70.00 71.00 73.00 71.00 71.00 73.00 74.00 71.00
[100] 67.00 71.00 74.00 67.00 68.00 67.00 67.00 66.00 66.00 68.00 72.00
[111] 63.00 64.00 78.00 65.00 65.00 65.00 71.50 66.00 64.00 62.00 66.00
[122] 62.00 70.00 72.00 68.90 68.00 68.00 70.50
> N <- length(pop$Height)
> N  # population size
[1] 128
> pop.mean <- mean(pop$Height)
> pop.mean  # population mean for Height
[1] 67.90609
> summary(pop$Height)  # other population parameters for Height
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.00   65.00   67.00   67.91   71.00   78.00
```

(b) Question: What is the population size and population mean height?

- Population size: 128
- Population mean height: 67.91 inches

The simulation

We will next use an R simulation to compare how the sample mean behaves (i.e. the sample mean's "sampling distribution") under two different sampling designs: random sampling **with** replacement vs. random sampling **without** replacement (aka Simple Random Sample). The basic idea behind this simulation is:

1. pick sample size n
2. draw a random sample of size n (either with or without replacement) from the population
3. compute the sample mean
4. repeat steps 2-3 a large number of times ("`reps`") and compare how the sample mean behaves under the different sampling designs

Below are these basic steps translated into R. Keep the number of `reps` set at 1 (for now) to answer the questions below (c).

```
> n <- 10  # sample size
> reps <- 1  # number of repetitions
> est.repl <- rep(NA, reps)  # vector for with replacement
> est.norepl <- rep(NA, reps)  # vector for without replacement
>
> set.seed(77)
> for (i in 1:reps) {
+     # with replacement:
+     samp.repl <- sample(1:N, n, replace = T)  # units sampled
+     data1 <- pop$Height[samp.repl]  # height of units sampled
+     est.repl[i] <- mean(data1)  # sample mean
+     # without replacement:
+     samp.norepl <- sample(1:N, n, replace = F)  # units sampled
+     data2 <- pop$Height[samp.norepl]  # height of units sampled
+     est.norepl[i] <- mean(data2)  # sample mean
+ }
```

Comment: The `set.seed()` command will make your random simulation results reproducible - i.e. the same random samples will be produced when using the same seed. If it wasn't set prior to using the `sample` command you would get different results each time you run the simulation (or knit this document). Changing the seed value, or deleting the command, will produce a different random sample. Feel free to change the seed to some other integer value but keep the `set.seed` command to ensure your written answers below match the results shown in the R code.

(c) Questions: Your population was enumerated 1 to N. What are the numbers and heights of the units that you sampled for your with replacement sample? What is the average height in this sample? Repeat these two questions using your without replacement sample. Are your two sample means the same? Why or why not?

With replacement sampling design allow inclusion of the same person in the sample, while without replacement sampling design doesn't. For instance, unit 109 appears twice in the sample with replacement. Therefore, the drawn samples vary depending on the sampling design, resulting in difference in the two sample means.

With Replacement

- Numbers: 38, 92, 111, 122, 95, 59, 109, 109, 73, 59
- Heights: 65, 70, 63, 62, 71, 64, 68, 68, 75, 64
- Average Height: 67

Without Replacement

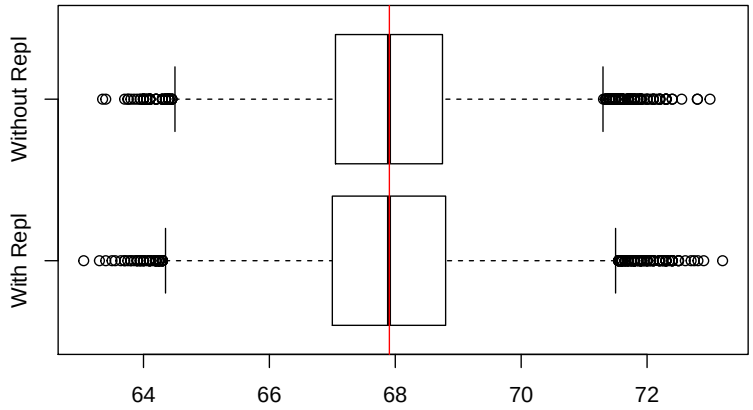
- Numbers: 112, 126, 21, 1, 56, 45, 69, 86, 128, 2
- Heights: 64, 68, 72, 63, 65, 60, 64, 66, 70.5, 65
- Average Height: 65.75

Next we need to generate lots of samples (using each design) to explore the distribution of the sample mean. We will run the simulation above again but change the values of `reps` to 50000:

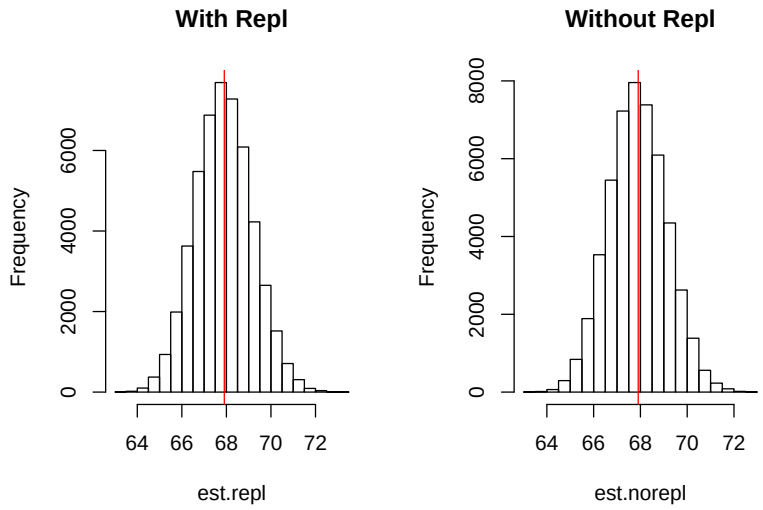
```
> n <- 10  # sample size
> reps <- 50000  # number of repetitions
> est.repl <- rep(NA, reps)  # vector for with replacement
> est.norepl <- rep(NA, reps)  # vector for without replacement
>
> set.seed(77)
> for (i in 1:reps) {
+     # with replacement:
+     samp.repl <- sample(1:N, n, replace = T)  # units sampled
+     data1 <- pop$Height[samp.repl]  # height of units sampled
+     est.repl[i] <- mean(data1)  # sample mean
+     # without replacement:
+     samp.norepl <- sample(1:N, n, replace = F)  # units sampled
+     data2 <- pop$Height[samp.norepl]  # height of units sampled
+     est.norepl[i] <- mean(data2)  # sample mean
+ }
```

Then run the commands below to compare the sampling distributions of the sample mean under the two sampling designs.

```
> # compare sampling distributions of the sample mean for each
> # type of design
> boxplot(est.repl, est.norepl, names = c("With Repl", "Without Repl"),
+     horizontal = TRUE)
> abline(v = pop.mean, col = "red")
```




```
> par(mfrow = c(1, 2))
> hist(est.repl, main = "With Repl")
> abline(v = pop.mean, col = "red")
> hist(est.norepl, main = "Without Repl")
> abline(v = pop.mean, col = "red")
```




```
> par(mfrow = c(1, 1))
>
> summary(est.repl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  63.05   67.00   67.90   67.91   68.80   73.20
> summary(est.norepl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  63.35   67.05   67.90   67.91   68.75   73.00
>
> # Bias: average estimate value - pop.mean
> mean(est.repl) - pop.mean  # with replacement bias
[1] 0.00707393
> mean(est.norepl) - pop.mean  # without replacement bias
[1] 0.00262581
>
> # SE: standard deviation of estimates
> sd(est.repl)  # with replacement: SE of sample mean
[1] 1.282308
> sd(est.norepl)  # without replacement: SE of sample mean
[1] 1.236475
```

(d) Questions: Are the sampling distributions skewed or symmetric? Are they roughly normal? (you can check histograms too for modes) Do both designs yield (roughly) unbiased estimates of the population mean? What are the SEs of the sample mean for each design? Which design offers you more precision when estimating the population mean?

Sampling distributions are symmetric and roughly normal. The bias of estimator yielded from both designs are close to 0, suggesting that they both produce unbiased estimates, yet the with replacement bias (0.0071) is slightly higher than without replacement bias (0.0026). SEs of the sample mean for both designs are around 1.2, but SE for the sample mean with replacement is slightly higher (1.2823) than SE for the sample mean without replacement (1.2365). This suggests that without replacement offers more precise estimation of the population mean.

The Finite Population Correction (FPC)

The multiplicative factor that makes the without replacement SE smaller than the with replacement SE is the square root of the **finite population correction (FPC)**. The ratio of variances (*squared* SEs) from

the simulation approximates the FPC, but the exact formula for the FPC is $1 - n/N$.

```
> (sd(est.norepl)/sd(est.repl))^2  # without SE < with SE
[1] 0.9297932
> 1 - n/N  # why? the Finite Pop. Correction (FPC)
[1] 0.921875
```

(e) Questions: How much (by what factor) is the SE of the sample mean reduced by using the without replacement random sample for $n = 10$ when compared to a with replacement random sample? Will this factor increase or decrease if we change the sample size to $n = 50$? Give an intuitive (not formulaic!) explanation for why there is less variability in the sample mean when you use a without replacement sampling design compared to a with replacement design.

When n = 10, FPC = 1-10/128 = 118/128. This means the SE of the sample mean using the without replacement random sample is $\sqrt{118/128}$ (around 96.01%) of the SE of sample mean using with replacement random sample. When n = 50, FPC = 1-50/128 = 78/128, suggesting that the factor decreases as sample size increases. In other words, the SE of the sample mean decreases as sample size increases. This makes sense because there is more information about the population as the sample size (relative to the population size) increases, thus reducing the variability of the estimates.