Lab 6: Introducing the Central Limit Theorem

(Optional) Relevant Textbook Exercises

The following questions in your textbook are helpful practice for understanding today's material:

Baldi and Moore: Ex. 13.5, 13.8, 13.9, 13. 10, 13.12, 13.14

Introduction

This is a collaborative assignment that the entire lab section will conduct together. We will all contribute to a central data source (a Google sheet) and the GSI will summarize the results as the class progresses. The purpose is to develop a very concrete idea of sampling distributions, and to see the central limit theorem in (live) action.

The underlying population

Suppose you had a data frame containing the **entire population** of all residents of Alameda County. You have data on three variables:

1. Born either out (=1) versus in (=0) the county.
2. Number of siblings (integer)
3. Number of visits to the hospital last year
4. Read in the data, L06_Alameda.csv, and assign it to the name `alameda_pop`. Calculate the true (population) mean, and make histograms or bar charts of the distribution for each variable.

```
library(dplyr)
library(ggplot2)

#### YOUR CODE GOES HERE ####
```

Notice that the distribution of number of siblings and number of hospital visits are discrete distributions and both skewed right. Today, we will focus on the distribution of the number of siblings `num_sibs`. Remember, we know the population mean exactly, because we have all the data. You calculated the underlying population mean in the code chunk above.

Calculating the sampling distribution

We are now going to look at the approximate **sampling distribution for the sample mean** of the `num_sibs` in (live) action. Remember from earlier lectures that a **sampling distribution** is a distribution for a statistic, like the sample proportion or the sample mean.

Each student will be tasked with repeatedly taking a random sample of the population. Once you take your sample you will compute the sample mean and add it to a shared Google sheet. As the data is added to the Google sheet, your GSI will graph the results to illustrate how the sampling distribution varies for increasingly larger sample sizes.

The GSIs will provide you the link to the communal google sheet. The columns in the sheets are `n (Sample_size)` `mean(numSibs)` `Name (your sign in)`.

Your task

1. Randomly generate 10 simple random samples of size $n = 5$ from the population from each of the three target populations. Get the average of each of these to get 10 averages. You can simply re-run the random sampling procedure 5 times, get the average each time and paste it. Paste it into the provided google sheet making sure to type in the sample size the average is based upon in the appropriate column.

```
#### YOUR CODE GOES HERE ####
```

After you've calculated 10 sample means using the above code, copy and paste your data into the google sheet for your lab section. The links to all the google sheets are:

- 101B (Thursday 5-7pm): https://docs.google.com/spreadsheets/d/130FOigRrcdzyMpxSmzxEcdx2TOthUeKZFBpfJgHpb
  edit?usp=sharing
- 102B (Friday 3-5pm 30 wheeler): https://docs.google.com/spreadsheets/d/1sVx2Vd_57DWNHRvQ1UP_BhaMuRUnc-H-L-f5Ujuv-fs/edit?usp=sharing
- 103B (Friday 4-6pm 151 barrows): https://docs.google.com/spreadsheets/d/1sJpPRIyl83CcFwOyfwJCgyM1LubDZx-RvA
  edit?usp=sharing
- 104B (Friday 10-12am): https://docs.google.com/spreadsheets/d/1iOUF5bohUL3_tEHkjOjXC_bueFVLm-sA-lWtxHFLgS0/edit?usp=sharing
- 105B (Thursday 11-1pm):
  https://docs.google.com/spreadsheets/d/1PwcGfK4dKbybxFfrwXm8-H4yCzKVBzCenDaf902Btlg/edit?usp=sharing
- 106B (Friday 3-5pm 185 Barrows): https://docs.google.com/spreadsheets/d/1AP3oewsuzDDpG6oPnMEqXXjzYNY1Q2E
  edit?usp=sharing
- 107B (Friday 3-5pm 587 Barrows): https://docs.google.com/spreadsheets/d/15JdSP4V3-K5BY76sxRXL3wQcDWhNkuA
  edit?usp=sharing
- 108B (Friday 4-6pm 175 Barrows): https://docs.google.com/spreadsheets/d/151RrhBvC33sb_4mJDl5ibNOPJRKvKQK4LlCCHg-1pDA/edit?usp=sharing
- 109B (Thursday 6-8am): https://docs.google.com/spreadsheets/d/1bR2eXPNrdKIi_Cxw5nllQFPwR3-KWcvhA93a6IfRx
  edit?usp=sharing
- 110B (Friday 4-6pm 155 Barrows): https://docs.google.com/spreadsheets/d/17Td3U8t4MoH0nCOTevhJjsiENPaYAWT_egRclsxM9IM/edit?usp=sharing

Once the sheet is full, look at the plot of the **sampling distribution** for the mean number of siblings when $n = 5$.

- What is the range of the sampling distribution of the mean?

Once the class has examined the sampling distribution when $n = 5$, repeat the same steps for n=50.

2. Repeat for a sample size of $n = 50$

```
#### YOUR CODE GOES HERE ####
```

After you calculated your 10 sample means, navigate to the google sheet from before, but switch to the sheet with $n = 50$ (you can switch tabs in the bottom left). Add your data for $n = 50$.

Once this is done, look at the plot to the right; now with $n = 50$

- What is the range of the sampling distribution of the mean? How does it compare to when $n = 5$?

3. Repeat for sample size $n = 500$.

```
#### YOUR CODE GOES HERE ####
```

- What is the range of the sampling distribution of the mean? How does it compare to when $n = 5$ and $n = 50$?

4. Suppose you have 500 classmates, and they have done this lab and added their data to this sheet: https://docs.google.com/spreadsheets/d/1AXStOd618raoWvrBequxOh5CDwgisFJoHo50fmcKb_E/edit?usp=sharing

Open the link, and look at the resulting sampling distributions for $n = 5$, $n = 50$, and $n = 500$. This is what happens when you repeat the sampling 5,000 times.

5. For which sample size should the sampling distribution of the mean be most normal?

(a) n=5
(b) n=50
(c) n=500

Assign your letter choice as a string. Example: sampleSize_answer<-“b”

```
sampleSize_answer<- "REPLACE WITH a,b, or c. Keep the quotes"
sampleSize_answer
```

```
## [1] "REPLACE WITH a,b, or c. Keep the quotes"
```

```
check_problem1()
```

```
## [1] "Checkpoint 1 Error: Wrong"
##
## Problem 1
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```
# Just run this chunk.
total_score()
```

```
##             Test Points_Possible       Type
## Problem 1 FAILED               1 autograded
```

Submission

For assignments in this class, you'll be submitting using the **Terminal** tab in the pane below. In order for the submission to work properly, make sure that:

1. Any image files you add that are needed to knit the file are in the `src` folder and file paths are specified accordingly.
2. You **have not changed the file name** of the assignment.
3. The file is saved (the file name in the tab should be **black**, not red with an asterisk).
4. The file knits properly.

Once you have checked these items, you can proceed to submit your assignment.

1. Click on the **Terminal** tab in the pane below.
2. Copy-paste the following line of code into the terminal and press enter.

```
cd; cd ph142-sp20/lab/lab06; python3 turn_in.py
```

3. Follow the prompts to enter your Gradescope username and password. When entering your password, you won't see anything come up on the screen–don't worry! This is just for security purposes–just keep typing and hit enter.
4. If the submission is successful, you should see "Submission successful!" appear as output.
5. If the submission fails, try to diagnose the issue using the error messages–if you have problems, post on Piazza.

The late policy will be strictly enforced, **no matter the reason**, including submission issues, so be sure to submit early enough to have time to diagnose issues if problems arise.