Math 154, Exam 1 - take home portion

Ra-Zakee Muhammad

due: Sunday, Oct 24, 11:59pm

PROCEDURE: READ FIRST BEFORE STARTING!!!

The take-home part of the midterm is due Sunday, October 24 at midnight. Your assignment should be turned in to the GitHub repo you created based on the assignment + Gradescope, just like the HW assignments.

You are to do these problems ENTIRELY ON YOUR OWN - with no human help. You may consult your textbook, class handouts, solutions to homeworks, and any notes which you wrote. You may NOT use any other books or other online materials (see below). You may use a calculator. You can take as long as you like to do this exam. You may want to read through the problems and cogitate on them before starting the fun activity.

The only **person** from whom you may solicit help is ME. You can do this by email (`jo.hardin@pomona.edu`), DM in Discord, or in person (Estella 2351). Be sure to ask if you don't understand something.

Please remember to explain your reasoning, when appropriate. Don't write extraneous things in your answers - keep them concise and clear. Your work should be done using Markdown in R Studio and compiled as a single reproducible document (including code). However, there should be a narrative surrounding the code and explaining the results of the output.

Internet: You may not use the internet for anything except: the R software program (okay to use stack exchange to figure out programming errors; okay to Google to figure out R commands, etc.)); the HW solutions on Sakai; the class notes; and the course website. The main idea here: no Googling how to **do** the problem; it is okay to Google to figure out the R syntax.

If you use the internet include the URL(s) from which you found the help (if the website is not from the course materials).

If you use another person Some of you may know that I've dealt with cheating recently. If there is any evidence of collaboration between two or more individuals, you will be immediately reported to the Dean of Students. I will be looking very carefully through the code for unexplained similar patterns to your code solutions.

Obligatory Honor Pledge Notice: Please don't cheat. It isn't fun for any of us.

You should post both .Rmd and .pdf files to your GitHub repository + Gradescope. The .Rmd file you post should compile to the .pdf file. Before posting to Gradescope, check your GitHub repository to confirm that the .pdf file you posted is the one you wanted to post.

1. (+3 pts) Send me an email with a reproducible example (`reprex()`). The email should have the following items:

Note: an error message of can't find data / can't find function / etc. will not receive full credit.

(a) A question related to the code.
(b) A description of the output you see when you run the code.
(c) Reproducible code
(d) I will copy and paste your code into my R console, it must compile.
(e) The compiled R code must produce the same output that you describe in (b).
(f) The compiled R code must address the question that you pose in (a).

Feel free to email the `reprex()` at any point, you do not need to wait until you are turning in the exam! Also, feel free to Google "how do I create a reproducible example?" You can use any non-human resource you can find to do this part of the problem.

2. Spoiler: this is a **new** idea that hasn't been introduced in class. Don't worry, you can do it! You have all the background to solve the problem. We are going to use bootstrapping to address a hypothesis test. It is going to be awesome.

In 1912 in the US it seems as though women were, on average, 63.75 inches.[1]

We want to evaluate whether the average height of women in the US has increased since 1912.

The null hypothesis states "there is nothing going on", i.e., no change since 1912: $H_0 : \mu = 63.75$ inches.

The alternative hypothesis reflects the research question, i.e., women are taller now than in 1912. Since the research question does state a direction for the change, use a one-sided alternative hypothesis: $H_A : \mu > 63.75$ inches.

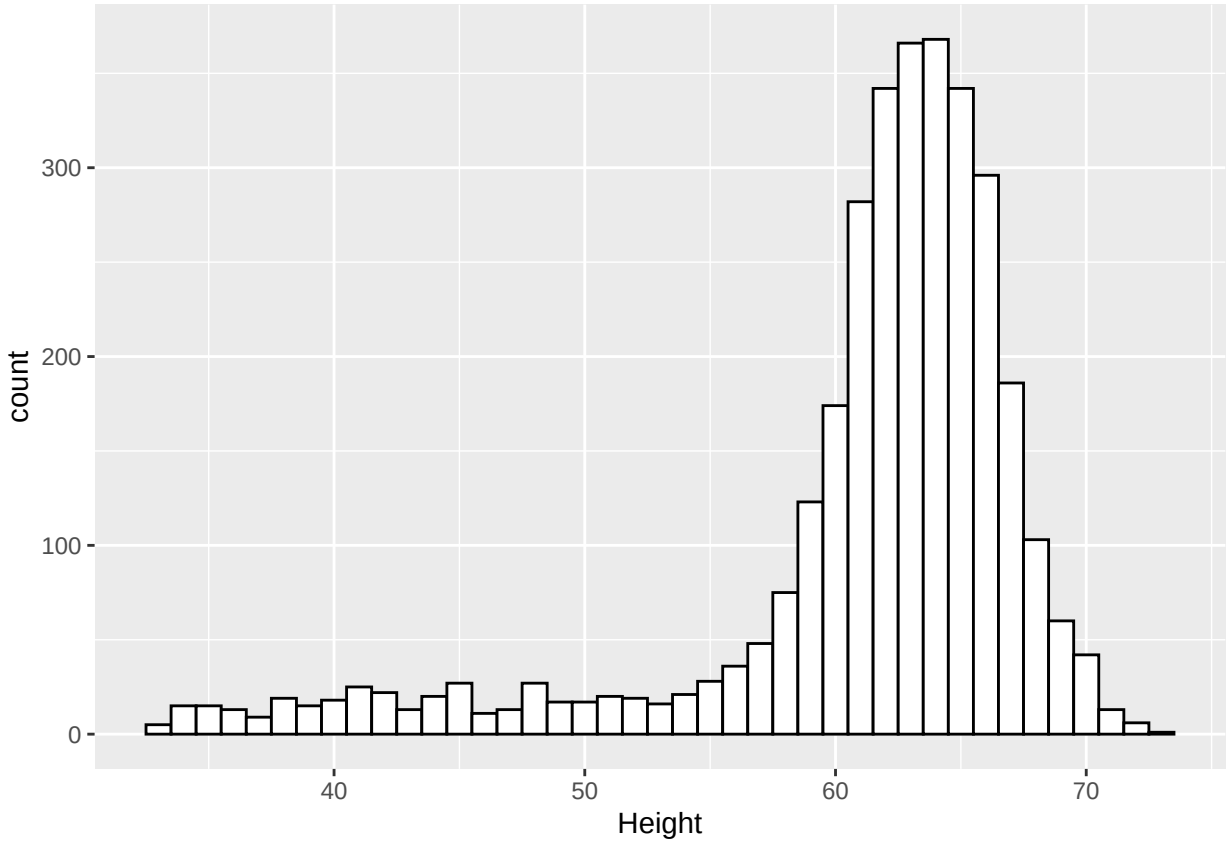
(a) (+3 pts) Use the NHANES data (information from roughly 100 years later). Filter the NHANES data so that the data analysis is inline with the hypotheses (hint: there are two variables needed to filter and make sure the units (inches or cm?) are correct). Create a histogram of the height variable.

```
library(NHANES)
 data(NHANES)
```

```
  womens_heights <- NHANES %>%
  filter(Gender == "female", !is.na(Height)) %>%
  mutate(Height = 0.393701 * Height) %>%
  distinct(ID, Height)
```

```
  womens_heights %>%
  ggplot(mapping = aes(x = Height)) + geom_histogram( color = "black", fill = "white", binwidth = 1)
```

[1] Not a reliable source: https://ahundredyearsago.com/2012/02/06/average-height-for-males-and-females-in-1912-and-2012/



(b) (+1 pt) What is the average height of women in the modified dataset?

```
 womens_heights %>%
 summarise( Average_womens_height = mean(Height))
```

```
## # A tibble: 1 x 1
##   Average_womens_height
##                   <dbl>
## 1                61.144
```

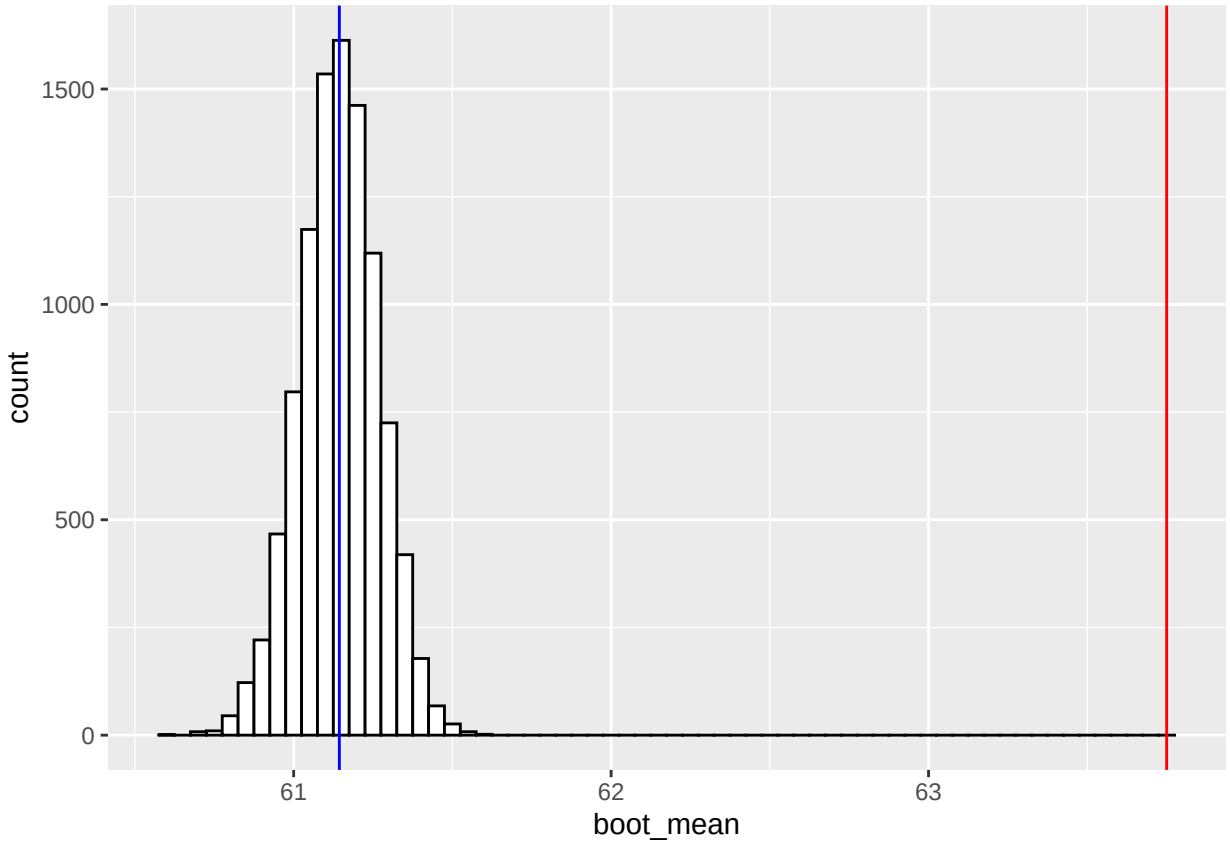
(c) (+5 pts) Create a bootstrap distribution (histogram) of the sample mean. On the plot add vertical lines at both the observed mean as well as the hypothesized mean. Color the observed line blue and the hypothesized line red. [Hint: follow the R code given in the slides on bootstrapping for 10/5, the function that bootstraps and outputs a statistic will work nicely here.]

```
boot_stat_func <- function(df){
    df %>%
    mutate(obs_mean = mean(Height)) %>%
    sample_frac(size=1, replace=TRUE) %>%
    summarize(boot_mean = mean(Height),
              obs_mean = mean(obs_mean))}
```

```
set.seed(4747)
bootstrap_data <- map_df(1:10000,~boot_stat_func(womens_heights))
bootstrap_data
```

```
## # A tibble: 10,000 x 2
##    boot_mean obs_mean
##        <dbl>    <dbl>
##  1    61.187   61.144
##  2    61.118   61.144
##  3    61.044   61.144
##  4    61.092   61.144
##  5    61.198   61.144
##  6    61.038   61.144
##  7    61.098   61.144
##  8    61.068   61.144
##  9    61.090   61.144
## 10    61.098   61.144
## # ... with 9,990 more rows
```

```
bootstrap_data %>%
  ggplot(aes(x = boot_mean))+
  geom_histogram(color = "black", fill = "white", binwidth = .05)+
  geom_vline(aes(xintercept = mean(obs_mean)), colour="blue")+
  geom_vline(aes(xintercept = 63.75), colour="red")
```

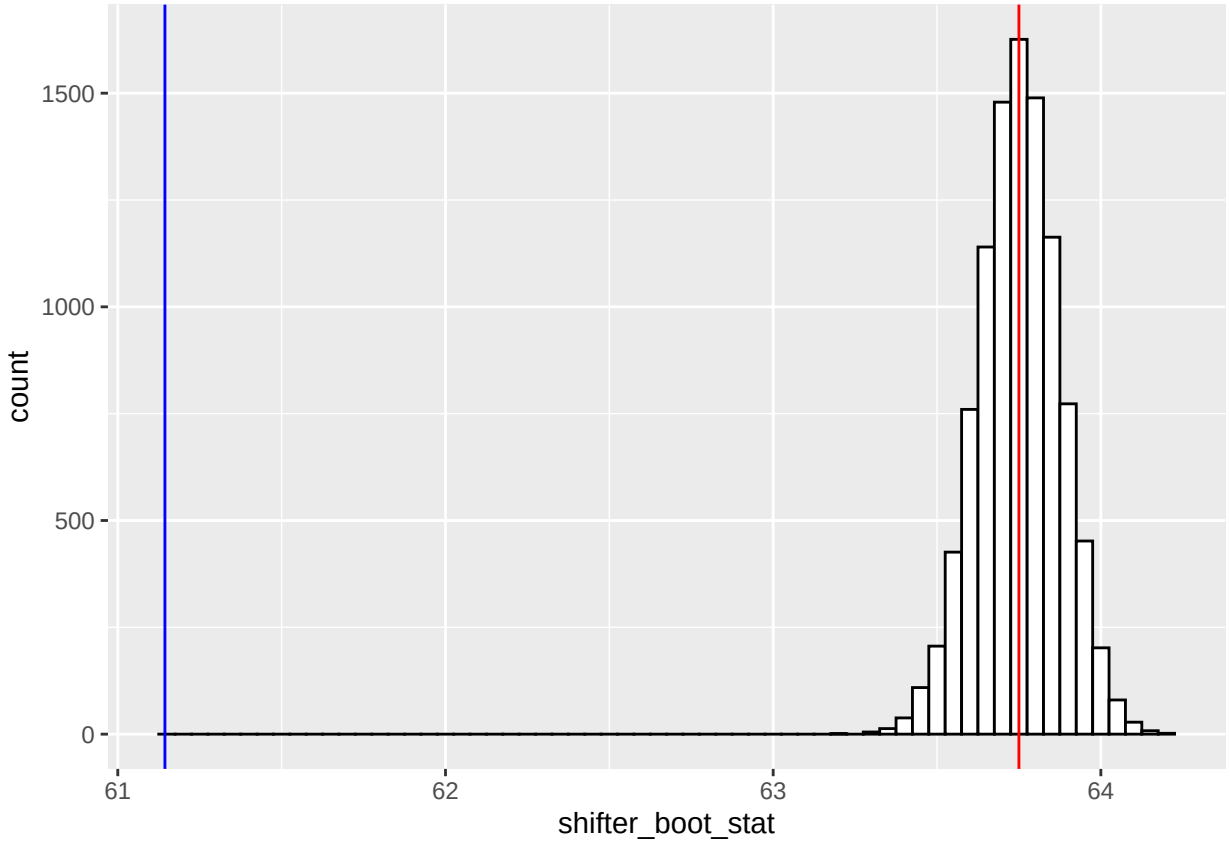


(d) (+2 pts) As you know, the bootstrap distribution describes the shape and spread of the sampling distribution (without any hypotheses). If you are interested in the *null* sampling distribution, where should the center of the data distribution be?
Shift the bootstrap distribution by the appropriate amount and re-plot the histogram as the best estimate (assuming you don't know any central limit theorem theory) for the null sampling distribution of the sample mean under the assumption that the null hypothesis is true.

```
shifted_dist <- bootstrap_data %>% mutate(shifter_boot_stat = boot_mean + (63.75- 61.144))
shifted_dist
```

```
## # A tibble: 10,000 x 3
##    boot_mean obs_mean shifter_boot_stat
##        <dbl>    <dbl>             <dbl>
##  1    61.187   61.144            63.793
##  2    61.118   61.144            63.724
##  3    61.044   61.144            63.650
##  4    61.092   61.144            63.698
##  5    61.198   61.144            63.804
##  6    61.038   61.144            63.644
##  7    61.098   61.144            63.704
##  8    61.068   61.144            63.674
##  9    61.090   61.144            63.696
## 10    61.098   61.144            63.704
## # ... with 9,990 more rows
```

```
shifted_dist %>%
  ggplot(aes(x = shifter_boot_stat))+
  geom_histogram(color = "black", fill = "white", binwidth = .05)+
  geom_vline(aes(xintercept =  mean(obs_mean)), colour="blue")+
  geom_vline(aes(xintercept = 63.75), colour="red")
```



(e) (+2 pts) Using the null-shifted-bootstrapped distribution from the previous question, calculate a p-value to address the research question at hand.

```
shifted_dist %>%
  summarize(p_val =
      sum(shifter_boot_stat > 61.144) /
      10000)
```

```
## # A tibble: 1 x 1
##   p_val
##   <dbl>
## 1     1
```

(f) (+2 pts) Using the number calculated for the p-value, conclude the problem in context using words like 1912 and 2012 and height.

Because our p-val of 1 doesn't fall below the .05 statistical significance threshold, we fail to reject the null hypothesis meaning we do not have significant reason to believe average height of women increased from the year of 1912 to 2012. We maintain the null hypothesis that average women's height remained the same from the year of 1912 to 2012.

(g) (+2 pts) Provide 1-2 sentences describing how the null sampling distribution was created – which part of the process above created each of the center, spread, and shape of the sampling distribution?

The null sampling distribution was created by first bootstrapping the sample mean of data regarding present heights of women 10000 different times and using this bootstrapped data to approximate a sample mean distribution of women's heights in the present time. This gave us shape and spread of our null sampling distribution, but not center because the distribution centered around the observed mean of the provided data; in order to get the center of the null sampling distribution we then had to shift all sample means in our bootstrap distribution by the difference between the null parameter of 63.75 and our observed mean since the null sample distribution should be centered around the true population parameter, and H0 assumes that to be 63.75.