

## HW3

● Graded

Student

Sangwon Ji

Total Points

45 / 45 pts

Question 1

Question 1

5 / 5 pts

✓ - 0 pts Plot A corresponds to the bootstrap and plot B corresponds to the permutation test.

The distribution of bootstrapped statistics should be centered around the median difference in the sample and the permuted statistics should be centered around zero. Correct

- 2 pts Incorrect answer, correct reasoning

- 2 pts Correct answer, incorrect reasoning

Question 2

Question 2

10 / 10 pts

✓ - 0 pts The first method bootFun1 will separate the two groups and sample from them separately, so that the bootstrap samples for group1 and group2 have the same sample sizes as the original group1 and group2.

The second method bootFun2 will not separate the two groups and will sample from both combined. Bootstrap samples from method 2 will have different sample sizes for group1 and group2.

- 8 pts Vague answer

- 4 pts Correct answer for bootFun1 and incorrect answer for bootFun2 or vice-versa

- 10 pts Missing answer

## Question 3

## Question 3

30 / 30 pts

3.1 a

5 / 5 pts

```
✓ - 0 pts ## subsetting data into two groups
group1 <- subset(heart, cp==1 | cp==2)
group2 <- subset(heart, cp==3)

#running t-test on chol and getting confidence intervals
t.test(group1$chol, group2$chol)$conf.int
`

CI: (-18,13)
```

- 1 pt Incorrect groups.
- 1 pt Has t-test, but does not report the CI
- 3 pts Did not t-test chol
- 5 pts Missing Answer
- 3 pts Did not use t-test

3.2 b

15 / 15 pts

```
✓ - 0 pts bootOnce <- function(x,y){
  x = sample(x, replace = T)
  y = sample(y, replace = T)
  diff = mean(x) - mean(y)
  return(diff)
}

bootOnce(group1$chol[!is.na(group1$chol)], group2$chol[!is.na(group2$chol)])

diffs <- replicate(1000, bootOnce(group1$chol[!is.na(group1$chol)],
  group2$chol[!is.na(group2$chol)]))

LB = quantile(diffs,0.025)
UB = quantile(diffs,0.975)

print(paste("CI: (",round(LB,2),"",round(UB,2),"")`)

CI ~ (-19,14)
```

- 5 pts Mistake in BootOnce function
- 2 pts Mistake in calculating the right quantiles
- 1 pt Incorrect groups
- 15 pts Missing Answer
- 2 pts Missing bootstrap

```
✓ - 0 pts `BootSerum <- function(group1, group2){
  sample1 <- sample(1:nrow(group1), replace = T)
  sample2 <- sample(1:nrow(group2), replace = T)

  gp1 <- group1[sample1,]
  gp2 <- group2[sample2,]


  diff = mean(gp1$chol > 200) - mean(gp2$chol > 200)
  return(diff)
}

diffs = replicate(1000,BootSerum(group1,group2))
LB = quantile(diffs,0.025)
UB = quantile(diffs,0.975)

print(paste("CI: (",round(LB,2),",",round(UB,2),")"))`

CI ~ (-0.04, 0.19)
or
CI ~ (-4%,19%)
```

- 3 pts Error in sampling the right variable
- 1 pt Mistake in calculating the right quantile
- 10 pts Missing Answer

 Regrade Request

Submitted on: Mar 25

Hi,

I thought I had to multiply 100 as it was asking for the difference in percent of the two groups ?

points adjusted

Reviewed on: Apr 04

## Question 4

## Late Penalty

0 / 0 pts

✓ - 0 pts On time

- 4.5 pts 1 day
- 9 pts 2 days late
- 13.5 pts 3 days late

Question assigned to the following page: [1](#)

# HW3

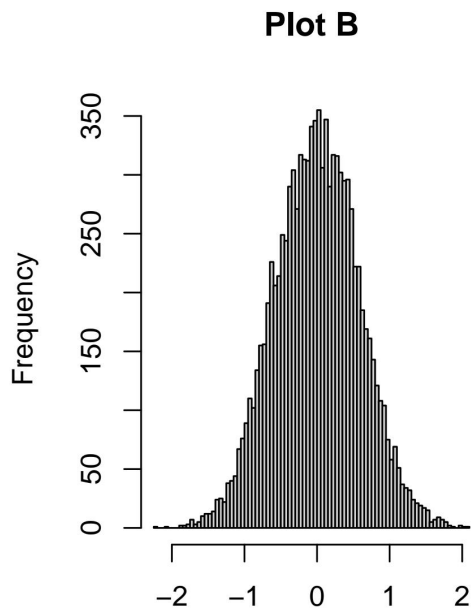
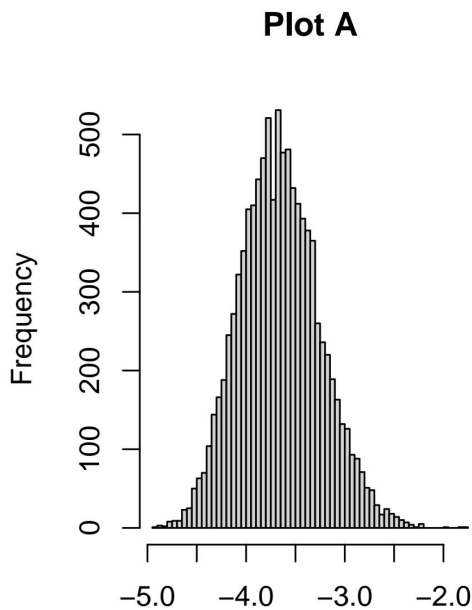
STAT 131A Spring 2023

Due March 6

## Question 1 (5 points)

Consider the question of comparing the difference between two groups. We decide to use the difference in the medians to evaluate whether there is any significant difference between the two groups. One of the histograms below is the result of a permutation test: the histogram of difference in the medians from the permutations in the permutation test comparing these two groups. The other histogram is the result of the bootstrap: a histogram of the difference in the medians from bootstrap resampling of the data to create confidence intervals. Which is which? (explain your reasoning)

```
blindData<-read.table("blindData.txt",sep="\t",header=FALSE)
par(mfrow=c(1,2))
hist(blindData[,1],breaks=100,main="Plot A",xlab="")
hist(blindData[,2],breaks=100,main="Plot B",xlab="")
```



Questions assigned to the following page: [3.1](#) and [2](#)

My answer The big difference between them is where they are concentrated mostly on. Looking at the values on the x-axis, for Plot A, we can assume that it's the bootstrap since it's concentrated on the mean, while Plot B is concentrating on zero. Therefore, Plot A is the bootstrap of a histogram and Plot B is the histogram resulted of the permutation test.

### Question 2 (10 points)

Assume you have a data.frame, `df` with a column `y` and a column `group`. `y` is the response variable, and `group` is a factor variable that gives which group the observation is in; there are two possible groups labeled `group1` and `group2` (i.e. levels of the factor).

Consider the two following functions that both do the following:

1. create a *single* bootstrap sample when given `df`
2. applies the function `FUN` to the bootstrap sample
3. returns the result of `FUN`. `FUN` can be any statistic that takes the data from two groups.

```
bootFun1<-function(df, FUN){
  group1<-df$y[df$group=="group1"]
  group2<-df$y[df$group=="group2"]
  sample1 <- sample(x=group1, size=length(group1),replace = TRUE)
  sample2 <- sample(x=group2, size=length(group2),replace = TRUE)
  return(FUN(sample1,sample2))
}
bootFun2<-function(df, FUN){
  whObs <- sample(x=1:nrow(df), size=nrow(df),replace = TRUE)
  sampleDf<-df[whObs,]
  sample1<-sampleDf$y[sampleDf$group=="group1"]
  sample2<-sampleDf$y[sampleDf$group=="group2"]
  return(FUN(sample1,sample2))
}
```

Describe the difference in the two strategies for re-sampling from the data.

My answer that for the first function, it takes in two subsets, then shuffles it then gives the replacement and it's using the sample with the same length of each of the subsets. For the second function, it takes in the replacement first, then it takes the two samples from the shuffled dataframe. The length of the second one is also equivalent to the number of rows in the dataframe. Also, the sampled groups for this function is created on based on those with group 1 and others in group 2, with the difference in the statistic that is used.

Hint: You will likely want to create a simple toy dataset to try this out and play around with the code in each one.

### Question 3

We are going to finish off the data from HW1 containing information on predicting heart disease in patients. We have provided another copy of the data with this HW, to avoid having to find the data from last time, but it is the same dataset. Read the data in again,

```
heart<-read.csv("heartDisease.csv",header=TRUE)
head(heart)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal num
```

Questions assigned to the following page: [3.1](#) and [3.2](#)



```
## 1 63 1 1 145 233 1 2 150 0 2.3 3 0 6 0
## 2 67 1 4 160 286 0 2 108 1 1.5 2 3 3 2
## 3 67 1 4 120 229 0 2 129 1 2.6 2 2 7 1
## 4 37 1 3 130 250 0 0 187 0 3.5 3 0 3 0
## 5 41 0 2 130 204 0 2 172 0 1.4 1 0 3 0
## 6 56 1 2 120 236 0 0 178 0 0.8 1 0 3 0
```

In HW2, you performed hypothesis testing for serum cholesterol in mg/dl (`chol`) – now we will consider creating confidence intervals.

- a. (5 points) Create a parametric confidence interval for the difference in the mean values serum cholesterol in mg/dl (`chol`) between those patients with chest-pain type any type of angina (i.e. `cp` either `typical angina` or `atypical angina`) and those patients with non-anginal pain. Describe carefully what the CI tells you. How does it compare to results of your t-test from HW2.

```
# code for CI
angina <- subset(heart, cp == 1 | cp == 2)
non_angina <- subset(heart, cp == 3)

diff.ci <- t.test(angina$chol, non_angina$chol, var.equal = TRUE)$conf.int
diff.ci

## [1] -19.34835 13.93838
## attr("conf.level")
## [1] 0.95
```

My answer is The result of the t-test on homework 2 evaluated to be -18.91046 13.50048. Given the values, it's somewhat similar to each other. However, since the estimated values are similar, they were to have similar values.

- b. (15 points) Repeat the above, using a bootstrap confidence interval instead and give your conclusions. How does it differ from the parametric bootstrap?

```
# code for running bootstrap CI

boot <- function(){
  boot.non_angina <- sample(na.omit(angina$chol), replace = TRUE)
  boot.angina <- sample(na.omit(non_angina$chol), replace = TRUE)
  diff <- mean(boot.non_angina) - mean(boot.angina)
  return(diff)
}

replicate <- replicate(1000, boot())

boot.ci <- quantile(replicate, probs = c(0.025, 0.975))
boot.ci

##      2.5%      97.5%
## -19.92777 12.55316
```

My answer is that comparing it with the parametric bootstrap, the result typically comes out from -20~17 for the lower bound, and for upper bound it shows from 12 to 13. Therefore, I think the results are similar to the one we've got from the parametric bootstrap. It's not only similar in the numbers, but also in the bounds, which was quite interesting.

Question assigned to the following page: [3.3](#)

- c. (10 points) What if you wanted to consider the percent of patients that have serum cholesterol above the normal range (125–200 mg/dL is a normal range of serum cholesterol). Create bootstrap confidence intervals for that difference in that percentage between the two groups.

```
# code for running bootstrap CI
angina_chol <- angina$chol
non_angina_chol <- non_angina$chol

boot_chol <- function () {
  boot.angina_high <- sample(na.omit(angina_chol), replace = TRUE)
  boot.non_angina_high <- sample(na.omit(non_angina_chol), replace = TRUE)
  difference <- mean(boot.non_angina_high > 200) * 100 - mean(boot.angina_high > 200) * 100
  return(difference)
}

replicate2 <- replicate(1000, boot_chol())

boot.cic <- quantile(replicate2, probs = c(0.025, 0.975))
boot.cic

##          2.5%          97.5%
## -16.967871    6.562082
```

My answer is for this exercise, unlike the previous confidence intervals we've got from the parametric or using bootstrap, for the one's above chol with 200, which is regarded high, out of the normal range, the range seemed to be bigger than the previous one's, that is the difference was bigger, and it's unlike the range that we've got from the previous one's as well which were similar to each others.