Woodburn Module 1 Homework

Woodburn Team

9/1/2020

##Problem 1

Our first step in beginning the assignment is to set up our environment and loading the library to include the necessary “College” data set.

library(ISLR)  
dim(College)

## [1] 777 18

help(College)

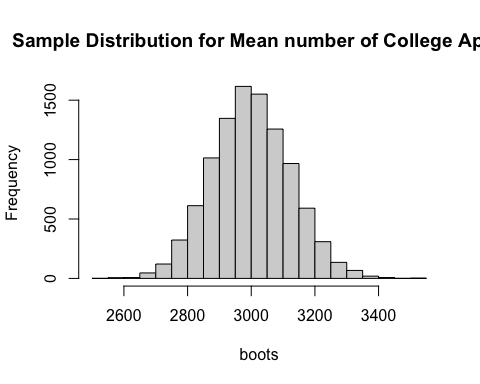
First, we need to create a name that refers to the Apps vector (column) within the College data set and find the mean.

collegeApps <- College$Apps  
mean(collegeApps)

## [1] 3001.638

Before moving forward with our analysis, it is important that we bootstrap the collegeApps vector named above. To do this, we will set a seed that can be used by later analysts to reproduce our analysis. For our first seed, that will act as a “starting point” for our random sample distribution, we use “2020.” The next chunk will include setting that seed and running code to bootstrap the sample to include 1000 observations by random sampling built off of our initial population. We will generate 10,000 observations and limit the inclusion to only 1000. We will also create a histogram to visualize the information represented in the data set. This new sample will be named “boots.”

set.seed(2020)  
boots <- NULL  
for (i in 1:10000) {  
meanCollegeApps <- mean(sample(collegeApps,1000,replace = TRUE))  
boots<- c(boots,meanCollegeApps)  
}  
hist(boots, main = "Sample Distribution for Mean number of College Apps")



Next we will calculate the Confidence Interval (CI) of this new sample of 1000. We will calculate this with a CI of 80%.

quantile(boots,c(.1,.9))

## 10% 90%   
## 2841.396 3159.211

After calculating the CI, we can say with confidence that the mean of CollegeApps will fall between 2841 and 3159 80% of the time.

Before further analysis, we will determine how the data set is organized. Is the data set a matrix or a data.frame?

mean(boots)

## [1] 2999.567

is.matrix(College)

## [1] FALSE

is.data.frame(College)

## [1] TRUE

## College is a dataframe, which reads like a list but prints like a matrix. We will be using a list (option3)

As seen above, “College” is a data.frame, which reads like a list but prints like a matrix. We will be using a list (option3), and utilizing the sapply function. We will rename our “boots” to “mylist.”

mylist<-boots  
  
sapply(mylist, mean)

(The resulting matrix represented by the console print out was omitted for length and ease of the reader.)

We will also examine a CI of 80% of the new “mylist.”

quantile(mylist,c(.1,.9))

## 10% 90%   
## 2841.396 3159.211

With the above mean of 3002, and confidence interval of 2825-3191, we are confident that the true population mean will fall between 2825 and 3191 80% of the time.

##Problem 2

For this problem we will add two additional seeds and execute 2 additional bootstraps to generate additional CI’s set for 80%. This will give us a better understanding of the true population mean. Seed 2 will be “0830”, seed 3 will be “3190”. We will discuss results after each code chunk execution.

#Second seed (0830)  
  
set.seed(0830)  
boots <- NULL  
for (i in 1:10000) {  
meanCollegeApps <- mean(sample(collegeApps,1000,replace = TRUE))  
boots<- c(boots,meanCollegeApps)  
}  
mylist<-boots  
  
# CI = 80% with new seed  
  
quantile(mylist,c(.1,.9))

## 10% 90%   
## 2846.634 3163.158

Using the second seed of “0830” to bootstrap 10,000 random samples, limiting inclusion to 1,000, we can confidently say that true population mean will fall between 2847 and 3163 80% of the time.

Now to look at a third seed.

#third seed (3190)  
  
set.seed(3190)  
boots <- NULL  
for (i in 1:10000) {  
meanCollegeApps <- mean(sample(collegeApps,1000,replace = TRUE))  
boots<-c(boots,meanCollegeApps)  
}  
mylist<-boots  
  
#CI = 80% with third seed  
  
quantile(mylist,c(.1,.9))

## 10% 90%   
## 2850.162 3161.126

After initiating a third random sample distribution using the seed “3190”, and calculating another CI at 80%, we can confidently say that the true population mean will fall between 2850 and 3161 80% of the time.

Until now we have limited the random sample to only include 1000 observations. Next we will examine CI of 80% with “5000” observations included for all three seeds (2020,0830,3190). We will adress the results at the end of analysis.

Seed 1 (2020)

set.seed(2020)  
boots <- NULL  
for (i in 1:10000) {  
meanCollegeApps <- mean(sample(collegeApps,5000,replace = TRUE))  
boots<- c(boots,meanCollegeApps)  
}  
  
mylist2020<-boots  
#CI=80%  
quantile(mylist2020,c(.1,.9))

## 10% 90%   
## 2930.169 3071.458

Seed 2 (0830)

set.seed(0830)  
boots <- NULL  
for (i in 1:10000) {  
meanCollegeApps <- mean(sample(collegeApps,5000,replace = TRUE))  
boots<- c(boots,meanCollegeApps)  
}  
  
mylist0830<-boots  
#CI=80%  
quantile(mylist0830,c(.1,.9))

## 10% 90%   
## 2931.789 3073.843

Seed 3 (3190)

set.seed(3190)  
boots <- NULL  
for (i in 1:10000) {  
meanCollegeApps <- mean(sample(collegeApps,5000,replace = TRUE))  
boots<- c(boots,meanCollegeApps)  
}  
  
mylist3190<-boots  
#CI =80%  
quantile(mylist3190,c(.1,.9))

## 10% 90%   
## 2932.284 3072.757

After executing code for all three seeds, we observed that there was a difference in time necessary for the analysis. We also observed that the bounds of our confidence intervals varied by only difference of 2 from CI of seed 1 to CI of seed 3.

#Problem 3 We want to explore if there is a significant difference between the mean number of applications received among public and private schools within the “College” data.frame. We think the mean number of applications for public schools is higher.

This can be understood as: Public > Private. Null Hypothesis: Public-Private <= 0 Alternative hypothesis is Public-Private >0.

We have chosen a p-value significance indicator of .05.

For our first step, we must establish a new sample distribution. We will also define terms to be used in later analysis.

newSamp <- sample(College$Private,777)  
numPubPriv <- data.frame(numApps=College$Apps, PubPriv=College$Private)  
numPubPriv

tapply(numPubPriv$numApps, numPubPriv$PubPriv, mean)

## No Yes   
## 5729.920 1977.929

(The above step included a console print out that has been omitted for length and ease to the reader.)

According to the above code, we have calculated the means of: Public colleges applications across all public colleges as 5730. Private colleges applications across all private colleges as 1978

In our next step, we will label the means listed above as “newPubMeans” and “newPrivMeans” representing public and private respectively.

newMeans <- tapply(numPubPriv$numApps,numPubPriv$PubPriv, mean)  
newPubMeans <- newMeans[1]  
newPrivMeans <-newMeans[2]

Next we will calculate the difference of these means.

newPubMeans - newPrivMeans

## No   
## 3751.991

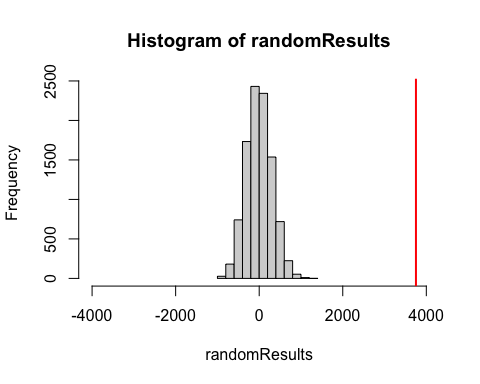
We have identified the mean difference between public and private college applications as 3751.991 or 3752.

Next we created another random sample based on our true population using the seed of 42, generating 10,000 observations and including 777 of them. This bootstraping will be named “randomResults”.

set.seed(42)  
randomResults <- NULL  
for(i in 1:10000){  
newSample <- sample (College$Private,777)  
numPubPriv <- data.frame(numApps=College$Apps, PubPriv=newSample)  
newMeans <-tapply(numPubPriv$numApps,numPubPriv$PubPriv, mean)  
randomResults <- c(randomResults, newMeans[1]-newMeans[2])  
}

After generating the randomResults sample, we will generate a histogram with a line to indicate the difference of means in comparrison to the randomResults histogram.

hist(randomResults, xlim=c(-4000,4000))  
abline(v=newPubMeans - newPrivMeans, col="red", lwd=2)

 We observe that the observations included in the “randomResults” sample are displayed in a normal distribution (similar to a bell curve) centered around zero. The red line indicates the difference in means (3752) of the two subsets of public college applications and private college applications. Based on the visual representation, and understanding that our p-value should indicate the probability of occurrences above (or to the right of) the representation of mean differences demarcation line (in red), we can make the assumption that the p-value will be zero.

However, making an assumption is not a best practice or good habit for data scientists, therefore we will test the p-value below.

greaterThanOrig <- sum(randomResults > (newPubMeans - newPrivMeans))  
sum(greaterThanOrig)/10000

## [1] 0

As expected, the results of the above formula, wherein occurrences within the randomResults sample occur greater than the different of the mean for public college applications and the mean of private college applications. As such: p=0 p=0<.05 Because the p value is less than the reasonable (——) of .05, we have no choice but to “reject” the Null Hypothesis that public colleges receive more applications than private colleges.