Mini Project #5

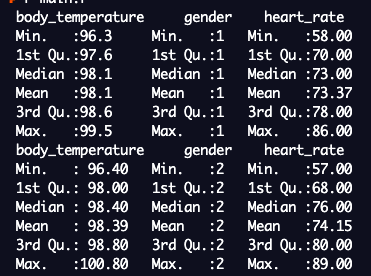
Sudarshana Jagadeeshi

Contribution of each group member: I completed the project in full

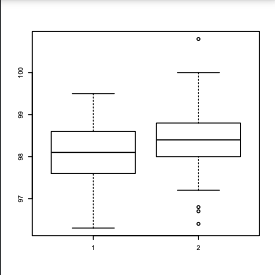
**Section 1**

**1.**

a) To determine the answer to this, we will gather summary statistics, and do a boxplot.



The top values are for the males. It seems that they have a slightly lower temperature for all members of the 5 member summary. The difference is around 0.1-0.3, centered 0.2. We then plot the boxplot:



The males are on the left. We see right away thathis mean is slightly lower than the females. We can see that there is a greater IQR for the male, but the number of outliers for the female are greater. This could mean that the female distribution has a greater SD.

We can then run a test in R to determine whether the two means are really different from each other. If we find sufficient evidence, we can reject the null hypothesis that the two means are the same.

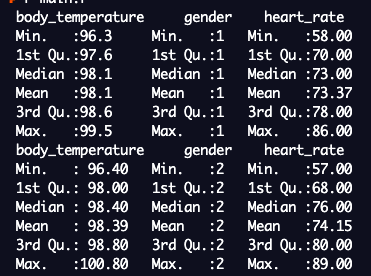
The confidence interval is:

(-0.53727651, -0.04118503)

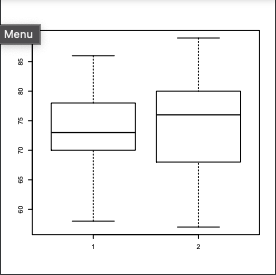
The p-value is found by looking up z= -0.2892308/0.1264=-2.288, which is 0.011. This is significant for a 95% confidence, and we reject the null hypothesis. Yes, male temperatures are lower than females on average.

b.

To determine the answer to this, we once again gather summary statistics, and do a boxplot.



The top values are for the males. While the males have a smaller mean, their other quartiles are not less than the females. The higher median than mean for the females indicate there is some left skew in the distribution dragging the mean down. We then plot the boxplot:



The males are on the left. Once again his mean is slightly lower than the females. We also confirm what we noted above, that the mean bar is not centered within the box, indicating skew. There are no outliers in either set.

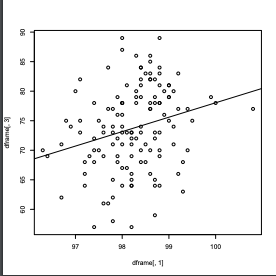
We can then run a test in R to determine whether the two means are really different from each other. If we find sufficient evidence, we can reject the null hypothesis that the two means are the same.

The confidence interval is:

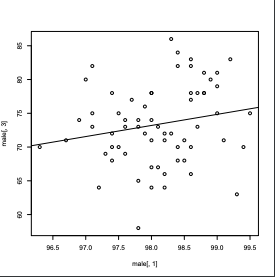
(-3.218278 , 1.649047)

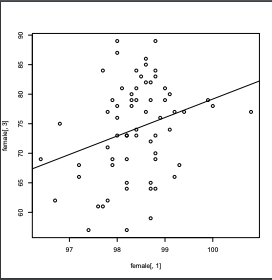
Which does includes 0. The p-value is found by looking up z= -0.7846154/1.241665=-0.63190586833, which is .528041. This is definitely not significant for a 95% confidence, and we fail to reject the null hypothesis. We do not have enough evidence that male heart rates are lower than females on average.

c.



The scatterplot, with temperature on the x and heartrate on the y, seems to have no real correlation. There is a very weak correlation, and we compute it to be 0.2536564. We can make two additional plots showcasing the data just for each gender, to see if this changes.





The first plot is for the male. We can see a bit more slope in the female plot. This is confirmed by comparing the correlation coefficients of 0.28(female) and 0.19(male). In the female case, temperature predicts heartrate a little bit better, but the association is still relatively weak. We would want an coefficient of 0.7 or better.

**2.**

a), b)

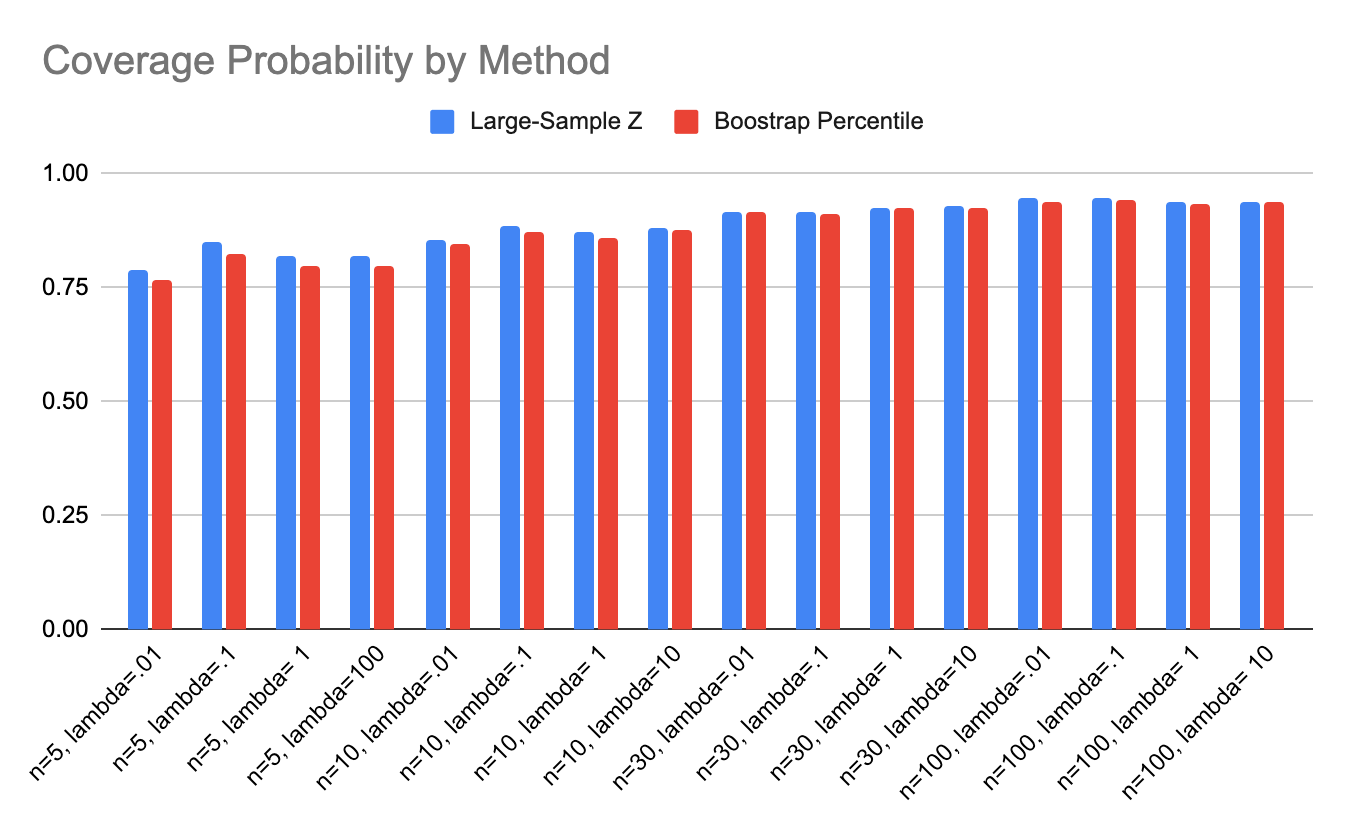
Large-sample z

|  | lambda=0.01 | lambda=0.1 | lambda=1 | lambda=10 |
| --- | --- | --- | --- | --- |
| n=5 | 0.79 | 0.852 | 0.818 | 0.818 |
| n=10 | 0.856 | 0.886 | 0.872 | 0.882 |
| n=30 | 0.914 | 0.916 | 0.926 | 0.93 |
| n=100 | 0.946 | 0.948 | 0.94 | 0.938 |

Bootstrap Percentile

|  | lambda=0.01 | lambda=0.1 | lambda=1 | lambda=10 |
| --- | --- | --- | --- | --- |
| n=5 | 0.768 | 0.822 | 0.798 | 0.796 |
| n=10 | 0.848 | 0.874 | 0.858 | 0.878 |
| n=30 | 0.914 | 0.912 | 0.924 | 0.924 |
| n=100 | 0.94 | 0.942 | 0.934 | 0.936 |

In graph form it is:



c)

> The clear trend is observed that the greater the value of n, the more accurate the interval becomes. For a 95% confident interval, it seems that upwards of n=100 is required. However, if we are satisfied with a 90% interval, we can choose n=30 as sufficient. This is in-line with the established statistics recommendation. > A bootstrap interval is about the same- n=30 is enough for 90% accuracy, but 100 is needed for typical 95% accuracy.

>It seems in all cases, large-sample z is a mildly better estimator. For n<30, it seems that large-sample z actually outperforms bootstrap, but past n=30, the difference goes to the third decimal point, and the advantage of one method over the other becomes less clear. It is possible that if n is increased even further, bootstrap may overtake large-sample z in coverage probability.

> Changing lambda has no effect on the coverage probability when fixing n. Going row by row, we cannot detect a trend of even mild increase or decrease, meaning the lambda term does not matter at all.

d)

No, the fixed values of lambda do not make a significant impact on our conclusions in part c. Observe that I have highlighted the maximum value in each row of the table to illustrate the fact that there is no trend to be observed.

This makes sense, because we are only altering the function- and larger function values will indeed result in a larger mean, and larger interval bounds, but it will not affect the coverage probability, because the CI will always contain 1.96\*2 standard errors.

**Section 2**

**Question 1**

**dframe <- read.csv('bodytemp-heartrate.csv', header = TRUE) #first row is NOT data**

**#take subsets that will help us later**

**male <- subset(dframe, gender == 1)**

**female<- subset(dframe, gender == 2) #data, condition**

**maletemps <- subset(dframe, gender == 1)[,1]**

**femaletemps <- subset(dframe, gender == 2)[,1]**

**maleheartrate <- subset(dframe, gender == 1)[,3]**

**femaleheartrate <- subset(dframe, gender == 2)[,3]**

**#print summary statistics to compare male and female**

**print(summary(male))**

**print(summary(female))**

**#PART A: does male temperature differ from female?**

**boxplot(maletemps, femaletemps)**

**#SE FORMULA, 1.96 for 95% interval**

**SE <- (sd(maletemps)\*\*2) /length(maletemps) + (sd(femaletemps)\*\*2) / length(femaletemps)**

**SE <- SE\*\*0.5**

**CI <- c(mean(maletemps) - mean(femaletemps) - 1.96\* SE,mean(maletemps) - mean(femaletemps) + 1.96\* SE)**

**#look up in the z-table to get a p-value**

**print((mean(maletemps) - mean(femaletemps))/ SE)**

**#PART B: does male heart rate differ from female?**

**boxplot(maleheartrate, femaleheartrate)**

**#Once again the same SE formula**

**SE2 <- (sd(maleheartrate)\*\*2) /length(maleheartrate) + (sd(femaleheartrate)\*\*2) / length(femaleheartrate)**

**SE2 <- SE2\*\*0.5**

**CI2 <- c(mean(maleheartrate) - mean(femaleheartrate) - 1.96\* SE2, mean(maleheartrate) - mean(femaleheartrate) + 1.96\* SE2)**

**#look up in the z-table to get a p-value**

**print((mean(femaletemps) - mean(femaletemps))/ SE)**

**#PART C: does temperature affect heartrate?**

**plot(male[,1], male[,3])**

**abline(lm(male[,3] ~ male[,1])) #line of best fit**

**print(cor(male[,1], male[,3]))**

**plot(female[,1], female[,3])**

**abline(lm(female[,3] ~ female[,1])) #line of best fit**

**print(cor(female[,1], female[,3]))**

**Question 2**

**library(boot)**

**#parameters to test**

**n=100**

**lambda=1**

**hits1 <- vector() #initialize empty vector**

**hits2 <- vector()**

**getmean <- function(values, indices) {**

**return(mean(values[indices])) #not just mean(values), need only the mean of the bootstrapped indices**

**}**

**dothis <- function() {**

**values <- rexp(n, lambda)**

**#1.96 for 95% interval**

**CI <- c(mean(values) - 1.96 \* sd(values)/(sqrt(n)), mean(values) + 1.96 \* sd(values)/(sqrt(n))) #standard CI formula, SE is sd/sqrt(n)**

**if(CI[1]< (1/lambda) && CI[2] > (1/lambda)){**

**hits1 <<- c(hits1, CI)#append**

**}**

**bootstrapdata <- boot(values, getmean, R=1000)**

**lower <- sort(bootstrapdata$t)[.025 \* 1000] #2.5th percentile**

**higher <- sort(bootstrapdata$t)[.975 \* 1000]**

**CI2 <- c(lower, higher) #combine lower and upper bounds**

**if (CI2[1]< (1/lambda) && CI2[2] > (1/lambda)){**

**hits2 <<- c(hits2, CI2)**

**}**

**}**

**numtimes <- 5000 #this was given in the handout**

**replicate(numtimes, dothis())**

**cov1 <- length(hits1)/(numtimes\*2) #because each CI we push to the array has two elements**

**cov2 <- length(hits2)/(numtimes\*2)**

**print(cov1)**

**print(cov2) #the final coverage values**