CS 6313 STATISTICAL METHODS FOR DATA SCIENCE

Mini Project 5

NOVAMBER 19, 2021

AFRAA ALSHAMMARI
PRAVALIKA DOMAL

Contribution of Each Member:

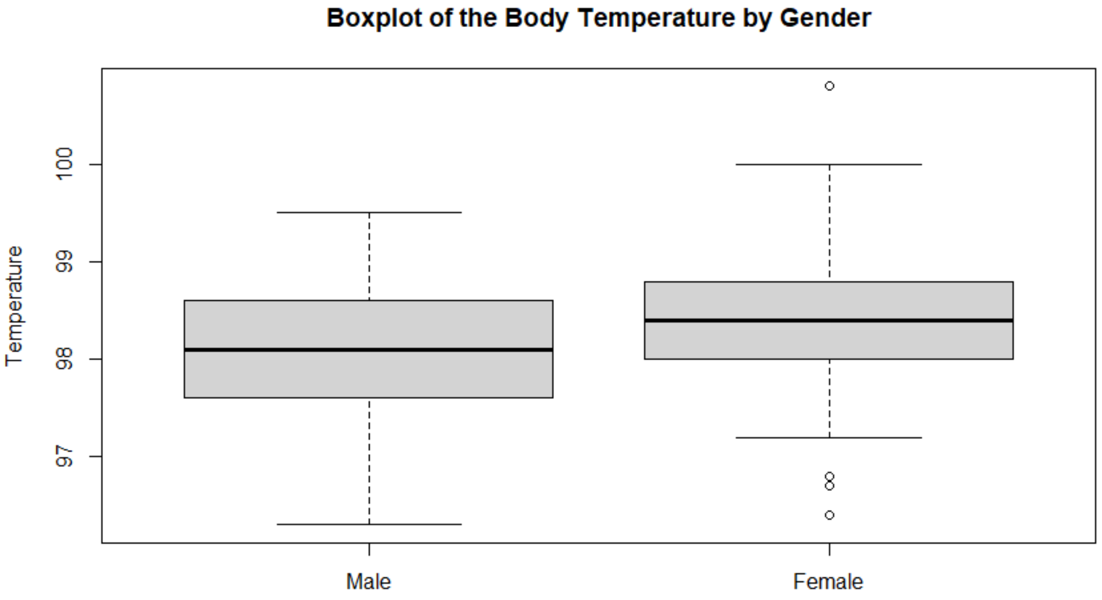
Both worked together to finish and submit this mini project. Starting with the R and conducting the needed statics to solve Q1 and Q2. Afraa did the documentation of the experiments and the analytical solution needed, and Pravalika worked on analytical solution scripts for finding accuracy of scripts.

Section 1:

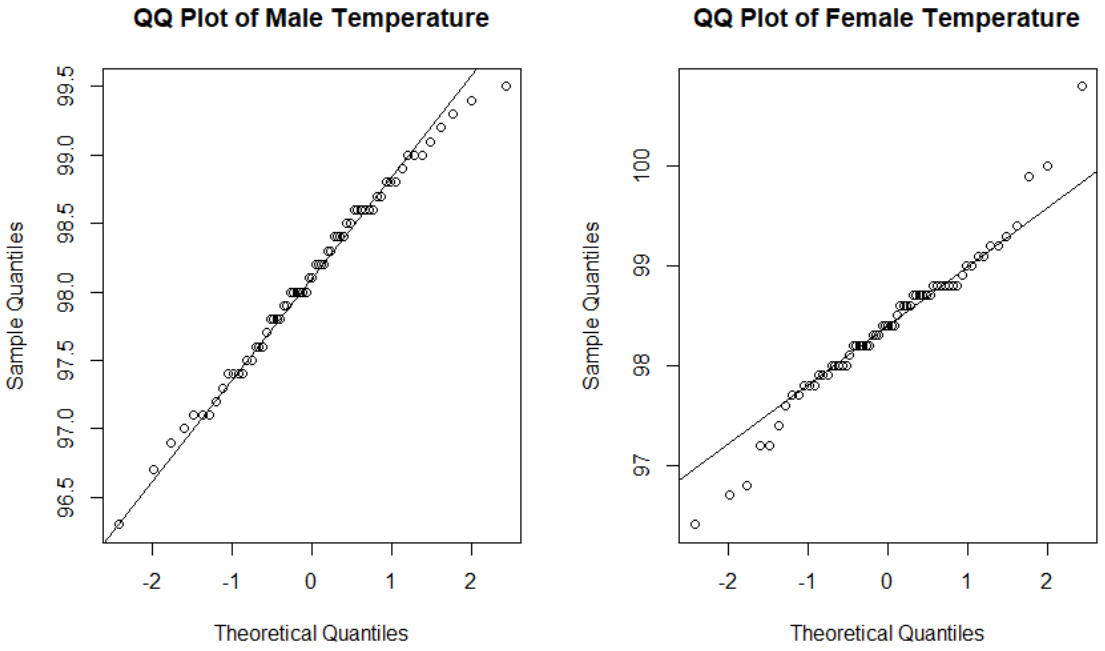
Q1. Consider the data stored in bodytemp-heartrate.csv on eLearning, containing measurements of body temperature and heart rate for 65 male (gender = 1) and 65 female (gender = 2) subjects.

(a) Do males and females differ in mean body temperature? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

First step here is to read the file bodytemp-heartrate using read function, and then separate the data according to gender to perform the data analysis by using subset function. For the analysis a boxplot and QQ plot is generated as represented below in Graph 1 and 2 respectively. From Graph 1 it is obvious that the female temperature has higher values than the male as the median in the female is higher than the median in the male sample. Also, the mean value for the female is observed higher than the male. Moreover, In the female sample the outliers exists indicating variability in the data in contrast to the male data which implies the inequality of variance between these samples.



Graph 1. Boxplot of the Body Temperature by Gender



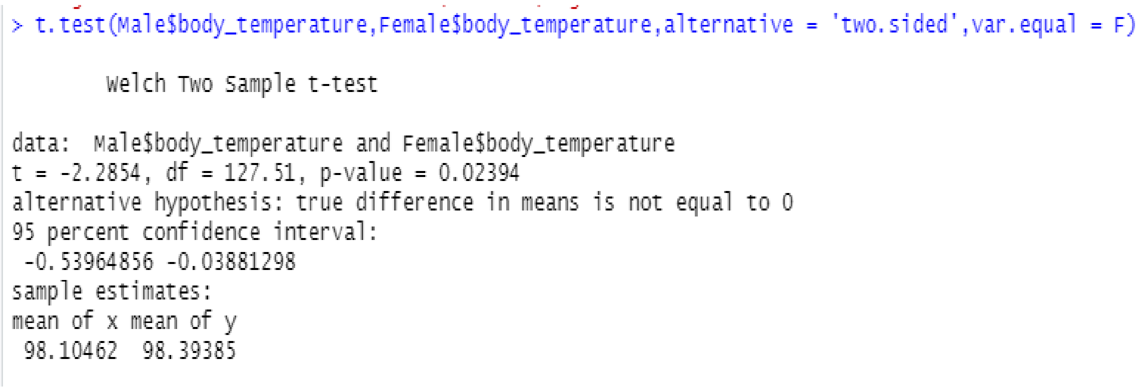
Graph 2. QQ Plot of the Body Temperature by Gender

In Graph 2 the distribution of the two sample is approximately normal. To explore the data a hypothesis testing approach with Male mean noted as Mm and Female mean as Fm will be followed.

H0: difference mean = 0 => Mm-Mf=0.

H1: difference mean ≠ 0 => Mm-Mf≠0.

From the above the samples has an approximate normal distribution and unequal variance, therefore, a T test will be used with the Satterthwaite approximation to calculate the confidant interval. The T test will be two-sided since the exploration analysis is on the equality feature of means of the 2 samples and not to explore if one is higher or lower than the other. Noting, the two samples are independent. The result of the test is in the sample code and output below:

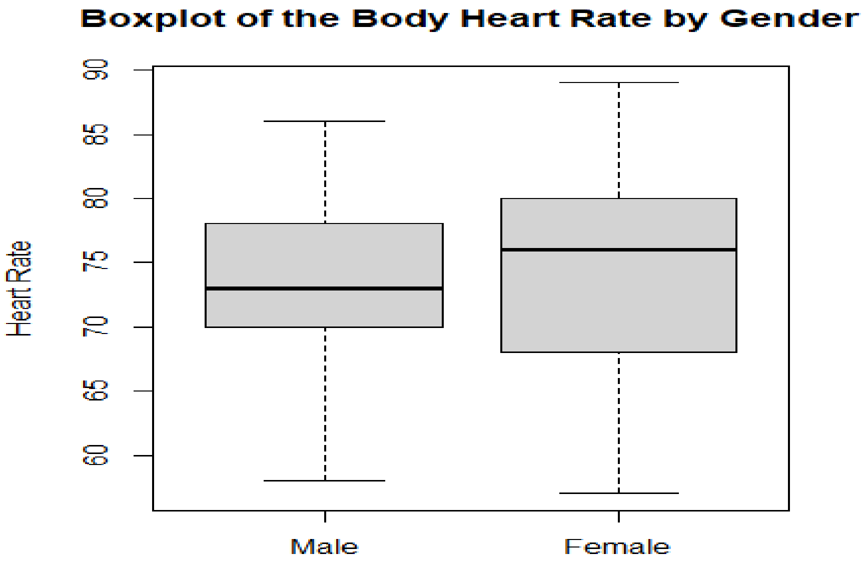


```
> t.test(Male$body_temperature,Female$body_temperature,alternative = 'two.sided',var.equal = F)

        Welch Two Sample t-test

data:  Male$body_temperature and Female$body_temperature
t = -2.2854, df = 127.51, p-value = 0.02394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean of x mean of y 
 98.10462  98.39385 
```

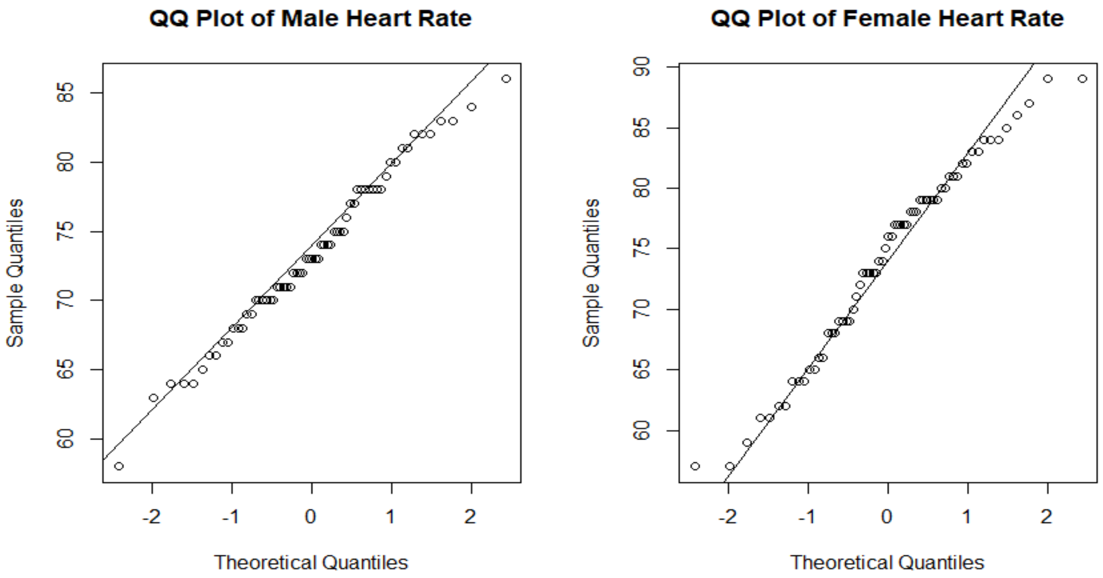
The p-value is 0.02394 which is less than 0.05 and it is in not in the interval [-0.53964856, -0.03881298] therefore, the null hypothesis is rejected and accept the existence of difference between the mean of temperature between male and female bodies. Noting, since the interval is small the existence of difference is very small the statement that the female bodies temperature is slightly higher than the man`s holds.

(b) Do males and females differ in mean heart rate? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

In this part of the question the same methodology that has been used in part a will be followed to explore the difference of the 2 samples heart rate mean. Graph 3 illustrates the boxplot of the 2 sample hear rate which the difference of the mean of the 2 samples hear rate can be clearly visible showing the female sample has higher values in Q3 and median, but lower in Q1.



Graph 3. Boxplot of the Body Hear Rate by Gender



Graph 4. QQ Plot of the Body Heart Rate by Gender

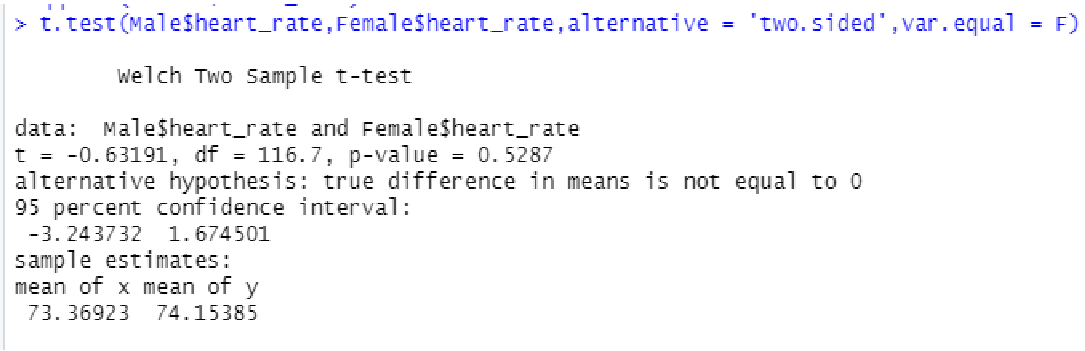
From Graph 4 the samples distributions are considered to be approximately normal.

To explore the data a hypothesis testing approach with Male hear rate mean noted as Mm and Female mean as Fm will be followed.

H0: Difference mean = 0 => Mm-Mf=0.

H1: Difference mean ≠ 0 => Mm-Mf≠0.

From the above the samples has an approximate normal distribution and unequal variance, therefore, a T test will be used with the Satterthwaite approximation to calculate the confidant interval. The T test will be two-sided since the exploration analysis is on the equality feature of means of the 2 samples and not to explore if one is higher or lower than the other. Noting, the two samples are independent. The result of the test is in the sample code and output below:



```
> t.test(Male$heart_rate,Female$heart_rate,alternative = 'two.sided',var.equal = F)

	Welch Two Sample t-test

data:  Male$heart_rate and Female$heart_rate
t = -0.63191, df = 116.7, p-value = 0.5287
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.243732  1.674501
sample estimates:
mean of x mean of y 
 73.36923  74.15385 
```

The p-value is 0.5287 which is higher than 0.05 and it is in the interval [-3.243732, 1.674501] therefore, the null hypothesis is accepted and reject the H1 that stated the existence of difference between the mean of heart rate between male and female bodies.

(c) Is there a linear relationship between body temperature and heart rate? Does this relationship depend on gender? Answer these questions by performing an appropriate analysis of the data, including an exploratory analysis.

In this part the existence of a linear relationship between the body temperature and heart rate is questioned. For a relationship testing the correlation between the two factors must be examine. First the scatter plot with the regression line will be created to examine the association between the 2 attributes in both samples. Graph 5 below shows the scatter plot and it indicate a positive relationship between the temperature and the heart rate in both samples with the slop >0 in both graphs and the association correlation can be viewed as week. For this the correlation is calculated and the findings are:

The correlation between body temperature and heart rate in male sample is 0.1955894.

The correlation between body temperature and heart rate in female sample is 0.2869312.

These values of correlation are not high, and this indicates a week relationship between body temperature and heart rate, and this consists with the findings from Graph 5. Furthermore, in the female sample the correlation value is higher than the male sample which indicates that the correlation between body temperature and heart rate in female is slightly stronger than of the male sample.

Section 2:

Q2: The goal of this exercise to see how large n should be for the large-sample and the (parametric) bootstrap percentile method confidence intervals for the mean of an exponential population to be accurate. To be specific, let X1, . . . , Xn represent a random sample from an exponential (λ) distribution. Note that this distribution is skewed and its mean is $\mu = 1/\lambda$. We can construct two confidence intervals for μ — one the large-sample z-interval (interval 1) and the other a (parametric) bootstrap percentile method interval (interval 2). We would like to investigate their accuracy, i.e., how close their estimated coverage probabilities are to the assumed nominal level of confidence, for various combinations of (n, λ). This investigation will focus on $1 - \alpha = 0.95$, $\lambda = 0.01, 0.1, 1, 10$ and $n = 5, 10, 30, 100$. Thus, we have a total of $4 * 4 = 16$ combinations of (n, λ) to investigate.

(a) For a given setting, compute Monte Carlo estimates of coverage probabilities of the two intervals by simulating appropriate data, using them to construct the two confidence intervals, and repeating the process 5000 times.

Solution:

To calculate the Monte Carlo estimates of coverage probabilities and confidence intervals, we created the following functions:

Checkz - simulates a sample, generates an interval, and reports if the true mean exists inside the confidence interval using n and lambda values as input parameters.

Zpro – uses the input parameters n and lambda to call the checkz function 5000 times and determine the coverage probability.

Mean.star - a function that takes a sample from a distribution and returns the mean.

Checkb - it calls the mean.star function 1000 times, constructs the confidence interval, and returns whether the true mean is present in the interval using the n and lambda input parameters.

Bpro- takes the input parameters n and lambda, creates a parametric initial bootstrap sample, and calls checkb 5000 times to compute the coverage probability.

Using these functions, the coverage probabilities for the (n,lambda) combination of (5,0.01) are as follows:

Z-interval: 0.8102

Bootstrap interval: 0.8960

(b) Repeat (a) for the remaining combinations of (n, λ). Present an appropriate summary of the results.

Solution:

We get the following results by repeating the technique for the remaining combinations:

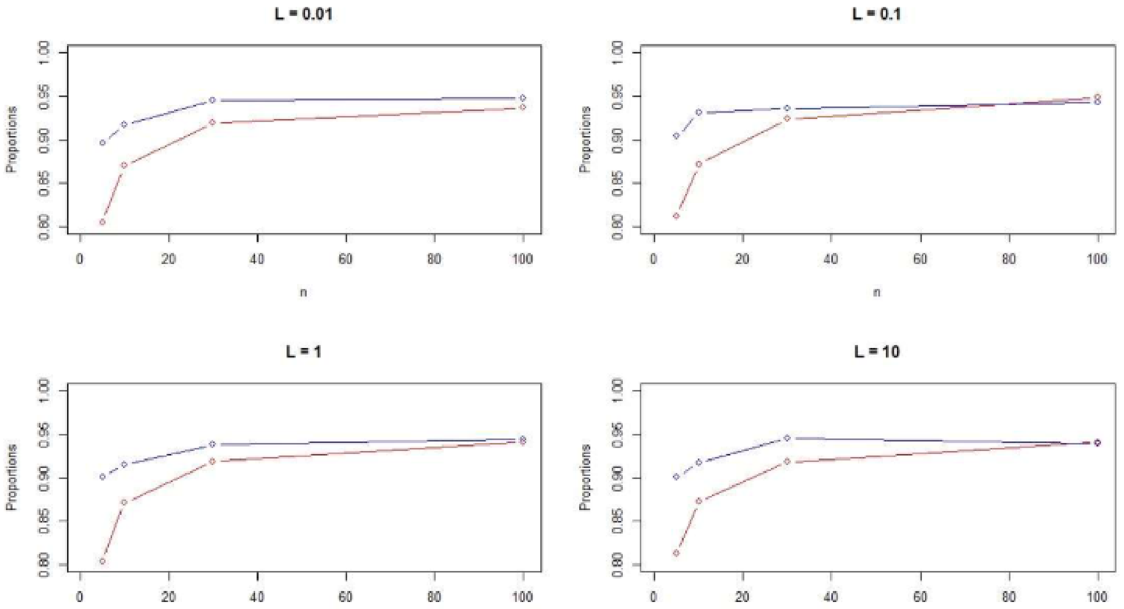
Z-proportions	L=0.01	L=0.1	L=1	L=10
N=5	0.8102	0.8124	0.8042	0.8132
N=10	0.8702	0.8716	0.8716	0.8728
N=30	0.9192	0.9236	0.9184	0.9178
N=100	0.9366	0.9482	0.9404	0.9408

B-proportions	L=0.01	L=0.1	L=1	L=10
N=5	0.8960	0.9038	0.9004	0.9002
N=10	0.9168	0.9304	0.9148	0.9172
N=30	0.9452	0.9356	0.9374	0.9448
N=100	0.9478	0.9430	0.9434	0.9388

Graphically representing the data, we get

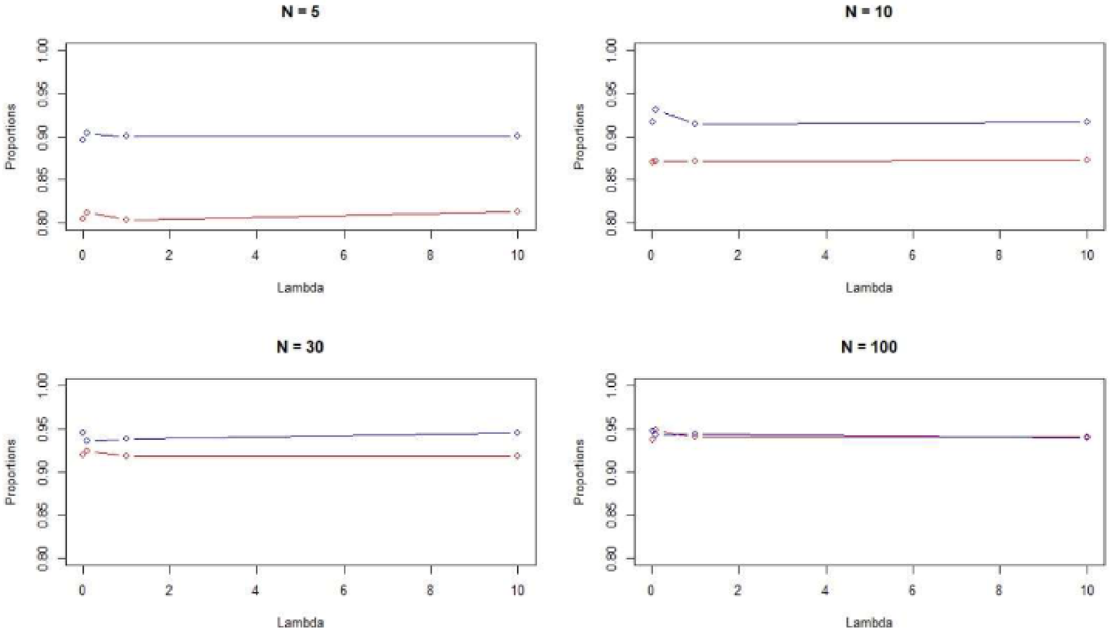
Graph 1: The colors red and blue show z-proportions and bootstrap proportions, respectively.

The values are plotted over n while lambda remains constant.



Graph 2: The colors red and blue show z-proportions and bootstrap proportions, respectively.

The values are plotted over lambda while n remains constant.



(c) Interpret all the results. Be sure to answer the following questions: In case of the large-sample interval, how large n is needed for the interval to be accurate? Likewise, in case of the bootstrap interval, how large n is needed for the interval to be accurate? Do these answers depend on λ? Can we say that one method is more accurate than the other? Which interval would you recommend? Provide justification for all your conclusions.

Solution:

The graphs in Graph 1 don't vary dramatically when lambda is adjusted; we may conclude that the coverage probabilities are independent of lambda. Also, we can observe that the coverage probabilities obtained using the bootstrap approach is higher than those obtained using the z interval method. We can deduce from Graph 2 that the coverage probabilities are influenced by n. When n is large (n=100), the coverage probabilities obtained using the large sample z-interval are as accurate as the coverage probability obtained using the bootstrap approach. From n=30 onwards, the bootstrap technique coverage probability are on the higher side (approx.). Considering all of the graphs, we can conclude that the coverage probabilities obtained using the bootstrap approach are greater for every combination of (n, lambda) than those obtained using the large-sample z-interval method, implying that the bootstrap method is more accurate even for low n values. As a result, the bootstrap method is advised.

(d) Do your conclusions in (c) depend on the specific values of λ that were fixed in advance? Explain.

Solution:

The output from the code in section 2 helps us to infer that the:

For n 5 lambda 0.1, the coverage probability for bootstrap is 0.61.

For n 5 lambda 0.1, the coverage probability for large sample z is 0.8102.

Also,

For n 10 lambda 0.1, the coverage probability for bootstrap is 0.61.

For n 10 lambda 0.1, the coverage probability for large sample z is 0.695.

For n 30 lambda 0.1, the coverage probability for bootstrap is 0.8758.

For n 30 lambda 0.1, the coverage probability for large sample z is 0.7134.

For n 100 lambda 0.1, the coverage probability for bootstrap is 0.7218.

For n 100 lambda 0.1, the coverage probability for large sample z is 0.9388.

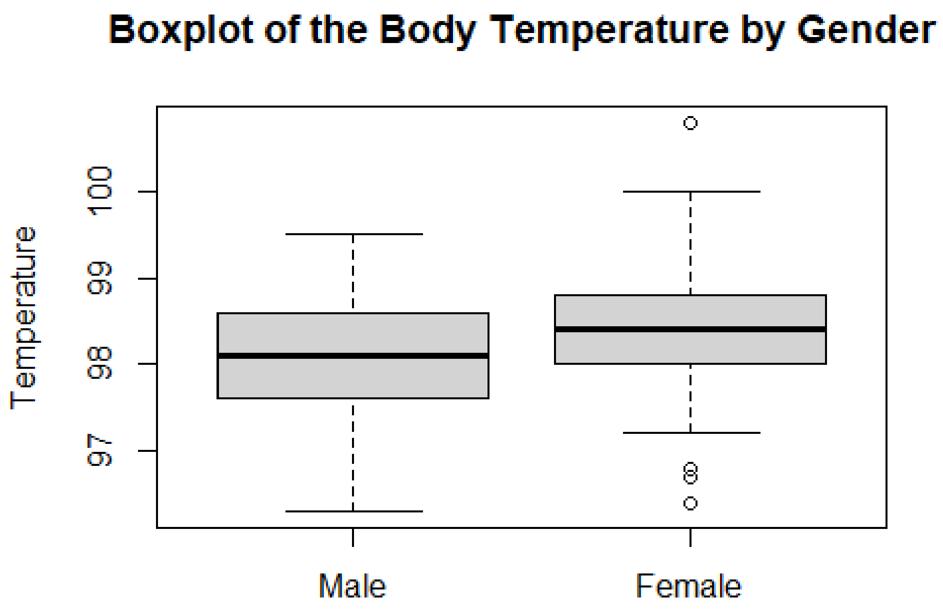
Therefore:

For specific values of lambda, the results in (c) hold true. Lambda = 0.1 in this scenario.

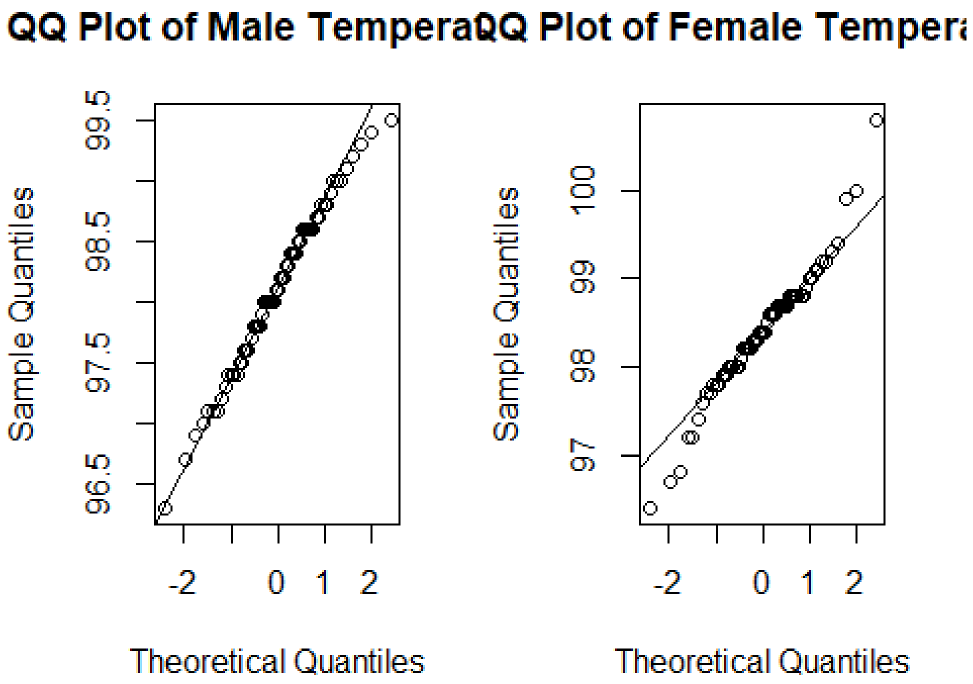
R code

Code1:

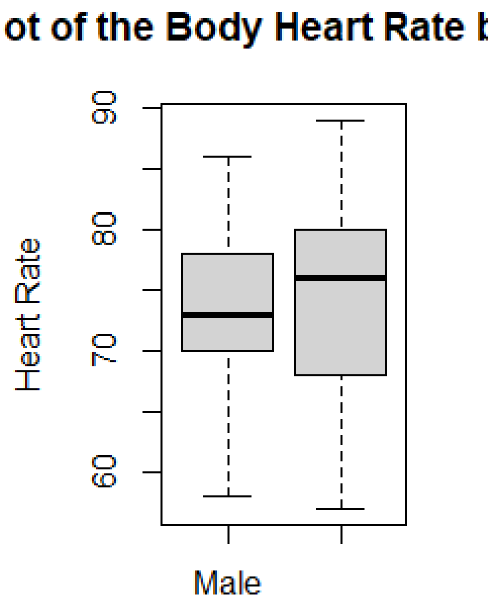
```
#Q1
#a.
#1.Read file bodytemp-heartrate using read function
BodyTempHeartrate= read.csv("bodytemp-heartrate.csv")
#2.Seperate the data according to gender to perform the analysis by using subset function
Male=subset(BodyTempHeartrate,BodyTempHeartrate$gender==1)
Female=subset(BodyTempHeartrate,BodyTempHeartrate$gender==2)
#3.Draw boxplot for the data
boxplot(Male$body_temperature, Female$body_temperature,main="Boxplot of the Body Temperature by Gender", names=c('Male','Female'),ylab="Temperature")
```



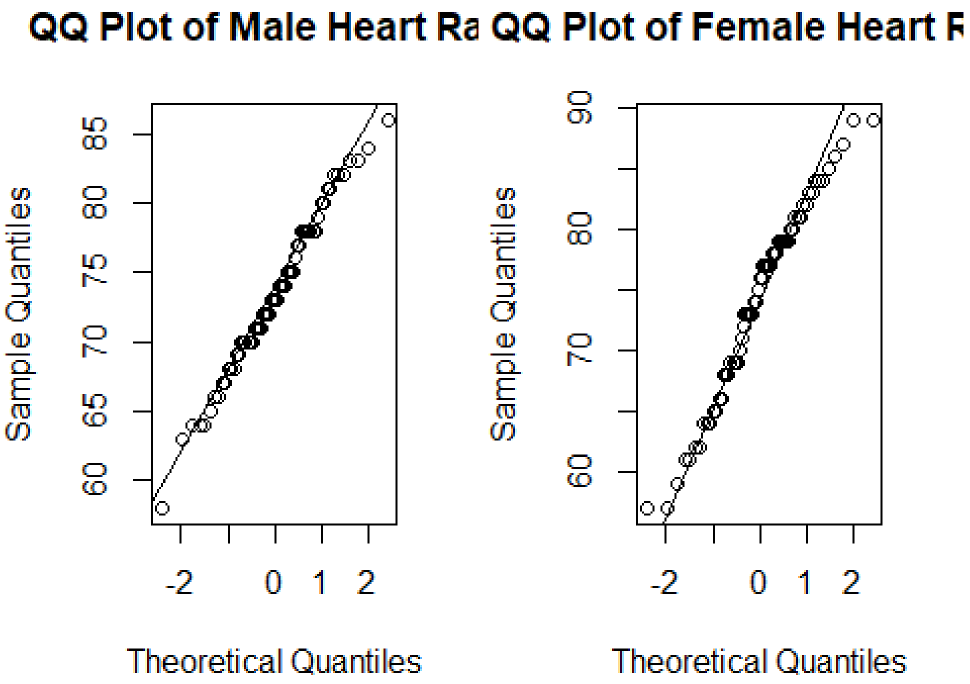
```
#4. Draw the QQ plot for the data
par(mfrow=c(1,2))
qqnorm(Male$body_temperature,main = "QQ Plot of Male Temperature")
qqline(Male$body_temperature)
qqnorm(Female$body_temperature,main = "QQ Plot of Female Temperature")
qqline(Female$body_temperature)
```



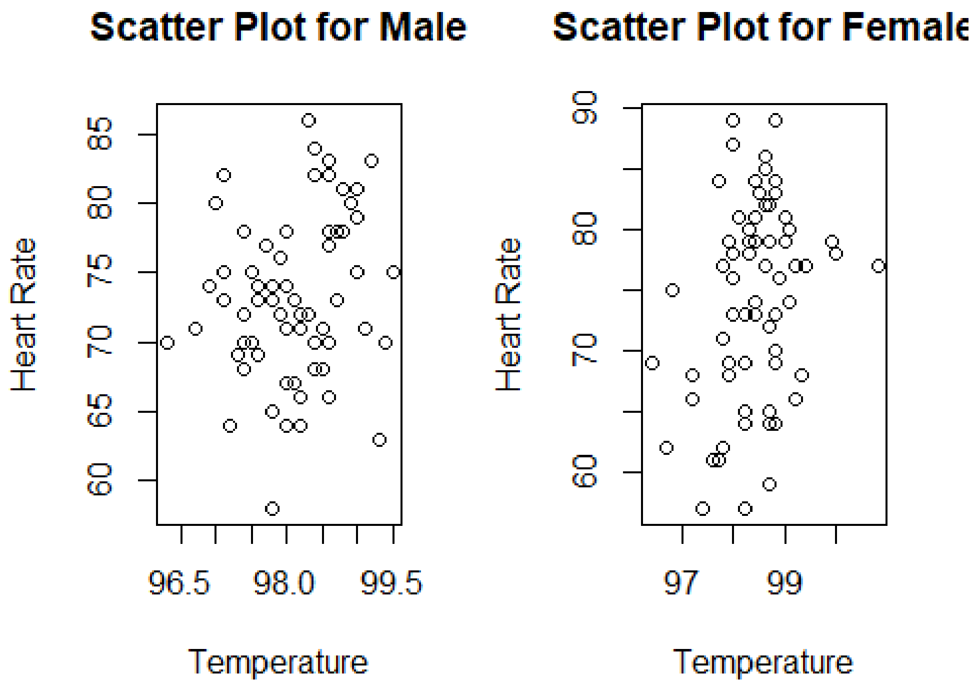
```
#5.T test twoside with unequal variance
t.test(Male$body_temperature,Female$body_temperature,alternative = 'two
.sided',var.equal = F)

##
##  Welch Two Sample t-test
##
## data:  Male$body_temperature and Female$body_temperature
## t = -2.2854, df = 127.51, p-value = 0.02394
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.53964856 -0.03881298
## sample estimates:
## mean of x mean of y
##  98.10462  98.39385

#b.
#1.Draw boxplot for the data
boxplot(Male$heart_rate, Female$heart_rate,main="Boxplot of the Body He
art Rate by Gender", names=c('Male','Female'),ylab="Heart Rate")
#2. Draw the QQ plot for the data
par(mfrow=c(1,2))
```



```
qqnorm(Male$heart_rate,main = "QQ Plot of Male Heart Rate")
qqline(Male$heart_rate)
qqnorm(Female$heart_rate,main = "QQ Plot of Female Heart Rate")
qqline(Female$heart_rate)
```



```
#3.T test two sided with unequal variance
t.test(Male$heart_rate,Female$heart_rate,alternative = 'two.sided',var.
equal = F)

##
##  Welch Two Sample t-test
##
## data:  Male$heart_rate and Female$heart_rate
## t = -0.63191, df = 116.7, p-value = 0.5287
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.243732  1.674501
## sample estimates:
## mean of x mean of y
##  73.36923  74.15385

#c
#1. Draw scatter plot with the regression line to examine the associati
on between the 2 attributes in both samples
par(mfrow=c(1,2))
plot(Male$body_temperature,Male$heart_rate,xlab="Temperature",yla ="Hea
rt Rate",main = "Scatter Plot for Male")
abline(lm(Male$body_temperature ~ Male$heart_rate))
plot(Female$body_temperature,Female$heart_rate, xlab="Temperature",ylab
="Heart Rate" ,main = "Scatter Plot for Female")
abline(lm(Female$body_temperature ~ Female$heart_rate))
```


Scatter Plot for Male

Scatter Plot for Female

```
#2. Using Cor function
cor(Male$body_temperature,Male$heart_rate)

## [1] 0.1955894

cor(Female$body_temperature,Female$heart_rate)

## [1] 0.2869312
```

Code 2:

```
#Reading the dataset

> bodytemp.heartrate <- read.csv("C:/Users/priya/Downloads/bodytemp-heartrate.csv")

>  View(bodytemp.heartrate)


# creating function checkz

> checkz <- function(n,lambda){u <- rexp(n,lambda)

+ lb<-mean(u)-qnorm(0.975)*sd(u)/sqrt(n)

+ ub<-mean(u)+qnorm(0.975)*sd(u)/sqrt(n)

+ tm=1/lambda
+ if(ub>tm & lb<tm){
+ return(1)
+ }
+ else {
+ return(0)
+ }
+ }

#creating function zpo
> zpo<-function(n,lambda){
+ values<-replicate(5000,checkz(n,lambda))
+ ones<-values[which(values == 1)]
+ return(length(ones)/5000)
+ }

#getting the value of n=5 and lambda=0.01 zpo
> zpo(5,0.01)
[1] 0.8102

#creating function mean.star
  > mean.star<-function(n,lambda){
+ u.star<-rexp(n,lambda)
+ return(mean(u.star))
+ }

# creating function checkb
> checkb<-function(n,lambda){
+ u<-rexp(n,lambda)
+ tm<-1/lambda
+ lambda1=1/mean(u)
+ V<-rerplicate(1000,mean.star(n,lambda1))
+ bound<-sort(V)[c(25,975)]
+ if(boundp[2]>tm & bound[1]<tm){
+ return(1)
+ }
+ else{
+ return(0)
+ }
+ }

#creating function bpro
> bpro<-function(n,lambda){
+ values<-replicate(5000,checkb(n,lambda ))
+ ones<-values[which(values == 1)]
+ return(length(ones)/5000)
+ }

#getting the value of n=5 and lambda=0.01 bpro
> bpro(5,0.01)
[1] 0.8960

#generating the proportion values for bootstrap and z interval for combinations of n and lambda
> zcimatrix<-matrix(c(zpo(5,0.01),zpo(10,0.01),zpo(30,0.01),zpo(100,0.01),zpo(5,0.1),zpo(10,0.1),zpo(30,0.1),zpo(100,0.1),zpo(5,1),zpo(10,1),zpo(30,1),zpo(100,1),zpo(5,10),zpo(10,10),zpo(30,10),zpo(100,10)),nrow = 4,ncol = 4)
> bcimatrix<-matrix(c(bpro(5,0.01),bpro(10,0.01),bpro(30,0.01),bpro(100,0.01),bpro(5,0.1),bpro(10,0.1),bpr
o(30,0.1),bpro(100,0.1),bpro(5,1),bpro(10,1),bpro(30,1),bpro(100,1),bpro(5,10),bpro(10,10),bpro(30,10),bpro(100,10)),nrow = 4,ncol = 4)


#drawing line graphs for all these values

> par(mfrow=c(2,2))

>plot(c(5,10,30,100),zcimatrix[,1],main="L=0.01",xlab='n',ylab='proportions',col='red',type='b',xlim = c(1,100),ylim = c(0,1))

> lines(c(5,10,30,100),bcimatrix[,1],col='blue',type='b')

>plot(c(5,10,30,100),zcimatrix[,2],main="L=0.1",xlab='n',ylab='proportions',col='red',type='b',xlim = c(1,100),ylim = c(0,1))

> lines(c(5,10,30,100),bcimatrix[,2],col='blue',type='b')

>plot(c(5,10,30,100),zcimatrix[,3],main="L=1",xlab='n',ylab='proportions',col='red',type='b',xlim = c(1,100),ylim = c(0,1))

> lines(c(5,10,30,100),bcimatrix[,3],col='blue',type='b')

>plot(c(5,10,30,100),zcimatrix[,4],main="L=10",xlab='n',ylab='proportions',col='red',type='b',xlim = c(1,100),ylim = c(0,1))

> lines(c(5,10,30,100),bcimatrix[,4],col='blue',type='b')

>plot(c(0.01,0.1,1,10),zcimatrix[,1],main="N=5",xlab='Lambda',ylab='proportions',col='red',type='b',xlim = c(0.01,10),ylim = c(0,1))

> lines(c(0.01,0.1,1,10),bcimatrix[,1],col='blue',type='b')

>plot(c(0.01,0.1,1,10),zcimatrix[,2],main="N=10",xlab='lambda',ylab='proportions',col='red',type='b',xlim = c(0.01,10),ylim = c(0,1))

> lines(c(0.01,0.1,1,10),bcimatrix[,2],col='blue',type='b')

>plot(c(0.01,0.1,1,10),zcimatrix[,3],main="N=30",xlab='lambda',ylab='proportions',col='red',type='b',xlim = c(0.01,10),ylim = c(0,1))

> lines(c(0.01,0.1,1,10),bcimatrix[,3],col='blue',type='b')

>plot(c(0.01,0.1,1,10),zcimatrix[,4],main="N=100",xlab='lambda',ylab='proportions',col='red',type='b',xlim = c(0.01,10),ylim = c(0,1))

> lines(c(0.01,0.1,1,10),bcimatrix[,4],col='blue',type='b')
```