Project 5

Name of group members: 1. Ajay Danda (AXD180068) 2. Satyam Bhikadiya (SXB180124)

Contribution of each group member:

Ajay Danda:

- Equal Contribution in solving Q1
- Equal Contribution in solving Q2
- Equal Contribution in Documenting the Report

Satyam Bhikadiya:

- Equal Contribution in solving Q1
- Equal Contribution in solving Q2
- Equal Contribution in Documenting the Report

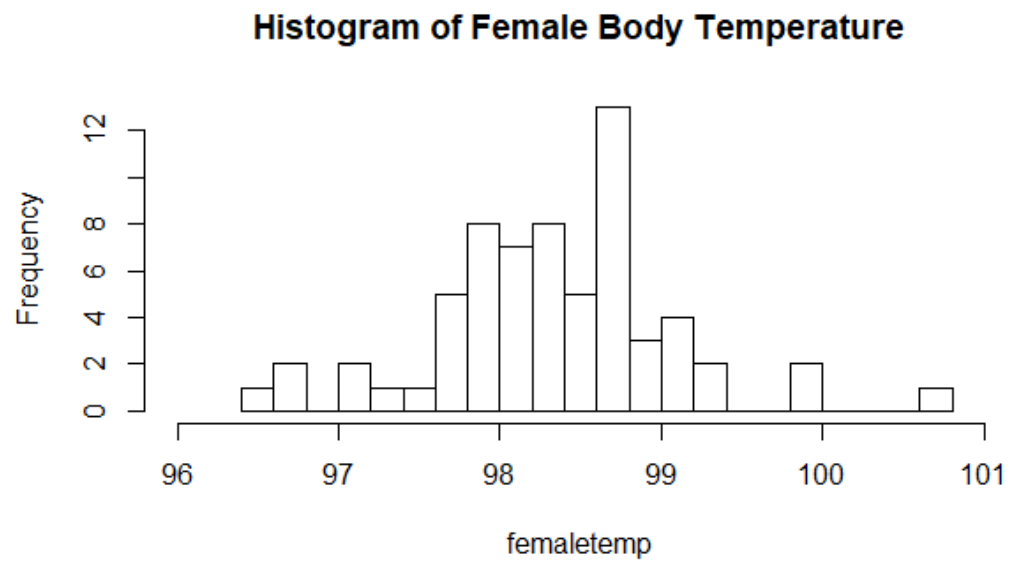
Question1:

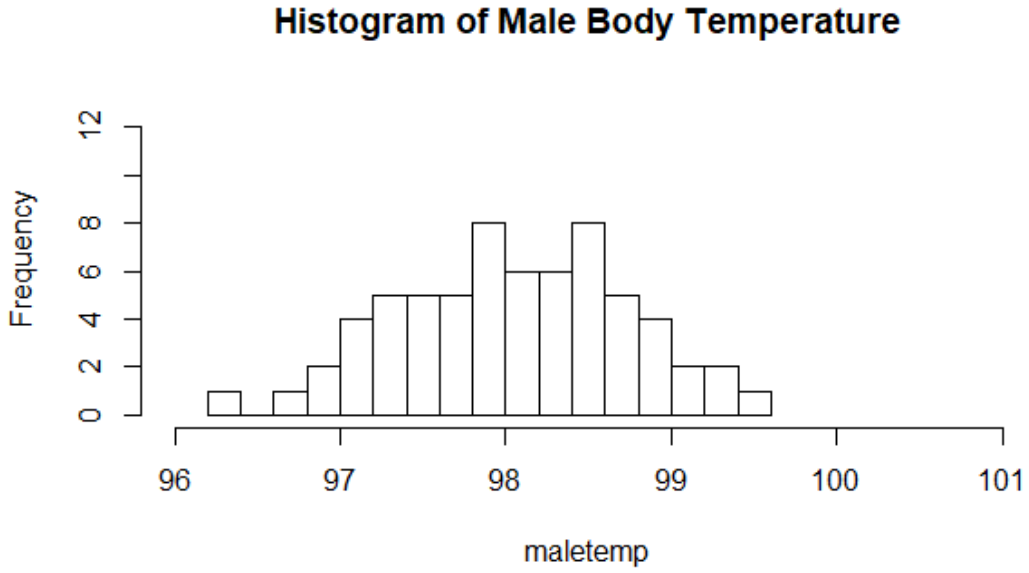
Consider the data stored in bodytemp-heartrate.csv on eLearning, containing measurements of body temperature and heart rate for 65 male (gender = 1) and 65 female (gender = 2) subjects.

(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.

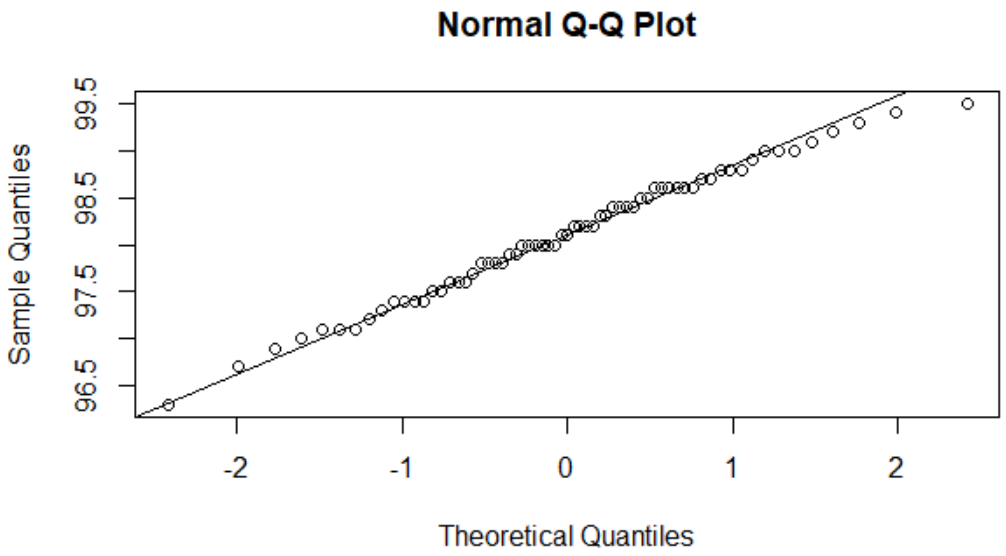
Answer:

In order to understand the mean body temperatures of all male and female data, we first need to separate the male data from the female data. We know that for males, gender==1, therefore after finding the locations of all male data, we observe that all male data is stored from location 1 to 65 and hence the next 65 tuples represent female data. From the male and female data, we only need the body temperature for each of them in order to understand their mean body temperature.
For understanding the distribution of both the data, we plot histograms for both male and female body temperatures on the same scale.
The following are the histograms that were observed:

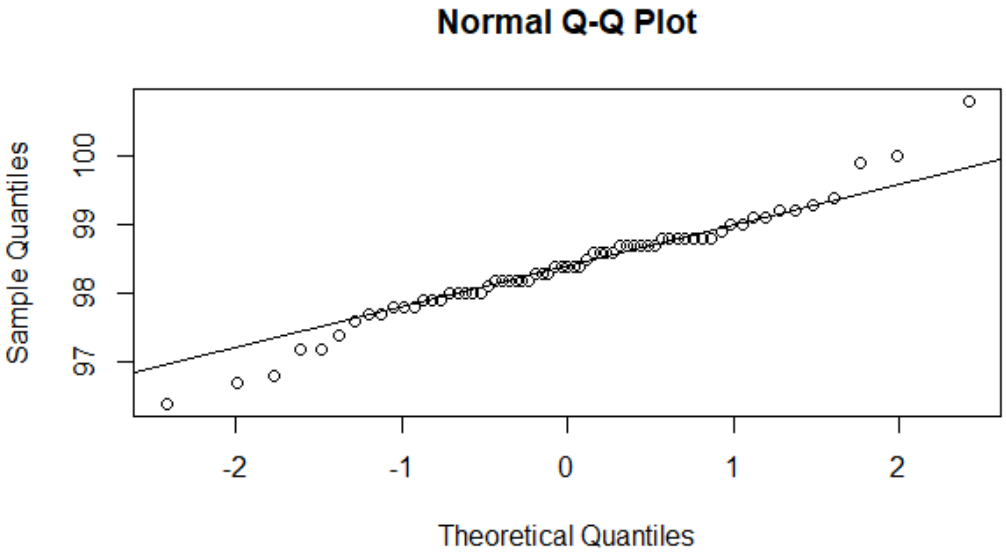




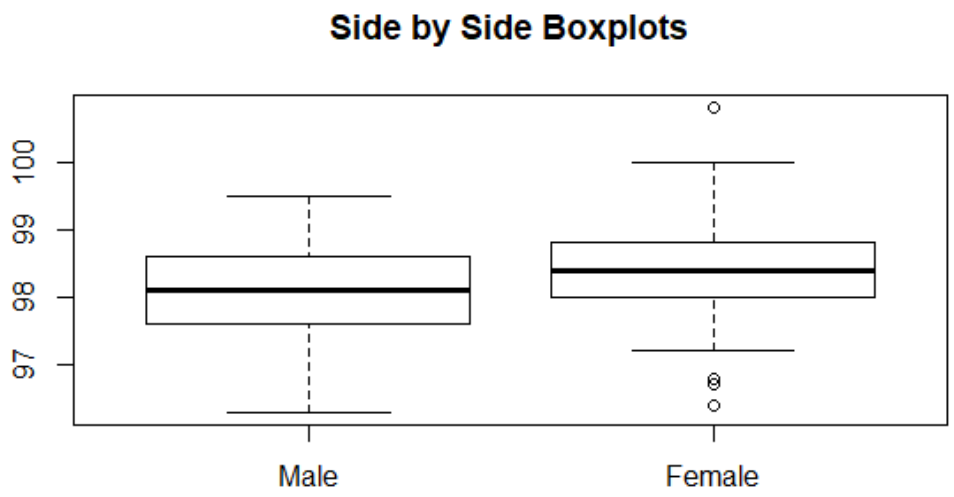
Both the histograms show that the distribution can be approximated as a Normal distribution. In order to make sure that the distributions are Normal, a QQ Plot for both the distributions is plotted.
The QQ Plot for Male body temperature is as follows:



The QQ Plot for Female body temperature is as follows:



Both the QQ Plots clearly suggests that the distributions can be approximated as Normal. Since for each distribution the value of n=65 is large enough and from the results of the QQ plot, we can say that the two distributions are approximately Normal.
Now in order to understand the means of each distributions and the difference between them, we can plot a Side by Side Boxplot.



The boxplot clearly suggests that there is a significant but a small difference between the means of the two sample distributions.
In order to perform the appropriate analysis, we need to find whether we can assume the variance of the two distributions is equal or unequal. We can find this from the Interquartile ranges of the two sample distributions.

The boxplot suggests that there is a significant difference between the Interquartile Ranges of the two sample distributions. We compute the IQR for each distribution from R and found that the IQR for male body temperature is 1, whereas the IQR of female body temperature is 0.8. The difference between them is 0.2. Hence, we cannot assume that the variances of the two distributions are equal.

In order to find that the two populations differ in their mean value or not, we use Hypothesis testing.
Our Null hypothesis is that the "Means of the two populations are equal" and hence our Alternate hypothesis becomes "Means of the two populations are not equal". We use Welch Two sample t-test to do the same.

The results obtained are:

```
Welch Two Sample t-test

data:  maletemp and femaletemp
t = -2.2854, df = 127.51, p-value = 0.02394
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean of x mean of y
 98.10462  98.39385
```

From this, we can see that the p-value is 0.02394 which is smaller than 0.025 for 95% level of significance of a two-sided test. This is sufficient evidence to show that the Null Hypothesis can be rejected. Therefore, we can say that Means of the two populations differ in their values.

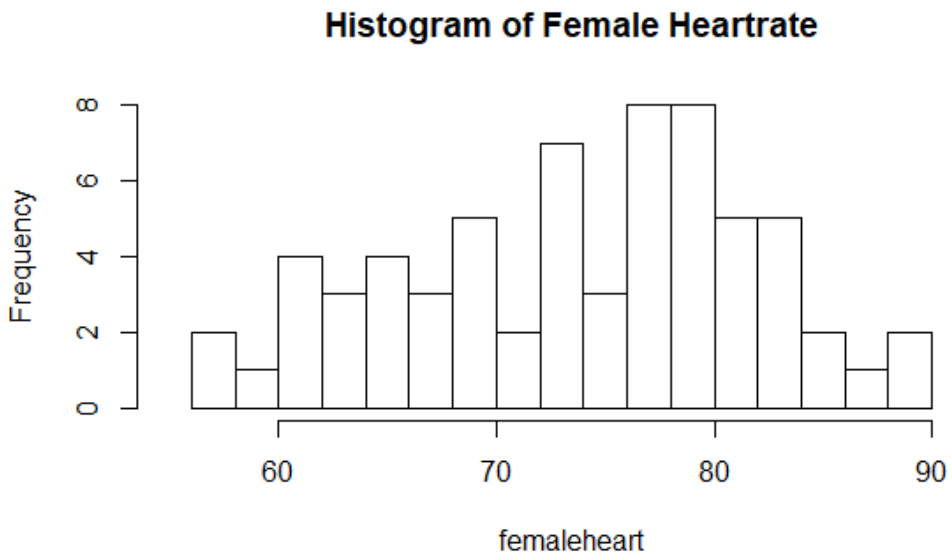
(b) Do males and females differ in mean heart rate? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

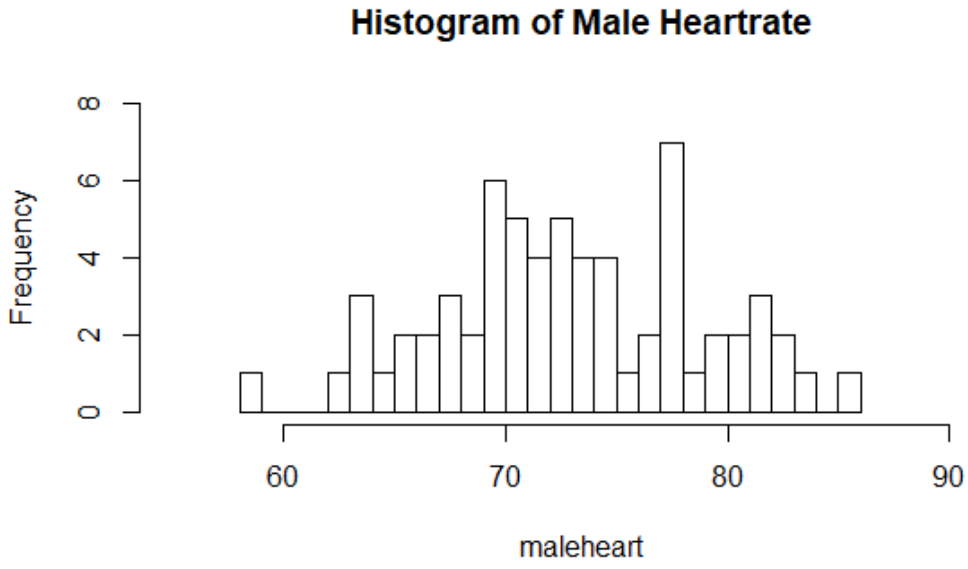
Answer:

Part b is very similar to what we have done in Part a. From the male and female data, we only need the heartrates for each of them in order to understand their mean heartrate.

For understanding the distribution of both the data, we plot histograms for both male and female heartrate on the same scale.

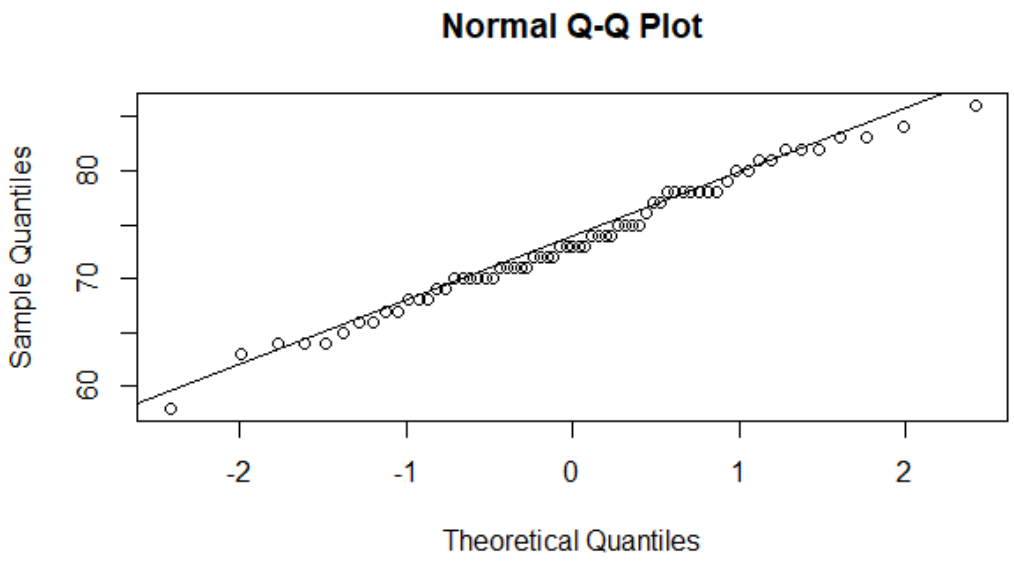
The following are the histograms that were observed:

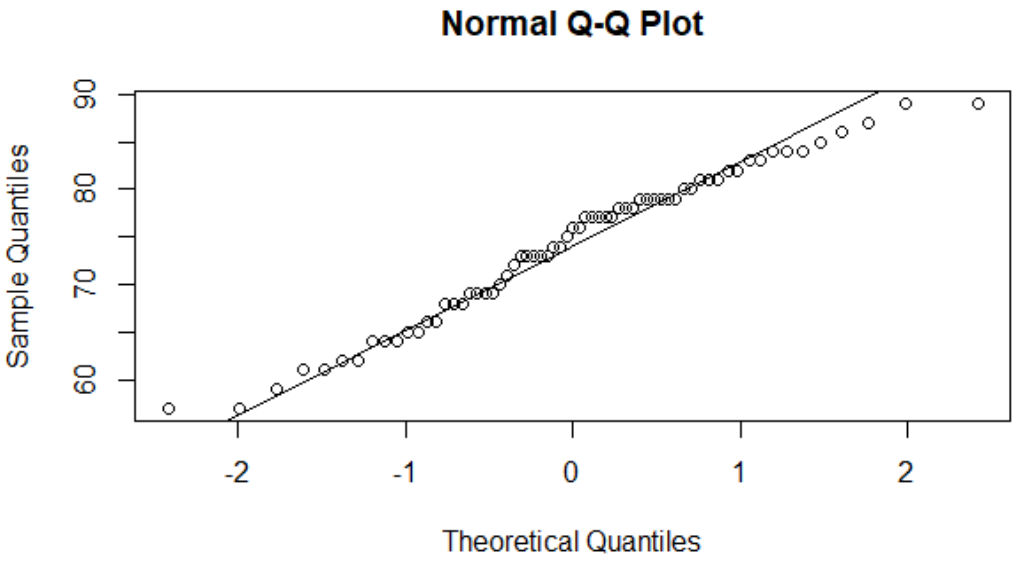




Both the histograms show that the distribution can be approximated as a Normal distribution. In order to make sure that the distributions are Normal, a QQ Plot for both the distributions is plotted.

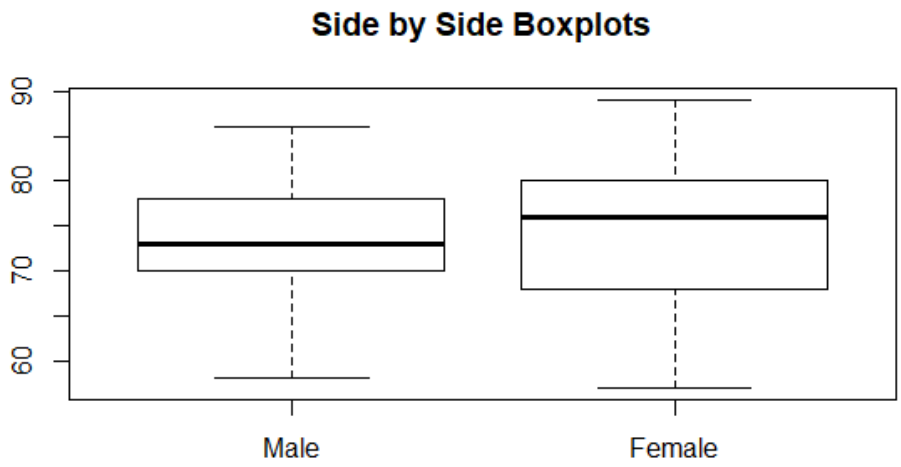
The QQ Plot for Male heartrate is as follows:





Both the QQ Plots clearly suggests that the distributions can be approximated as Normal. Since for each distribution the value of n=65 is large enough and from the results of the QQ plot, we can say that the two distributions are approximately Normal.

Now in order to understand the means of each distributions and the difference between them, we can plot a Side by Side Boxplot.



The boxplot clearly suggests that there is a significant but a small difference between the means of the two sample distributions.
In order to perform the appropriate analysis, we need to find whether we can assume the variance of the two distributions is equal or unequal. We can find this from the Interquartile ranges of the two sample distributions.

The boxplot suggests that there is a significant difference between the Interquartile Ranges of the two sample distributions. We compute the IQR for each distribution from R and found that the IQR for male heartrate is 8, whereas the IQR of female heartrate is 12. The difference between them is 4. Hence, we cannot assume that the variances of the two distributions are equal.

In order to find that the two populations differ in their mean value or not, we use Hypothesis testing.
Our Null hypothesis is that the "Means of the two populations are equal" and hence our Alternate hypothesis becomes "Means of the two populations are not equal". We use Welch Two sample t-test to do the same.

The results obtained are:

```
Welch Two Sample t-test

data:  maleheart and femaleheart
t = -0.63191, df = 116.7, p-value = 0.5287
95 percent confidence interval:
 -3.243732  1.674501
sample estimates:
mean of x mean of y
 73.36923  74.15385
```

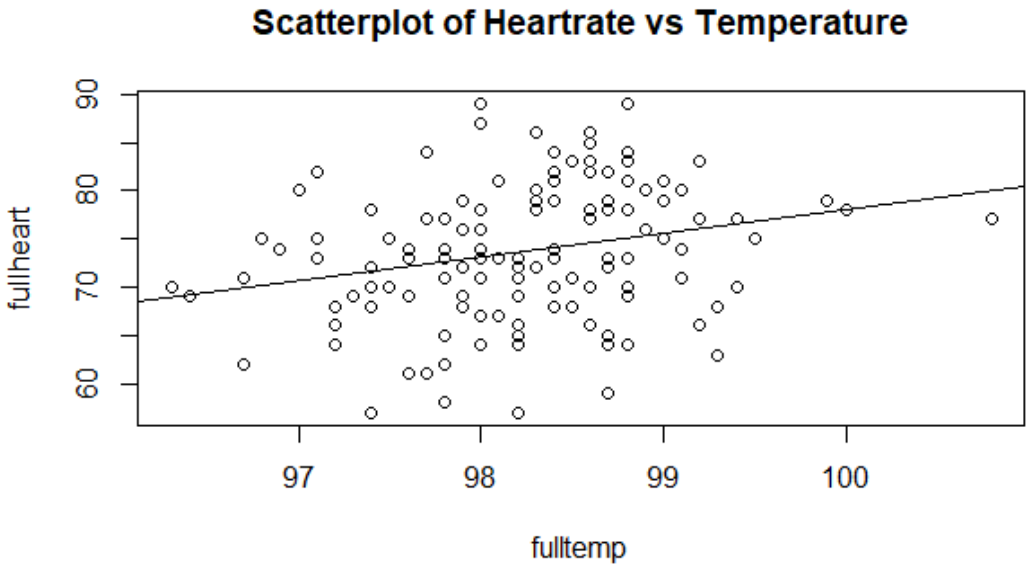
From this, we can see that the p-value is 0.5287 which is larger than 0.025 for 95% level of significance of a two-sided test. This is sufficient evidence to show that the Null Hypothesis cannot be rejected.
The p-value does not lie within the rejection region and therefore, we can say that we fail to reject the Null hypothesis suggesting that the Means of the two populations are equal.

(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.

Answer:

In order to understand the relationship between body temperature and heartrate, we have to plot a scatterplot and fit a regression line to understand how they are correlated. The correlation between body temperature and heartrate can also be found using cor() function in R. This however will give us the sample correlation which is an estimator of the population correlation.

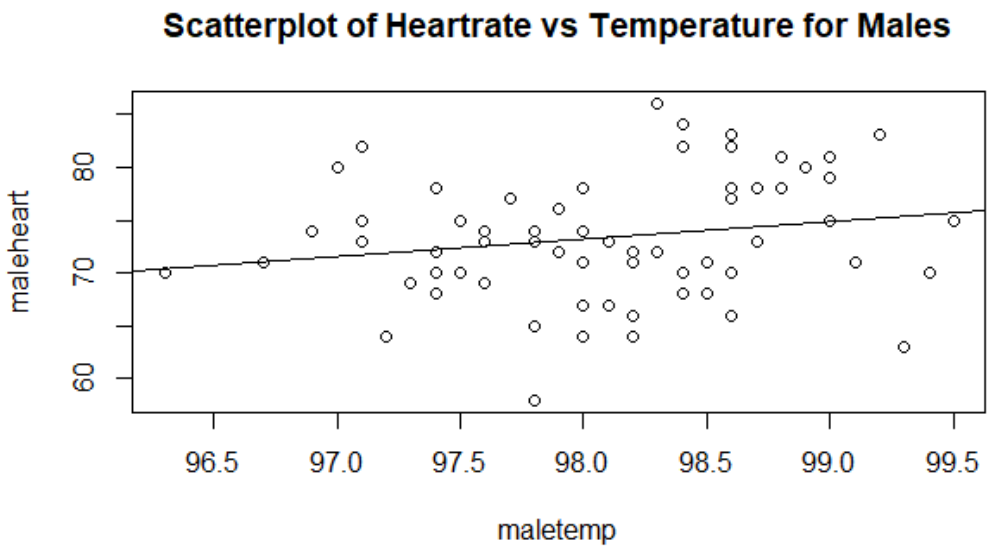
Plotting a scatterplot of Heartrate vs Body Temperature, we get:



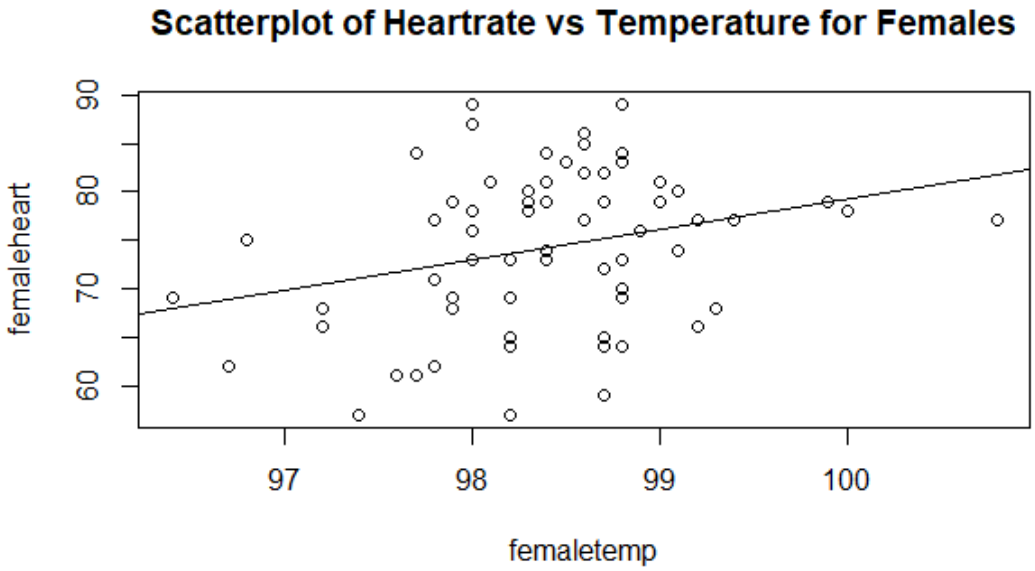
The scatterplot suggests that there is no strong linear relationship between the body temperature and heartrates. There is a linear relationship that exists between them, but it is weak. This was further proved when we compute the correlation between body temperature and heartrate in R using the cor() function. The sample correlation is 0.2536564. This further proves that there is weak correlation between body temperature and heartrate.

To answer whether this weak linear relationship depends on gender, we plot two scatterplots, one for each gender. These scatterplots show linear relationship between body temperature and heartrate for males and females separately.

Plotting a scatterplot of Heartrate vs Body Temperature Males, we get:



The scatterplot suggests that there is no strong linear relationship between the body temperature and heartrates in males. There is a linear relationship that exists between them, but it is weak. The sample correlation for Males is 0.1955894 indicating weak correlation.
Similarly, we plot a scatterplot for Females, we get:



The scatterplot suggests that there is no strong linear relationship between the body temperature and heartrates in females. There is a linear relationship that exists between them, but it is weak. The sample correlation for Females is 0.2869312 indicating weak correlation.
However, we see that the correlation between body temperature and heartrate in females is stronger than in males. The difference between sample correlations is also large enough to say that relationship between body temperature and heartrate depend on gender. But one of the things to note is that this conclusion is drawn based on the given sample only and within the range of the fitted regression line. It will not be appropriate to say that relationship between body temperature and heartrate depends on gender for the entire two populations.

Question2:

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 distribution. Note that this distribution is skewed and its mean is 1/λ. 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 - α = 0:95, λ= 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.

Answer:

To compute the coverage probabilities create a function probcoverage(). This function takes the value of n and λ as parameters and gives 2 estimated coverage probabilities. In order to do so, we calculate two confidence intervals, one for large sample Z-interval and other a parametric percentile bootstrap interval. The value of 1-α=0.95. The function first generates random exponential distribution for computation of confidence interval. The function then computes the large sample Z and parametric bootstrap confidence intervals. Once the intervals are computed then it computes their coverage probabilities. This process is repeated 5000 time for each (n,λ) value pair.
The first (n,λ) is (5,0.01). The coverage probabilities that were computed for this (n,λ) pair for large sample Z CI and parametric bootstrap CI are : 0.876 and 0.903 respectively.
This has to be done for all (n, λ) pairs.

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

Answer:

n	λ	Coverage probability (Large sample Z CI)	Coverage probability (Parametric Percentile Bootstrap CI)
5	0.01	0.8760	0.9030
5	0.1	0.8674	0.8976
5	1	0.8650	0.8968
5	10	0.8656	0.8930
10	0.01	0.9028	0.9250
10	0.1	0.9060	0.9242
10	1	0.9032	0.9252
10	10	0.9072	0.9230
30	0.01	0.9304	0.9352
30	0.1	0.9294	0.9350

30	1	0.9390	0.9476
30	10	0.9302	0.9370
100	0.01	0.9454	0.9468
100	0.1	0.9450	0.9454
100	1	0.9428	0.9448
100	10	0.9428	0.9452

(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.

Answer:

We see from the table above that as the value of n increases, the coverage probability for both the confidence intervals also improves. However, from the two confidence intervals whose coverage probability is computed, we see that for all cases the coverage probability is better for parametric bootstrap CI. For the large sample interval, the low values of n like 5,10 give coverage probability within a range from 0.8650 to 0.9072. However, we observe that as the value of n increases the coverage probability also increases. For large sample interval, when n is 30 or 100, the coverage probability converges more towards 0.95 that is 95% confidence interval. We see that for even n=100, the coverage probability for large sample interval does not reach 0.95. Therefore, we can say that for large sample interval, n should be as large as possible for the interval to be accurate.

For parametric bootstrap interval, the coverage probabilities observed are better than the large sample intervals in all cases. You can see for all values of (n, λ) the coverage probability for parametric bootstrap is always greater than large sample interval which shows that the bootstrap interval gives a more accurate interval for various values of (n, λ). But however, this is true only for small values of n like 5,10. As value of n increases, we see that the difference between the coverage probabilities of bootstrap interval and large sample interval decreases. Both are approximately, equally accurate for large values of n like 30,100. From the table we can see, that for n=100, λ=10, the coverage probability for parametric bootstrap is 0.9452 and that of large sample interval is 0.9428. There is not much difference in their coverage probabilities as n increases. This follows the Central Limit Theorem. Hence, we can say that, the bootstrap interval gives a better coverage probability which converges more towards 0.95 and hence should be the choice when n is small. As n increases, n is large enough for Normal approximation to hold. So similar to large sample interval, the bootstrap interval requires large value of n to be more accurate.

The answers do not depend on the value of λ. From the table, we can see that the coverage probabilities for both the intervals does not depend on the value of λ. For so many different combinations of λ with n, we still can't see any specific trend that can be seen because of values in λ. The trend is only observed for changes in value of n. The variability that is seen in the table for various values of (n, λ) is because of Monte Carlo simulations. The value of λ, does not affect the coverage probabilities of the two intervals in any way.

From all the observations in the table, we can say that for smaller values of n, the bootstrap interval gives better coverage probability and hence should be the choice when n is small. For large values of n, the bootstrap interval and large sample interval give approximately the same coverage probabilities and hence no method has an upper hand over the other. Hence one cannot recommend an interval because it

all depends on the value of n. For smaller n, bootstrap is better, but for large n, both are approximately the same.

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

Answer:

The conclusions that are drawn from the observations do not depend on the specific values of λ. This is because, λ does not affect the coverage probabilities that are obtained from the two intervals. Also. For the estimating mean of the population, we used Maximum Likelihood Estimator. We estimated the population mean from the sample mean, but since MLE was used, we get an unbiased estimator of mean that converges towards λ even if the value of λ is determined beforehand. Therefore, to conclude we can say that the value of λ does not affect any part of the conclusions that are made.

R codes

Question 1:

a.

```
#Reading the csv file for computation
> fulldata=read.csv('bodytemp-heartrate.csv', header = TRUE, sep = ",")
> maledata=which(fulldata$gender==1) #Checking which positions are male data
> maledata
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
[27] 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
[53] 53 54 55 56 57 58 59 60 61 62 63 64 65
> maledata=fulldata[1:65, ] #Storing maledata with their bodytemp & heartrate
> femaledata=fulldata[66:130, ] #Similarly storing femaledata
> maletemp=maledata[,1]   #Storing male body temperature data
> femaletemp=femaledata[,1]  #Storing female body temperature data


> hist(femaletemp, main="Histogram of Female Body Temperature", breaks = 20, xlim = c(96,
101), ylim = c(0,13)) #Plotting histogram for Female bodytemp


> hist(maletemp, main="Histogram of Male Body Temperature", breaks = 20, xlim = c(96,101)
, ylim = c(0,13)) #Plotting histogram for Male bodytemp


> qqnorm(maletemp) #Plotting QQ Plot for Male body temperatures
> qqline(maletemp)

> qqnorm(femaletemp) #Plotting QQ Plot for Female body temperatures
> qqline(femaletemp)


> boxplot(maletemp,femaletemp, main="Side by Side Boxplots", names=c('Male','Female')) #P
lotting side by side boxplot to understand means


#Computing IQR of both female & male bodytemp to check variance are equal or #not
> IQR(maletemp)
[1] 1
> IQR(femaletemp)
[1] 0.8


#Confirming what was obtained from boxplots that there is difference in means
> t.test(maletemp,femaletemp, alternative = "two.sided", conf.level = 0.95, var.equal = F
ALSE)

        Welch Two Sample t-test

data:  maletemp and femaletemp
t = -2.2854, df = 127.51, p-value = 0.02394
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean of x mean of y
 98.10462  98.39385
```

b.

```
> maleheart=maledata[,3] #Storing male heartrate data
> femaleheart=femaledata[,3] #Storing female heartrate data

> hist(femaleheart, main="Histogram of Female Heartrate", breaks = 20, xlim = c(55,92), y
lim = c(0,8)) #Plotting histogram for Female heartrates

> hist(maleheart, main="Histogram of Male Heartrate", breaks = 20, xlim = c(55,92), ylim
= c(0,8)) #Plotting histogram for Male heartrates

> qqnorm(maleheart) #Plotting QQ Plot for Male heartrates
```

```
> qqline(maleheart)

> qqnorm(femaleheart) #Plotting QQ Plot for Female heartrates
> qqline(femaleheart)

> boxplot(maleheart,femaleheart, main="Side by Side Boxplots", names=c('Male','Female'))
#Plotting side by side boxplots to understand the means


#Computing IQR of both female and male heartrates to check variance are equal #or not
> IQR(maleheart)
[1] 8
> IQR(femaleheart)
[1] 12
#Confirming what was obtained from boxplots that there is difference in means
> t.test(maleheart,femaleheart, alternative = "two.sided", conf.level = 0.95, var.equal =
FALSE)

        Welch Two Sample t-test

data:  maleheart and femaleheart
t = -0.63191, df = 116.7, p-value = 0.5287
95 percent confidence interval:
 -3.243732  1.674501
sample estimates:
mean of x mean of y
 73.36923  74.15385
```

c.

```
> fulltemp=fulldata[,1] #Storing all the body temperatures
> fullheart=fulldata[,3] #Storing all the heartrates
> cor(fulltemp,fullheart) #Computing correlation b/w the two
[1] 0.2536564

#Plotting Scatterplot to understand the linear relationship b/w them
> plot(fulltemp,fullheart,main = "Scatterplot of Heartrate vs Temperature")
> abline(lm(fullheart~fulltemp))


> cor(maletemp,maleheart) # Correlation b/w male heartrates and bodytemp
[1] 0.1955894

#Plotting scatterplot to understand whether gender affects the relationship
> plot(maletemp,maleheart,main = "Scatterplot of Heartrate vs Temperature for Males")
> abline(lm(maleheart~maletemp))

> cor(femaletemp,femaleheart) #Correlation b/w female heartrates and bodytemp
[1] 0.2869312

#Plotting scatterplot to understand whether gender affects the relationship
> plot(femaletemp,femaleheart,main = "Scatterplot of Heartrate vs Temperature for Females
")
> abline(lm(femaleheart~femaletemp))
```

Question 2:

```
probcoverage=function(n,lambda)
    {
    sample1<-rexp(n,lambda)
    actualmean=1/lambda
    estimatedmean=mean(sample1)
    estimatedlambda=1/estimatedmean
    estimatedsd=mean(sample1)

    Zci=estimatedmean+c(-1,1)*qnorm(1-(1-.95)/2)*(estimatedsd/sqrt(n))

    meanboot=function(n,estimatedlambda){
      bootsample=rexp(n,estimatedlambda)
```

```
      meanofbootsample=mean(bootsample)
      return(meanofbootsample)
    }

    library(boot)
    bootdist=replicate(999,meanboot(n,estimatedlambda))
    percCI=sort(bootdist)[c(25,975)]

    if(Zci[1]<=actualmean && Zci[2]>=actualmean){
      acceptzci=1
      }
    else{
      acceptzci=0
    }
    if(percCI[1]<=actualmean && percCI[2]>=actualmean){
      acceptCI=1
    }
    else{
      acceptCI=0
    }
    finalcoverage=c(acceptzci,acceptCI)
    return(finalcoverage)

  }

   nval=c(5,10,30,100)
  lambdavalues=c(0.01,0.1,1,10)
  for (n in nval) {
    for (lambda in lambdavalues) {
      cat("(n,lambda)",n,lambda)
      coverageprop=replicate(5000, probcoverage(n,lambda))
      cat("\n Coverage probabilities for Z CI and Parametric Bootstrap respectively are:"
)
      meancoverage=rowMeans(coverageprop)
      cat("\n",meancoverage)
      cat("\n")

    }
  }



 (n,lambda) 5 0.01
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.876 0.903
(n,lambda) 5 0.1
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.8674 0.8976
(n,lambda) 5 1
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.865 0.8968
(n,lambda) 5 10
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.8656 0.893
(n,lambda) 10 0.01
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9028 0.925
(n,lambda) 10 0.1
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.906 0.9242
(n,lambda) 10 1
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9032 0.9252
(n,lambda) 10 10
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9072 0.923
(n,lambda) 30 0.01
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9304 0.9352
(n,lambda) 30 0.1
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9294 0.935
```

```
(n,lambda) 30 1
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.939 0.9476
(n,lambda) 30 10
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9302 0.937
(n,lambda) 100 0.01
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9454 0.9468
(n,lambda) 100 0.1
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.945 0.9454
(n,lambda) 100 1
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9428 0.9448
(n,lambda) 100 10
 Coverage probabilities for Z CI and Parametric Bootstrap respectively are:
 0.9428 0.9452
```