Mini Project 5

Names of group members: Karan Risbud(KSR190005), Shubham Vartak(SXV200115)

Contribution of each group member: Both the Project group members worked together on the project. Collaborated to solve the problem and implementation of R programming.

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.

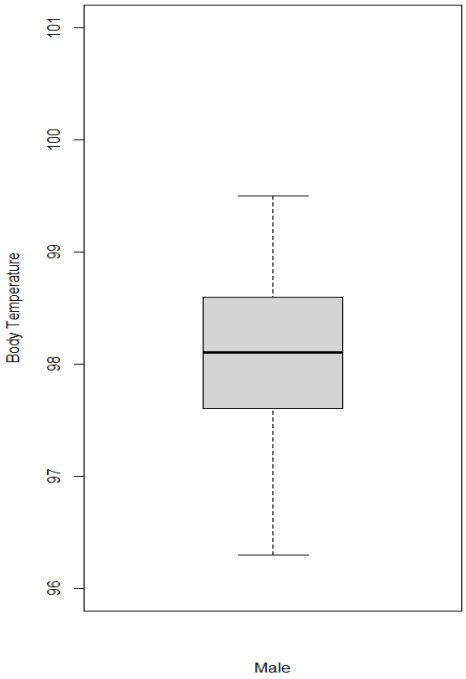
(a) Do males and females differ in mean body temperature? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

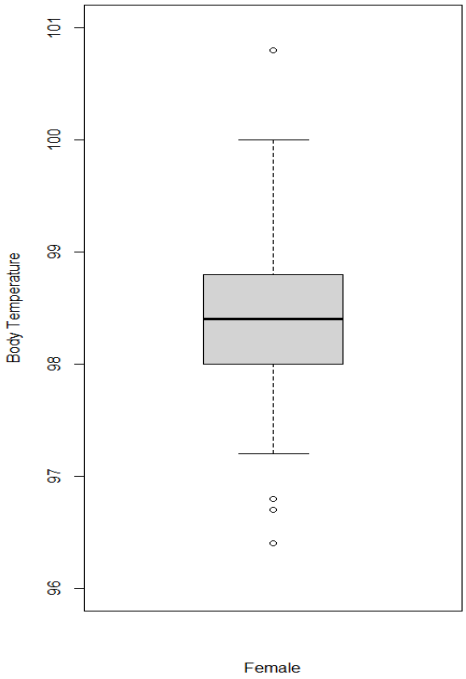
R code:

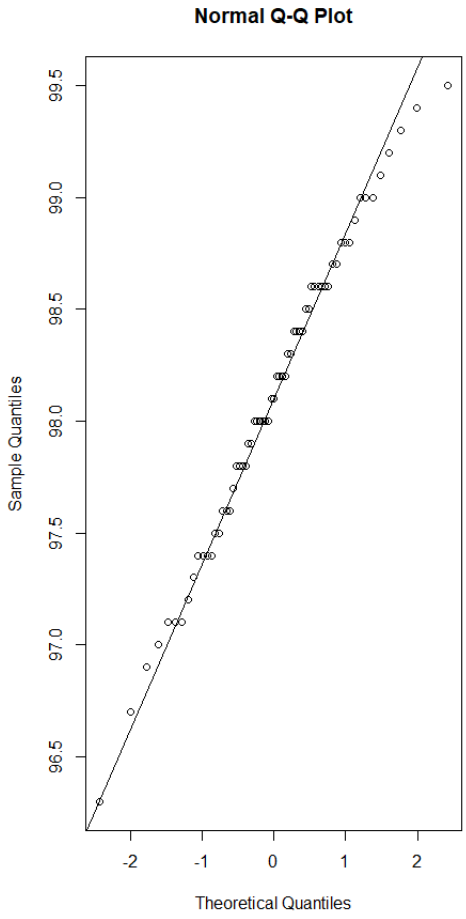
```
dataset = read.csv("D:/Fall'21/STATS/mini_project_5/bodytemp-heartrate.csv")
male = dataset[dataset$gender == 1,]
female = dataset[dataset$gender == 2,]

mean(male$body_temperature)
mean(female$body_temperature)

par(mfrow=c(1,2))
boxplot(male$body_temperature,ylim=c(96,101),xlab="Male",ylab="Body Temperature")
boxplot(female$body_temperature,ylim=c(96,101),xlab="Female",ylab="Body Temperature")

par(mfrow=c(1,2))
qqnorm(male$body_temperature)
qqline(male$body_temperature)

qqnorm(female$body_temperature)
qqline(female$body_temperature)

summary(male$body_temperature)
summary(female$body_temperature)
```

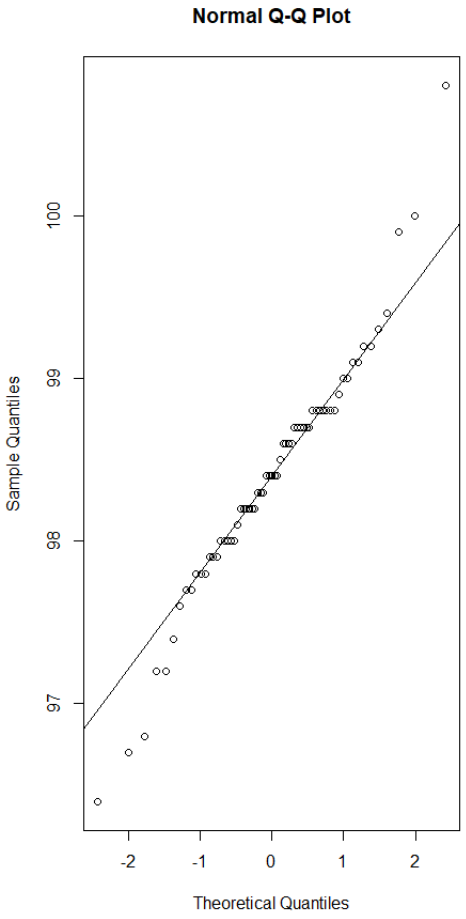
```
> summary(male$body_temperature)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   96.3    97.6    98.1    98.1    98.6    99.5
> summary(female$body_temperature)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  96.40   98.00   98.40   98.39   98.80  100.80
```

Plots:


Body Temperature
96
97
98
99
100
101
Male


Body Temperature
96
97
98
99
100
101
Female


Normal Q-Q Plot
Sample Quantiles
96.5
97.0
97.5
98.0
98.5
99.0
99.5
-2
-1
0
1
2
Theoretical Quantiles


Normal Q-Q Plot
Sample Quantiles
97
98
99
100
-2
-1
0
1
2
Theoretical Quantiles

Observations:

1. From visualizing the QQ plot, we can assume male and female body distributions to be approximately normal(also number of samples are greater than 30).
2. Based on summary and box plot we cannot assume equal variances between the two distributions. Also, from summary we observe that mean body temperature in female is slightly higher than that in male.
3. We perform hypothesis testing to determine difference in mean body temperature between male and female. We will use t test with unequal variance and treat them as independent samples.
4. Null hypothesis h0: mean(male.body_temp) - mean(female.body_temp) = 0
5. Alternate hypothesis h1: mean(male.body_temp) - mean(female.body_temp) != 0

R Code:

```
t.test(male$body_temperature,female$body_temperature,alternative = "two.sided",var.equal = FALSE)
```

```
        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
```

Observations:

1. Based on the t test we obtain the p value as 0.02394 which is much less than 0.05
2. Also, the CI does not contain 0.
3. Based on these 2 observations we reject the null hypothesis.
4. Thus, our conclusion is that there is difference between mean body temperature of male and female and the females mean body temperature is higher.

Q1b) Do males and females differ in mean heart rate? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

R code:

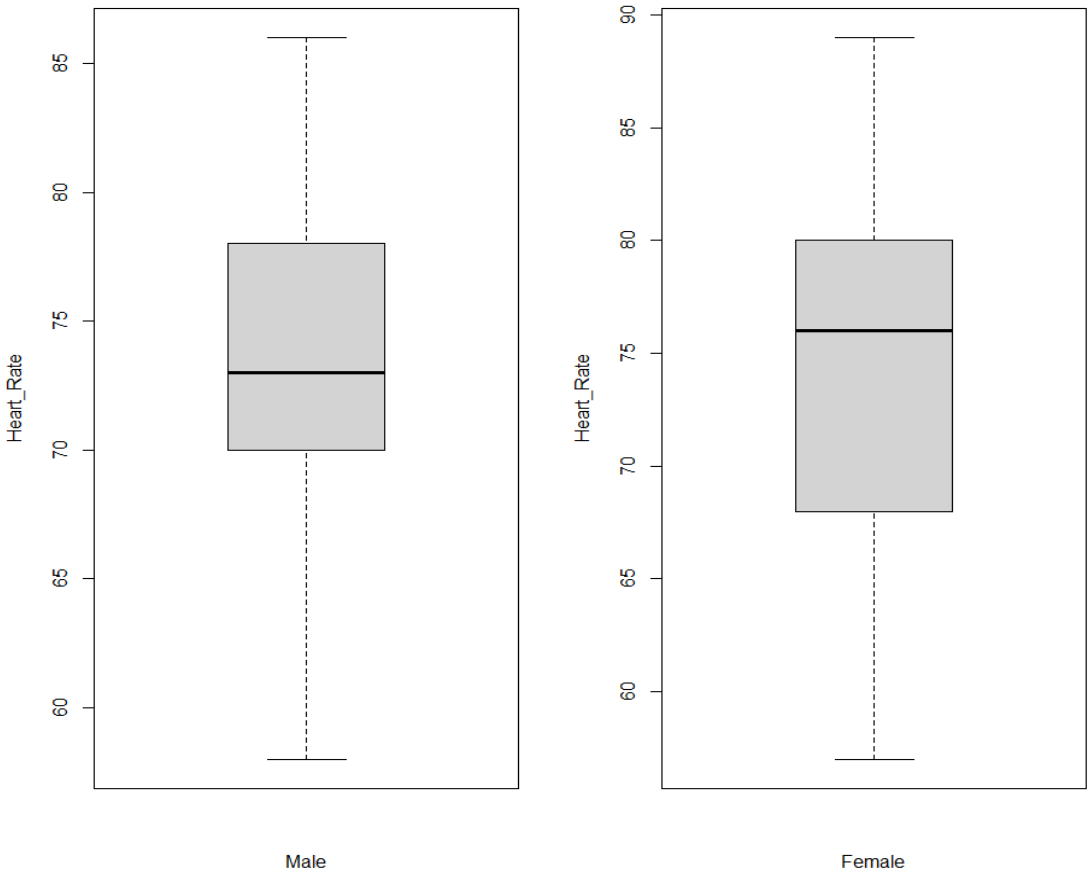
```
male_heartrate = dataset[dataset$gender == 1,]
female_heartrate = dataset[dataset$gender == 2,]

mean(male_heartrate$heart_rate)
mean(female_heartrate$heart_rate)

par(mfrow=c(1,2))
boxplot(male_heartrate$heart_rate,xlab="Male",ylab="Heart_Rate")
```

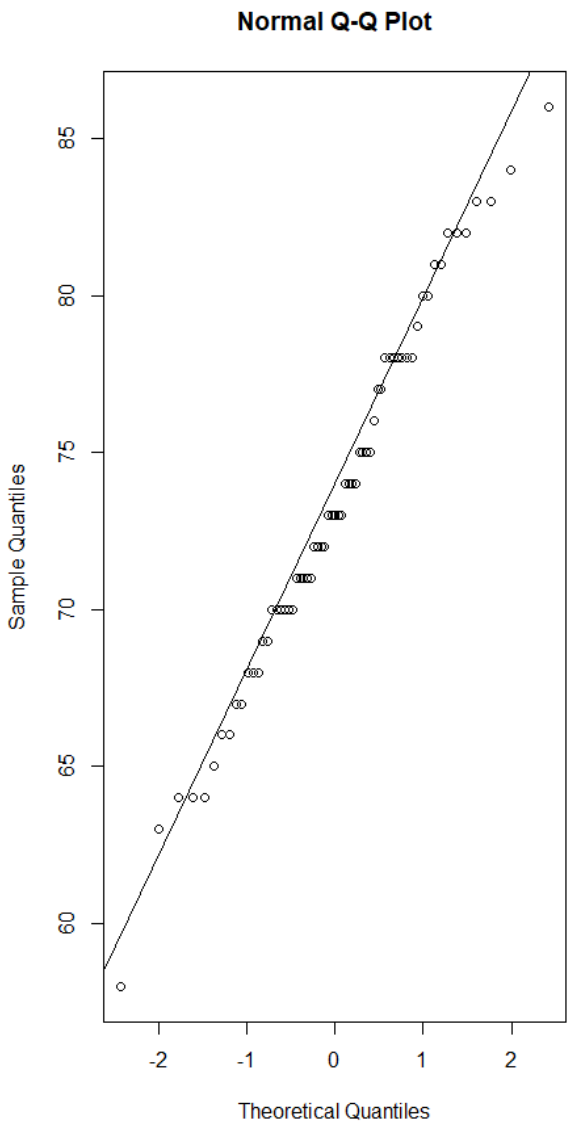
```
boxplot(female_heartrate$heart_rate,xlab="Female",ylab="Heart_Rate")

par(mfrow=c(1,2))
qqnorm(male_heartrate$heart_rate)
qqline(male_heartrate$heart_rate)

qqnorm(female_heartrate$heart_rate)
qqline(female_heartrate$heart_rate)

summary(male_heartrate$heart_rate)
summary(female_heartrate$heart_rate)
```

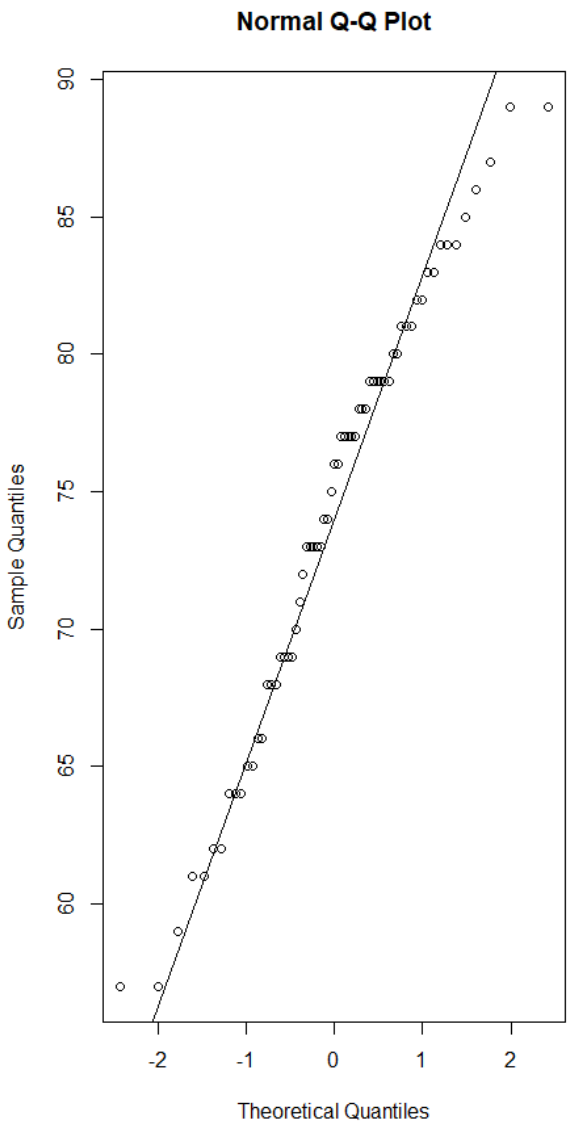
Summary:

```
> summary(male_heartrate$heart_rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  58.00   70.00   73.00   73.37   78.00   86.00
> summary(female_heartrate$heart_rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  57.00   68.00   76.00   74.15   80.00   89.00
```

Plots:







Observations:

1. From visualizing the QQ plot, we can assume male and female body distributions to be approximately normal. Slight skewness visible via boxplot.
2. Based on summary and box plot we cannot assume equal variances between the two distributions. Hence, we will use t test with unequal variance and treat the two samples as independent sample.
3. We perform hypothesis testing to determine difference in mean Heart Rates between male and female.
4. Null hypothesis h0 : mean(male.heart_rate) - mean(female.heart_rate)=0
5. Alternate hypothesis h1: mean(male.heart_rate) - mean(female.heart_rate)!=0

R Code:

```
t.test(male_heartrate$heart_rate,female_heartrate$heart_rate,alternative = "two.sided",var.equal = FALSE)
```

```
        Welch Two Sample t-test

data:  male_heartrate$heart_rate and female_heartrate$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
```

Observations:

1. Based on the t test we obtain the p value as 0.5287 which is much higher than 0.05
2. Also, the CI does contain 0.
3. Based on these 2 observations we accept the null hypothesis and reject alternate hypothesis as there is insufficient evidence.
4. Thus, our conclusion is that there is no difference between mean body heart_rate of male and female.

Q1c) 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.

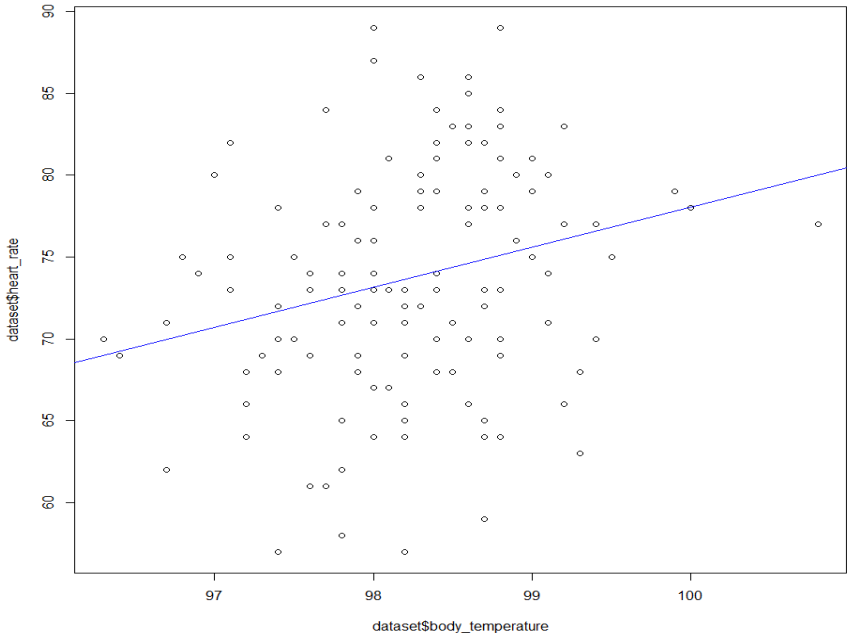
R Code:

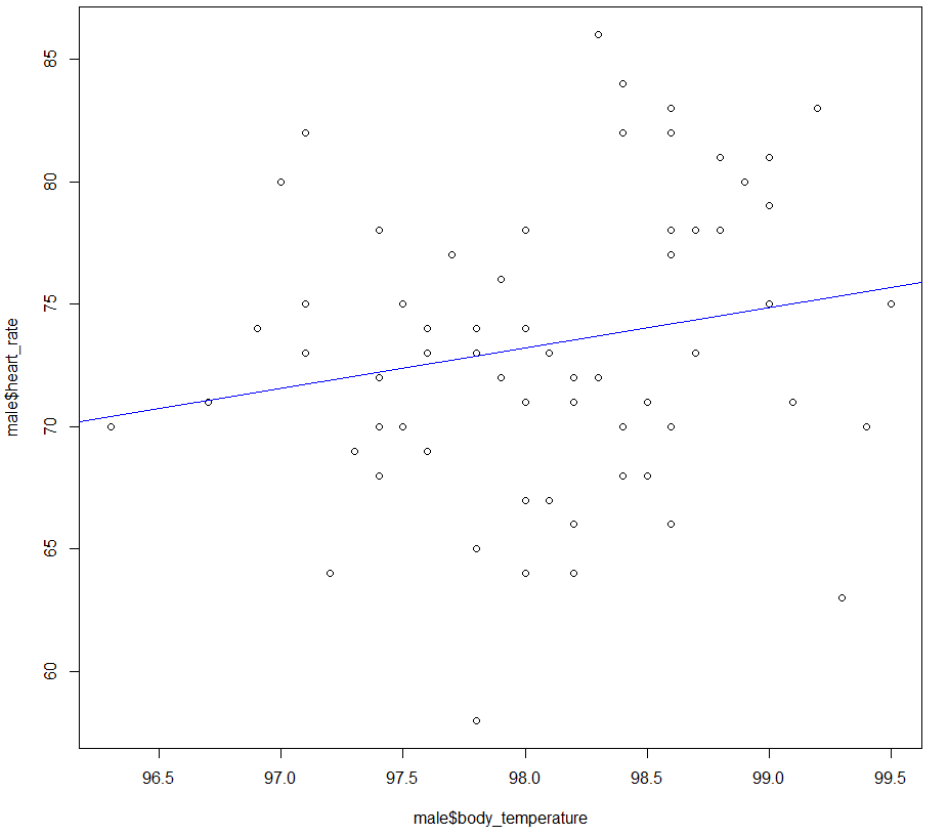
```
plot(dataset$body_temperature,dataset$heart_rate)
abline(lm(dataset$heart_rate~dataset$body_temperature),col="blue")
cor(dataset$body_temperature,dataset$heart_rate)
lm(dataset$body_temperature~dataset$heart_rate)

par(mfrow=c(1,1))
plot(male$body_temperature,male$heart_rate)
abline(lm(male$heart_rate~male$body_temperature),col="blue")
cor(male$body_temperature,male$heart_rate)
lm(male$body_temperature~male$heart_rate)

plot(female$body_temperature,female$heart_rate)
abline(lm(female$heart_rate~female$body_temperature),col="blue")
cor(female$body_temperature,female$heart_rate)
lm(female$body_temperature~female$heart_rate)
```

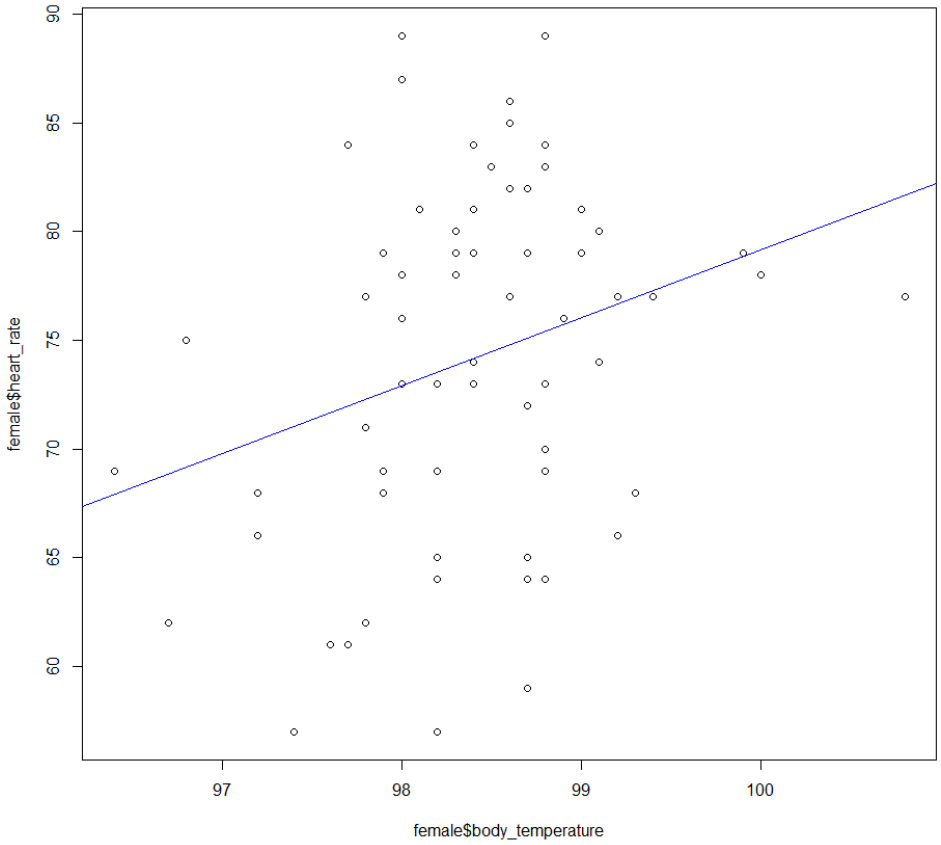
Output:

```
> plot(dataset$body_temperature,dataset$heart_rate)
> abline(lm(dataset$heart_rate~dataset$body_temperature),col="blue")
> cor(dataset$body_temperature,dataset$heart_rate)
[1] 0.2536564
> lm(dataset$body_temperature~dataset$heart_rate)

Call:
lm(formula = dataset$body_temperature ~ dataset$heart_rate)

Coefficients:
       (Intercept)  dataset$heart_rate
          96.30675             0.02633

>
> par(mfrow=c(1,1))
> plot(male$body_temperature,male$heart_rate)
> abline(lm(male$heart_rate~male$body_temperature),col="blue")
> cor(male$body_temperature,male$heart_rate)
[1] 0.1955894
> lm(male$body_temperature~male$heart_rate)

Call:
lm(formula = male$body_temperature ~ male$heart_rate)

Coefficients:
    (Intercept)  male$heart_rate
       96.39789          0.02326

>
> plot(female$body_temperature,female$heart_rate)
> abline(lm(female$heart_rate~female$body_temperature),col="blue")
> cor(female$body_temperature,female$heart_rate)
[1] 0.2869312
> lm(female$body_temperature~female$heart_rate)

Call:
lm(formula = female$body_temperature ~ female$heart_rate)

Coefficients:
      (Intercept)  female$heart_rate
         96.44211            0.02632
```

Plots:




male$heart_rate
male$body_temperature
60
65
70
75
80
85
96.5
97.0
97.5
98.0
98.5
99.0
99.5


female$heart_rate
female$body_temperature
60
65
70
75
80
85
90
97
98
99
100

Observations:

1. Based on plot which does not consider gender, the observed line has a positive slope and the value of correlation coefficient is 0.253.
2. This indicates that there is "weak positive linear relationship" between body temperature and heart rate.
3. Based on the 2 plots that considers gender. We observe a line with positive slope and positive value of correlation that indicates positive linear relationship.
4. The value of correlation when gender is male is 0.195 and that of female is 0.2869.
5. We can say that female's body temperature and heart rate are more strongly connected than that in male.
6. Also, significant difference in correlation values suggest that it does depend on gender.

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 μ = 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 $1 - \alpha = 0.95$, 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.

Q2a) 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.

R Code:

```
z.proportion = function(x,lambda){
  SE = sd(x)/sqrt(length(x))
  ci = (mean(x) + c(-1,1)*qnorm(0.975)*SE)
  population.mean = 1/lambda
  if(ci[2]>population.mean & ci[1]< population.mean)
    return(1)
  else
    return(0)

}

boot.proportion = function(x,n,lambda){
  #calculated 999 means
  b = replicate(1000,mean(rexp(n,1/mean(x))))
  ci = sort(b)[c(25,975)]
  population.mean = 1/lambda
  if(ci[2]>population.mean & ci[1]<population.mean)
    return(1)
  else
    return(0)
}

mc.sim = function(n,lambda){
  x.sample = rexp(n,lambda)
  z = z.proportion(x.sample,lambda)
  p = boot.proportion(x.sample,n,lambda)
  return(c(z,p))


}

monte.Carlo.sim = function(n, lambda){
  #MC Trials
  mc = replicate(5000,mc.sim(n, lambda))
  #output results
```

```
  return(c(sum(mc[1,])/length(mc[1,]),sum(mc[2,])/length(mc[2,])
         ))
}


print(monte.Carlo.sim(5, 0.01))
```

Output:

```
> print(monte.Carlo.sim(5, 0.01))
[1] 0.8062 0.8970
```

Observations:

1. Here We take values of n and lambda and use monte Carlo simulations to construct 2 CI (z interval and percentile parametric bootstrapping)
2. We simulate this 5000 times and check each time whether the CI contains the true value of population mean.
3. Based on the results obtained, we calculate the coverage probabilities.
4. We got the probabilities as:
5. z-interval = 0.8062
6. p-interval = 0.8970

2b) Repeat (a) for the remaining combinations of (n, λ). Present an appropriate summary of the results.

R Code:

```
#make 2 empty matrices
z_matrix = matrix(nrow=4, ncol=4)
p_matrix = matrix(nrow=4, ncol=4)
#all values for lambda and n to test for
lambda.vector = c(0.01, 0.1, 1, 10)
n.vector = c(5, 10, 30, 100)
#for each value n and lambda, run 5000 MC trials and estimate the coverage

row = 1
for(lambda in lambda.vector){
  col = 1
  for(n in n.vector){
    proportion.vector = monte.Carlo.sim(n, lambda)
    print(row)
    z_matrix[row, col] = proportion.vector[1]
    p_matrix[row, col] = proportion.vector[2]
    col = col + 1
  }
  row = row + 1
}
z_matrix
p_matrix
#output both matrices to a csv for ease of access for reporting
write.csv(data.frame(z_matrix), "z_out.csv")
write.csv(data.frame(p_matrix), "p_out.csv")
```

Observations:

1. We repeat the above process for the remaining observations
2. We calculate for every value of n and lambda and store it in excel file.
3. We obtain the following values represented in table as shown:

Output:

```
> z_matrix
       [,1]   [,2]   [,3]   [,4]
[1,] 0.8196 0.8736 0.9138 0.9336
[2,] 0.8126 0.8668 0.9260 0.9400
[3,] 0.8224 0.8766 0.9164 0.9382
[4,] 0.8200 0.8618 0.9230 0.9332
> p_matrix
       [,1]   [,2]   [,3]   [,4]
[1,] 0.8988 0.9258 0.9354 0.9408
[2,] 0.8958 0.9228 0.9446 0.9472
[3,] 0.8980 0.9244 0.9396 0.9460
[4,] 0.8966 0.9158 0.9452 0.9432
```

Z-Table:

	N=5	N=10	N=30	N=100
λ=0.01	0.8196	0.8736	0.9138	0.9336
λ=0.1	0.8126	0.8668	0.926	0.94
λ=1	0.8224	0.8766	0.9164	0.9382
λ=10	0.82	0.8618	0.923	0.9332

P-Table:

	N=5	N=10	N=30	N=100
λ=0.01	0.8988	0.9258	0.9354	0.9408
λ=0.1	0.8958	0.9228	0.9446	0.9472
λ=1	0.898	0.9244	0.9396	0.946
λ=10	0.8966	0.9158	0.9452	0.9432

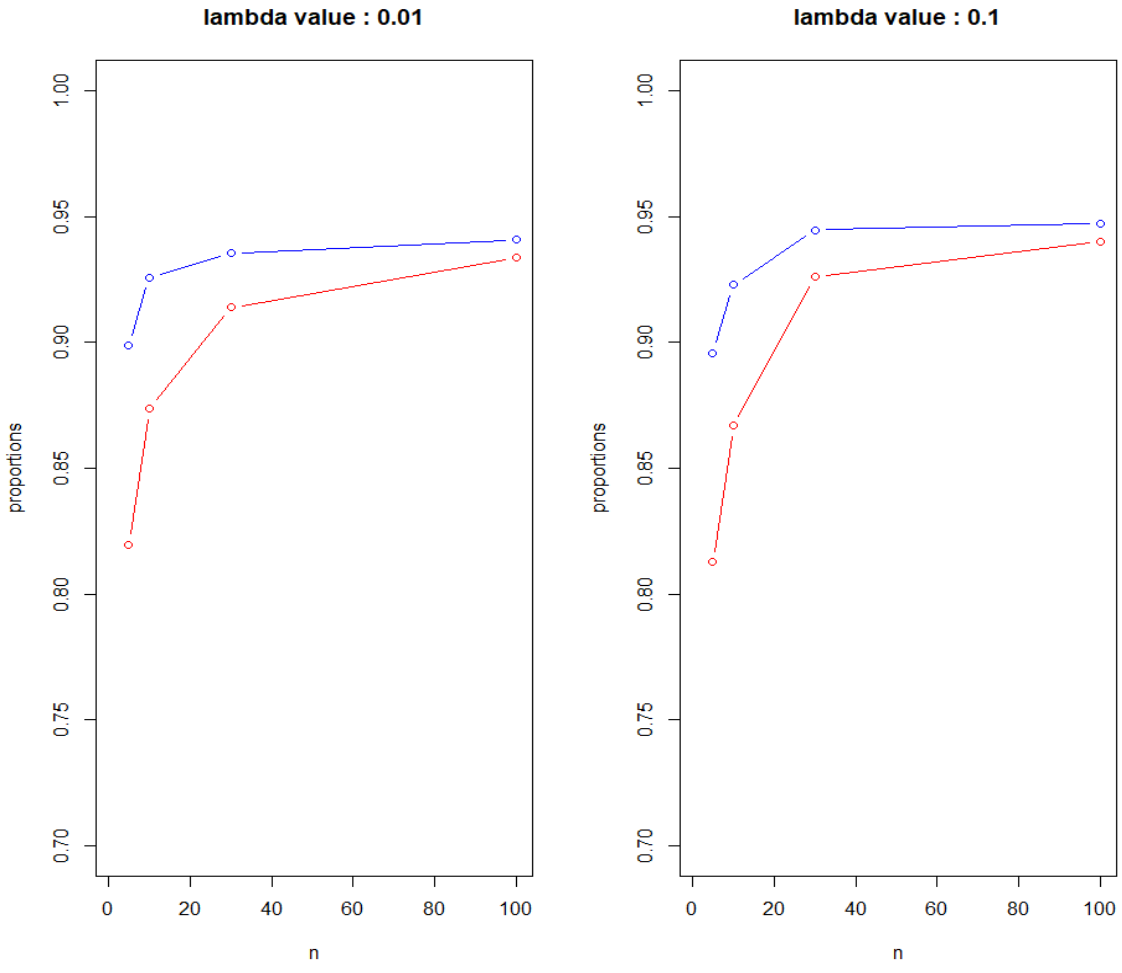
Q2c) Interpret all the results. Be sure to answer the following questions: In case of the large-sample interval, how large n is needed for the interval to be accurate? Likewise, in case of the bootstrap interval, how large n is needed for the interval to be accurate? Do these answers depend on λ? Can we say that one method is more accurate than the other? Which interval would you recommend? Provide justification for all your conclusions.

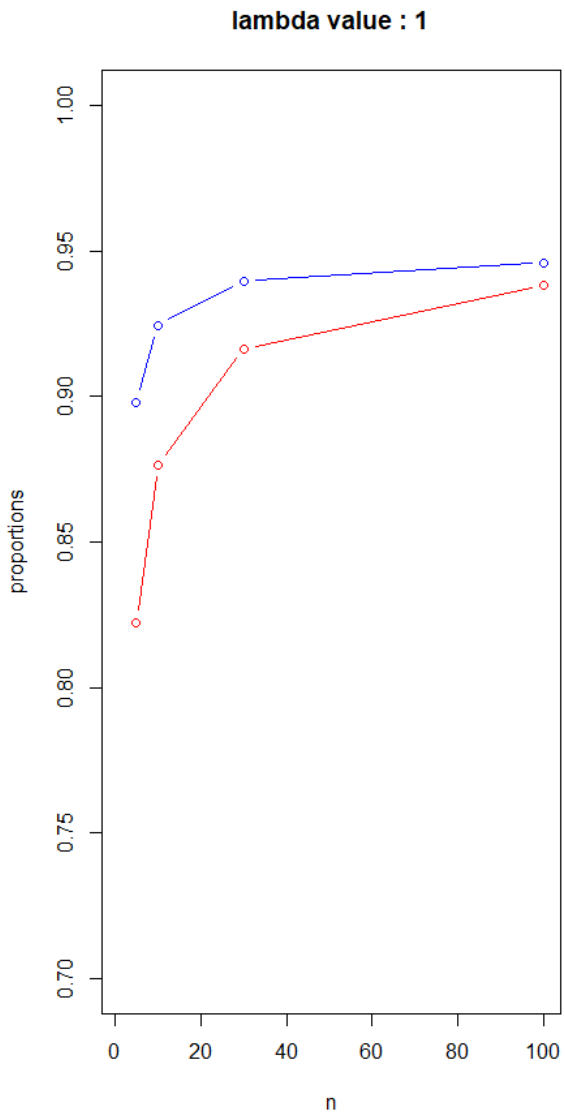
R Code:

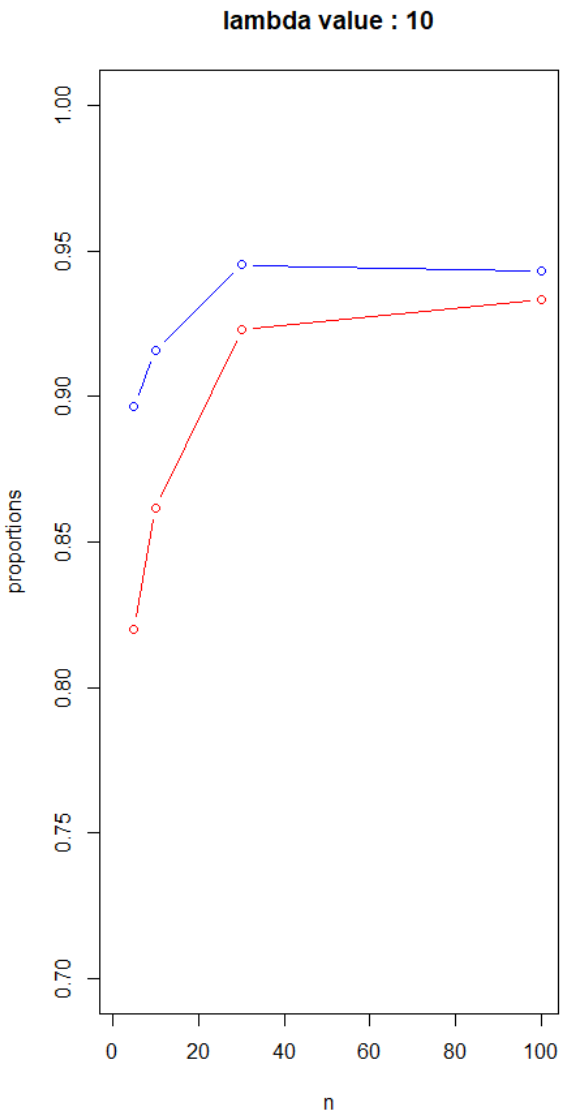
```
par(mfrow=c(1,2))
for (i in c(1,2,3,4)) {
  plot(c(5,10,30,100),z_matrix[i,],main = paste("lambda value :", lambda.vector[i]),xlab =
'n',ylab='proportions',type='b',col='red',xlim = c(1,100),ylim = c(0.7,1))
  lines(c(5,10,30,100),p_matrix[i,],col='blue',type='b')
}

for (i in c(1,2,3,4)) {
  plot(c(0.01,0.1,1,10),z_matrix[,i],main = paste("n value :", n.vector[i]),xlab =
'Lambda',ylab='proportions',type='b',col='red',xlim = c(0.01,10),ylim = c(0.7,1))
  lines(c(0.01,0.1,1,10),p_matrix[,i],col='blue',type='b')
}
```

Plots:

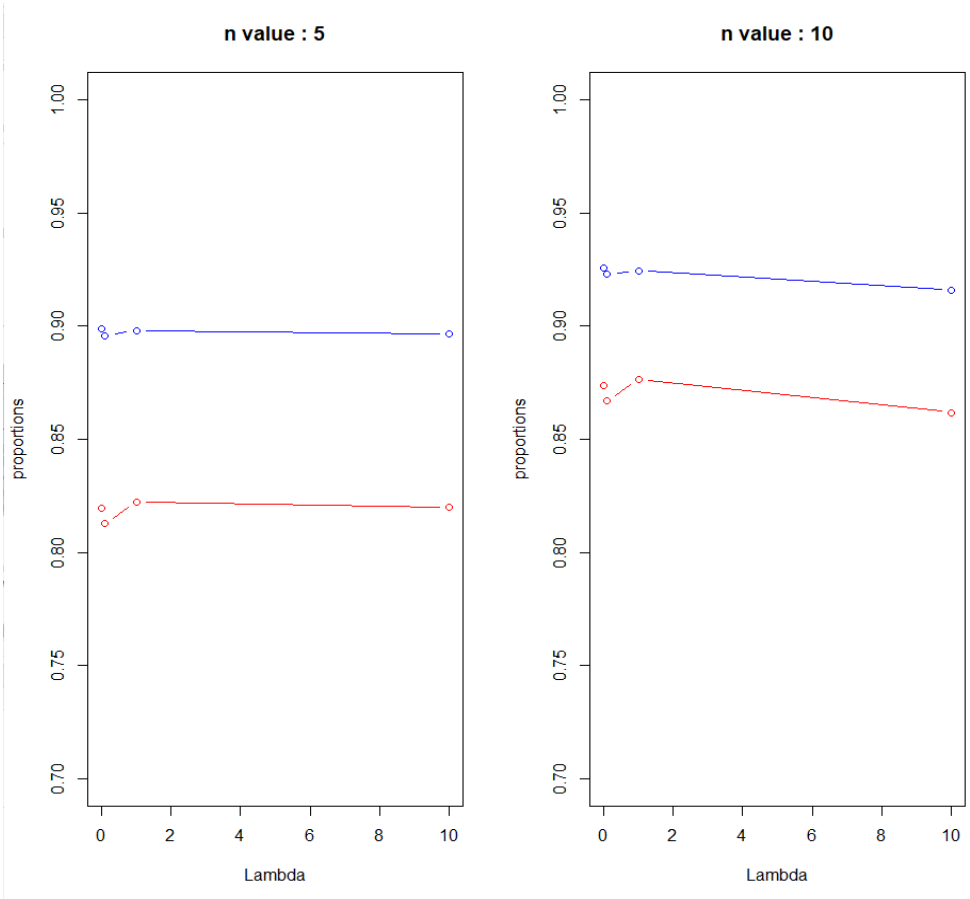
Keeping "λ" constant:

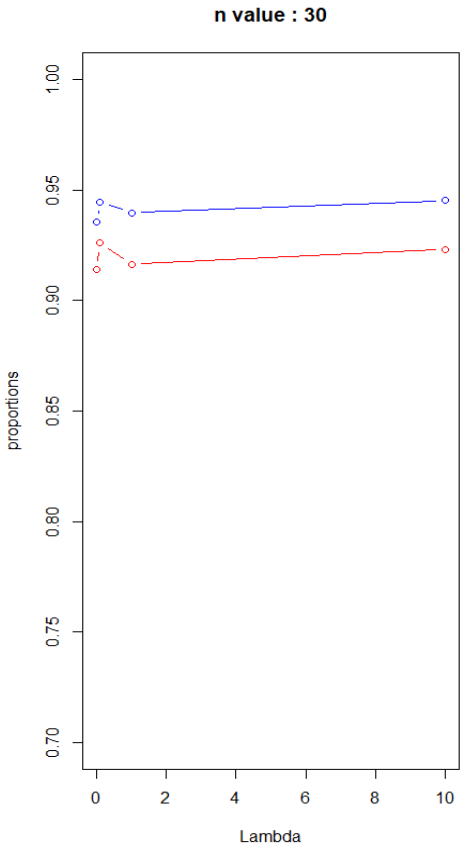


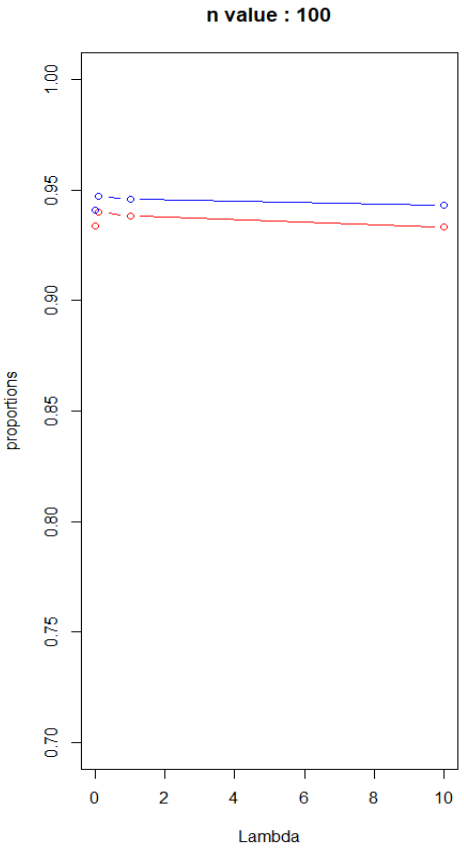

lambda value : 1
proportions
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
20
40
60
80
100
n


lambda value : 10
proportions
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
20
40
60
80
100
n

Keeping “n” constant:







Observations:

1. We drew 2 different types of graphs.
2. Initially we kept lambda constant and tried to plot the coverage probabilities based on the different values of n.
3. Secondly, we kept n as constant and plotted the probabilities for different values of lambda.
4. Based on type 1 graphs we can say that there is no significance difference based on different values of lambda. Also, from the table we observe the values of coverage probabilities for different values of lambda and keeping n constant. We can see no such significant change.
5. Hence, we can say that coverage probabilities do not depend on the values of lambda.
6. For large sample interval we can observe that the probability reaches close to 0.95 as n reaches 100. So, n should be 100 or above for large sample interval.
7. For percentile bootstrap interval we can observe that the probability reaches close to 0.95 as n is 30. So, n should be 30 or above for percentile bootstrap interval.
8. Based on the graphs and the results of coverage probabilities for different values of n we can say that percentile bootstrap performs better.
9. I would recommend percentile bootstrap as it performs good for small values of n. But if n value is sufficiently large. we can observe from the graph that lines are nearly same. The gap between lines start to reduce as value of n increases. So, when n is large, large sample interval provides similar results as percentile bootstrap with less computation.
10. Hence when n is sufficiently large it would be better to use large sample interval.

Q2d) Do your conclusions in (c) depend on the specific values of λ that were fixed in advance? Explain.

Observations:

1. The conclusion does not depend on any specific value of lambda.
2. This is observed from the type 1 graph where when we keep n constant and try to find probabilities based on different values of lambda, there is no significant change observed.
3. Also, it can be seen from the table that for large sample interval and parametric percentile bootstrapping ,the coverage probabilities do not change for different values of lambda when value of n is constant. Hence, we can conclude that the observations obtained in 2.c does not depend on any specific value of lambda.