**STATISTICAL METHODS FOR DATA SCIENCE CS 6313.001 FALL 2019**

**Mini Project #5**

**Participants :**

**Reetish Chand Guntakal Patil (RXG190006)**

# Question 1:

1. 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.

(b) Do males and females differ in mean heart rate? Answer this question by per-

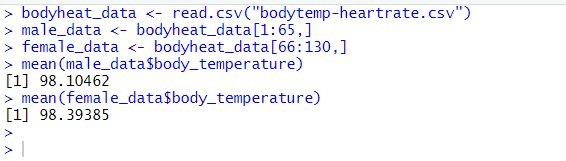
forming an appropriate analysis of the data, including an exploratory analysis.

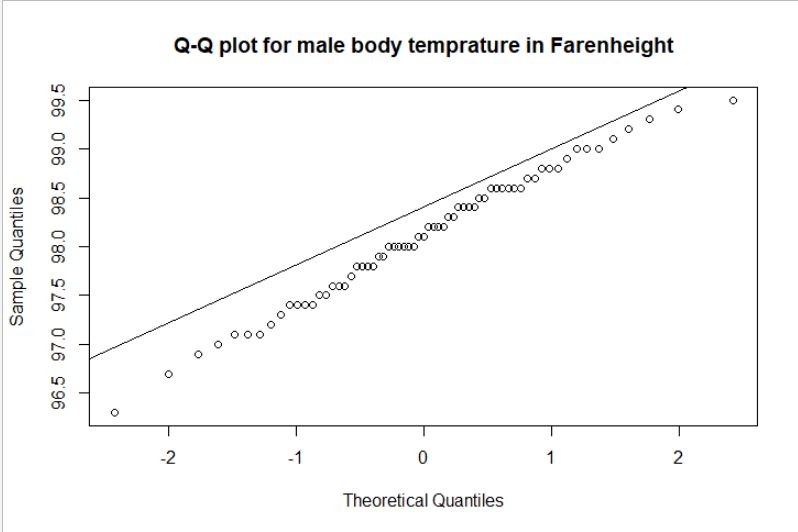
(c) Is there a linear relationship between body temperature and heart rate? Does

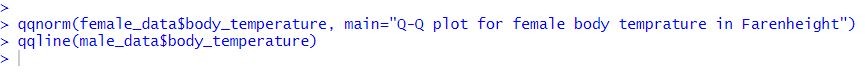
this relationship depend on gender? Answer these questions by performing an

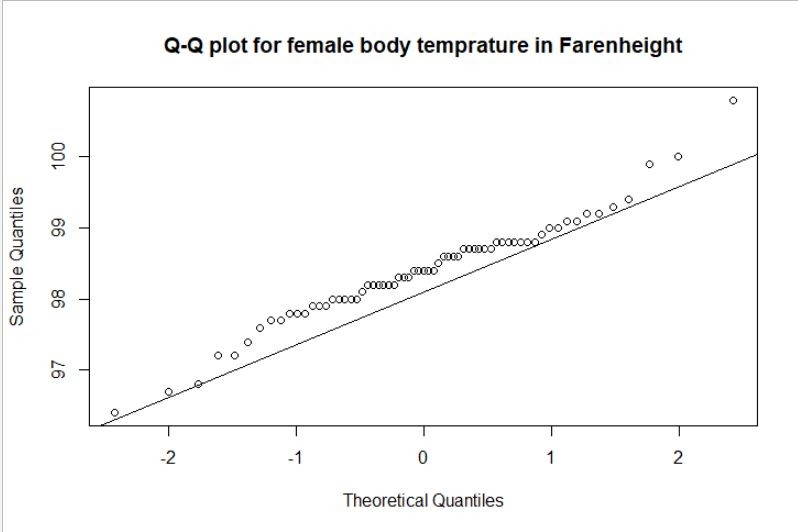
# appropriate analysis of the data, including an exploratory analysis.

# A.



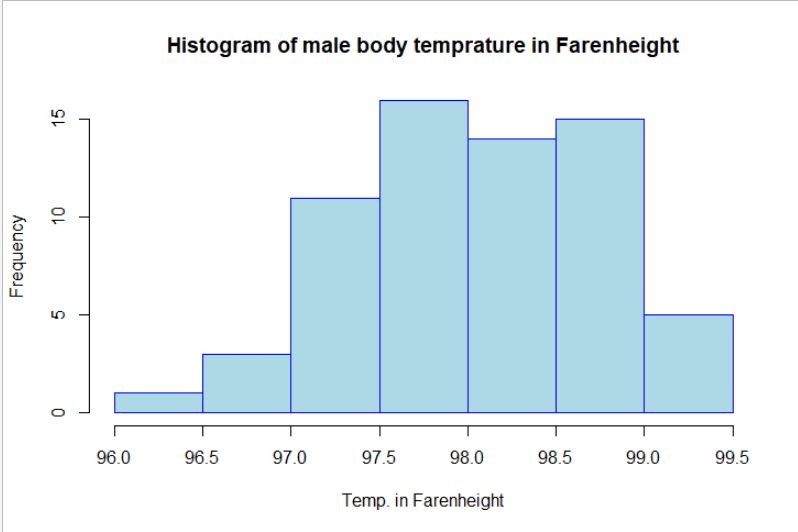


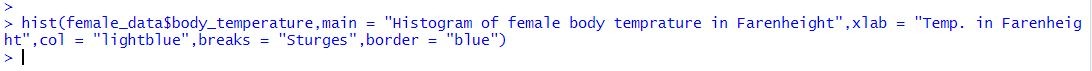


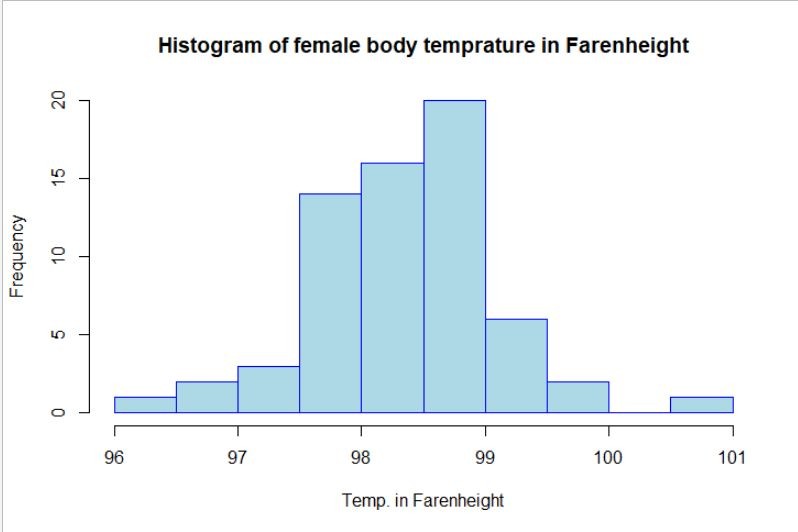


**Observation**:

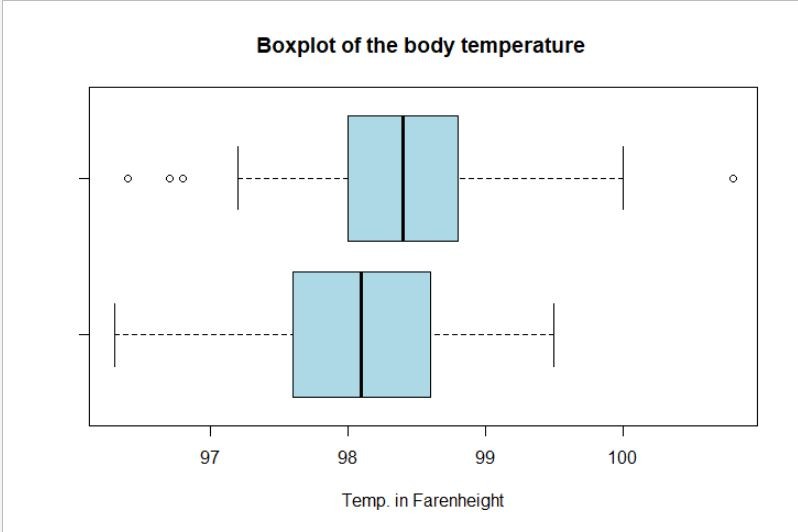
Both types of data have approximately normal distribution











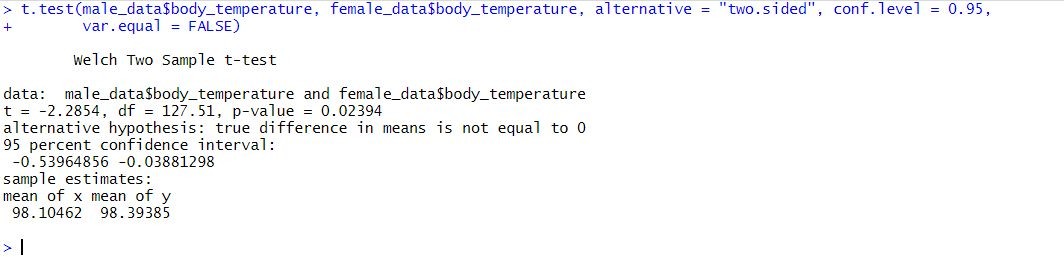
## Observation:

* The data comes from an approximately normal distribution
* The variance for both the group is not the same

Now performing the Hypothesis testing(Inferential Statistics) to validate the results:

**Null Hypothesis**: The difference in mean body temperature between the male and female group is zero.

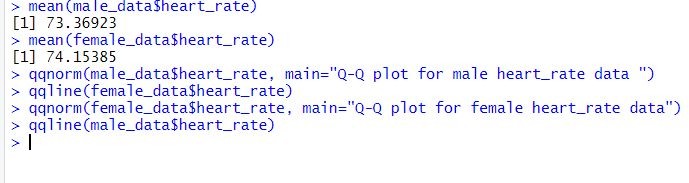
**Alternate Hypothesis** : The difference in mean body temperature between male and female group is not zero.

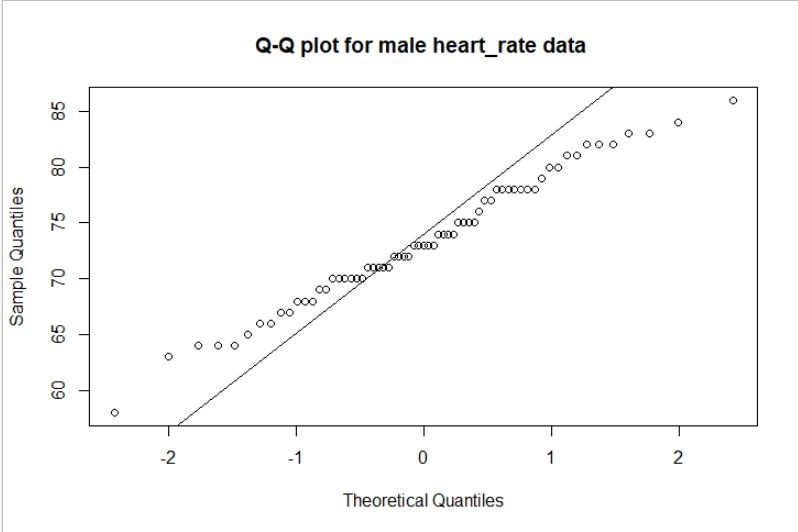


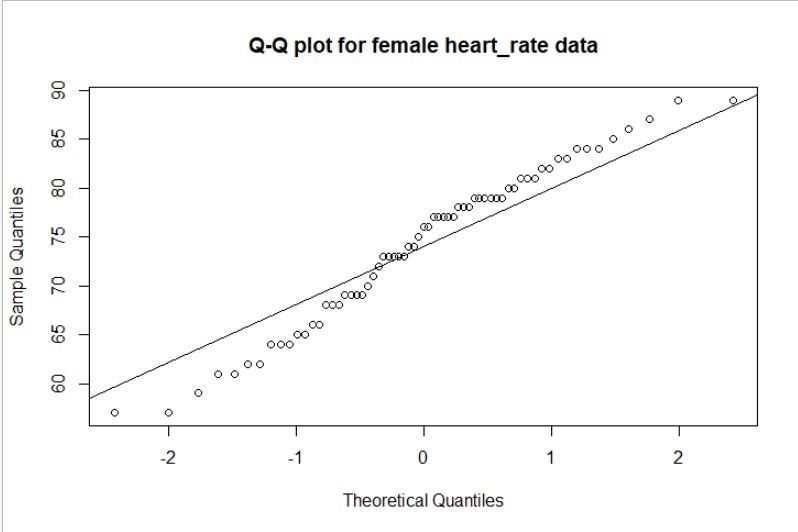
# Conclusion:

From the above result we can say that the C.I comes out to be [-0.53964856 ,-0.03881298]. As, zero does not falls between [-0.53964856 ,-0.03881298] we here infer that the Null Hypothesis is rejected that is the mean body temperature of male and the female group is not equal.

## B.







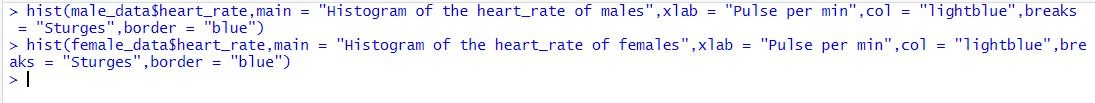
**Observation**:

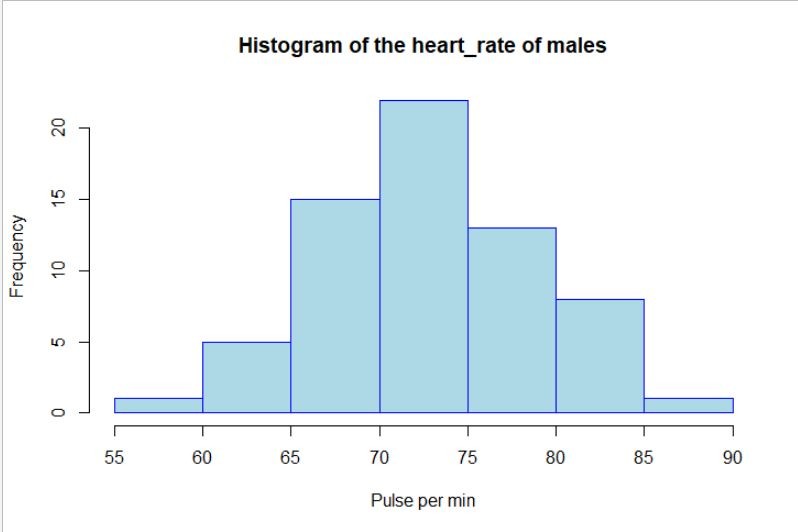
Both the data follow normal distribution

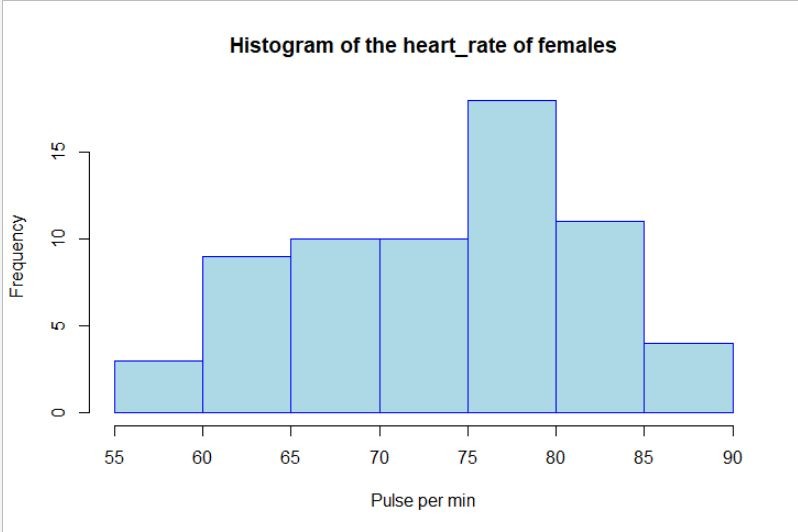


## Observation:

* The data comes from an approximately normal distribution
* The variance for both the group is not the same



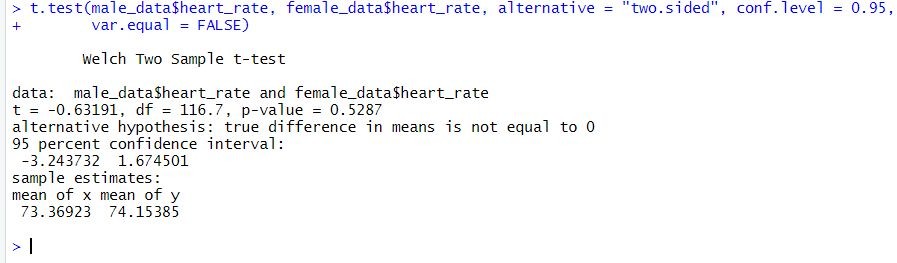




Now performing the Hypothesis testing(Inferential Statistics) to validate the results:

**Null Hypothesis**: The difference in mean heart\_rate between the male and female group is zero.

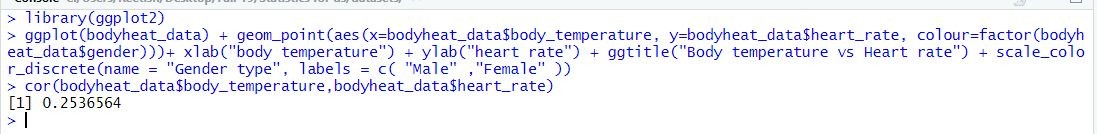
**Alternate Hypothesis** : The difference in mean heart\_rate between male and female group is not zero.

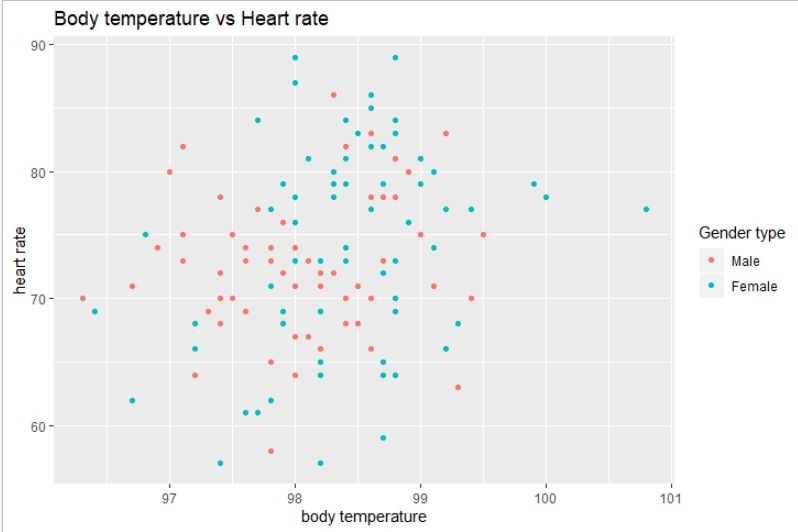


## Conclusion:

From the above result we can say that the C.I comes out to be [-3.243732, 1.674501]. As, zero does falls between [-3.243732, 1.674501] we here infer that the Null Hypothesis is accepted that is, the mean heart\_rate of male and the female group is equal.

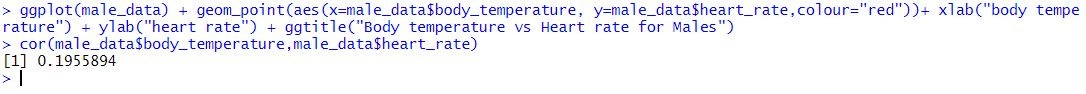
## C.

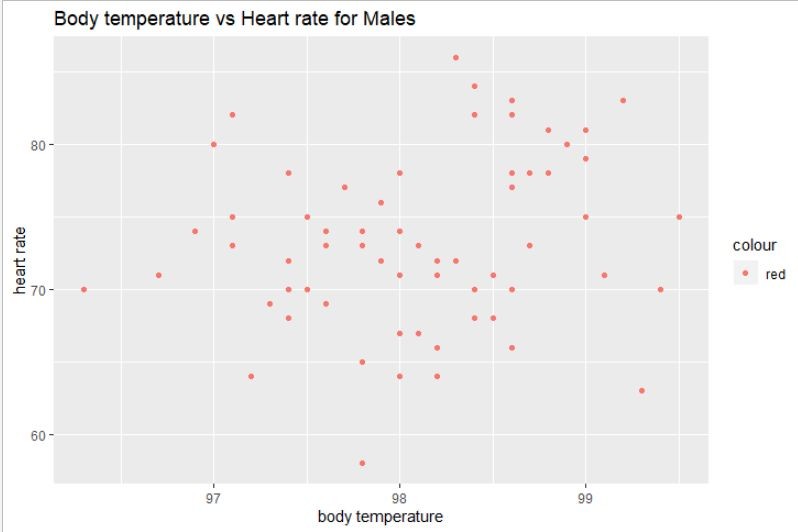




Observation:

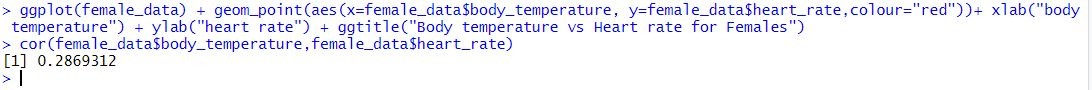
very poor linear relationship between body temperature and heart rate i.e. approximately 0.253

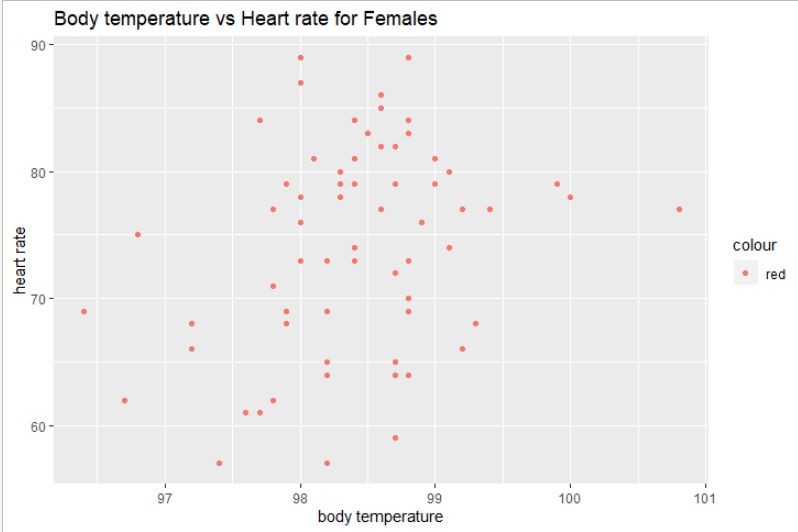




Observation:

very poor linear relationship between body temperature in males and heart rate in males i.e. approximately 0.196





## Observation:

very poor linear relationship between body temperature in females and heart rate in females i.e. approximately 0.288

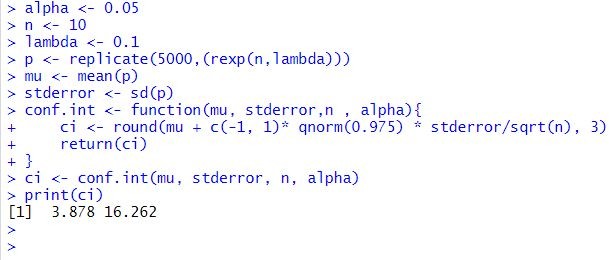
## Conclusion:

linear correlation of heart rate vs body temperature for females(0.288) is a little bit better in comparison to the linear correlation of heart rate vs body temperature for males (0.196)

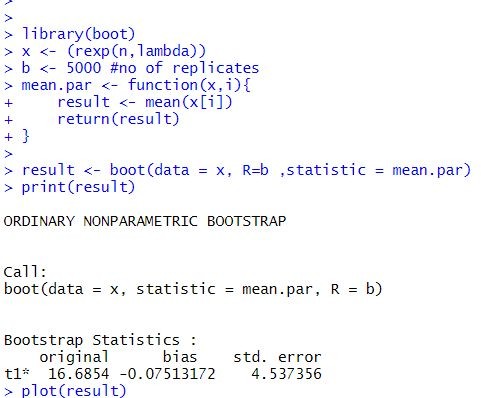
## Question 2:

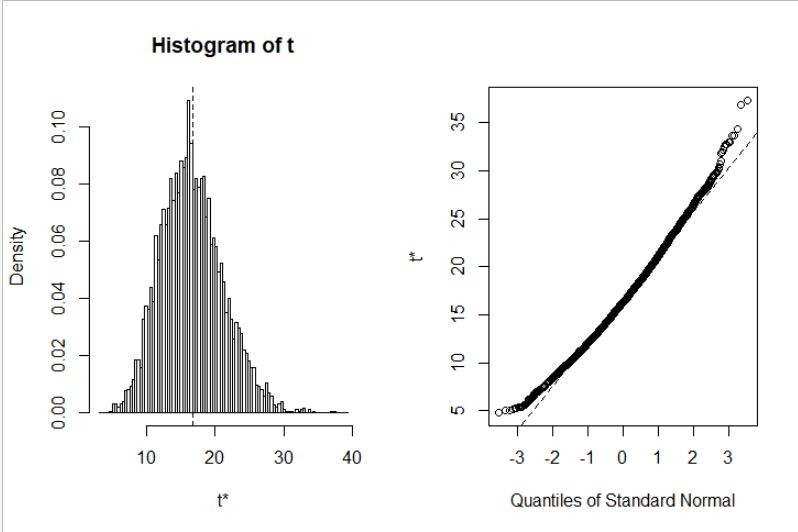
## 

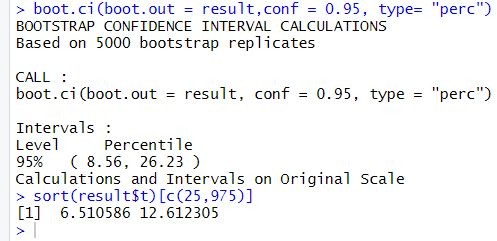
**A.**



The confidence interval for n=10 and lambda=0.1 comes out to be [3.771 ,16.241] . As the mean for exponential distribution is 1/lambda here it comes out to be 1/0.1 = 10 . Since 10, lies inside the C.I this test approves the approximation of C.I to be correct via large-sample interval.



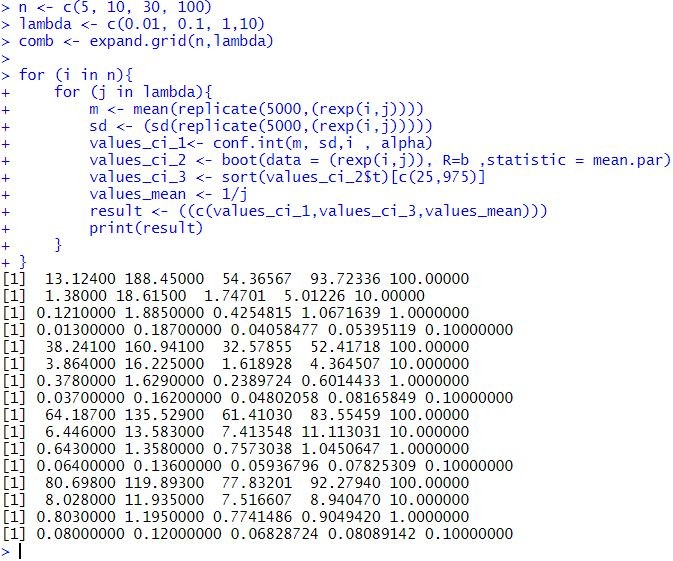




The confidence interval for n=10 and lambda=0.1 comes out to be ( 5.17, 16.47 ) . As the mean for exponential distribution is 1/lambda here it comes out to be 1/0.1 = 10 . Since 10, lies inside the C.I this test approves the approximation of C.I to be correct via the Percentile bootstrap interval.

The confidence interval for n=10 and lambda=0.1 comes out to be ( 3.935285, 8.002999 ) . As the mean for exponential distribution is 1/lambda here it comes out to be 1/0.1 = 10 . Since 10 does not lies inside the C.I this test does not approves the approximation of C.I to be correct one via the Percentile bootstrap interval.

## B.

Repeating (a) for remaining given n,lambda combinations:

## Observation:

every value of lambda and n satisfies the mean C.I for the large-sample interval

In case of the bootstrap interval only two cases n=30 and lambda=10,n=100 and lambda=0.01 satisfies the mean C.I as appropriate result/inference

## C.

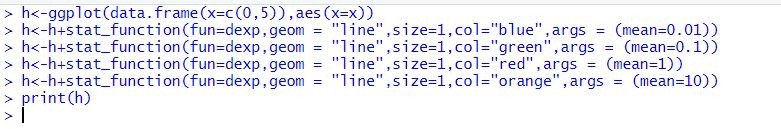
For different values of n and lambda from the C.I results we can see that:

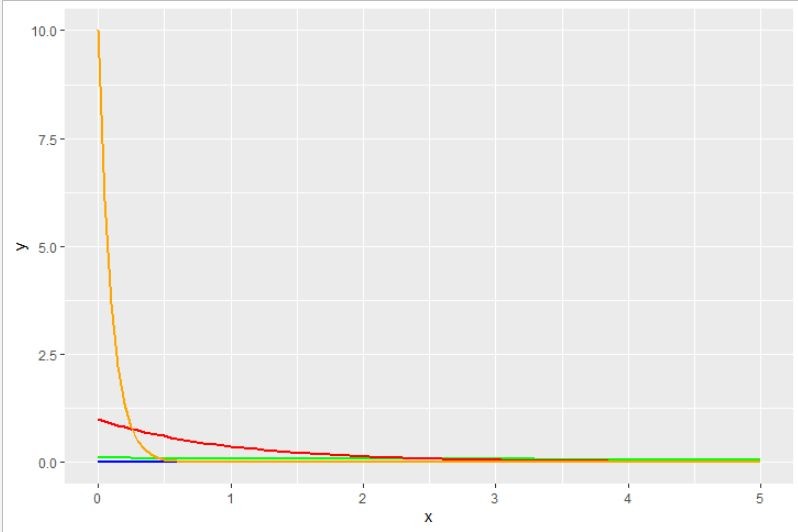
In case of the large-sample interval n=5 sufficies to be large enough to give appropriate outcome irrespective of the value of lambda.

In case of the bootstrap interval for n=30 and lambda=10 the C.I comes out to be [0.09995918, 0.13837950] for which mean is 0.1, as it falls within the C.I this is supposed to be the right approximation. Also, for n=100 and lambda=0.01 the C.I comes out to be [88.43387 ,104.33904] for which mean is 100,as it falls within the C.I so this is supposed to be the right approximation for C.I.

## D.

Yes, the conclusion in (c) depends on the specific values of lambda that were fixed in advance for the percentile bootstrap interval.





## Observation:

the value of lambda increases the initial value of the exponential density function between (0,1) increases.This doesn't make any difference for The large-sample intervals but makes a significant difference incase of The percentile bootstrap interval. As the second case, bootstraps the value of x from the distribution it can pick most of extreme values from the distribution that can lead to significant interval of C.I. So, for this case even n=5 could lead to a significant C.I for a particular value of lambda and not even 1000 can give you appropriate C.I.

Coming back to the original 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 answer this question:

1. The large-sample interval : Even n=5 sufficies to be large enough because this method takes 95% of the data distribution excluding the variational part(which is between 0 to 1).
2. The percentile bootstrap interval: Can’t tell any appropriate large enough value of n. n=5 could be large enough to give significant result and n=1000 couldn't be large enough to give significant inference.