# Mini Project 5

Names: Sathya Pooja Ramireddy(sxr176830)

Sashidhar Donthiri (sxd173730)

Contributions: Both of us have individually solved it, later discussed it and put it together.

Q1)

(a) Given heart rate for 65 male(gender = 1) and 65 female (gender = 2) subjects.

Initially lets consider the data being read into the dataframe "data" and convert male->1 and female->2

```
data= read.csv(file="C:/Users/shash/Documents/R for Stats/Mini Projects/Mini Project 5/data.csv")

str(data) #to Get dataframe info like no of observations and variables

# data.frame':   130 obs. of  3 variables:

# $ body_temperature: num  96.3 96.7 96.9 97 97.1 97.1 97.1 97.2 97.3 97.4 ...

# $ gender          : int  1 1 1 1 1 1 1 1 1 1 ...

# $ heart_rate      : int  70 71 74 80 73 75 82 64 69 70 ...


#Given gender as factor to consider and also male->1 , female->2

data$gender=as.factor(data$gender) #all observations of gender is saved in data$gender

data$gender

levels(data$gender)=c("male","female")

#assign names to factor data like male and female

#Seperated dataframes as data.m (1) to  male and data.f(2) to female

data.m=subset(data,gender=="male")

data.f=subset(data,gender=="female")


#Exploratory analysis of body temperature
```

Do males and females differ in mean body temperature?

Lets consider the analysis step by making box plots of body temperatures of male and female observations.

```
#boxplots and QQplots on body temperature based on gender

par(mfrow=c(2,2))

boxplot(body_temperature ~ gender, data=data, main="Boxplot using body_temp observations")

m.bt=data.m$body_temperature

f.bt=data.f$body_temperature

#QQplots of bodytemp based on gender

qqnorm(m.bt)

qqline(m.bt)

qqnorm(f.bt)

qqline(f.bt)
```

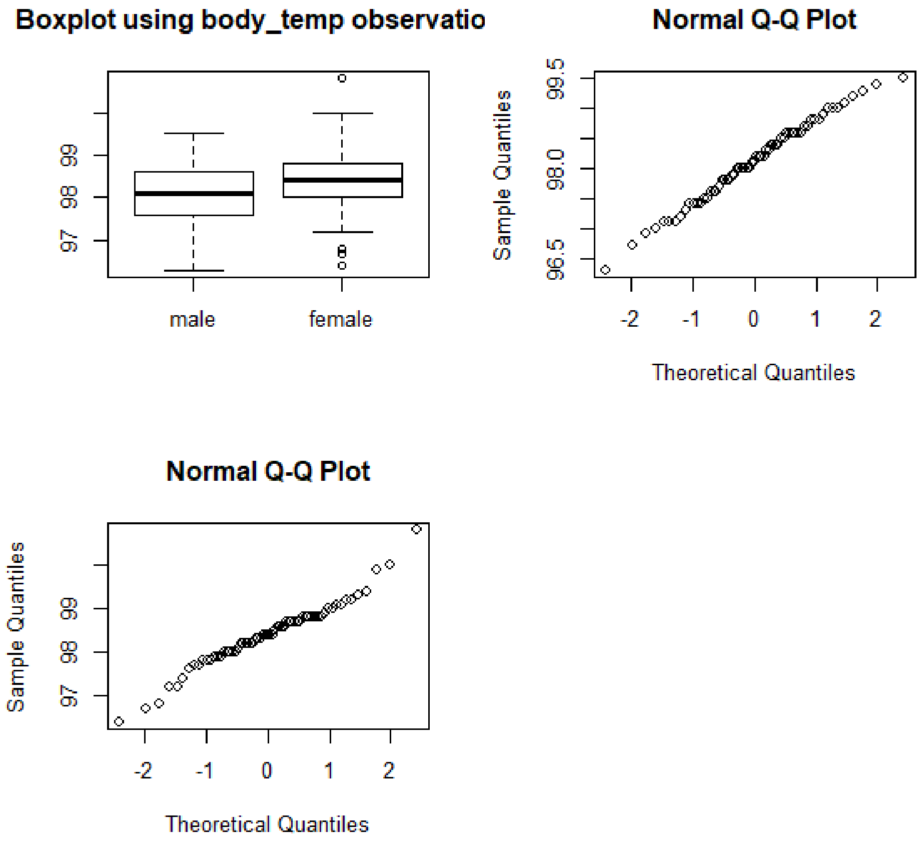


**Fig 1 :** Side-by-side boxplots of body temperatures for male and females and normal Q-Q plots for the same data for males(top right) and females (bottom left panel)

- From the above figure, we see that the three quartiles -Q1,median as highlighted, Q3 for females are comparably larger than that of males.
- This implies that the distribution for females may have slightly larger mean that that of the males.
- There are outliers in female boxplot that implies there is more variability compared to male.
- The QQ plot as shown in above figure shows that the distribution is almost normal when QQnorm is drawn that suggests that we can consider t-test.
- We may not consider equal variances assumption.
  Let us perform null hypothesis on $H_0$: $\mu_{male} = \mu_{female}$ with an alternate hypothesis of $H_1$ : $\mu_{male} \neq \mu_{female.}$ From above QQ normal plot we can consider that the distribution is normal.
- The appropriate test for the above distribution is t-test with Scatterthwaite's approximation.

```
#Welch two-sample test
t.test(m.bt,f.bt,alternative="two.sided",var.equal = F)
#Apply Welch two-sample test
Welch Two Sample t-test
data:  m.bt and f.bt
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 above test has given us with p-value= 0.02394, that implies null hypothesis can be rejected since p-value < 0.05 .
- From this we can conclude that there is statistically significant difference in the mean body temperature for males and females. Alternate hypothesis is accepted.
  The 95% confidence Interval for $\mu_{male} - \mu_{female}$ is [-0.53964856 -0.03881298], that implies mean temperature of females is more than that of males very slightly, by an amount in the range of 0.04 and 0.54 F.

(b) #Exploratory analysis of heart rate
Do males and females di_er in mean heart rate?

Lets consider the analysis step by making box plots of heart rate of male and female observations.

```
#boxplots and QQplots on body temperature based on gender
par(mfrow=c(2,2))
boxplot(heart_rate~gender,data=data, main="Boxplot using heart_rate observations")
m.rate=data.m$heart_rate
f.rate=data.f$heart_rate
#QQplots of heart_rate based on gender
qqnorm(m.rate)
qqline(m.rate)
```

```
qqnorm(f.rate)
qqline(f.rate)
```

- Exploratory analysis of heart rate of males and females is similar to that of body temperature in the above problem.

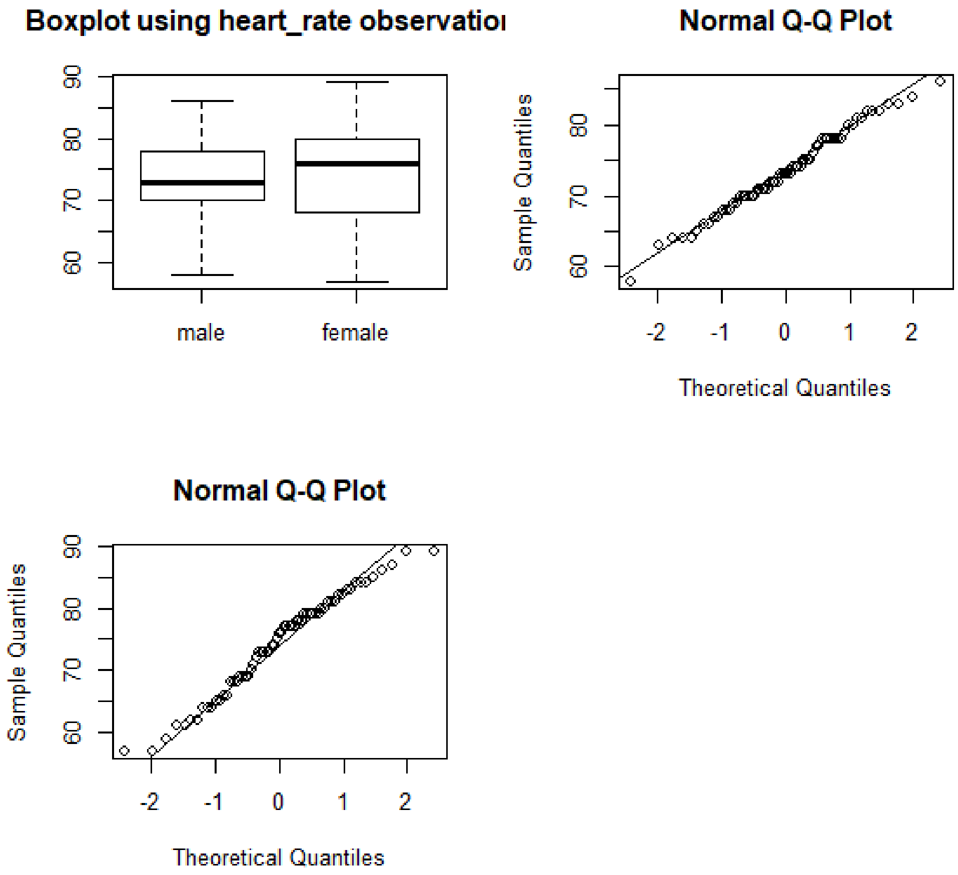


**Fig 2 :** Side-by-side boxplots of heart rate for male and females and normal Q-Q plots for the same data for males(top right) and females (bottom left panel)

- The above boxplot explains, consider median and  Q2 for females are larger than that of males but the Q1 is smaller for females to that of males.
- So we cannot conclude the mean comparisons as we have concluded to that of males above.
- The females are stretched longer than males, and females seem to have more variablitiy than that of males.
- The Normal QQ-plots for the data as shown above, show that the assumption of normality may be appropriate.
- As above, let us perform t-test with Scatterthwaite's approximation in order to test the null hypothesis like:
    - $H_0$: $\mu_{male}$= $\mu_{female}$ against the alternate hypothesis, $H_1$:  $\mu_{male} \neq \mu_{female}$.
    - #Apply welch twosided sampletest

```
welchtest=t.test(m.rate,f.rate,alternative="two.sided", var.equal = F)

welchtest


Welch Two Sample t-test

data:  m.rate and f.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
```

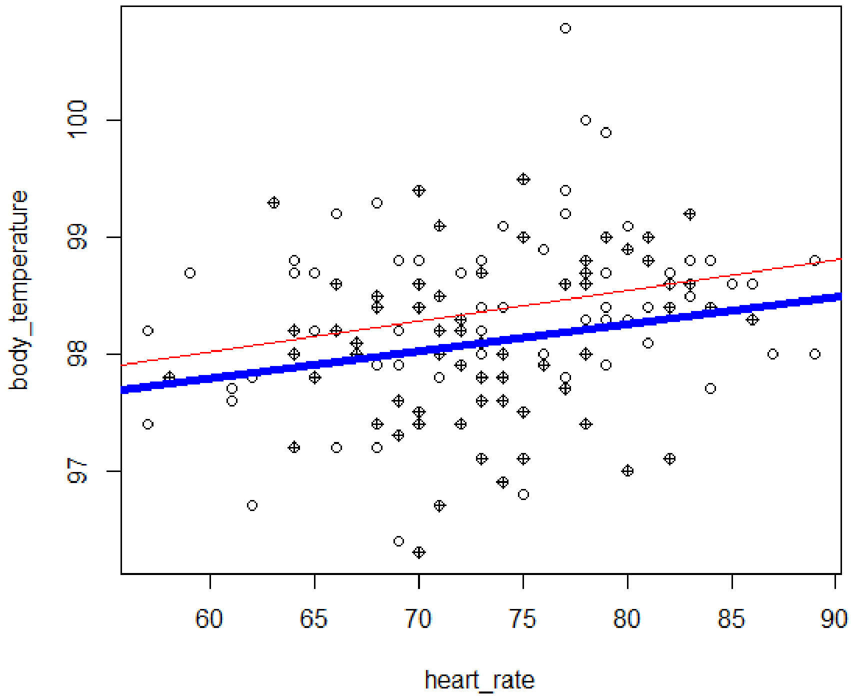
- The test gives p-value=0.528. Thus, we can accept null hypothesis, that indicates there Is statistically significant difference in mean heart rates of males and females.
- The 95% Confidence interval for difference in means $\mu_{male}$ - $\mu_{female}$ is [-3.24,1.67] that implies the difference in two means is estimated to fall in range of -3.24 and 1.67 number of beats per minute.
- But the difference is not statistically significant

(c) Linear Relationship between body temperature and heart rate:

Let us draw and consider a scatter plot and further plot a regression line that reflects the linear relationship between them and also if its dependent on gender.

- In Scatter plot- body temperature against heart rate using different types of points- “*” for male and “o”for female.
- From the scatter plot below, with regression line, the relationship between two variables appear linear but weak one.

```
#Linear relationship between body temperature and heart rate
#scatter plot that plots datapoints of heartrate vs body_temp where
#red corresponds to female and blue to male
par(mfrow=c(1,1))
plot(body_temperature~heart_rate, data=data.m,pch=10,
ylim=range(data$body_temperature),xlim=range(data$heart_rate))
points(body_temperature~heart_rate, data=data.f,pch=1)
```



```
#males
#correlation between body_temp and heart_rate of males
cor(data.m$body_temperature,data.m$heart_rate)
# [1] 0.1955894
fit.m=lm(body_temperature~heart_rate, data=data.m)
fit.m

#reuslt
Call:
  lm(formula = body_temperature ~ heart_rate, data = data.m)
Coefficients:
  (Intercept)   heart_rate
96.39789      0.02326
#females
#correlation between body_temp and heart_rate of females
cor(data.f$body_temperature,data.f$heart_rate)
#[1] 0.2869312

fit.f=lm(data.f$body_temperature~heart_rate,data=data.f)
fit.f
```

```
Call:
  lm(formula = data.f$body_temperature ~ heart_rate, data = data.f)

Coefficients:
  (Intercept)   heart_rate
96.44211      0.02632

abline(fit.m,lwd=5) #thick line for male
abline(fit.f,lwd=1) #thinner one for female
```

- The sample correlations observed for male and females are 0.20 and 0.29. The larger the value of correlation metric the stronger is the relationship.
- Both the correlation metric values are small and weak. This shows that the relationship is slightly weaker for males compared to females.
- Although the Regression line fit, that assumes that the heart rate is measured without error, which may not be true.
- Also the estimated intercept and slope for females are slightly larger than that of males.
  ($\hat{\beta}_{0,female}$ , $\hat{\beta}_{1,female}$ ) = (96.44,0.026)
  ($\hat{\beta}_{0,male}$ $\hat{\beta}_{1,male}$ ) = (96.39,0.023)

Q2)

(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.

The estimate for coverage probabilities for Z0test and Bootstrap(n=5, lambda=0.01) methods is given by

**Z-test coverage probabilities**
N=5
Lambda=0.01
CP=0.8164

**Boot-Strap Coverage Probabilities**
N=5
Lambda=0.01
CP=0.787

(b) As mentioned, repeat the above step (a), for remaining combinations of (n,lambda).

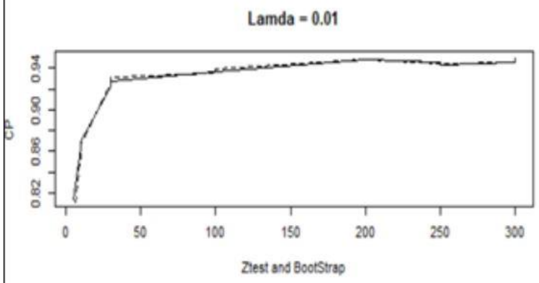
Results:

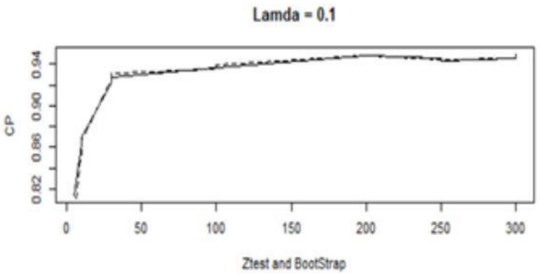
The estimate for coverage probabilities for Z-test and BootStrap for different combinations of n and lambda are given by:

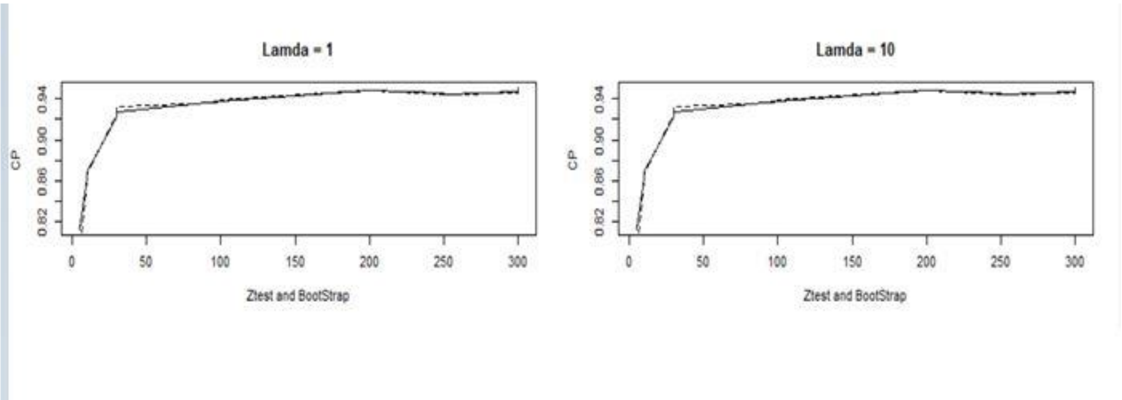
**Z-Test coverage Probabilities:** **Bootstrap Coverage Probabilities:**

| n | Lambda | CP_z-test | CP_bootstrap |
|---|---|---|---|
| 5 | 0.01 | 0.8162 | 0.7896 |
| 5 | 0.1 | 0.8126 | 0.7899 |
| 5 | 1 | 0.8126 | 0.7899 |
| 5 | 10 | 0.8126 | 0.7899 |
| 10 | 0.01 | 0.8632 | 0.8571 |
| 10 | 0.1 | 0.8694 | 0.8662 |
| 10 | 1 | 0.8694 | 0.8662 |
| 10 | 10 | 0.8694 | 0.8662 |
| 30 | 0.01 | 0.9221 | 0.9258 |
| 30 | 0.1 | 0.9276 | 0.9316 |
| 30 | 1 | 0.9276 | 0.9316 |
| 30 | 10 | 0.9276 | 0.9316 |
| 100 | 0.01 | 0.937 | 0.9356 |
| 100 | 0.1 | 0.9372 | 0.9392 |
| 100 | 1 | 0.9372 | 0.9392 |
| 100 | 10 | 0.9372 | 0.9392 |
| 300 | 0.01 | 0.9462 | 0.9454 |
| 300 | 0.1 | 0.9502 | 0.9508 |
| 300 | 1 | 0.9502 | 0.9508 |
| 300 | 10 | 0.9502 | 0.9508 |

Now, let us consider trying to plot the values on n against the obtained coverage probabilities for Z-test and BootStrap simulation of Cis where in the plot, Z-test(**Solid Line**) and Dotted Line(*BootStrap*)







(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.

Now from the obtained results we can infer that:

- From the above table we can say that at Low values of n(5,10) the Coverage probabilities of Z-test seemed to be a bit higher than that of Bootstrap.
- As the value of n increased, from n=30 the Bootstrap method seems to have high Coverage Probabilities compared to Z-test
- Overall, it seems like the Coverage Probabilities of Bootstrap are slightly higher than that of Z-test.
- As n increases, the Coverage Probabiliites for both Z-test and BootStrap got closer to 0.95.
- When observed Coverage Probability of both BootStrap and Z-test were almost equal 0.95 and very slightly higher than 0.95 at n=300. But BootStrap had slightly higher Coverage Probabilities than Z-test for each value of n >=30.
- This clearly implies that the Coverage Probability is more accurate for larger values of n for both the tests.
- From above results in table, we found that the coverage probabilities were the same for some values of lambda and again when we looked at it overall for each value on n, they were not significantly different from one another.
- At the same time for smaller n values, especially with this problem statement and simulation, BootStrap for n=30,100 is better where as for larger n values, both Z-test and BootStrap have almost similar values for Coverage Probabilities.
- BootStrap is a better choice for given values of n

(d) From the obtained plots in (b) which has the coverage probabilities for BootStrap and Z-test plotted against different values of n(5,10,30,100,300) with fixed Lambda values(0.01, 0.1, 1, 10), the observations/results for each value of Lambda seemed to be consistent.

## R-Code for Problem 2:

```
#Import/Add bootstrap library
library(boot)
#replication val=5000
m=5000
#sample sizes and Lambda vector declarations
N=c(5,10,30,50)
Lambda=c(0.01,0.1,1,10)

#consider matrix for storing info of CI and means from Z-test and BootStrap
ztestMatrix=matrix(0,length(N)*length(Lambda),3)
bsMatrix=matrix(0,length(N)*length(Lambda),3)
#variable declaration to find whether its true or not for Z-test and
#Boot CI coverage probabilities
ll=numeric(m)
ul=numeric(m)
bsll=numeric(m)
bsul=numeric(m)

#Data replication
replicatefun=function(obs,lambda, rep_val=m){
  rep_data=replicate(rep_val,rexp(n=obs,rate=lambda))
  return(rep_data)
}


#generate BootStrap mean values
```

```
BS_mean.npar=function(x,indices){
  res=mean(x[indices])
  return(res)
}


#Declare a variable that has to be used to store Final Coverage probability values in the vector
fcp=1
for (i in 1:length(N))
{
  for (j in 1:length(Lambda))
  {
    #matrix to store CI and Coverage (true or false) for Ztest and Boot
    x_mat <- matrix(NA, nrow = m, ncol = 3)
    colnames(x_mat) <- c("lcl", "ucl", "CP")
    y_mat <- matrix(NA, nrow = m, ncol = 3)
    colnames(y_mat) <- c("lcl", "ucl", "CP")
    #to find the true mean value
    meanval <- 1/Lambda[j]
    #Replicated data
    data <- replicatefun(obs=N[i],lambda=Lambda[j],rep_Val=m)
    #store the resulting vector of means from sampled data for Z-test
    xbar_mean <- apply(data, 2, meanval)
    #store the resulting vector of Standard Deviations from sampled data for Z-test
    stdev <- apply(data, 2, sd)
    #now let us compute Z-test CI for 5000 replications for eachvalue of N
    for (a in 1:dim(data)[2])
    {
      ll[a] <- xbar_mean[a]-(1.96*(stdev[a]/sqrt(N[i])))
      ul[a] <- xbar_mean[a]+(1.96*(stdev[a]/sqrt(N[i])))

    }
    #Assigning CI and Coverage (True or False) for Ztest Test
    x_mat[, 1] <- ll
    x_mat[, 2] <- ul
    x_mat[, 3] <- (x_mat[, 1] < meanval) & (x_mat[, 2] > meanval)
    #Final Matrix for Ztest that has Coverage Probabilities for each value of N and Lambda
    ztestMatrix[k,1]=N[i]
    ztestMatrix[k,2]=Lambda[j]
    ztestMatrix[k,3]=mean(x_mat[, 3])
    #Computing CI using Bootstrap sampling on the Replicated dataset data
    for (b in 1:dim(data)[2])
    {
      set.seed(1234)
      BootSampleMean.npar.boot <- boot(data[,b],BootSampleMean.npar, R=999, sim="ordinary", stype="i")
      CI <- boot.ci(BootSampleMean.npar.boot,type="perc")
      #print(str(CI))
      bsll[b] <- CI$percent[1,4]
      bsul[b] <- CI$percent[1,5]

    }
    #Assigning CI and Coverage (True or False) for BootStrap Test
    y_mat[, 1] <- bsll
    y_mat[, 2] <- bsul
    y_mat[, 3] <- (y_mat[, 1] < meanval) & (y_mat[, 2] > meanval)
    #Final Matrix for Bootstrap test that has Coverage Probabilities for each value of N and Lambda
    bsMatrix[k,1]=N[i]
    bsMatrix[k,2]=Lambda[j]
    bsMatrix[k,3]=mean(y_mat[, 3])
    #result= print(bsMatrix)
   fcp=fcp+1
  }
}


#Finally to show the Matrix of Coverage Probabilities for Z-test and BootStrap test
ztestMatrix
bsMatrix
#Plotting the values of Coverage Probabilities Ztest and Bootstrap against n for each values of Lamda
par(mfrow = c(1, 2))
for (i in 1:length(Lambda))
{
  plot(ztestMatrix[,1], ztestMatrix[,3], type = "l", lty = 1,main = paste0("Lambda = ", Lambda[i]), ylab =
"CP")
  lines(bsMatrix[,1], bsMatrix[,3], lty = 2)
}
#With a X-Axis label
par(mfrow = c(1, 2))
for (i in 1:length(Lamda))
{
  plot(ztestMatrix[,1], ztestMatrix[,3], type = "l", lty = 1,main = paste0("Lambda = ", Lambda[i]), ylab =
"CP",xlab="Ztest and BootStrap")
  lines(bsMatrix[,1], bsMatrix[,3], lty = 2)
}
```