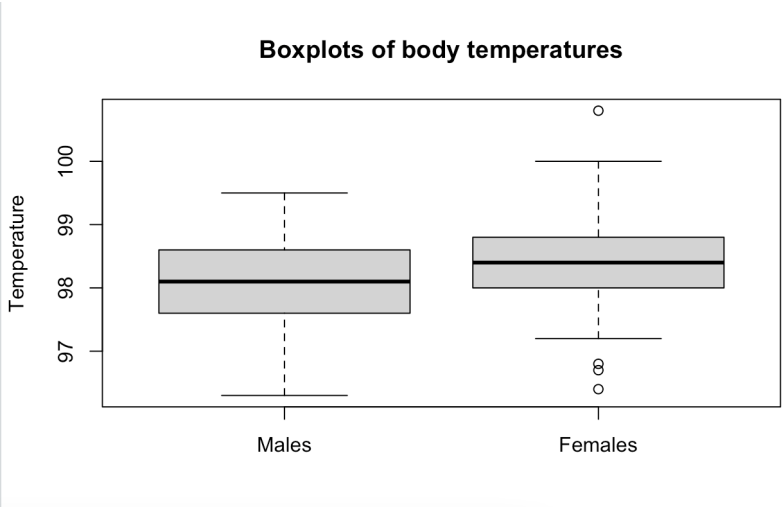
Mini Project : 5

Members :
1. Puru Jaiswal (NetID : PXJ200018)
2. Sara Tabassi (NetID : SXT200083)

Both the team members worked together to finish the project. Collaborated to learn R and then write the code. Both of us worked on each of the problems independently and then reviewed each other's answers to determine which solution was optimal.
Both members worked efficiently to complete the project requirements.

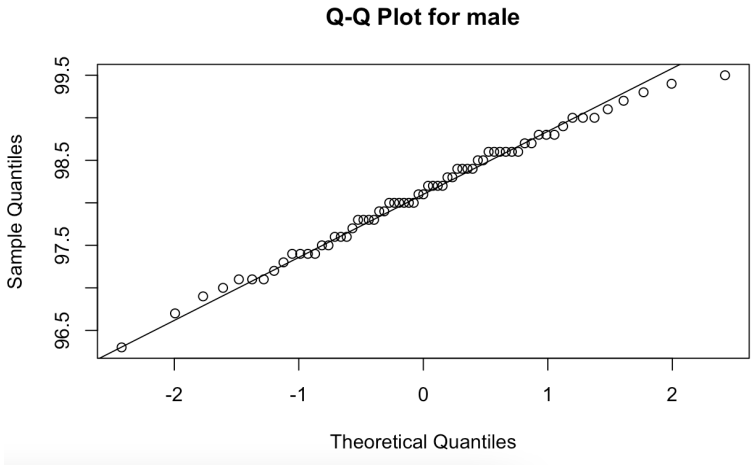
1. Consider the data stored in bodytemp-heartrate.csv on eLearning, containing measurements of body temperature and heart rate for 65 male (gender = 1) and 65 female (gender = 2) subjects.

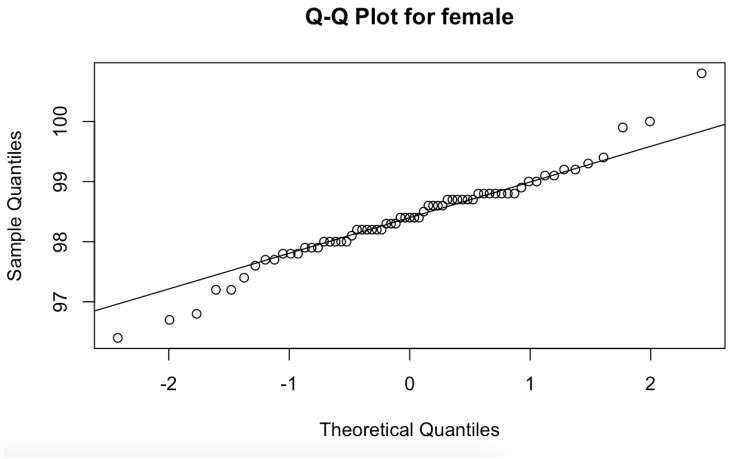
(a) Do males and females differ in mean body temperature? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.



Q1, Q2(Median), Q3 are higher for females than males, so females have slightly higher mean values than males. There exists outliers in the female box plot more than male box plot hence we cannot assume equal variances.

QQ Plot





Looking at the QQplots for both males and females we can say that the heart rate distribution is approximately normal.

M -> body temperatures for males, sample mean $\overline{m}$ estimates the population mean μ_m

F -> body temperatures for females, sample mean $\overline{f}$ estimates the population mean μ_f

Null hypothesis H0 : difference between means $\overline{m} - \overline{f} = 0$

Alternative hypothesis H1: $\overline{m} - \overline{f} \neq 0$

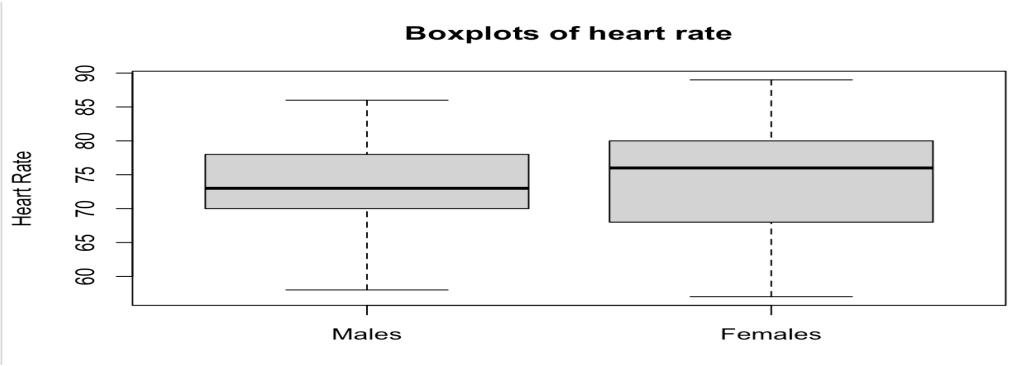
Samples are independent samples with unequal variances coming from approximately normal distributions, therefore we can use t-distribution with Satterthwaite approximation to get CI.

Confidence Interval : -0.53964856 -0.03881298

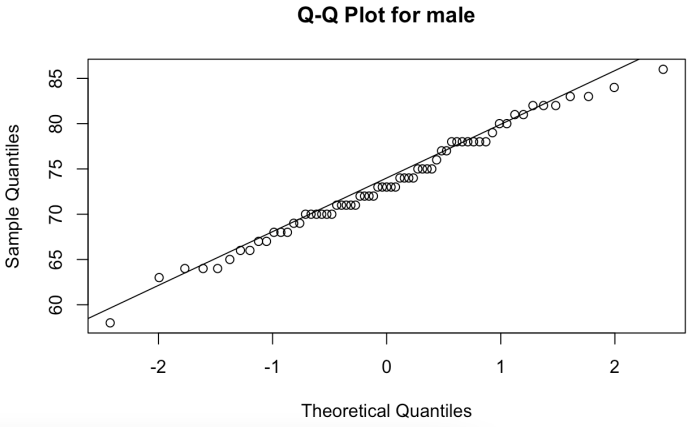
p-value : 0.02394

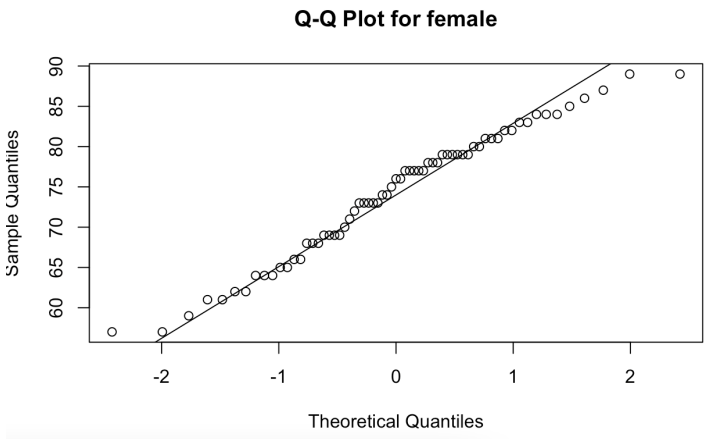
Since p value < 0.05 and 0 does not exist in the CI, we can say reject the null hypothesis and conclude that the means are not equal.

(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.



As the range is more in female boxplot as compared to the male boxplot, therefore variability is more in female.





Based on the QQ plots we can say that the distribution of both male and female heart rates are approximately normal.

M -> heart rates for males, sample mean $\overline{m}$ estimates the population mean μ_m

F -> heart rates for females, sample mean $\overline{f}$ estimates the population mean μ_f

Null hypothesis H0 : difference between means $\overline{m} - \overline{f} = 0$

Alternative hypothesis H1: $\overline{m} - \overline{f} \neq 0$

Samples are independent samples with unequal variances coming from approximately normal distributions, therefore we can use t-distribution with Satterthwaite approximation to get CI.

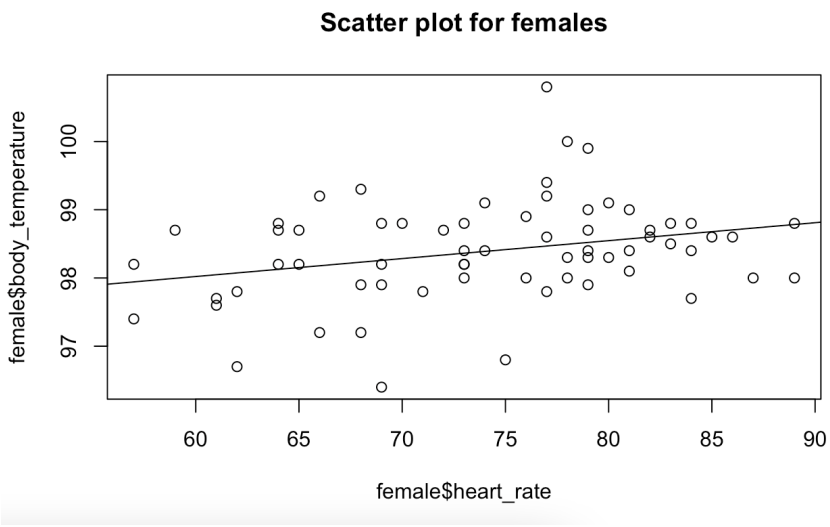
Confidence Interval : -3.243732 1.674501

p-value : 0.5287

p value is greater than 0.05 and the value 0 lies between CI, therefore we accept the null hypothesis, and we can conclude that heart rate means of both male and female 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.

Scatter plot for males



Scatter plot for females

Based on the graph, the slope is greater than zero. Therefore there is a positive correlation between body temperature and heart rate values. The strength is weak.

```
cor(male$body_temperature,male$heart_rate)

cor(female$body_temperature,female$heart_rate)
```

Males -> Correlation between body temperature and heart rate : 0.1955894
Females -> Correlation between body temperature and heart rate : 0.2869312

The relationship between heart rate and body temperature is weak as the values are smaller, comparing males and females we can conclude that correlation values for females is higher than of the males, so the correlation between body temperatures and heart rate is stronger among females in comparison with males.

2.

The goal of this exercise is 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 $X_1, \ldots, X_n$ represents 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.

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.**

   Using the data from below code we get the coverage probabilities as :
   Z-interval: 0.804
   Bootstrap interval: 0.892

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

Z - proportions	L = 0.01	L=0.1	L=1	L=10
N=5	0.804	0.818	0.8108	0.8052
N=10	0.8712	0.8666	0.8702	0.8758
N=30	0.9154	0.9166	0.9186	0.9158
N=100	0.9412	0.937	0.9392	0.9402

B - proportions	L = 0.01	L=0.1	L=1	L=10
N=5	0.892	0.8996	0.899	0.8986
N=10	0.9234	0.926	0.919	0.9256
N=30	0.9428	0.9376	0.9392	0.9374
N=100	0.9526	0.9428	0.9458	0.9434

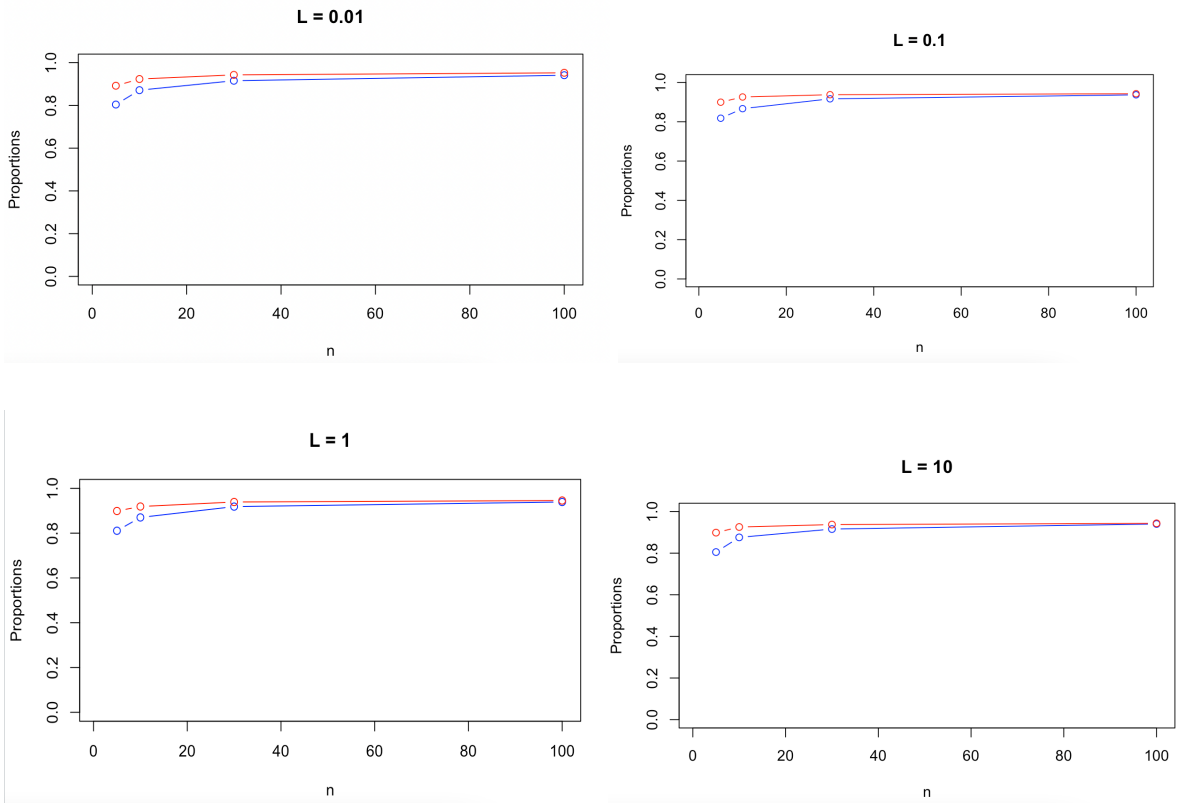
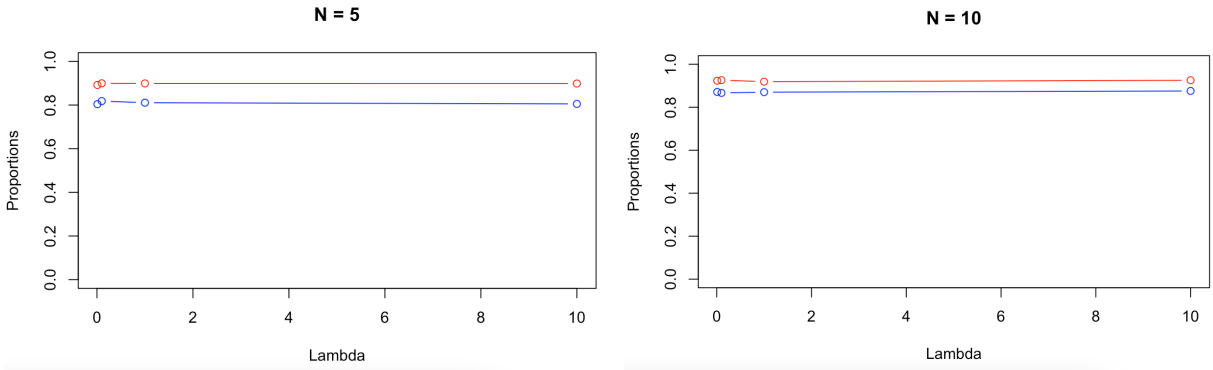


Figure 1

For the above graph Blue represents the z-proportions and Red represents the bootstrap proportions.



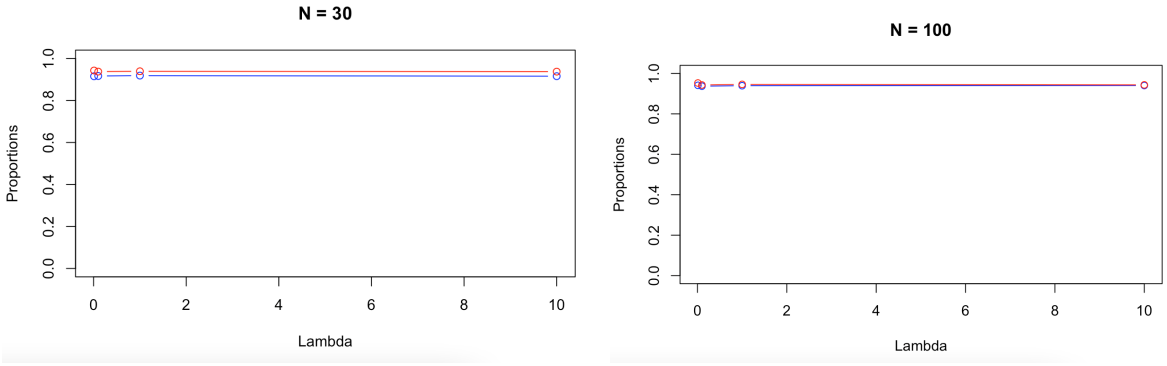


Figure 2

In the above figure blue represents the z proportions and red represents the bootstrap proportions.

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.**

   In Figure 1, as λ changes the graph doesn't change therefore we can conclude that the coverage probabilities do not depend on λ.
   In Figure 2 , we can say that the coverage probabilities depend on n.
   For bootstrap, n = 30. For Large sample, n = 100. As the data show, as n gets bigger the significance level gets closer and closer to alpha (1-.95). Bootstrap is generally more accurate than large samples, especially considering how it converges to .95 much faster than large sample.Therefore, we would recommend bootstrap, because it works much better in bootstrap, since it requires smaller sample size. As the data show, these answers do not depend on lambda, only sample size.

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

   The data shows that no, lambda being fixed does not impact the accuracy. Our conclusions are based on the sample size, not lambda. The accuracy does not change with a change in lambda.

Section 2 : Contains R Code

1. R-Code for question 1

a.

```
body.temp.heart.rate = read.csv("bodytemp-heartrate.csv", header=T)
#male and female separate datasets
male = subset(body.temp.heart.rate, body.temp.heart.rate$gender==1)
female = subset(body.temp.heart.rate, body.temp.heart.rate$gender==2)

# boxplots
boxplot(male$body_temperature, female$body_temperature,
        main = "Boxplots of body temperatures", names = c('Males', 'Females'), ylab =
"Temperature")

# QQ Plot
qqnorm(male$body_temperature, main='Q-Q Plot for male')
qqline(male$body_temperature)
qqnorm(female$body_temperature, main='Q-Q Plot for female')
qqline(female$body_temperature)
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.

```
boxplot(male$heart_rate, female$heart_rate,
        main = "Boxplots of heart rate", names = c('Males', 'Females'), ylab = "Heart
Rate")

# QQ Plot
qqnorm(male$heart_rate, main='Q-Q Plot for male')
qqline(male$heart_rate)
qqnorm(female$heart_rate, main='Q-Q Plot for female')
qqline(female$heart_rate)
```

10. 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.

cor(male$body_temperature,male$heart_rate)
cor(female$body_temperature,female$heart_rate)

2. R-Code for question 2

a and b:

```
#interval 1
nboot <- 5000
nVals = c(5, 10, 30, 100)
lambdaVals = c(0.01, 0.1, 1, 10)
alpha = 0.05
for (n in nVals) {
   for (l in lambdaVals) {
      x = replicate(5000,sort(rexp(n, l)))
      mean.x = colMeans(x)
      sd.x =apply(x, 2, sd)
      conf.int.lower = mean.x + -1 * qnorm(1 - (alpha/2)) * sd.x/sqrt(n)
      conf.int.higher = mean.x + 1 * qnorm(1 - (alpha/2)) * sd.x/sqrt(n)
      coverage.prop= (1/l >= conf.int.lower & 1/l <= conf.int.higher)
      cat(n, l, mean(coverage.prop==TRUE), "\n")
   }
}
```

5 0.01 0.804

5 0.1 0.818
5 1 0.8108
5 10 0.8052
10 0.01 0.8712
10 0.1 0.8666
10 1 0.8702
10 10 0.8758
30 0.01 0.9154
30 0.1 0.9166
30 1 0.9186
30 10 0.9158
100 0.01 0.9412
100 0.1 0.937
100 1 0.9392
100 10 0.9402

#interval 2

```
lambda.star <- function(x){
   n <- length(x)
   xbar <- 1/mean(x)
   xstar <- replicate(1000, mean(rexp(n, xbar)))
   return (xstar)
}
nboot <- 5000
nVals = c(5, 10, 30, 100)
lambdaVals = c(0.01, 0.1, 1, 10)
for (n in nVals) {
  for (l in lambdaVals) {
    x <- rexp(n,rate=l)
    lambda.boot.dist <- replicate(5000, sort(lambda.star(rexp(n,rate=l))))
    conf.ints.lower = lambda.boot.dist[ceiling(25),]
    conf.ints.higher = lambda.boot.dist[floor(975),]
    coverage.prop= (1/l >= conf.ints.lower & 1/l <= conf.ints.higher)
    cat(n, l, mean(coverage.prop==TRUE), "\n")
  }
}
```

5 0.01 0.892
5 0.1 0.8996
5 1 0.899
5 10 0.8986
10 0.01 0.9234

10 0.1 0.926
10 1 0.919
10 10 0.9256
30 0.01 0.9428
30 0.1 0.9376
30 1 0.9392
30 10 0.9374
100 0.01 0.9526
100 0.1 0.9428
100 1 0.9458
100 10 0.9434

Code for plot
Column Wise - Figure 1
plot(c(5,10,30,100), c(0.8052, 0.8758, 0.9158, 0.9402), main = "L = 10", xlab = 'n', ylab = 'Proportions', col = 'blue', type = 'b', xlim = c(1,100), ylim = c(0,1))
lines(c(5,10,30,100), c(0.8986,0.9256,0.9374,0.9434), col = 'red', type = 'b')

Row Wise - Figure 2
plot(c(0.01,0.1,1,10), c(0.9412, 0.937, 0.9392, 0.9402), main = "N = 100", xlab = 'Lambda', ylab = 'Proportions', col = 'blue', type = 'b', xlim = c(0.01,10), ylim = c(0,1))

lines(c(0.01,0.1,1,10), c(0.9526,0.9428,0.9458,0.9434), col = 'red', type = 'b')