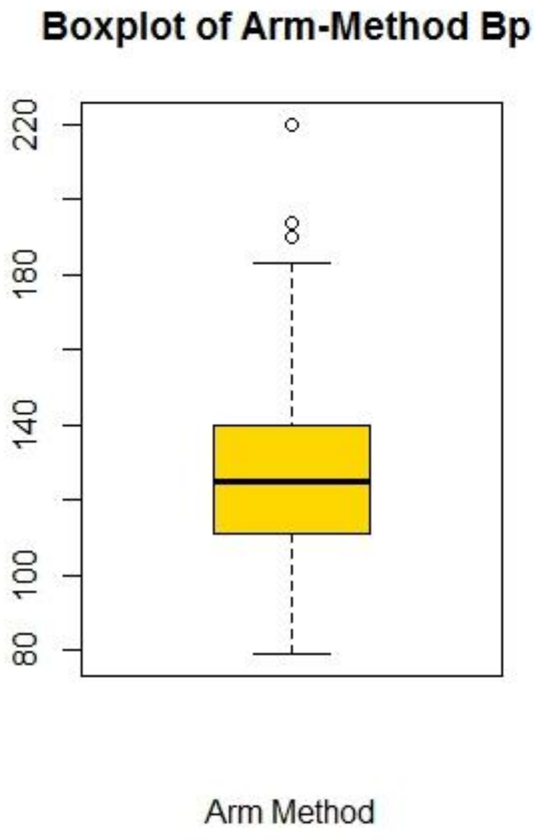
MINI PROJECT 3

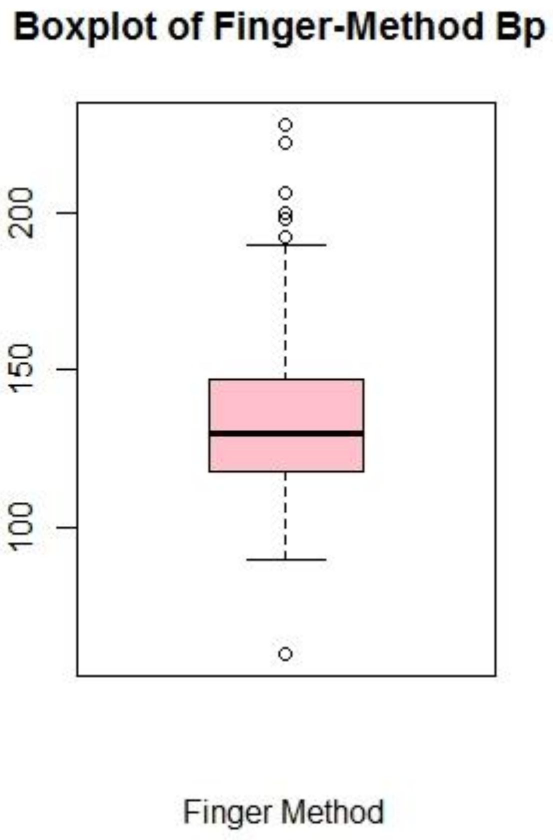
By- Raghav Mathur (rxm162130), Deepak Shanmugam (dxs161930)

Contribution to the Project: Each member contributed equally to the project. We worked together to analyse the distributions, determine CI, and calculating (n) for acceptable accuracy.

EXERCISE 1:

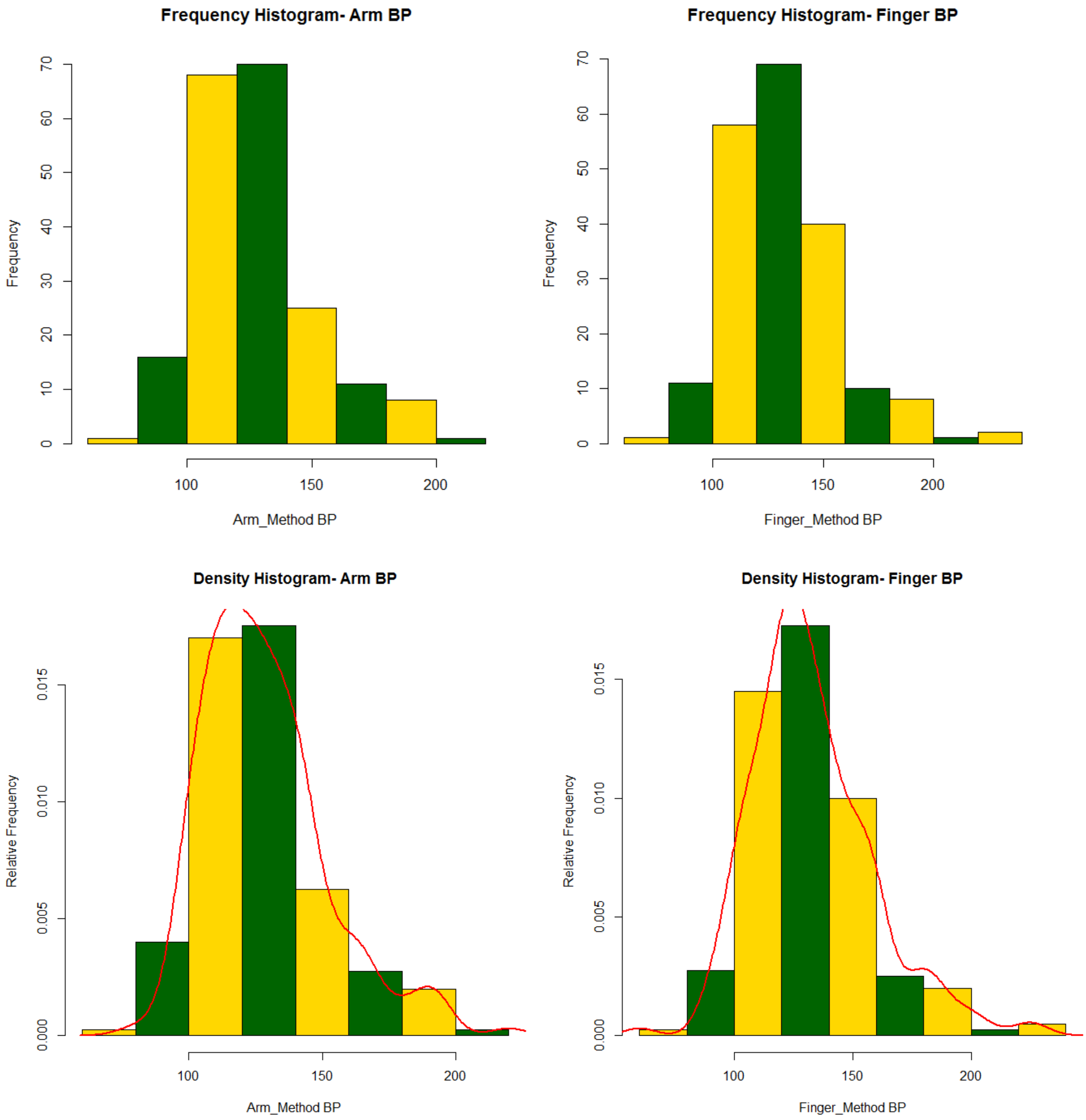
Section 1:
(a)

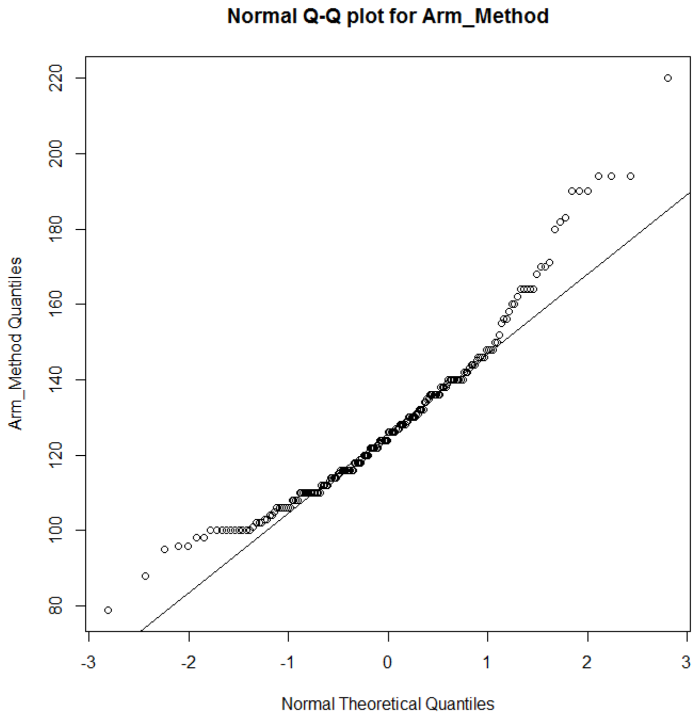


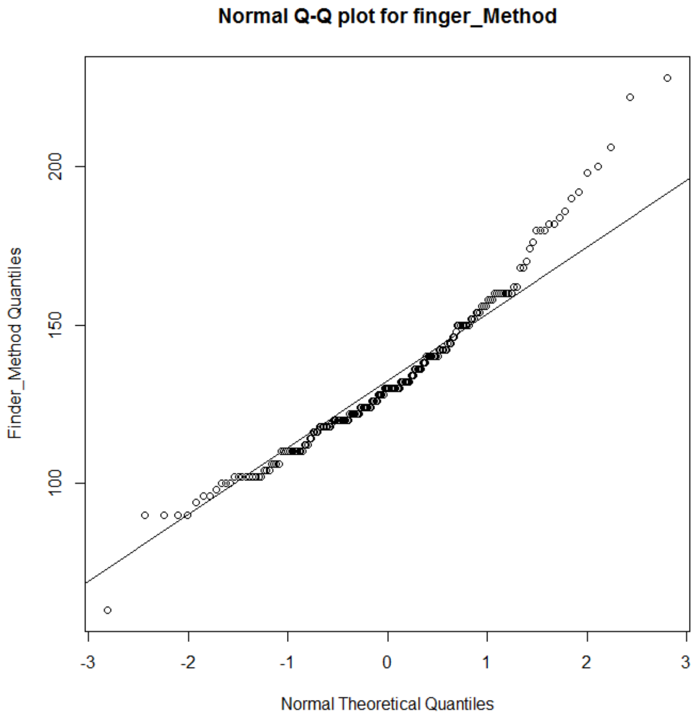


Arm Method	Finger Method
• The difference between the Mean and Median is small as the distribution is slightly right skewed. • We can see outliers above the upper whiskers as they fall above Q_3 + 1.5(IQR) limit	• The mean of the distribution seems to be greater than the median of the distribution as the distribution seems right skewed. • There are outliers above and below the upper and lower whisker respectively as they fall outside the interval from Q_1 -1.5(IQR) to Q_3 + 1.5(IQR)
Similarity	
Looking the upper and lower quantiles, we can say that the IQR for both the distributions are similar, hence their variance. Also, the spread of both distributions look similar considering the range from Q_1 - 1.5(IQR) to Q_3 + 1.5(IQR).	

(b)

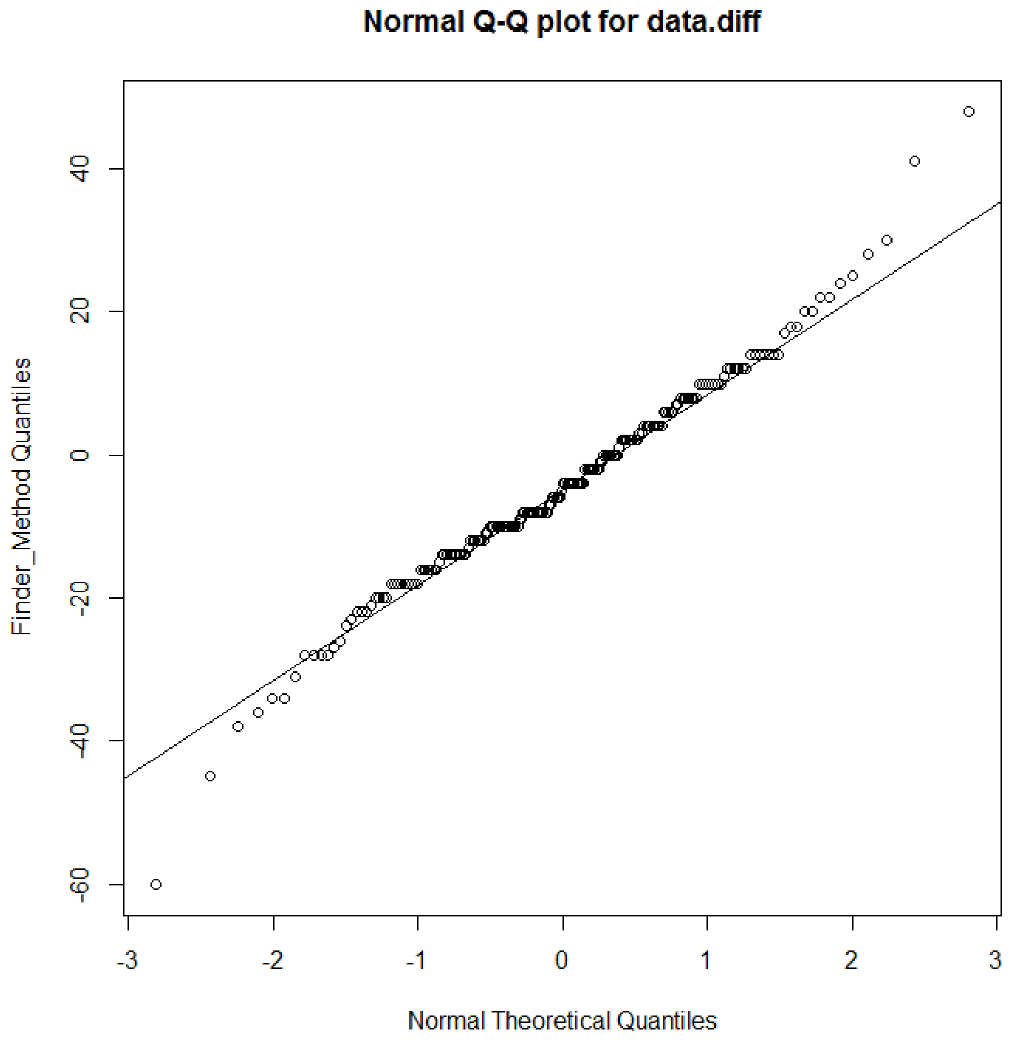

Frequency Histogram- Arm BP
Frequency
0
10
20
30
40
50
60
70
100
150
200
Arm_Method BP
Frequency Histogram- Finger BP
Frequency
0
10
20
30
40
50
60
70
100
150
200
Finger_Method BP
Density Histogram- Arm BP
Relative Frequency
0.000
0.005
0.010
0.015
100
150
200
Arm_Method BP
Density Histogram- Finger BP
Relative Frequency
0.000
0.005
0.010
0.015
100
150
200
Finger_Method BP





<table>
<tr><th>Arm Method</th><th>Finger Method</th></tr>
<tr><td>• The histogram showing the frequency of range of blood pressure appears to have a normal distribution.
• The Mean is slightly greater than the Median as observed from the shape of histogram (slightly right skewed).
• Also, we can observe from the histogram that some outliers in the lower and upper tail as their frequency is close to 1, which is clear from its corresponding QQplot.
• Majority of points in QQplot have reasonable normal distribution as they lie very close to the QQline. The upper and lower tails points have greater variability which is reasonable as it is a Real dataset.</td><td>• The histogram showing the frequency of range of blood pressure appears to have a normal distribution.
• The Mean is greater than the Median as observed from the shape of histogram (right skewed).
• Also, we can observe from the histogram that some outliers in the upper tail as their frequency is close to 1, which is clear from its corresponding QQplot.
• Except the outliers, the upper tail points do not have high variability considering most of the points that are lying in proximity to the QQline.</td></tr>
<tr><td colspan="2">Similarity</td></tr>
<tr><td colspan="2">The density curves in histogram show a similarity and both distributions are likely to be normal as the Mean and Median values seem to lie close to each other. The assumption of normality is reasonable as normal distribution provides a good fit to majority of data as seen from the QQplot. Apart from upper and lower tail points, maximum of the points are tightly bound to the QQline showing that normal distribution is a good fit for this real data.</td></tr>
</table>

(c)



The 95% Confidence Interval for the difference in the mean of the 2 methods comes out to be- [-4.580887 -4.009113]. This means that difference in means of 2 methods will have a very small width of confidence interval resulting in a high probability of it being lying within this interval. The data comes from paired sample where independent Mean of 'arm method' data is 128.52 and 'finger method' data is 132.81 which have a difference of -4.295. Observing the CI, we can say that they don't have identical means as 0 does not lie in this interval. Assumptions made: the data is paired data as it comes from the same population and assume that data has normal distribution. Based on QQplot for difference in Means we observe that normal distribution is a good fit for this real data.

Section 2: (R Code)

(a)

```
#package for coloured histograms and boxplots
install.packages("reshape2")
# read the dataset
asetwd("C:/Users/mastr/OneDrive/Documents/UTD NOTES/UTD-STATISTICS/mini project 3")
data= read.table("bp.txt", header = TRUE)
data=data.frame(data)
arm_method= (data[,1])
finger_method= (data[,2])
```

```
# show 5 point summary
summary(arm_method)
summary(finger_method)
#print parallel boxplots
par(mfrow=c(1,2))
boxplot (arm_method, col=(c("gold","darkgreen")),range=1.5, main = "Boxplot of Arm-Method Bp ",xlab = "Arm Method")
boxplot(finger_method, col=(c("pink","darkgreen")),range=1.5,main = "Boxplot of Finger-Method Bp ",xlab = "Finger Method")
```

(b)

```
# frequency histogram by default
par(mfrow=c(1,2))
hist(arm_method, col=(c("gold","darkgreen")),xlab="Arm_Method BP ", ylab="Frequency", main="Frequency Histogram- Arm BP")
hist(finger_method, col=(c("gold","darkgreen")), xlab="Finger_Method BP", ylab="Frequency", main="Frequency Histogram- Finger BP")
# relative frequency (density) histogram
hist(arm_method, freq=FALSE, xlab="Arm_Method BP", ylab="Relative Frequency", main="Relative Frequency Histogram of Arm BP")
hist(finger_method, freq=FALSE, xlab="Finger_Method BP", ylab="Relative Frequency", main="Relative Frequency Histogram of Finger BP")
# checking for outlier
iqr1 <- IQR(arm_method)
lower_arm <- quantile(arm_method, prob=0.25) - 1.5*iqr1
upper_arm <- quantile(arm_method, prob=0.75) + 1.5*iqr1
c(lower_arm, upper_arm)
iqr2 <- IQR(finger_method)
lower_finger <- quantile (finger_method, prob=0.25) - 1.5*iqr2
upper_finger <- quantile(finger_method, prob=0.75) + 1.5*iqr2
c(lower_finger, lower_finger)
#QQplots and QQlines
par(mfrow=c(1,2))
qqnorm(arm_method,main = "Normal Q-Q plot for Arm_Method", xlab = "Normal Theoretical Quantiles", ylab = "Arm_Method Quantiles")
qqline(arm_method)
qqnorm(finger_method,main = "Normal Q-Q plot for finger_Method", xlab = "Normal Theoretical Quantiles", ylab = "Finder_Method Quantiles")
qqline(finger_method)
```

(c)

```
#calculating difference of means in 2 methods
data.diff=data$armsys-data$fingsys
dbar=mean(data.diff)
#plotting QQplot for the paired set of data
qqnorm(data.diff,main = "Normal Q-Q plot for Difference of Means", xlab = "Normal Theoretical Quantiles", ylab = "Finder_Method Quantiles")
qqline(data.diff)
#calculte confidence interval
x=sd(data.diff)*(1/(length(data.diff)*(1/2)))
CI= dbar +c(-1,1)*(qnorm(1-0.025)*x)
```

EXERCISE 2:

Section 1:
(a)

n

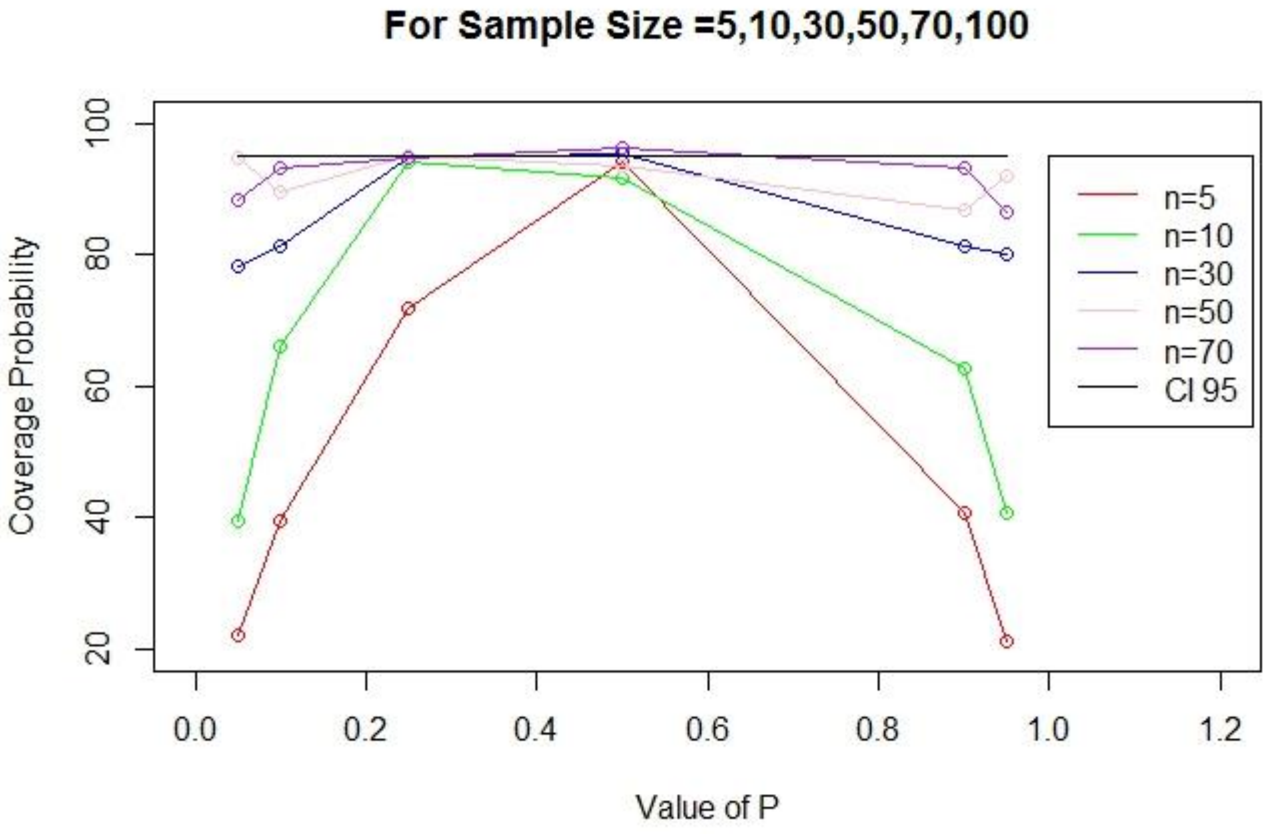
p

	0.05	0.1	0.25	0.5	0.9	0.95
5	21.8	40.0	72.4	94.2	39.2	21.4
10	40.2	64.2	94.0	91.8	64.0	39.4
30	78.2	81.4	94.6	95.2	81.6	79.6
50	94.8	89.2	95.0	93.4	86.8	92.2
100	88.2	93.0	94.8	96.6	93.2	86.6

The probability of having an acceptable accuracy increases with the value of n. For n=5, we see that the coverage probability is varying for different values of ‘p’ and with a high range (from 21.4 to 94.2). As we increase the sample size(n) the varied range decreases and probability of having an acceptable accuracy gets close to 95%.

The probability of having an acceptable accuracy increases with the value of ‘p’ up to (p = 0.5) after which there is a fall in the coverage probability. From this we can observe that the coverage probability is closer to 95% when p=0.5 for regardless of ‘n’.

Based on our observations, we can say that at **p= 0.5** we need **n>=30** to have an acceptable accuracy for a 95% confidence interval. Hence, our value will depend on ‘p’.



Section 2: (R Code)

```r
#Put all the n values in an array
n_array=c(5, 10, 30, 50, 100)
n_count=length(n_array)
#Put all the p values in an array
p_array=c(0.05, 0.1, 0.25, 0.5, 0.9, 0.95)
p_count=length(p_array)
#Initialize a matrix to store the coverage probabilities
result=matrix(data = NA, nrow = 5, ncol = 6)
#Dimension names for the matrix
dimnames(result) = list(c("5", "10","30", "50","100"),c("0.05", "0.1", "0.25","0.5", "0.9", "0.95")) # column names
replications = 500 #Number of times to replicate

for(x in 1:n_count){
  n=n_array[x]
  for(y in 1:p_count){
    count=0
    p=p_array[y]
    #Generate the random samples from binomial distribution
    random_samples=replicate(replications,rbinom(n,1,p))
    for(i in 1:replications){
      #Estimate p.hat for the given n and p
      p.hat= mean(random_samples[,i]==1)
        #Find confidence interval for the estimated p.hat
      CI_interval = p.hat +c(-1,1)*(qnorm(0.975) * sqrt((p.hat*(1-p.hat))/n))
        #Check whether p(probability) lies within the confidence interval
      if(CI_interval[1] <= p && p <= CI_interval[2]){
             count=count+1
        }
    }
    #Find the coverage probability
    coverage_probability = (count/replications)*100
    #Store the coverage probability in the matrix
    result[x,y]=coverage_probability
  }
}
#plotting ‘p’ against ‘coverage probability’ for each value of Sample Size ‘n’
CI=c(95,95,95,95,95,95)
plot(p_array,result[1,],type="o",col="red",ylim = c(20, 100),xlim = c(0,1.2), xlab = "Value of P",
     ylab = "Coverage Probability", main = "For Sample Size =5,10,30,50,70,100")
lines(p_array,result[2,],type="o",col="green")
lines(p_array,result[3,],type="o",col="blue")
lines(p_array,result[4,],type="o",col="pink")
lines(p_array,result[5,],type="o",col="purple")
lines(p_array,CI,col="black")
legend(1, 95, legend=c("n=5","n=10","n=30","n=50","n=70","CI 95"),
       col=c("red", "green","blue","pink","purple","black"),lty = 1)
```