STAT 6313

Mini Project 3

Spring 2017

Name of the Members:

Krupali Patel

Ankita Patil

Contribution of each member:

Both of us contributed equally to the project.

Each and every step is done with proper discussion.

Section 1

Exercise 1 (10 points)
Consider the dataset stored in the file bp.txt. This dataset contains one measurement of systolic blood pressure (in mmHg) made by each of two methods—a finger method and an arm method—from the same 200 patients.

(a) Perform an exploratory analysis of the data by examining the distributions of the measurements from the two methods using boxplots. Comment on what you see. Do the two distributions seem similar? Justify your answer

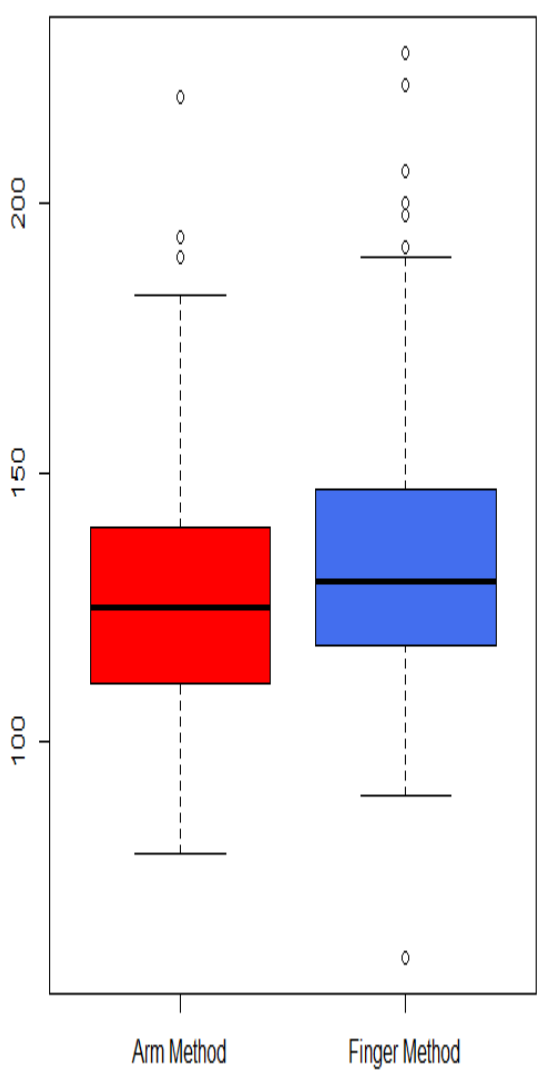
Analysis and comments on the two boxplots:

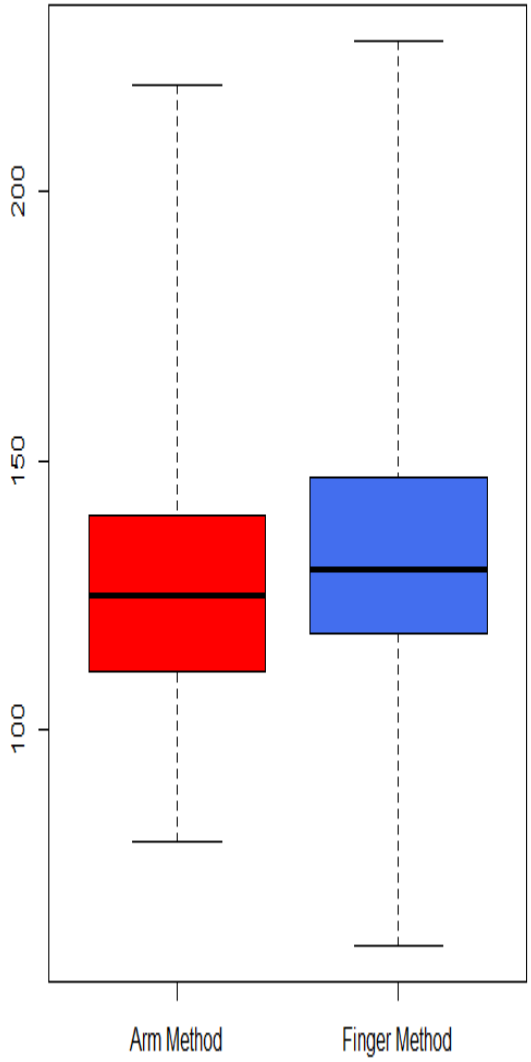
- Observations taken by Arm method has less number of outliers in comparison to Finger Method. (as can be seen from the points in the two box plots)
- Futhermore the above analysis can be numerically justified as below:

```
> summary(bloodpressure$armsys)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   79.0   111.5   125.0   128.5   140.0   220.0
> summary(bloodpressure$fingsys)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   118.0   130.0   132.8   146.5   228.0
```

- IQR and mean of the two box plots are nearly same.

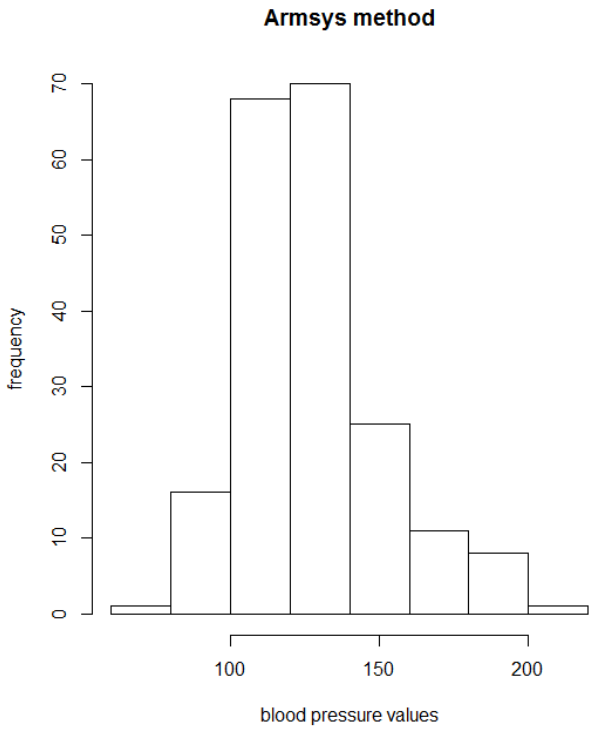
But if we pay close attention to the two boxes of the plot, we find that distribution of the data collected by arm method is **almost symmetric** whereas for the finger method it is **right skewed** (as top half of the box is larger than the bottom half)

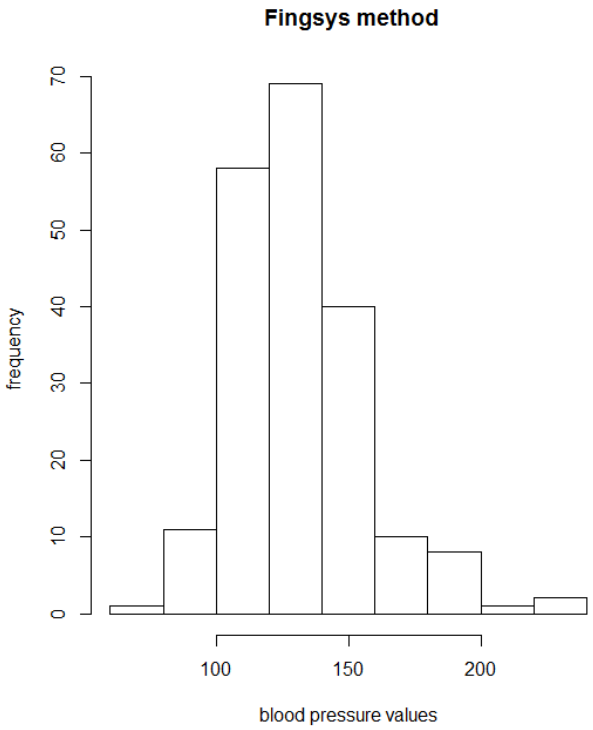

200
150
100
Arm Method
Finger Method

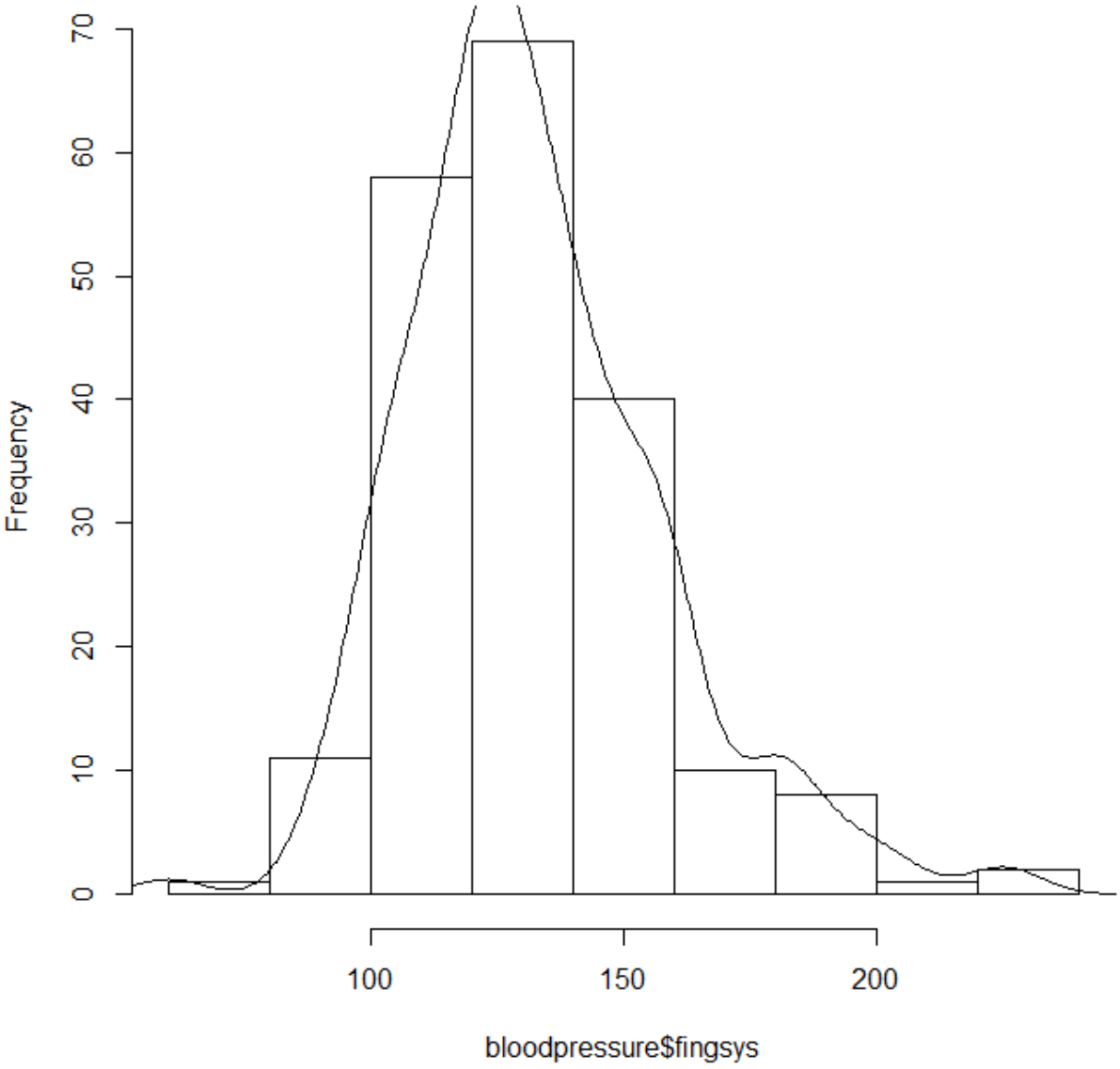

200
150
100
Arm Method
Finger Method

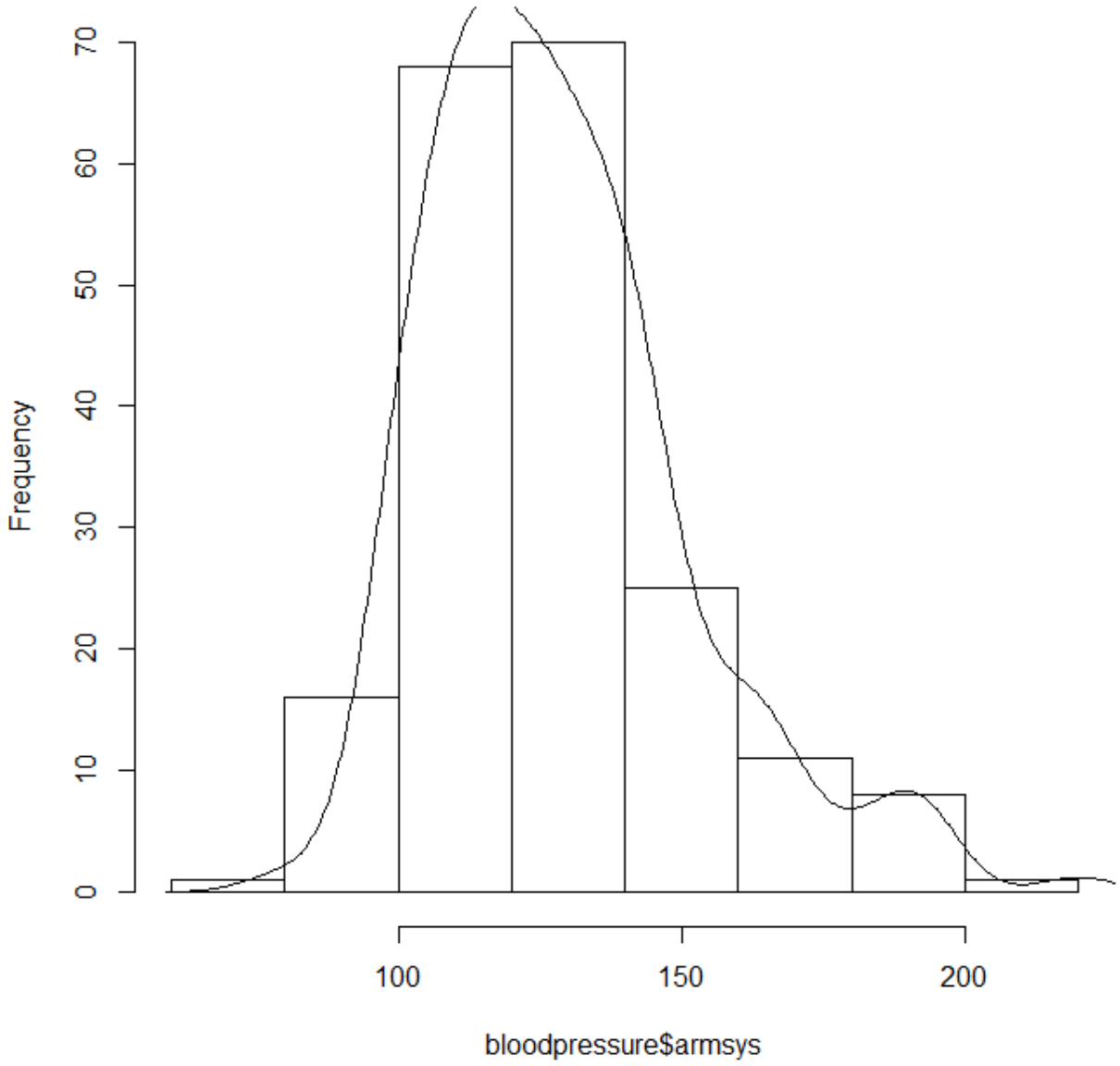
(b) Use histograms and QQ plots to examine the shapes of the two distributions. Comment on what you see. Does the assumption of normality seem reasonable? Justify your answer.

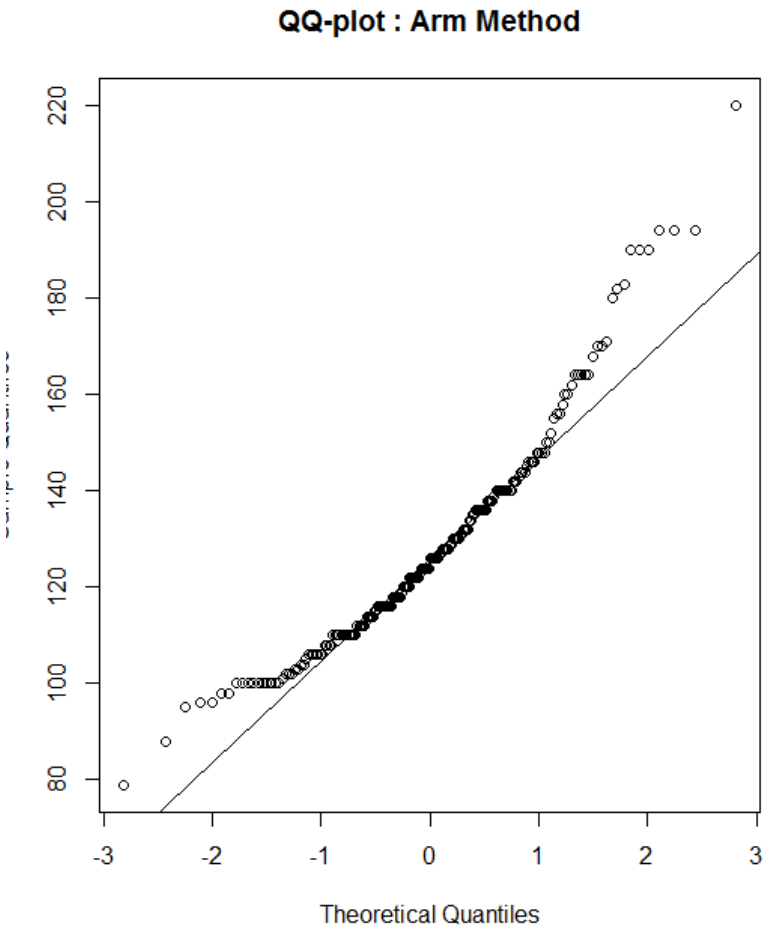
If we look at both the histograms, the arm method data distribution is very similar to normal distribution but histogram of finger method does not follow normal distribution. Finger method histogram suggests that it is little right skewed.

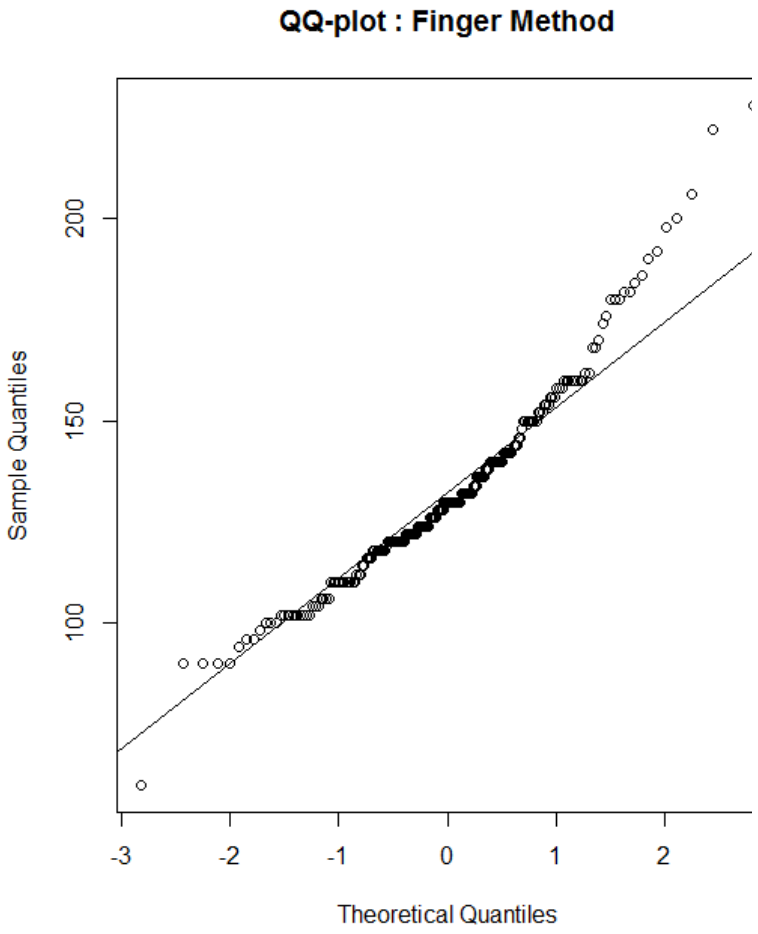





Histogram of bloodpressure$fingsys
Frequency
0
10
20
30
40
50
60
70
100
150
200
bloodpressure$fingsys


Histogram of bloodpressure$armsys
Frequency
0
10
20
30
40
50
60
70
100
150
200
bloodpressure$armsys





QQ- plot gives us a better picture of the data distribution in comparison to histogram above because QQ – plot of both the methods show that majority of the points does not lie on the normal distribution line and hence our assumption of normality does not seem that reasonable. If majority of the point lie on the line we could conclude that data is normally distributed which in our case it is not.

(c) Construct an appropriate 95% confidence interval for the difference in the means of the two methods. Interpret your results. Can we conclude that the two methods have identical means? What assumptions, if any, did you make to construct the interval? Do the assumptions seem to hold? Justify all your answers.

When we consider the data of the two methods as independent we find that the mean is not identical as we get means of the two as :

Mean of armsys method: 128.52
Mean of fingsys method 132.815
difference in Mean of two methods: -4.295
confidence interval: -9.0961939 0.5061939

Therefore, the two data is not identical we need to consider them as paired data.

We assumed initially that the two data follow normal distribution, which if they did we would have got identical means and mean should have lied in the confidence interval. But since our means are not identical we can say that our data is not normally distributed.

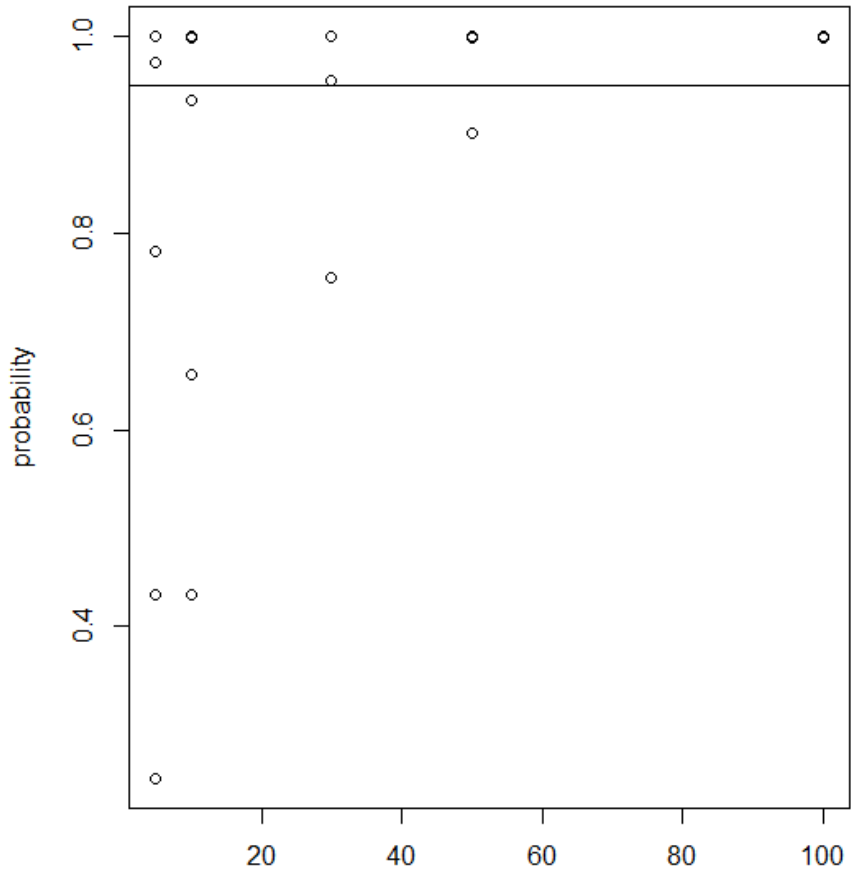
Exercise 2 (10 points)
We know how to construct a large sample confidence interval for a population proportion p. How large n should be for this interval to have acceptable accuracy? Answer this question by computing the coverage probability of this interval using Monte Carlo simulation, and examining how close the probability is to the nominal confidence level. Take level of confidence to be 95% but use a variety of values for n and p, e.g., n = 5, 10, 30, 50, 100, and p = 0.05, 0.1, 0.25, 0.5, 0.9, 0.95. Summarize your results graphically. Comment on any patterns you see in the results. Based on your findings, what n would you recommend for the use of this confidence interval? Would your answer depend on p? Explain.

It can be seen from the table below that as value of n increases, probability that 'phat' will lie in confidence interval will increase. As a 'p' increases even for small values of 'n' but high values of 'p ' probability that 'phat' will lie in confidence interval will increase.

n \ p	0.05	0.1	0.25	0.5	0.9	0.95
5	0.232	0.396	0.776	0.962	1.000	1.000
10	0.400	0.672	0.934	0.996	1.000	1.000
30	0.788	0.962	0.98	1.000	1.000	1.000
50	0.938	0.97	1.000	1.000	1.000	1.000
100	0.996	1.000	1.000	1.000	1.000	1.000

Here, in the below graph straight line is a line with probability value= 0.95, since we want to find the value of n for which confidence interval is 95% all the points above this fall into desired category. But these values are affected by probability value as well, as can be seen from the graph. But for n =100 all the values with different probability fall into category of confidence int erval greater than 95 hence n=100 would be the safest value to consider.

Our answer depends on p as well(as seen from table above), so for small values of n high values of p can be taken.



Section 2

R CODE

```
#the following line reads the data from bp.txt file and stores it in bloodpressure

bloodpressure <- read.table(file="bp.txt",header = TRUE)


#=========================
# Question 1 a
#=========================

par(mfrow=c(1,2)) # 2 plots in 1 row
boxplot(bloodpressure, range=1.5,col=c('red','royalblue2'),names=c( 'Arm Method','Finger Meth
od')) # draws box plot for the data = bloodpressure

#=========================
# Question 1 b - Histogram
#=========================


# frequency histogram by default
par(mfrow=c(1,2))
hist(bloodpressure$armsys, xlab="blood pressure values", ylab="frequency", main="Armsys me
thod")
hist(bloodpressure$fingsys, xlab="blood pressure values", ylab="frequency", main=" Fingsys me
thod")


#=========================================================================
#the lines below will help to draw the curve by joining mid-point of the bar length

myhist <- hist(bloodpressure$armsys)
  # this gives multiplier which when we multiply with y coordinate of the curve returns scaled
#distance as per our histogram
multiplier <- myhist$counts / myhist$density
mydensity <- density(bloodpressure$armsys)
mydensity$y <- mydensity$y * multiplier[1]

plot(myhist)
lines(mydensity)
```

```
myhist <- hist(bloodpressure$fingsys)
multiplier <- myhist$counts / myhist$density
mydensity <- density(bloodpressure$fingsys)
mydensity$y <- mydensity$y * multiplier[1]

plot(myhist)
lines(mydensity)
#=================================================================

#=========================
# Question 1 b - QQ Plot
#=========================
par(mfrow=c(1,2))
qqnorm(bloodpressure$armsys,main = 'QQ-plot : Arm Method')
qqline(bloodpressure$armsys) #this will draw normal distribution line
qqnorm(bloodpressure$fingsys,main = 'QQ-plot : Finger Method')
qqline(bloodpressure$fingsys)




#=========================
# Question 1 c
#=========================

#the following code is to check the difference of the mean and to find if means are identical
#=================================================================
# Function below finds confidence interval when sigma is known and different
#=================================================================
conf.int <- function(mu, sigmax,sigmay, n,m, alpha) {

  ci <- mu + c(-1, 1) * qnorm(1 - (alpha/2)) * sqrt((sigmax^2/n)+(sigmay^2/m))
  return(ci)
}

armsys.m<- mean(bloodpressure$armsys)
armsys.sigma <- sd(bloodpressure$armsys)

fingsys.m<- mean(bloodpressure$fingsys)
fingsys.sigma <- sd(bloodpressure$fingsys)

alpha <- 0.05
diff.m <- armsys.m-fingsys.m
```

```
n<- nrow(bloodpressure)

ci <- conf.int(mu = diff.m , sigmax = armsys.sigma , sigmay = fingsys.sigma, n,n,alpha)
print("\n Assuming that both the random variables are independent")
cat("mean of arm:",armsys.m)
cat("\nmean of fing", fingsys.m)
cat("\ndiffre",diff.m)
print(ci)

#================================
# the following function uses t-distribution as variance is unknown
#================================

pairedconf.int <- function(mu, sigma,n,talpha_by_2) {

  ci <- mu + c(-1, 1) * qt(talpha_by_2, n-1) * sigma/sqrt(n)
  return(ci)
}
diff <- (bloodpressure$armsys- bloodpressure$fingsys)

alpha<-0.05
talpha_by_2<- 1-(alpha/2)
print(talpha_by_2)

n <- length(diff)
ci<- pairedconf.int(mean(diff),sd(diff),length(diff),talpha_by_2)
print(ci)
```

```
#================================
# Question 2
#================================


#====================================
# function to find confidence interval
#====================================
proportion.conf.int <- function(phat,n, alpha) {

  ci <- phat + c(-1, 1) * qnorm(1 - (alpha/2)) * sqrt(phat*(1-phat)/n)
  return(ci)
}

#====================================
#Function that uses monte-carlo simulation to find various values that follow binomial
#distribution
#====================================
simulate.data <- function(sim,i,j)
{
  data<- replicate(sim,rbinom(i,1,j))
  return (data)
}




#====================================
# Function to compute phat and confidence interval for various simulated value
#(here sim=500 so we will have 500 phat & confidence interval values one for each simulation)
#====================================
compute.phatAndCi <- function (sim,data,n)

{
  matrix_ci<-matrix(nrow=2,ncol=sim)
  counter<-0
  c<-1
 phat<-numeric()
```

```
  while(c<=sim)
  {
    index=1:n
    phat[c]<- sum(data[index,c]==1)/n
   # cat("\tphat \t",phat[c])
    if(phat[c]!=0)
    {
    ci<- proportion.conf.int(phat[c],i,alpha)

    matrix_ci[1,c] <- ci[1]
    matrix_ci[2,c] <-ci[2]
    }
    #if there is no success we will not  compute ci for that phat because if we do so we will get ci a
#s 0,0 and as per our condition of if statement above it will include failure events phat
    else
    {
      matrix_ci[1,c] <- -99
      matrix_ci[2,c] <- -99
    }
  #  cat("\t ci: ",matrix_ci[1,c], "\t",matrix_ci[2,c],"\n")
    # if p lies within confidence interval we will count on those 'p's and find the probability of gett
ing such 'p'
    if(phat[c]>=matrix_ci[1,c] && phat[c]<=matrix_ci[2,c])
      counter=counter+1
    c=c+1
  }
# cat("dimension",length(phat),dim(matrix_ci))
#  cat("counter for n",counter)
  return (counter/sim)
}

alpha=0.05
xaxis<-numeric()
yaxis<-numeric()

n<-c(5,10,30,50,100)
p<- c(0.05,0.1,0.25,0.5,0.9,0.95)

c<-1
for(i in n)
{
```

```
    for(j in p)
    {
      sim<- 500
      data<- simulate.data(sim,i,j)
      #print(data)
      prob<-compute.phatAndCi(sim,data,i)
  xaxis[c]<- i   # xaxis - this will store all values of n
yaxis[c]<-prob # yaxis - this will store values of probability for corresponding n values and p values
      c=c+1

    }
  }
  plot(x=xaxis,y=yaxis,xlab = "n",ylab = "probability" )
  y<- replicate(length(n),95)
  x<- replicate(length(n),0)
  abline(a=0.95,b=0) #normal confidence line
```