Mini Project -3

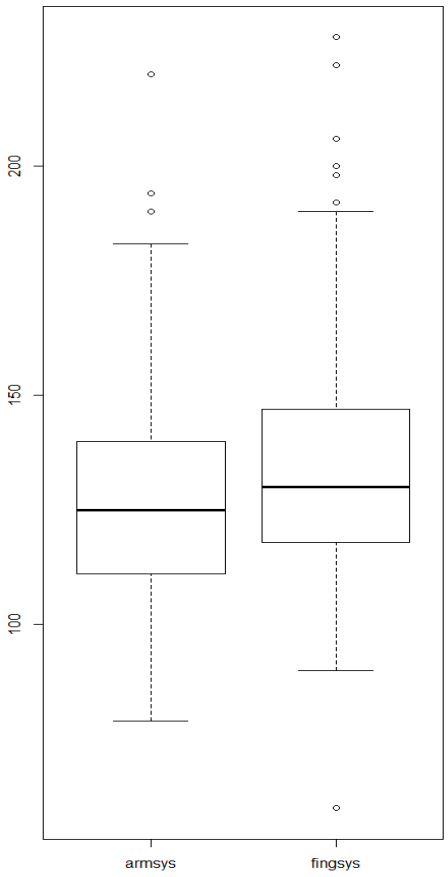
CS 6313.001 Statistics in Data Science

Falak Singhal (fxs161530@utdallas.edu)

Melvin James (mxj162130@utdallas.edu)

1.

a) The boxplots of the arm and finger systolic BP are follows –



Both the distributions are slightly right skewed as both have longer upper whiskers. Both the data sets also have equal spread as the interquartile range for both appears to be the same. Hence the two distributions seem similar.

From the boxplots it is evident that the median for arm data is about 125 mmHg and median value for finger data is about 130 mmHg. Also median for finger data is higher than median for arm data suggesting the systolic BP to be generally higher when measured from patient's finger.

Both the boxplots are plotted by using 1.5(IQR) rule and the outliers in the data are also shown. The finger systolic BP data appears to have more outliers than the arm systolic data suggesting that the finger systolic BP might not be as accurate.

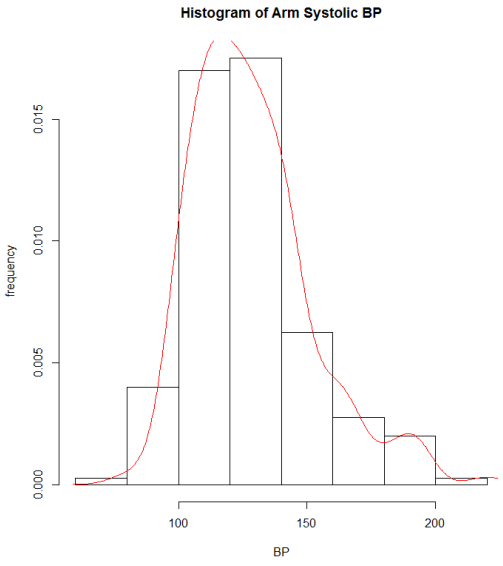
The five-point summary, as evident from boxplots is calculated to be –

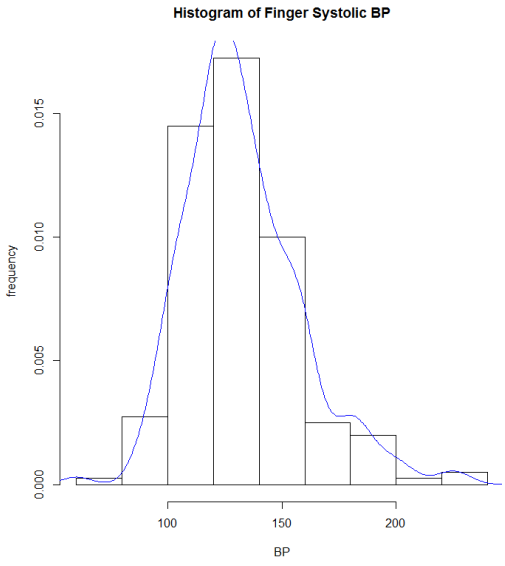
```
> summary(bp$armsys)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   79.0   111.5   125.0   128.5   140.0   220.0
> summary(bp$fingsys)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   118.0   130.0   132.8   146.5   228.0
```

And the interquartile range measuring the spread of the data is calculated as –

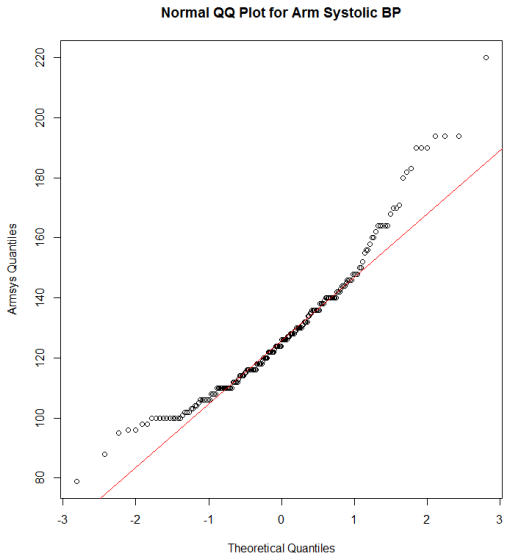
```
> IQR(bp$rmsys)
[1] 28.5
> IQR(bp$fingsys)
[1] 28.5
```

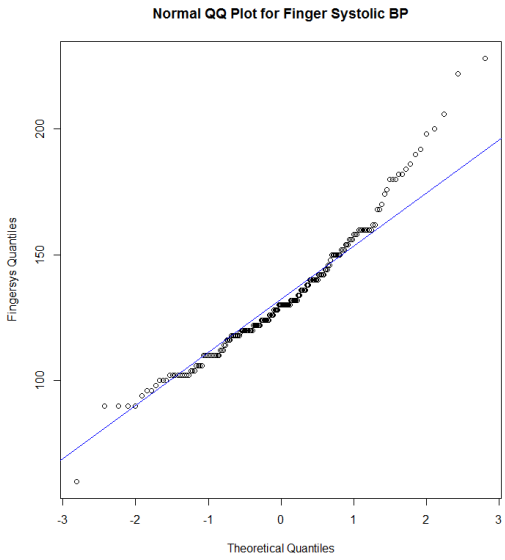
b) The histogram for the two data sets are shown below –





In both the distributions, the density curve is also shown. As predicted by the boxplots, the data in both data sets is right skewed. Both the histogram and the density curves suggest a normal distribution of the data which is further examined using the below QQ plots –





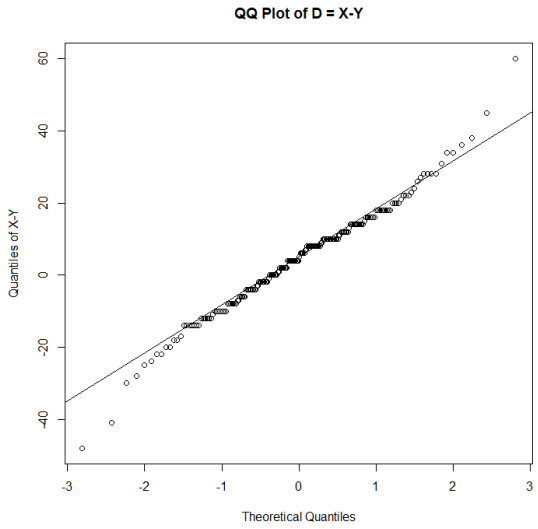
As the QQ plots suggest, the assumption of normailty seems resonable for both the data sets. The right skewednes s of the data sets is evident from the characteristic upward curvature in the plots.

c) The 95% confidence interval for the two data sets is given as

$$CI: \bar{D} \pm \mathbb{Z}_{\alpha/2} \frac{\sigma_D}{\sqrt{n}} ; where\ \sigma_D\ is\ the\ population\ std\ deviation$$

Since σ_D is unknown, it can be estimated using s_D is the sample standard deviation. In this case the CI is given as –

$$CI: \bar{D} \pm t_{n-1,\alpha/2} \frac{s_D}{\sqrt{n}} ; where\ s_D\ is\ the\ sample\ std\ deviation$$



Now, as the sample size is large (=200), $t_{n-1,\alpha/2} \approx \mathbb{Z}_{\alpha/2}$ and the CI c an be approximated by -

$$CI: \bar{D} \pm \mathbb{Z}_{\alpha/2} \frac{s_D}{\sqrt{n}}$$

Which is – [2.273471, 6.316529] interpretted as the plausible values of the difference in mean of the two data sets.

The normality assumption verified by plotting the quantiles of $D = X - Y$ with respect to the standard normal quantiles and genera ting a QQ plot.

Since the points closely follow the straight line, we can say that the as sumption of normality for D seems reasonable & $D \sim N(0,1)$

2.

Monte Carlo study of the data is performed by generating the Bernoulli(p) from Uniform(0,1) distribution. Then the proprti on of successes is measured as $\hat{p}$.

Based on the simulated $\hat{p}$, a 95% confidence interval is generated using the approximate CI for population proportion–

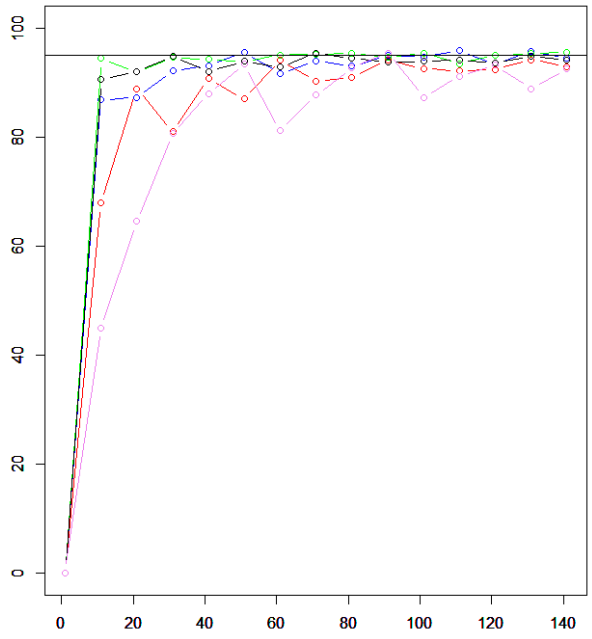
$$CI: \hat{p} \pm Z_{\propto/2} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

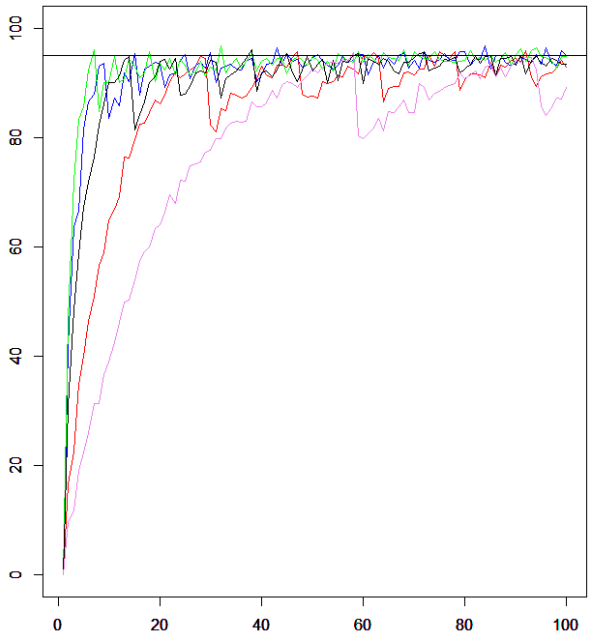
This experiment is replicated “nsim” number of times using various values of sample size “n” and the accuracy of the confid -ence interval is calculated as the mean number of the times the population proportion p is captured by the interval.

The experiment is repeated for different values of *p* and *n.*

An accuracy graph is plotted as follows, the values on the horizontal axis represents various sizes of the sample and the vert -ical axis representing the accuracy level for the given p value.

The following inferences can be drawn from the graph -





(Graph on right plots accuracies for larger number of size points)

Dependency on sample size-

1. The accuracy of the confidence interval does depend on the size of the sample. For a given probability value, it is ob served that for values of n below 20, the level of accuracy for the confidence intervals is low on average.
2. As the value of sample size n is increased, the accuracy increases achieveing the point of the nominal confidence le vel of 95%.

Hence for dependency on sample size, a higher value of n corresp-onds to a more accurate CI. Which is expected since the f ormula for the confidence interval for population proportion is an approxi-mation which holds good when the sample size is large.

Dependency on Population proportion p-

1. As a general trend it can be observed from the plots that for all different values of p, the accuracy of the confidenc e interval is increasing with the value of n.
2. Since this trend is common for all different values of p, these observations suggests the accuracy to be independen t of p.

Section-2

R- Code –

1.

```
bp<- read.csv(file = "D:/RWD Final/bp.txt", header = T, sep = "\t"); # reading the input file

boxplot(x = bp$armsys,range = 1.5, main="Arm Systolic BP") #plotting first data
boxplot(x = bp$fingsys,range = 1.5, main="Finger Systolic BP",add = T) #plotting second data set

summary(bp$fingsys) # 5 point summary for data set1
summary(bp$armsys) #5point summary for second data set

#calculating IQR
> IQR(bp$armsys)
# [1] 28.5
> IQR(bp$fingsys)
# [1] 28.5

#plotting Histogram and adding density curves for two data sets
hist(x = bp$armsys,main = "Histogram of Arm Systolic BP",xlab = "BP",ylab = "frequency", probability = T)
lines(x = density(bp$armsys),col="red")

hist(x = bp$fingsys,main = "Histogram of Finger Systolic BP",xlab = "BP",ylab = "frequency", probability = T)
lines(x = density(bp$fingsys),col="blue")

# QQ plot for the two data sets with normal distribution quantiles
qqnorm(y = bp$armsys,main = "Normal QQ Plot for Arm Systolic BP", ylab = "Armsys Quantiles");
qqline(y = bp$armsys, col="red")

qqnorm(y = bp$fingsys,main = "Normal QQ Plot for Finger Systolic BP", ylab = "Fingersys Quantiles");
qqline(y = bp$fingsys, col="blue")

# D = difference on two data points
D<-bp$fingsys-bp$armsys
Dbar<-mean(D)
Sd_D<-sd(D)

# constructing a CI for D
cp<-qt(p = 0.975, df = 199)
lower<-Dbar-(cp*Sd_D/sqrt(length(D)))
upper<-Dbar+(cp*Sd_D/sqrt(length(D)))
# CI for D
c(lower, upper)

#Normality assumption holds since D follows N(0,1) as n is large
qqnorm(D,main = "QQ Plot of D = X-Y", ylab = "Quantiles of X-Y")
qqline(D)
```

2.

```
# Construct a Single Confidence Interval for phat
conf.int <- function(sample.size, p, alpha=0.05) {
  U <- runif(sample.size); # generate Uniform (0,1)
  X <- 1*(U<=p); # Generate Bernoulli(p) from U
  phat<-mean(X); # estimated p = phat
```

```
  est.std.err <- sqrt(phat*(1-phat)/sample.size); # standard error in Phat
  ci <-phat + c(-1, 1) * qnorm(1 - (alpha/2)) * est.std.err; # CI
  return(ci);
}

#Repeat the experiment and calculate accuracy
Accuracy_Ci<-function(nsim=1000,size, phat){
  acc=rep(0,5); #accuracy data = 0
  for(i in 1:length(size)){
  cimat<-replicate(nsim,conf.int(size[i],phat)); # create nsim CIs for n=size[i] and phat
  acc[i]<-mean((phat>=cimat[1,])*(phat<=cimat[2,]))*100; # calculate the accuracy

  }
  return(acc);
}

# Sizes
size<-seq(10,200, by=15);

#Accuracy data sets fpr different p and n values
data1<-Accuracy_Ci(1000,size,0.2)
data5<-Accuracy_Ci(1000,size,0.95)
data4<-Accuracy_Ci(1000,size,0.9)
data3<-Accuracy_Ci(1000,size,0.25)
data2<-Accuracy_Ci(1000,size,0.1)

#PLotting the data sets
plot(x=size,y=data1, ylim=c(0,100),type="b",col="red",ann=F)
#par(new=TRUE) adds the next plot to the same graph
par(new=TRUE)
plot(x=size,y=data2, ylim=c(0,100),type="b",col="blue",ann=F)
par(new=TRUE)
plot(x=size,y=data3, ylim=c(0,100),type="b",col="green",ann=F)
par(new=TRUE)
plot(x=size,y=data4, ylim=c(0,100),type="b",col="black",ann=F)
par(new=TRUE)
plot(x=size,y=data5, ylim=c(0,100),type="b",col="violet",ann=F)
# Horizontal Line at 95%
abline(h = 95)
```