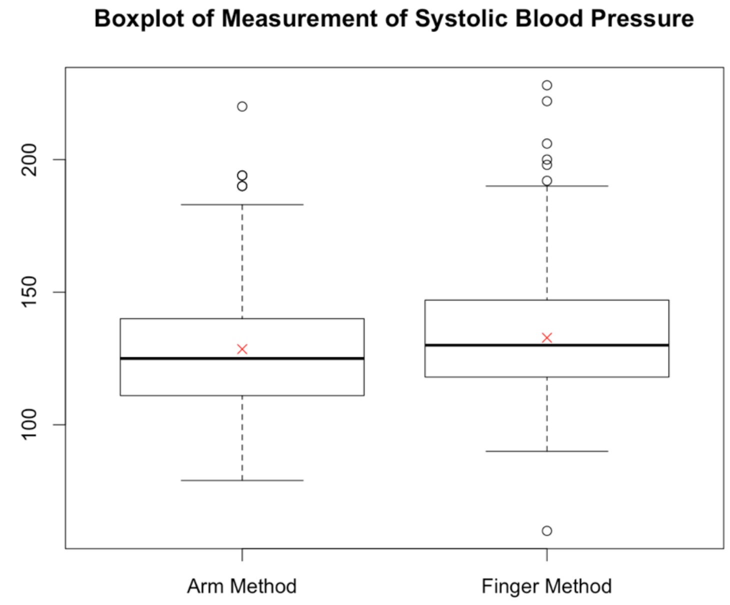
Exercise 1

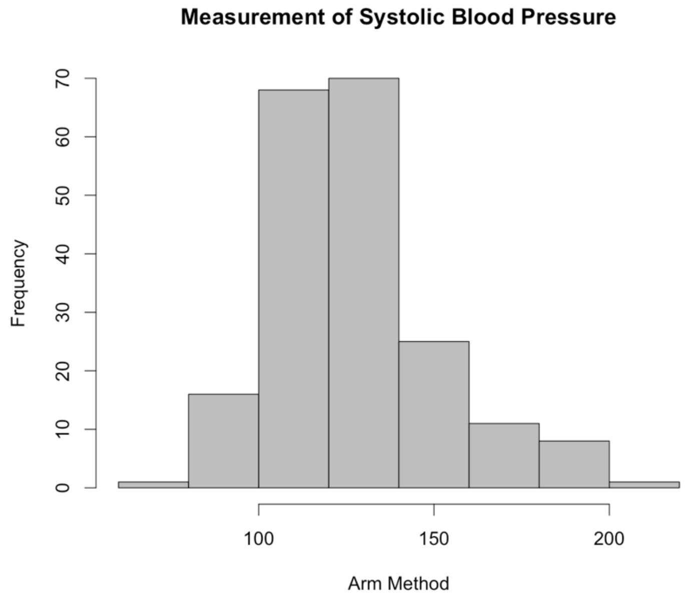
a)

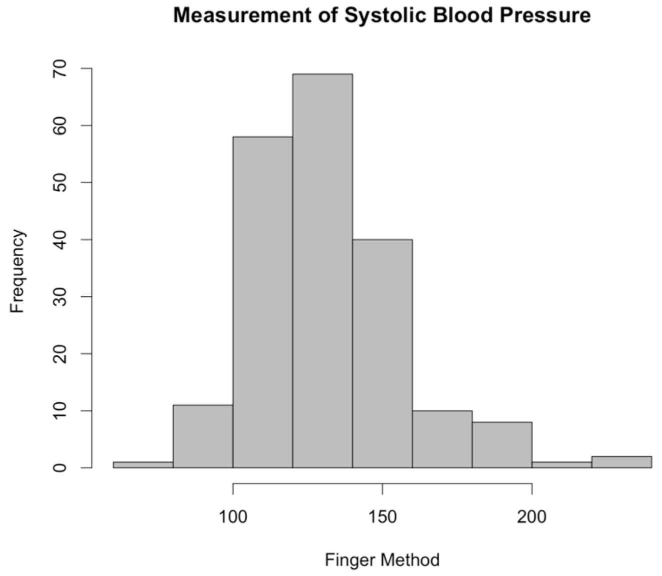


Observations:

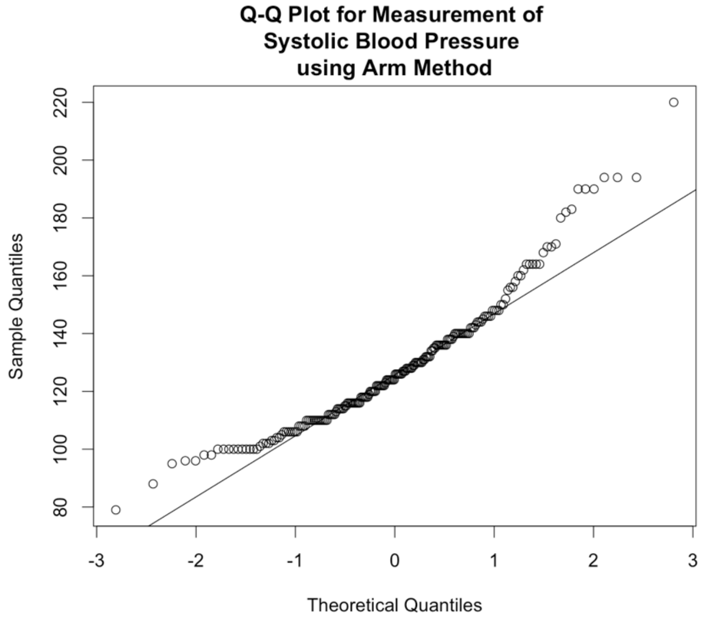
- The median of both the data sets are very close to each other, a little higher in the finger method (arm method=125, finger method=130)
- They have the same IQR.
- Finger method produces higher number of outliers.
- Both are slightly right skewed as mean is greater than the median.
- Both the data sets are distributed in a reasonably similar way but not exactly similar as they have the same skewedness, both have a similar variation, the only point being the finger method has more outliers and the median of finger method is a little higher.

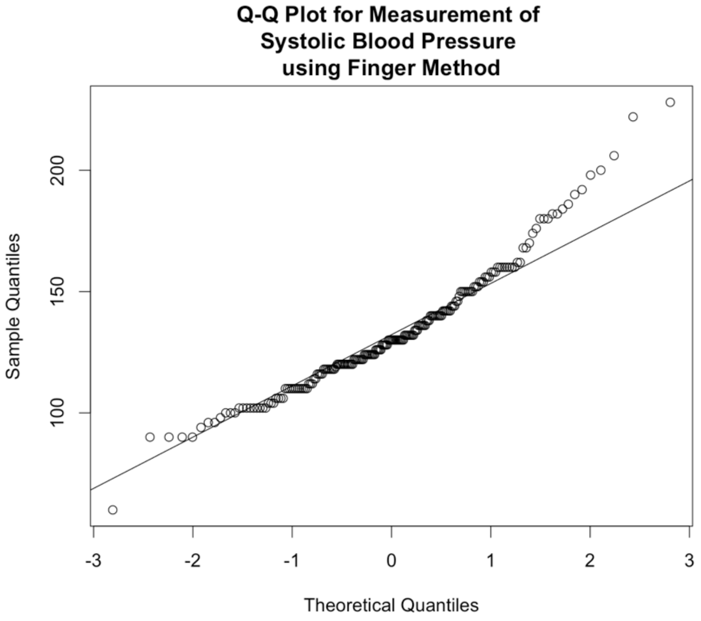
b)





Its clear from both the histograms that the data set for both, arm and finger method is right skewed. Most of the data points are in the lower half. Also, from the last two bars of the finger method we can tell there are outliers. Data from the q-q plot also shows that both the distributions are skewed. The data set follows the normal curve (represented by the straight line) very closely until the last couple of points on either extreme. The q-q plot for both the methods is right/positive skewed. The q-q plot confirms the non-normality.


Q-Q Plot for Measurement of Systolic Blood Pressure using Arm Method


Q-Q Plot for Measurement of Systolic Blood Pressure using Finger Method

c) 95% confidence interval for the difference in the means of the two methods-
Assumption: As both the distributions come from the same patient, we assume the distribution to be a paired distribution. Also, since the number of data points in the data set is 200, we assume that our paired distribution is approximately normal.
RESULT:
[1] -6.316529 -2.273471
Therefore, we can be 95% confident that the mean size for the measurement of blood pressure by arm method is between 6.31 and 2.27 mmHg lesser than the mean size by finger method.
As 0 is not included in the interval, we cannot conclude that the two methods have identical means.

d) Assumption: Since the number of data points in the data set is 200, we assume that our paired distribution is approximately normal. Therefore, we use the Z-test to compute the test statistics.
Null and alternative hypothesis:
H_0: μ(arm)- μ(finger) = 0 vs H_A: μ(arm)- μ(finger) ≠ 0
Alpha=0.05
Test Statistics:
RESULT:
[1] -4.164198
$Z\alpha/2 = 1.96$
With a two sided alternative, we reject H_0 if $|Z| \geq z_{\alpha/2}$

Here, $|Z| = 4.164198$, which is greater than $z_{\alpha/2}$, therefore we **reject** the null hypothesis. We conclude that the mean size for the measurement of blood pressure by arm method is not equal to the mean size by finger method.

e) Yes, the results are consistent. Its clear from the confidence interval that the means are not identical as 0 is not included in the interval and as $|Z| \geq z_{\alpha/2}$, we reject our null hypothesis which is the difference is mean is equal to 0, i.e., the mean of both measurements is not equal.

R code

```
sysBP <- read.table("/Users/charuarora/Documents/bp.txt",sep="\t",header=TRUE)
arm<-hist(sysBP$armsys,main="Measurement of Systolic Blood Pressure",
        xlab="Arm Method",border="black",col="grey")
fing<-hist(sysBP$fingsys,main="Measurement of Systolic Blood Pressure",
         xlab="Finger Method",border="black",col="grey")
boxplot(sysBP$armsys,sysBP$fingsys,names=c("Arm Method","Finger Method"),
      main="Boxplot of Measurement of Systolic Blood Pressure")
points(mean(sysBP$armsys),x=1,col="red",pch=4)
points(mean(sysBP$fingsys),x=2,col="red",pch=4)

qqnorm(sysBP$armsys,main="Q-Q Plot for Measurement of \nSystolic Blood Pressure \nusing Arm Method")
qqline(sysBP$armsys)
qqnorm(sysBP$fingsys,main="Q-Q Plot for Measurement of \nSystolic Blood Pressure \nusing Finger Method")
qqline(sysBP$fingsys)

metd<-sysBP$armsys-sysBP$fingsys
alpha=0.05
std.err<-sqrt(var(metd)/length(metd))
ci<-mean(metd)+c(-1,1)*qnorm(1-(alpha/2))*std.err
test.stat<-mean(metd)/std.err
p.val<-2*(1-pnorm(abs(test.stat)))
qnorm(1-(alpha/2))
```

EXERCISE 2

a) Null and alternative hypothesis:

H_0: μ=10 vs H_A: μ > 10

b) The population is normal and the sample size is 20. As the sample size is less than 30 and the population standard deviation is not known, we use a t-test with 19 (i.e., n-1) degrees of freedom to generate the test statistic.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Test Statistic formula:

RESULT:

[1] -1.974186

Null distribution of the test statistics is a t distribution with 19 degrees of freedom.

c) Test Statistic formula:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

RESULT:

```
[1] -1.974186
```

d) As our alternative is right tailed, the p value is generated using: $1 - \mathrm{pt}(\mathrm{t},19)$

RESULT:

```
[1] 0.9684606
```

e) Monte Carlo Simulations:

Approach:

First we generate random samples from a normal distribution with the given mean and standard deviation.

Compute the mean and standard deviation of the simulated data.

Calculate the tstat value for each of the simulated mean and standard deviation.

Generate the p value, which is the probability that the data is from the distribution when H0 is true.

```
nsim <- 10000
n <- 20
mu <- 10
sim_tstat <- function(n,s_mu,s_sd){
  x <- rnorm(n,s_mu,s_sd)
  tstat <- (mean(x) - mu) / (sd(x)/sqrt(n))
  pval <- 1 - pnorm(tstat,mean =10,sd=2.22)
  return (pval)
}
pval<-replicate(nsim, sim_tstat(n,9.02,2.22))
mean(pval)
```

```
p<-0
ex<-function()
{
  phat<-rt(10000,19)
  for(i in 1:length(phat))
  {
    if(phat[i]>=test.stat)
    {
      p<-p+1
    }

  }
  return(p)
}
m<-ex()/10000
```

m

The computed p value is: [1] 0.9684606
The simulated p value is: [1] 0.9999995 [1] 0.9693
As the value of nsim is increased, the simulated p value gets closer to the computed p value.

f) At 5% level of significance, our alpha is 0.05. The value generates is greater than the alpha value. Therefore, we **accept** the null hypothesis.

R code

```
n=20
smean<-9.02
ssd<-2.22
test.stat<-(smean-10)/(ssd/sqrt(n))
test.stat
pval1<-1-pt(test.stat,df=n-1)
pval1
```

EXERCISE 3
a) 95% confidence interval for the difference in mean credit limits of all credit cards issued in January 2011 and in May 2012 is:

RESULT:
[1] -302.8289 -201.1711

Therefore, we can be 95% confident that the mean credit limits of all the credit cards issued in January 2011 is between $302.8 and $201 lesser than the mean credit limits of all credit cards issued in May 2011.

b) Null and alternative hypothesis:
H_0: μ(May)=μ(Jan) vs H_A: μ(May) > μ(Jan)
Alpha=0.05

Test Statistics:
RESULT:
[1] 9.717132

Since it is a right tailed alternative, we reject H_0 if $Z \geq z_\alpha$
Here, z_α=1.644 which is less than Z, we **reject** the null hypothesis which states that the mean credit limits of all the credit cards issued in May 2011 is equal to the mean credit limits of all credit cards issued in January 2011. We do not need to perform a p test as the result is clear.

Since the number of data points in the data set is greater than 100, we assume that our distribution is approximately normal. Therefore, we use the Z-test to compute the test statistics.

R code

```
jan.mean<-2635
may.mean<-2887
jan.sd<-365
may.sd<-412
jan.n<-400
may.n<-500
alpha=0.05
std.err<-sqrt((jan.sd*jan.sd/jan.n)+(may.sd*may.sd/may.n))
std.err
ci<-jan.mean-may.mean+c(-1,1)*qnorm(1-(alpha/2))*std.err
ci
janmay.test<-(may.mean-jan.mean)/std.err
janmay.test
p.value<- (1-pnorm(janmay.test))
p.value
```