Mini Project #4
Name: Niveditha Varadha Chandrasekaran

Exercise 1

Problem Statement:

Consider the dataset stored in the file bp.xlsx. 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.

(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.

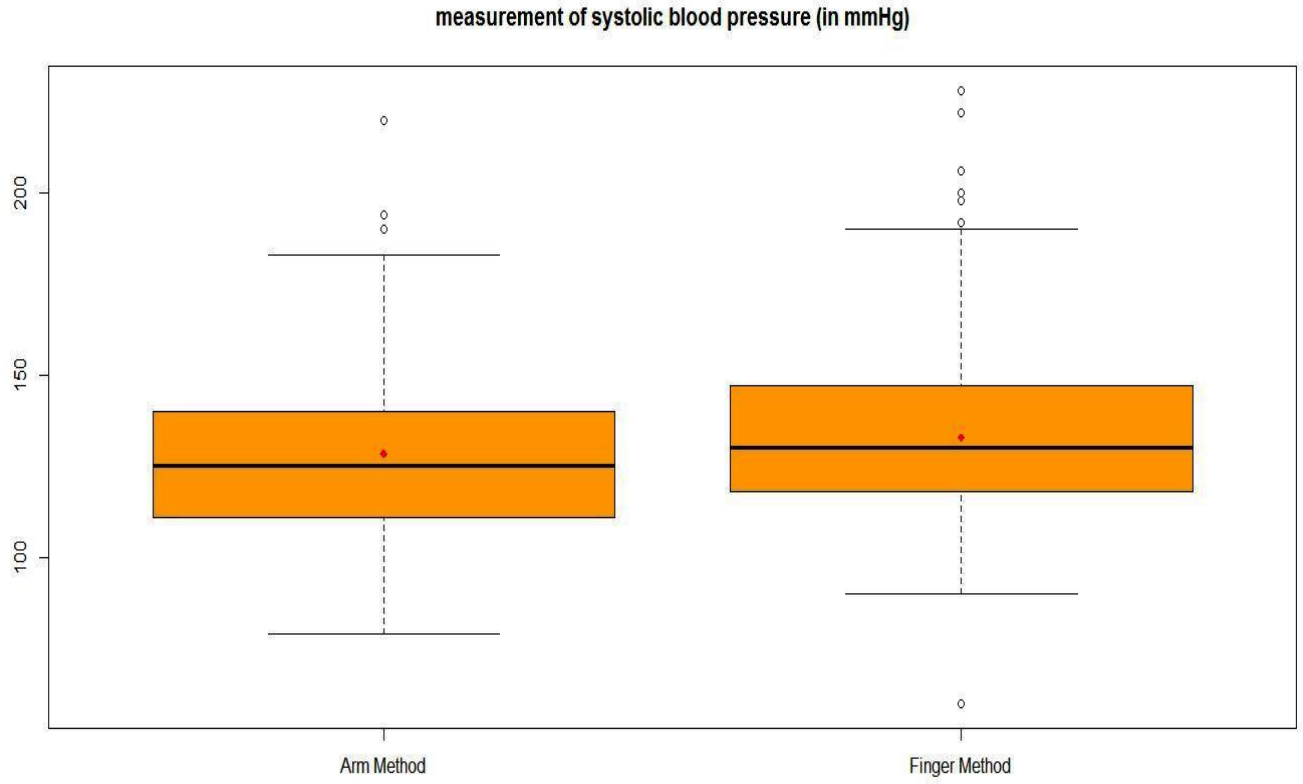
(c) Construct an appropriate 95% confidence interval for the difference in the means of the two methods. Interpret your results. Can we conclude that the two methods have identical means? Justify your answer. What assumptions, if any, did you make to construct the interval? Do the assumptions seem to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

(d) Perform an appropriate 5% level test to see if there is any difference in the means of the two methods. Be sure to clearly set up the null and alternative hypotheses. State your conclusion. What assumptions, if any, did you make to construct the interval? Do they seem to hold?

(e) Do the results from (c) and (d) seem consistent? Justify your answer.

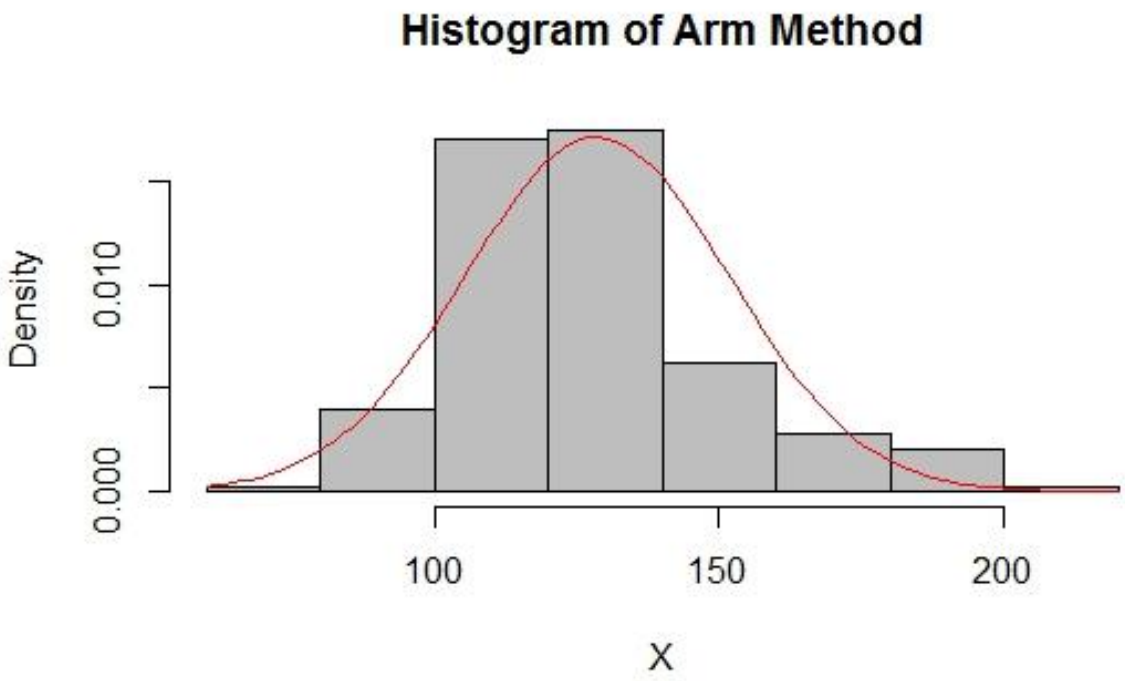
Solution:

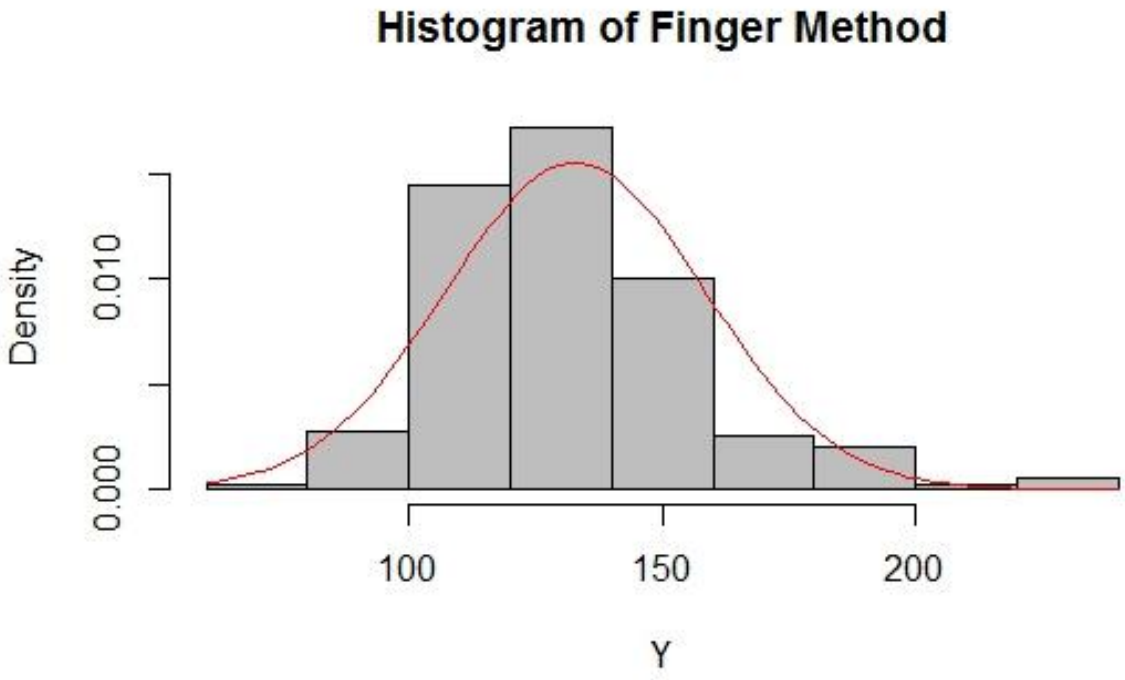
(a) Boxplot:



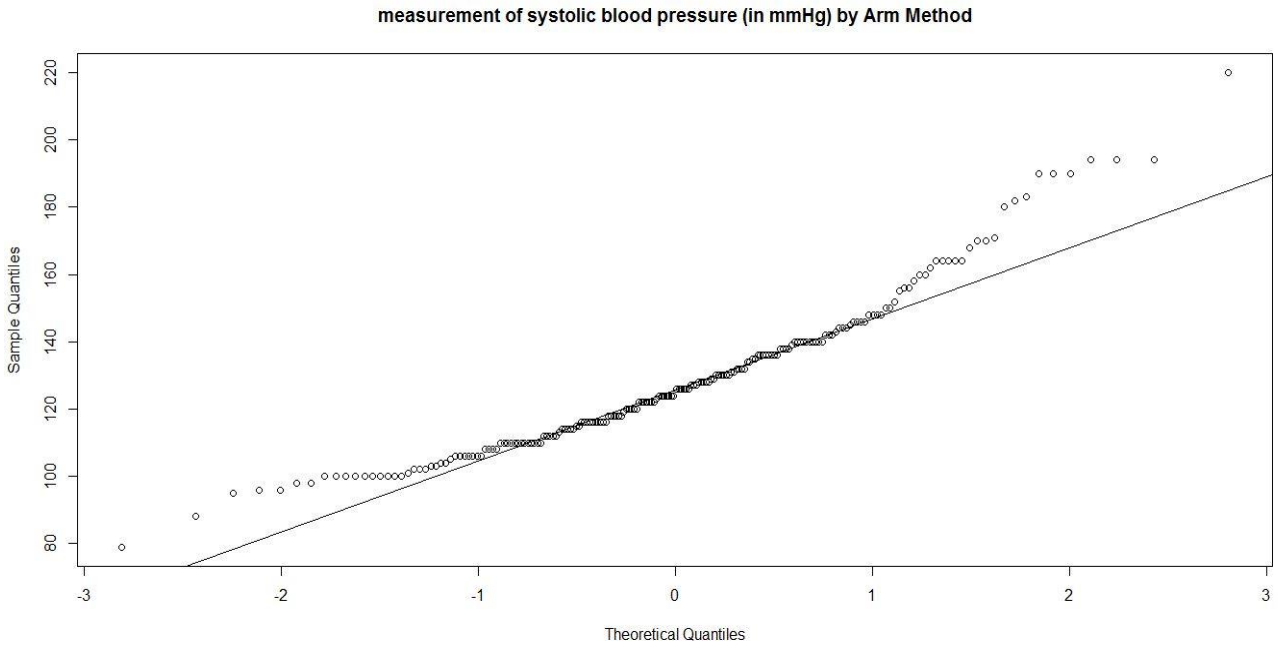
From the above boxplot, we see that for both arm and finger method the mean is slightly greater than the median. Hence we can say that the distribution of both **Arm method** and **Finger method** are **Right Skewed.** Therefore the two distributions seem to be approximately similar.

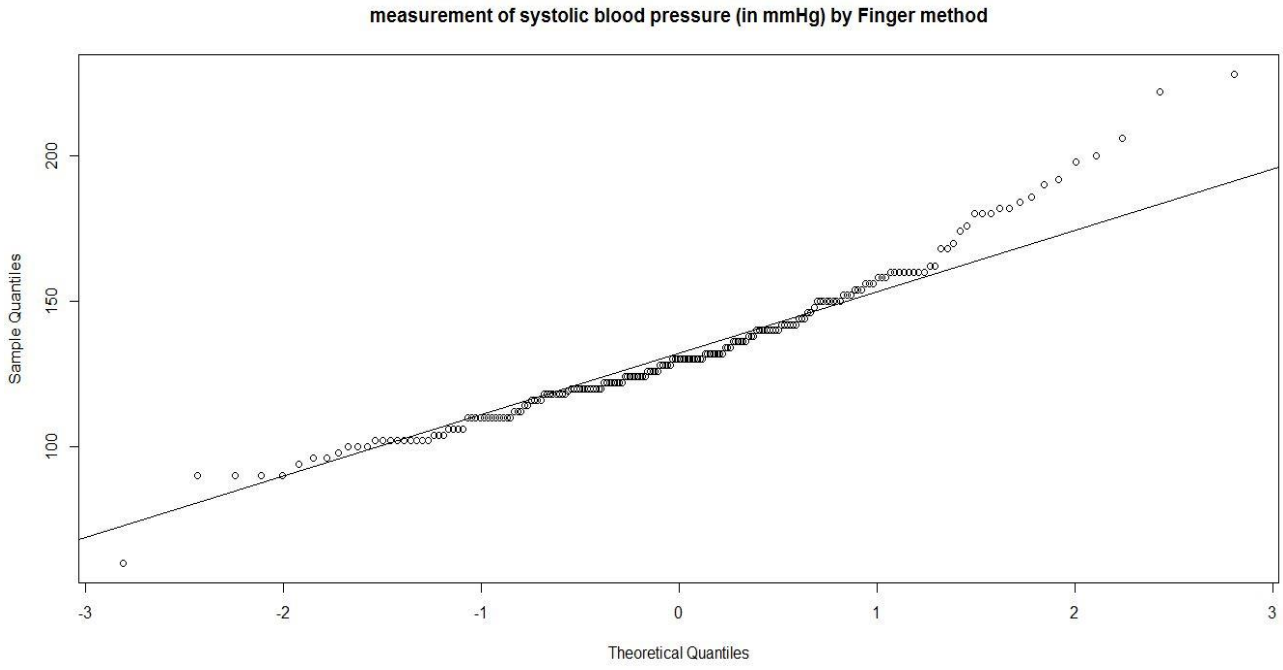
(b) Histogram:





Q-Q Plot:





From the histogram, we see that the distribution of both Arm and Finger method are **slightly right skewed.**
From Q-Q plot, we see that most of the points roughly fall on the straight line. Therefore, it seems reasonable to assume normality.

(c) Assumptions made to construct CI:

- We have paired sample and the parameter of interest is $\mu_x - \mu_y = \mu_D$ .Therefore apply one sample procedure to the differences.
- Here $n_x = n_y = n = 200$ are large (>=30). Therefore no normality assumption needed.
- σ_D is unknown, hence we use sample standard deviation (= s).
- CI is given by Dbar ± $t_{alpha/2,\ n-1}$(s/sqrt(n)).

n = 200
alpha = 1-0.95 = 0.05
Therefore the 95% CI is **(-6.328898, -2.261102)**. From the confidence interval we can see that the value 0 is not included. Therefore we cannot conclude that the two methods have identical means.

(d) Null hypothesis H_0: $\mu_x = \mu_y$
Alternative hypothesis H_1: $\mu_x \neq \mu_y$

The data are paired, and follow normal distributions with unknown variance. Therefore, we will do a paired t-test. Let $\mu_D = \mu_x - \mu_y$. Then, the above hypothesis can be reformulated as
Null hypothesis H_0: $\mu_D = 0$
Alternative hypothesis H_1: $\mu_D \neq 0$
p-value = 4.652e-05

```
        One Sample t-test

data:  D
t = -4.1642, df = 199, p-value = 4.652e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -6.328898 -2.261102
sample estimates:
mean of x
   -4.295
```

Since the p-value of 4.652e-05 is less than alpha, we reject the null hypothesis H_0. Therefore we can say that the two means are not identical.

(e) Yes the results from both (c) and (d) seem to be consistent because from the confidence interval we see that the means are not equal as the CI does not include the value 0 and from the t.test we get the p value which is less than **α** and hence we reject the null hypothesis H_0. Therefore in both cases the two means are not identical.

Exercise 2

Problem Statement:

Suppose we are interested in testing the null hypothesis that the mean of a normal population is 10 against the alternative that it is greater than 10. A random sample of size 20 from this population gives 9.02 as the sample mean and 2.22 as the sample standard deviation.

(a) Set up the null and alternative hypotheses.
(b) Which test would you use? What is the test statistic? What is the null distribution of the test statistic?
(c) Compute the observed value of the test statistic.
(d) Compute the p-value of the test using the usual way.
(e) Estimate the p-value of the test using Monte Carlo simulation. How do your answers in (d) and (e) compare? (8 points)
(f) State your conclusion at 5% level of significance.

Solution:

(a) Null hypothesis H_0: $\mu = 10$
Alternative hypothesis H_1: $\mu > 10$

(b) Given:
The distribution of the population is normal
N = 20 (< 30) and population standard deviation is unknown.
Null hypothesis H_0: $\mu = 10$
Alternative hypothesis H_1: $\mu > 10$

Therefore the test statistic is:

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

The null distribution of the test statistic is T – distribution with 19 (n-1) as the degrees of freedom.

(c) the test statistic is:

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

X_bar = 9.02, n = 20, μ_0 = 10, s = 2.22
Therefore the observed value of the test statistic is -1.974186

(d) p-value = 1- pt(t,n-1)
p-value = 0.9684606

(e) The p-value of the test using Monte Carlo simulation is: (Refer R code at the end)

S.No	P-value
1	0.9347639
2	0.9028192
3	0.9123771
4	0.9158883

From the p-values obtained using Monte Carlo simulation it is evident that:

- The p-values obtained from the simulation are nearly equal to value obtained from (d) which is greater than α(= 0.05). The p-value obtained from the simulation is consistent with the value obtained in (d).
- Hence we accept the null hypothesis.

(f) alpha = 0.05
From the previous step we see that p-value = 0.968460641
Therefore p-value is greater than **α**. Hence we accept the null hypothesis (H_0: μ = 10) and there is statistically no significant evidence that the mean is greater than 10.

Exercise 3

Problem Statement:

According to the credit rating agency Equifax, credit limits on newly issued credit cards increased between January 2011 and May 2011. Suppose that random samples of 400 credit cards issued in January 2011 and 500 credit cards issued in May 2011 had average credit limits of $2635 and $2887, respectively. Suppose that the sample standard deviations of these two samples were $365 and $412, respectively.

(a) Construct an appropriate 95% confidence interval for the difference in mean credit limits of all credit cards issued in January 2011 and in May 2011. Interpret your results. Be sure to justify your choice of the interval.
(b) Perform an appropriate 5% level test to see if the mean credit limit of all credit cards issued in May 2011 is greater than the same in January 2011. Be sure to specify the hypotheses you are testing, and justify the choice of your test. State your conclusion.

Solution:

(a) Given:
μ_x = 2635
μ_y = 2887
S_x = 365
S_y = 412
n_x = 400
n_y = 500

Assumptions made to construct CI:

- We have two independent samples and the parameter of interest is $\mu_x - \mu_y$.

- Here n_x and n_y are large (>=30). Therefore no normality assumption needed.
- σ_x and σ_y is unknown, hence we use sample standard deviation.
- No assumptions on σ^2_x and σ^2_y. They may be equal or unequal.

CI is given by:
Xbar- Ybar + c (-1, 1) * qnorm(1 - (α /2)) * sqrt((S_x^2/n_x) + (S_y^2/n_y))

Therefore 95% CI for the difference in mean credit limits of all credit cards issued in January 2011 and in May 2011 is **(-302.8289, -201.1711)**

Since the CI is below 0, and the difference in the sample mean (-252) is contained in the CI, this shows that the mean credit limits of all credit cards issued in May 2011 is greater than the mean credit limits of all credit cards issued in January 2011.

(b) Null hypothesis H_0: $\mu_x = \mu_y$
Alternative hypothesis H_1: $\mu_x < \mu_y$
As n_x and n_y value is greater than 30 we are choosing Z -test with unknown variances.
Therefore test statistic and p-value is given by:
zstat <- (xbar-ybar)/sqrt((s_x^2/nx) + (s_y^2/ny))
pval <- (pnorm(zstat))

zstat = -9.717132
p-value = 1.274297e-22

Since p-value = 1.274297e-22 is lesser than **α,** we reject the null hypothesis H_0 and accept the Alternative hypothesis (H_1: $\mu_x < \mu_y$) which means that that the mean credit limits of all credit cards issued in May 2011 is greater than the mean credit limits of all credit cards issued in January 2011.

R Code:

Exercise 1:

```
> library(readxl)
> #Read data from the excel.
> df<-read_excel("C:\\Users\\Niveditha\\Desktop\\bp.xlsx")
> X<-df$armsys
> Y<-df$fingsys
> means = c(mean(X),mean(Y))
> median(X)
[1] 125
> median(Y)
[1] 130
```

```
> #boxplots of the two methods
> boxplot(X,Y,names=c("Arm Method","Finger Method"),
+         main="measurement of systolic blood pressure (in mmHg)",
+         col="orange")
> points(means,col="red",pch=18)
>
> #Histogram for the two methods
> hist(X, probability = TRUE, col = "grey", main="Histogram of Arm Method")
> curve(dnorm(x, mean=mean(X), sd=sd(X)), ylab="Density", xlab="X",col = "red", add = TRUE)
> hist(Y, probability = TRUE, col = "grey", main="Histogram of Finger Method")
> curve(dnorm(x, mean=mean(Y), sd=sd(Y)), ylab="Density", xlab="X",col = "red", add = TRUE)
>
> #Q-Q plots for the two methods
> qqnorm(X,main="measurement of systolic blood pressure (in mmHg) by Arm Method")
> qqline(X)
> qqnorm(Y,main="measurement of systolic blood pressure (in mmHg) by Finger method")
> qqline(Y)
>
> #Finding the 95% CI
> D <- X-Y
> n <-length(D)
> alpha = 1-0.95
> CI<-mean(D)+c(-1,1)*qt(1-(alpha/2),df=n-1)*(sd(D)/sqrt(n))
> CI
[1] -6.328898 -2.261102
>
> #5% level test
> t.test(D, alternative = "two.sided",conf.level = (1 - alpha))

        One Sample t-test

data:  D
t = -4.1642, df = 199, p-value = 4.652e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -6.328898 -2.261102
sample estimates:
mean of x
   -4.295
```

Exercise 2:

```
> xbar<-9.02
> mu<-10
```

```
> sd<-2.22
> n<-20
> #test statistic
> teststat<-(xbar-mu)/(sd/sqrt(n))
> teststat
[1] -1.974186
> #p-value
> pval= 1-pt(teststat,n-1)
> pval
[1] 0.9684606
```

Monte carlo simulation:

```
> mu<-10
> sd<-2.22
> n<-20
> psim=function(){
+   x=rnorm(20,9.02,2.22)
+   tobs= (mean(x)- mu)/(sd(x)/sqrt(n))
+   pval<-1-pt(tobs,n-1)
+   return(pval)
+ }
> #Monte Carlo simulation
> pvalsim1=mean(replicate(10,psim()))
> pvalsim1
[1] 0.9347639
> pvalsim2=mean(replicate(100,psim()))
> pvalsim2
[1] 0.9028192
> pvalsim3=mean(replicate(1000,psim()))
> pvalsim3
[1] 0.9123771
> pvalsim4=mean(replicate(10000,psim()))
> pvalsim4
[1] 0.9158883
```

Exercise 3:

```
> #finding the 95% CI
> xbar= 2635
> ybar= 2887
> s_x= 365
> s_y= 412
> nx= 400
```

```
> ny= 500
> alpha =1-0.95
> ci <- xbar-ybar+ c(-1, 1) * qnorm(1 - (alpha/2)) * sqrt((s_x^2/nx) + (s_y^2/ny))
> ci
[1] -302.8289 -201.1711
>
> #test Statistic
> alpha= 0.05
> zstat <- (xbar-ybar)/sqrt( (s_x^2/nx) + (s_y^2/ny))
> zstat
[1] -9.717132
>
> #p-value
> pval <-  (pnorm(zstat))
> pval
[1] 1.274297e-22
```