Project 3 **Section: CS 6313.501**

Name : Panchami G. Rudrakshi

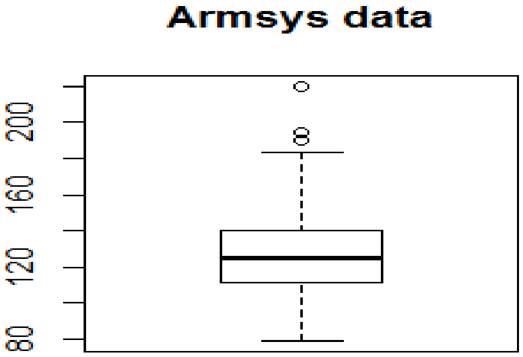
Exercise 1

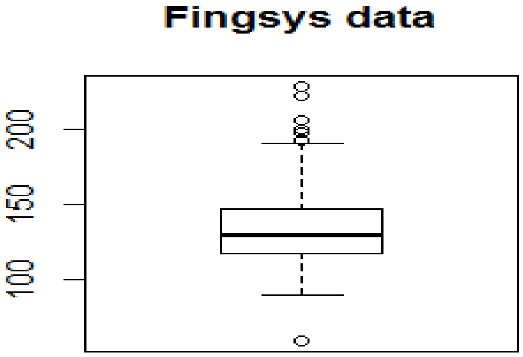
Consider the dataset stored in the file bp.txt. This dataset contains one measurement of systolic blood pressure (in mmHg) made by each of two methods—a finger method and an arm method—from the same 200 patients.

(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.

Steps:

1) Read armsys from bp.text and store it in armsys_data
2) Read fingsys from bp.text and store it in fingsys_data
3) Plot boxplots for the two.
4) Data can be analysed using summary.





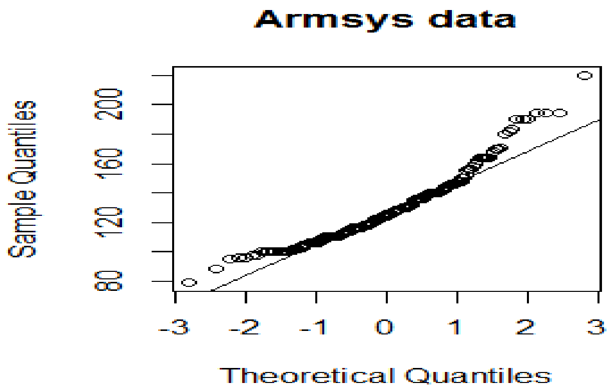
```
> summary(armsys_data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   79.0   111.5   125.0   128.5   140.0   220.0

> summary(fingsys_data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   118.0   130.0   132.8   146.5   228.0
```

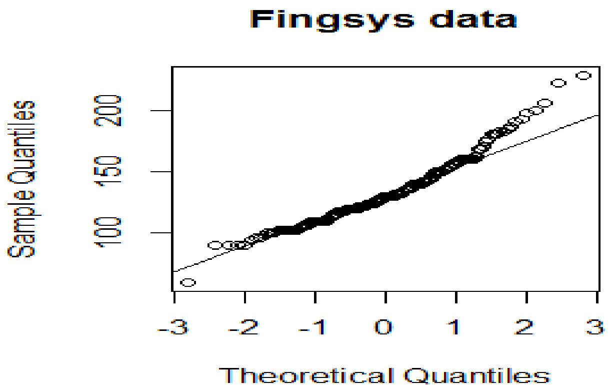
From the boxplot we observe that :

- Armsys has less outliers as compared to Fingsys.
- Also when calculated, the distributions have same inter-quartile range (28.5).
- Median of both the data sets are very close to each other, a little higher in fingsys method. Armsys method has 125 and Finger method has 130.0

Thus, it can be said that both the distributions are similar since both are right skewed.

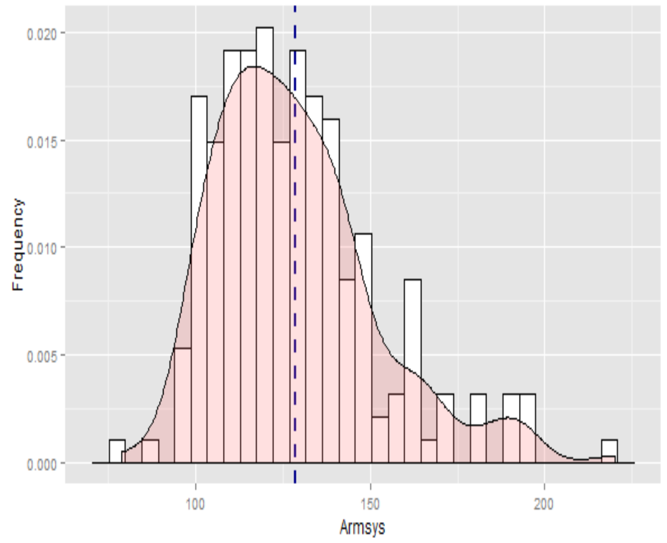
(b) Use histograms and QQ plots to examine the shapes of the two distributions. Comment on what you see. Does the assumption of normality seem reasonable? Justify your answer.

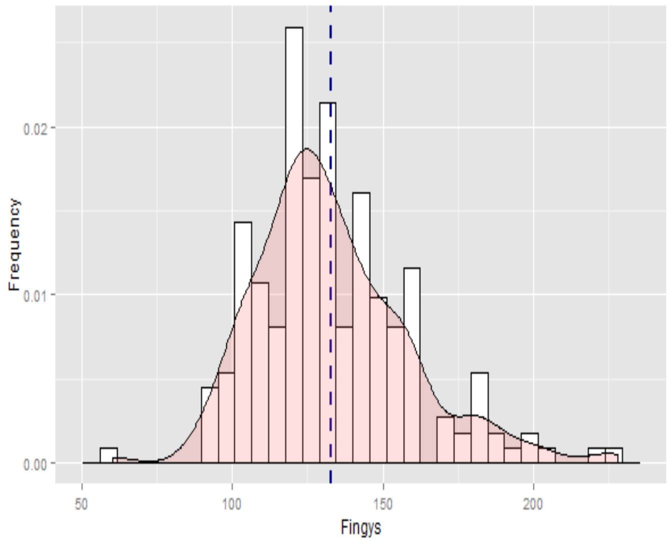




Observations from QQ-Plot :

- Data from the QQ-Plot shows that both the distributions are skewed as there are points from the data set that deviate from the normal curve at either of the extremes.
- Also we can observe that the data points are clustered at the initial part and hence darker and as we move towards the end of the curve the points are more distributed and hence it can assumed to be right skewed.





Observation from Histogram :

- From the histograms we can see that both the distributions are right skewed as most of the data points lies on the lower half.
- Also from the outmost bars in the histogram we can say that there are outliers .

Conclusion :
From the above observations of QQ-plot and Histogram we can conclude that the distributions are not normal.

(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?

CI: -9.0961939 0.5061939

- Since 0 is included in the interval, we can conclude that two methods have identical means.
- It is observed that the sample size is large i.e. greater than 30, and thus the distribution follows an approximately normal distribution according to law of large numbers.
- Yes, the assumption holds according to Law of Large Numbers.

(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?

Null Hypothesis : H_0 = armsys_mean - fingsys_mean =0
Alternate Hypothesis : H_A = armsys_mean - fingsys_mean != 0 (Not equal to)

```
zstat = (armsys_mean -fingsys_mean )/(sqrt((armsys_var/n.armsys_data) +
(fingsys_var/n.fingsys_data)))
```

```
zstat
[1] -1.753323
```

```
pval =  2*(1-pnorm(abs(zstat)))
[1] 0.07954652
```

The p-value obtained is 0.07 which is greater than alpha ie 0.05 thus we accept the null hypothesis.
Thus there is no significant difference in the means of the two methods.

Assumptions made : Since n is large we are performing a z-test

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

From (c) we know that the confidence interval is -9.0961939 0.5061939 which includes 0.
From (d) we know that the p-value is 0.07954652 which is greater than alpha.
Thus, we see that both the means are approximately same. So the Null Hypothesis is accepted and the results calculated are consistent.

Exercise 2 :

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.
Null hypothesis H_0 : mu = 10
Alternative hypotheses H_A : mu > 10

(b) Which test would you use? What is the test statistic? What is the null distribution of the test statistic?

Since variance is unknown and the size of population sample is small i.e. 20 in this case , we use T test.

Test statistics : $t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$ = 9.02 – 10 / (2.22 / √20)

where $\overline{x}$ is the sample mean
s is the sample standard deviation of the sample
n is the sample size
Degrees of freedom used in the test are $n - 1$.

Null distribution of the test statistic :
Since the population is normally distributed (given), the null distribution is normal with n-1 degrees of freedom.

(c) Compute the observed value of the test statistic.

```
   Test statistic =  9.02 - 10 / ( 2.22 / √20 ) = -1.974186
```

(d) Compute the p-value of the test using the usual way.

pval = 1 – pnorm (tstat) is used as we are checking the right side of the curve

```
p-value = 1 - pnorm (-1.974186)
        = 1- 0.02418029   =  0.9684606
```

(e) Estimate the p-value of the test using Monte Carlo simulation. How do your answers in (d) and (e) compare?

Monte Carlo simulation is done by generating 20 random number with mean : 9.02 and sd :2.22.
The simulation is replicated different no of times and the observation is made.
When run 100 times, p-val =   0.9298571
1000 times, p-val =   0.9104149
10000 times, p-val = 0.914396
The p-values obtained from the simulation are nearly equal to value obtained from (d) which is greater than 0.05. Thus, the null hypothesis holds true.
Also, the p-value obtained from the simulation is consistent with the value obtained from (d).

(f) State your conclusion at 5% level of significance.
Alpha = 0.05 (level of significance)
P-value = 0.9758197
If p-value > alpha, accept H$_0$
If p-value < alpha, reject H$_0$
In this case, `p-value > alpha (0.9684606 > 0.05 )`, therefore accept H$_0$ i.e. the mean of the given normal population is 10.

Exercise 3 :
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.

Given :
Since sample size is large the estimator **theta_hat = X_bar-Y_bar** is approximately Normal by the Central Limit Theorem.
1-alpha=0.95

N=400	X_bar=2635	sigma_X = 365
M=500	Y_bar=2887	sigma_Y = 412

Confidence interval for the difference of means with known standard deviations:

$$\boxed{\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

Since the difference in mean is negative (-252), credit limits on newly issued credit cards increased between January 2011 and May 2011.
Margin of error : **50.82888**
The CI obtained : **-302.8289 -201.1711**

- Thus there are 95% chances that difference in mean credit limits of all credit cards issued in January 2011 and in May 2011 i.e. -252 is within 1.96 times the standard deviation of the distribution of the true mean.
- Based on this interval, we can conclude that there is statistically significant difference in mean of the credit limits issued between January 2011 and May 2011 as the confidence interval does not include null value.
- The confidence interval is a range of likely values for the difference in means. Since the interval does not contain zero, we have sufficient evidence to conclude that there is a difference.

(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.

Given :
alpha=0.05

N=400	X_bar=2635	sigma_X = 365
M=500	Y_bar=2887	sigma_Y = 412

H_0 = x_mu - y_mu = D
H_A = x_mu - y_mu > D

Test statistics = $$\frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

Since the no. of data points in the data set is greater than 100, we assume that the distribution is approximately normal. Therefore we use the Z test to compute the test statistic.

Conclusion:
Since p-value(0) <0.05, we reject the null hypothesis i.e. mean credit limit of all credit cards issued in May 2011 is lesser than the same in January 2011.

CODE:

1)

```
# Read data (only column armsys) from bp.txt and store in armsys_data
armsys_data <- read.table(file.choose(), header=T)[,c('armsys')]
print(armsys_data)                                  #print armsys_data
# Read data  (only column fingsys) from bp.txt and store in fingsys_data
fingsys_data <- read.table(file.choose(), header=T)[,c('fingsys')]
print(fingsys_data)                                 #print fingsys _data

##Question a
par(mfrow=c(1,2))                          # plots boxplots side by side
boxplot(armsys_data)                       # plots boxplot for armsys
boxplot(fingsys_data)                      # plots boxplot for fingsys
summary(armsys_data)                # gives five_point summary of armsys
summary(fingsys_data)               # gives five_point summary of fingys

##Question b
qqnorm(armsys_data, main=’Armsys data');     #to plot QQ plot for Armsys
qqline(armsys_data)

qqnorm(fingsys_data, main='Fingsys data'); #to plot Q-Q plot for Fingsys
qqline(fingsys_data)

library(ggplot2)                              #import library for ggplot
data=read.table(file.choose(), header=T)      # read the bp.txt file

p<-ggplot(data,aes(armsys_data))+geom_histogram(aes(y=..density..),
colour="black", fill="white") + geom_vline(aes(xintercept=
mean(armsys_data, na.rm=T)),color="darkblue", linetype="dashed",size=1)+
geom_density(alpha=.2, fill="#FF6666")+ xlab("Armsys")+ylab("Frequency")
p

q<-ggplot(data,aes(fingsys_data ))+
geom_histogram(aes(y=..density..), colour="black", fill="white") +
geom_vline(aes(xintercept=mean(fingsys_data,na.rm=T)), color="darkblue",
linetype="dashed", size=1)+ geom_density(alpha=.2, fill="#FF6666") +
xlab("Fingys")+ylab("Frequency")
q

##Question c
n.armsys_data = length(armsys_data)            # length of Armsys's data
n.fingsys_data = length(fingsys_data)         # length of Fingsys's data
armsys_var=var(armsys_data)                  # variance of Armsys's data
fingsys_var=var(fingsys_data)               # variance of Fingsys's data


armsys_mean = mean(armsys_data)                  # mean of Armsys's data
armsys_mean                             # printing Armsys's dat mean
fingsys_mean = mean(fingsys_data)               # mean of Fingsys's data
fingsys_mean                          # printing Fingsys's data mean
se = sqrt((armsys_var/n.armsys_data)+(fingsys_var/n.fingsys_data))
se                                               # printing the SE value
# using qnorm function as n is 200 ->law of large numbers
ci = (armsys_mean - fingsys_mean) + c(-1,1)*qnorm(1-(alpha/2))*se
ci
```

## Question d

```
alpha=0.05                                                    #given
# 5% level alpha test
zstat = (armsys_mean -fingsys_mean )/(sqrt((armsys_var/n.armsys_data) +
(fingsys_var/n.fingsys_data)))
zstat
pval <-  2*(1-pnorm(abs(zstat)))
pval
```

2)

```
mu=10
n=20
sample_mean =9.02                                            # given
sample_sd = 2.22                                             # given

tstat = (sample_mean -mu )/(sample_sd/sqrt(n)    # finds test statistics
tstat                                           # prints test statistics

pval = 1 – pt(tstat,n-1) # finds p-value for right side of curve
pval                                                # prints test p-value
```

b)

```
pv=function(n)     # function for monte carlo simulation to find p-value
{
  a=rnorm(n,sample_mean,sample_sd)
  a
  b=mean(a)
  b
  c=sd(a)
  c
  tstat1=(b-mu)/(c/sqrt(n))
  pval=1-pt(tstat1,n-1)
  return(pval)
}
nsim=100                                  # no of simulations(varies)
pvalue=replicate(nsim,pv(n))     # replicates the value nsim no of times
pvalue                                                  # prints p-value
p_mean=mean(pvalue)                                         # finds mean
p_mean                                                      # print mean
```

3 a)

```
alpha=0.05                                                        # given
N=400                                                             # given
X_bar=2635                                                        # given
sigma_X = 365                                                     # given
M=500                                                             # given
Y_bar=2887                                                        # given
sigma_Y = 412                                                     # given
true_mean = X_bar-Y_bar                                     # to find mean
true_mean                                                    # prints mean
# to find confidence interval
ci=Y_bar-X_bar+ c(-1, 1) *qnorm(1 - (alpha/2)) * sqrt((sigma_X^2/N) +
(sigma_Y^2/M))
ci                                         # to print confidence interval
```

3 b)

```
alpha=0.05                                                          # given
N=400                                                               # given
X_bar=2635                                                          # given
sigma_X = 365                                                       # given
M=500                                                               # given
Y_bar=2887                                                          # given
sigma_Y = 412                                                       # given
# finds test statistics
zstat = (Y_bar -X_bar)/(sqrt((sigma_X^2/N) + (sigma_Y^2/M)))
zstat
# find p-value for the right side of the curve
pval = 1 - pnorm(abs(zstat))
pval                                                        # print p-value
```