Code

Name: Nikhil H

Load Packages

Hide

```
library(readxl)
library(dplyr)
library(car)
```

Question 1:

Data

Import the Average length of stay data and preparing it for analysis

Hide

```
# This is a chunk for your Data section.

##Import the dataset and skip the first 12 rows which contains description
alos <- read_excel("average-length-of-stay-multilevel-data.xlsx", skip = 12)

## Interesting variables -  Local Hospital Network (LHN), alos.
## Filter only the above coulmns and Local Hospital Network =South Western Sydney and store it in alos
_var
alos_var <- alos %>% dplyr::select('Local Hospital Network (LHN)','Average length of stay (days)') %>%
filter(alos$`Local Hospital Network (LHN)`== 'South Western Sydney')


##The datatype of attribue Average Length of Stay is character, we need to change the datatype
## to numeric
alos_var$`Average length of stay (days)`<- as.numeric(alos_var$`Average length of stay (days)`)
```

Summary Statistics

Calculate descriptive statistics (i.e., mean, median, standard deviation, first and third quartile, interquartile range, minimum and maximum values) of the Average Length of Stay for South Western Sydney hospitals .

Hide

```
# This is a chunk for your Summary Statistics section.
#Summarise  alos_var %>% summarise(Min = min(`Average length of stay
(days)`,na.rm = TRUE),                   Q1 =quantile(`Average length of stay
(days)`,probs = .25,na.rm = TRUE),
                  Median = median(`Average length of stay (days)`,na.rm = TRUE),
                  Q3 = quantile(`Average length of stay (days)`,probs = .75,na.rm = TRUE),
                  Max = max(`Average length of stay (days)`,na.rm = TRUE),
Mean = mean(`Average length of stay (days)`,na.rm = TRUE),
SD = sd(`Average length of stay (days)`,na.rm = TRUE),                   n =
n(),
                  Missing = sum(is.na(`Average length of stay (days)`)))
```

```
## The number of missing values in the summary is 128, we will remove the missing values
##remove the NA or missing values
alos_var <- alos_var[!(is.na(alos_var$`Average length of stay (days)`)),]
## check the sum of missing values it should be zero
sum(is.na(alos_var))
```

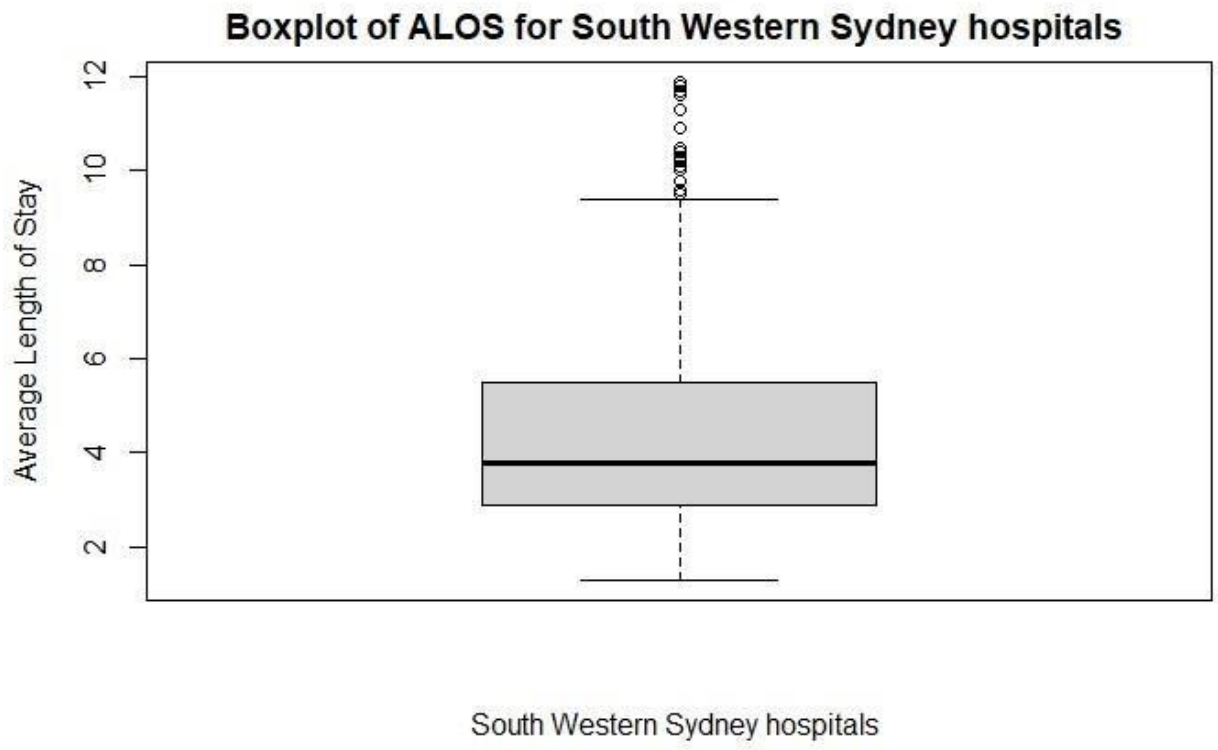
```
[1] 0
```

Check for Outliers

Plot a boxplot to check for outliers. There are some rare cases where the patients have stayed for longer time in the South Western Sydney hospitals. These cases will be the outliers as shown in the below boxplot. We will not remove the outliers as we want to consider these rare cases as well, which would help us in our analysis to determine the Average length of stay for South Western Sydney hospitals.

Hide

```
boxplot(alos_var$`Average length of stay (days)` ~ alos_var$`Local Hospital Network (LHN)`, ylab="Aver
age Length of Stay",xlab = "South Western Sydney hospitals " ,main = "Boxplot of ALOS for South Wester
n Sydney hospitals ")
```



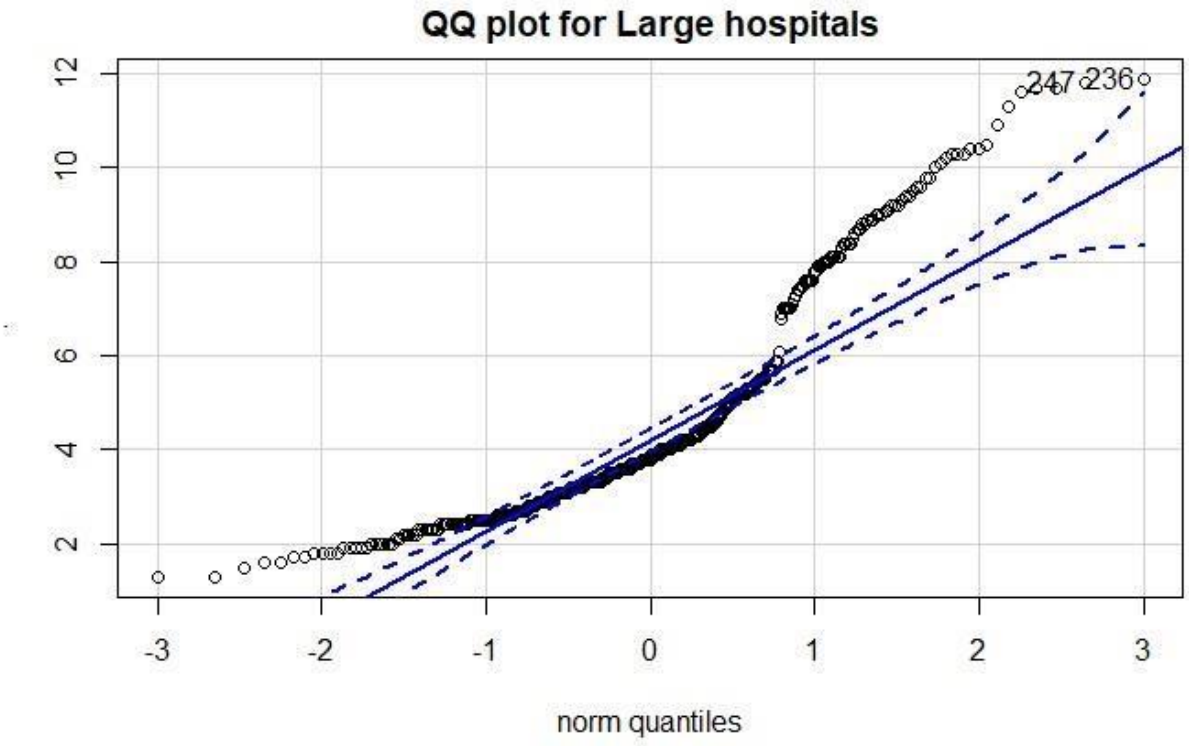
Check the normality of the data: Use the QQ plot to check the normality. The blue dashed lines of QQ plot correspond to 95% CI for the normal quantiles. The data points should fall inside the blue lines for the data to be normal.

Conclusion: From the below QQ plot We can conclude that the data is not normally distributed as the data points are falling outside the blue dashed line. But as the sample size (n=371) is greater than 30, we can assume that the sampling distribution of the mean will follow a normal distrinution based on the CLT(Central Limit Theorem).

Hide

```
alos_var$`Average length of stay (days)` %>% qqPlot(dist="norm",main="QQ plot for Large hospitals")
```

```
[1] 236 247
```



Hypothesis test - One-sample t-test:

The one-sample t-test is used to test whether there is evidence taken from a sample mean to suggest that the population mean is different to a previously assumed value. The one-sample t-test assumes the data are normally distributed and the population standard deviation is unknown.

Assumptions: Significance level = 0.05

H0 : mu = 4.5 HA

: mu != 4.5

The two tail test will be applied on the Sample as the sample can be either greater than or less than 4.5.

Conclusion: We assumed normality as n > 30. 95% CI of difference between means [4.394867 4.891926] p-value = 0.2573

Decision: Fail to Reject H0 The result of one-sample T-test found that the mean of Alos is not statistically significant as pvalue > 0.05 with confidence interval 95% and CI captures the expected mean = 4.5

Hide

```
t.test(alos_var$`Average length of stay (days)`, mu = 4.5, alternative = "two.sided")
```

==========

```
        One Sample t-test

data:  swshospital$`Average length of stay (days)`
t = 1.1346, df = 370, p-value = 0.2573
alternative hypothesis: true mean is not equal to 4.5
95 percent confidence interval:
 4.394867 4.891926
sample estimates:
mean of x
 4.643396
```

QUESTION 2)

Import the dataset using the read_csv function

Hide

```
tutorial <- read_csv("Assignment4b-3.csv")
```

Summary of the before tutorial data

Hide

```
tutorial %>%summarise(Min = min(`Score before tutorial`,na.rm = TRUE),
            Q1 =quantile(`Score before tutorial`,probs = .25,na.rm = TRUE),
            Median = median(`Score before tutorial`,na.rm = TRUE),
            Q3 = quantile(`Score before tutorial`,probs = .75,na.rm = TRUE),
            Max = max(`Score before tutorial`,na.rm = TRUE),
            Mean  =  mean(`Score  before  tutorial`,na.rm  =
TRUE),              SD = sd(`Score before tutorial`,na.rm =
TRUE),              n = n(),
            Missing = sum(is.na(`Score before tutorial`)))
```

```
  [38;5;246m# A tibble: 1 x 9 [39m
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  [3m [38;5;246m<dbl> [39m [23m [3m [38;5;246m<dbl> [39m [23m  [3m [38;5;246m<dbl> [39m [23m
  [3m [38;5;246m<dbl> [39m [23m [3m [38;5;246m<dbl> [39m [23m [3m [38;5;246m<dbl> [39m [23m [3 m
[38;5;246m<dbl> [39m [23m [3m [38;5;246m<int> [39m [23m   [3m [38;5;246m<int> [39m [23m
  [38;5;250m1 [39m    13    27     37    44    55  35.8  10.5  [4m1 [24m290       0
```

Hide

```
## There are no missing values and the mean is 35.8
```

Summary of the after tutorial data

Hide

```
tutorial %>%summarise(Min = min(`Score after tutorial`,na.rm = TRUE),
            Q1 =quantile(`Score after tutorial`,probs = .25,na.rm = TRUE),
            Median = median(`Score after tutorial`,na.rm = TRUE),
            Q3 = quantile(`Score after tutorial`,probs = .75,na.rm = TRUE),
            Max = max(`Score after tutorial`,na.rm = TRUE),
            Mean  =  mean(`Score  after  tutorial`,na.rm  =
TRUE),              SD = sd(`Score after tutorial`,na.rm =
TRUE),              n = n(),
            Missing = sum(is.na(`Score after tutorial`)))
```

```
□[38;5;246m# A tibble: 1 x 9□[39m
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  □[3m□[38;5;246m<dbl>□[39m□[23m □[3m□[38;5;246m<dbl>□[39m□[23m  □[3m□[38;5;246m<dbl>□[39m□[23m  □
□[3m□[38;5;246m<dbl>□[39m□[23m □[3m□[38;5;246m<dbl>□[39m□[23m □[3m□[38;5;246m<dbl>□[39m□[23m□[3 m  □
□[38;5;246m<dbl>□[39m□[23m □[3m□[38;5;246m<int>□[39m□[23m  □[3m□[38;5;246m<int>□[39m□[23m  □
□[38;5;250m1□[39m    33    37     41    44    55  41.2  4.95  □[4m1□[24m290       0
```

Find the difference between the scores and the Summary: Find the difference between the scores of before tutorial and after tutorial.The paired-samples t-test assumes that these differences, d, are normally distributed.

Hide

```
##mutate the columns
tutorial <- tutorial %>% mutate(d = `Score after tutorial` - `Score before tutorial`)
## summary for difference
tutorial %>% summarise(Min = min(d, na.rm = TRUE),
                 Q1 = quantile(d, probs = .25, na.rm = TRUE),
                 Median = median(d, na.rm = TRUE),
                 Q3 = quantile(d, probs = .75, na.rm = TRUE),
                 Max = max(d, na.rm = TRUE),
                 Mean = mean(d, na.rm = TRUE),
                 SD = sd(d, na.rm = TRUE),
                 n = n(),
                 Missing = sum(is.na(d)))
```

```
□[38;5;246m# A tibble: 1 x 9□[39m
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  □[3m□[38;5;246m<dbl>□[39m□[23m □[3m□[38;5;246m<dbl>□[39m□[23m  □[3m□[38;5;246m<dbl>□[39m□[23m
□[3m□[38;5;246m<dbl>□[39m□[23m □[3m□[38;5;246m<dbl>□[39m□[23m □[3m□[38;5;246m<dbl>□[39m□[23m □[3
m□[38;5;246m<dbl>□[39m□[23m □[3m□[38;5;246m<int>□[39m□[23m  □[3m□[38;5;246m<int>□[39m□[23m
□[38;5;250m1□[39m   -□[31m14□[39m     0      0    14    31  5.35  10.0  □[4m1□[24m290       0
```

Hide

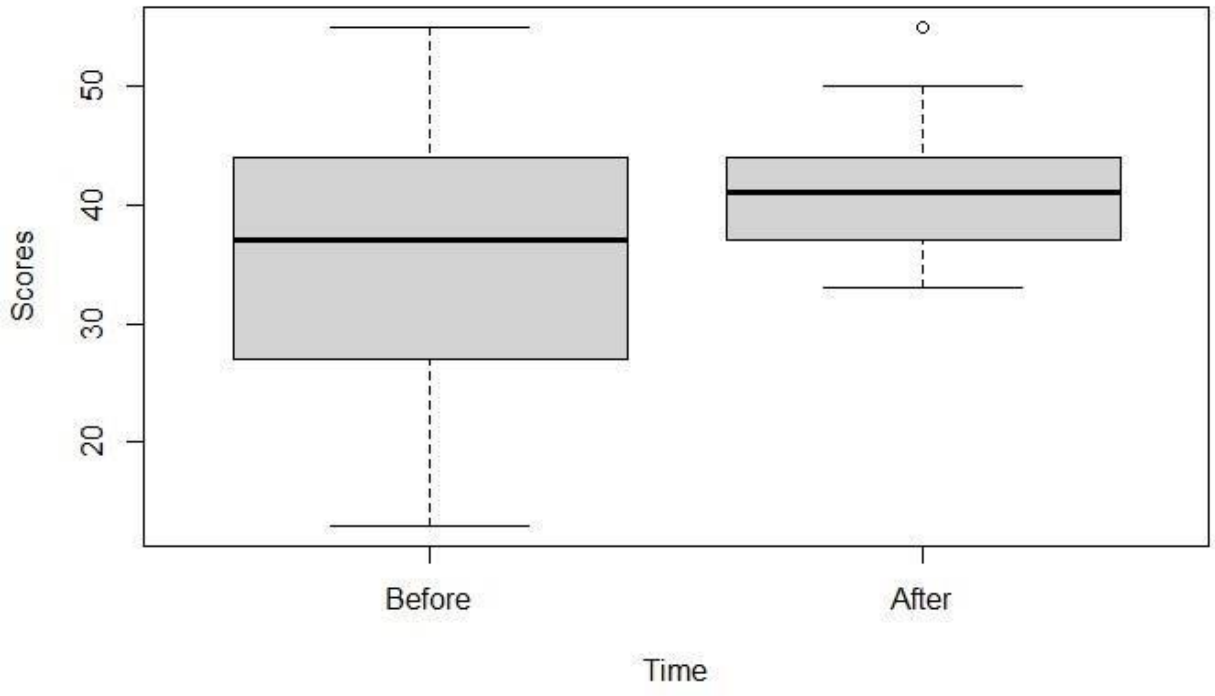
```
## There are no missing values
```

Paired Samples t-test Visulaisation: Plot a side by side boxplot to check for outliers and to check the dependency.

Conclusion: From the below boxlpot we can conclude that the there has been increase in the score of the students after the tutorial. We can get more information about this from the hypothesis test.

Hide

```
boxplot(
  tutorial$`Score before tutorial`,
  tutorial$`Score after tutorial`,
  ylab = "Scores",
  xlab = "Time"
)
axis(1, at = 1:2, labels = c("Before", "After"))
```



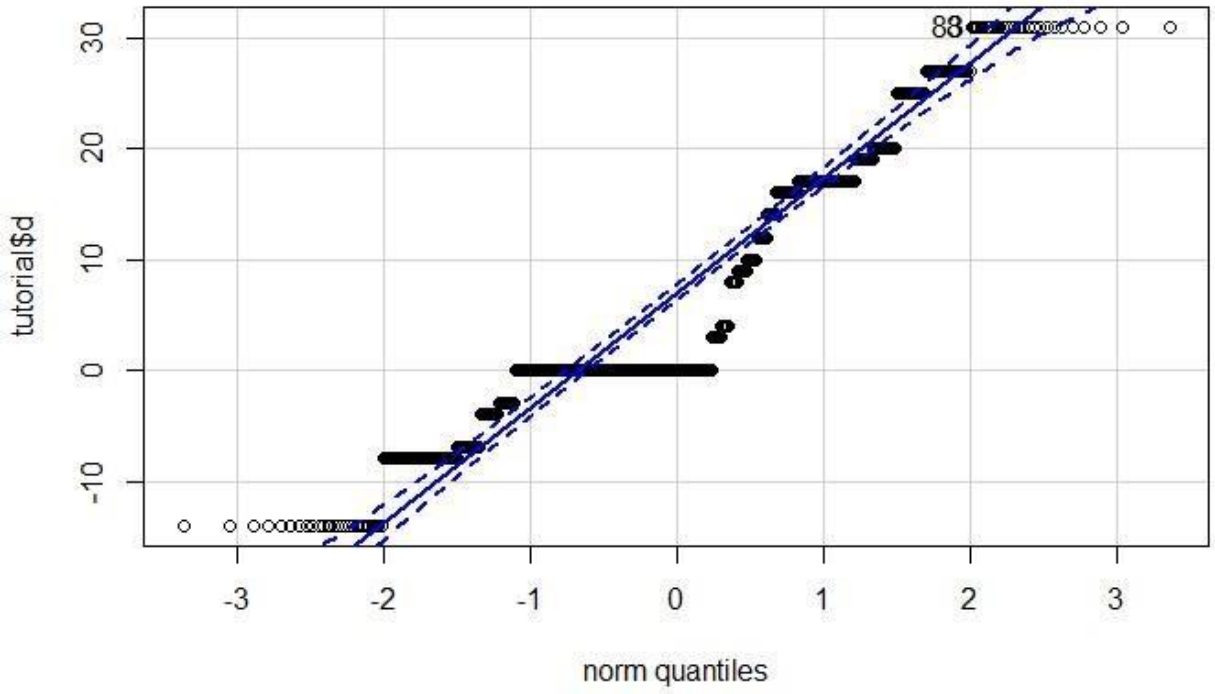
Plot the QQplot: Use the QQ plot to check the normality. The blue dashed lines of QQ plot correspond to 95% CI for the normal quantiles. The data points should fall inside the blue lines for the data to be normal.

Conclusion: From the below QQ plot We can conclude that the data is not normally distributed as the data points are falling outside the blue dashed line. But as the sample size (n=1290) is greater than 30, we can assume that the sampling distribution of the mean will follow a normal distrinution based on the CLT(Central Limit Theorem).

Hide

```
qqPlot(tutorial$d, dist = "norm")
```

```
[1]  8 83
```



Paired two Sample t-test:

Assumptions: Significance level = 0.05

H0 : mu = 0

HA : mu != 0

The two tail test will be applied on the Sample.

Hide

```
t.test(tutorial$`Score after tutorial` , tutorial$`Score before tutorial`, paired =TRUE,
       alternative = "two.sided")
```

```

    Paired t-test

data:  tutorial$`Score after tutorial` and tutorial$`Score before tutorial`
t = 19.144, df = 1289, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.802104 5.898671
sample estimates:
mean of the differences
              5.350388
```

Conclusion: We assumed normality as n > 30. 95% CI of difference between means [4.802104 5.898671] p-value = 2.2e16

Decision: Reject H0 Hence, there is a statistical significant difference between the Average length of stay for admitted patients at Large hospitals and medium hospitals.

Interpretation: A paired-samples t-test was used to test for a significant mean difference between scores of the students before and after tutorial.The mean difference of scores following tutorial was found to be 5.35 (SD = 10.0). Visual inspection of the Q-Q plot of the difference scores suggested that the data were approximately normally distributed. The paired-samples t-test found a statistically significant mean difference between the scores of students before and after tutorial, t(df=1289)=19.144, p<2.2e-16, 95% [4.802104 5.898671].