

Assignment 4 : Applied Analytics

Yogesh Haresh Bojja (s3789918)

Question 1:

INTRODUCTION

We have been provided with the data from Australian institute of health and welfare which calculates Average length of stay(ALOS). This data includes different 'Reporting unit' and its corresponding 'Reporting unit type' located across different states in Australia. Each reporting unit is categorised into a different 'Peer group'. This data shows the 'Number of overnight stays' and 'Total overnight patients bed' in the 'Time period' of 1 year in each 'Category' of the reporting unit. Average length of stay(ALOS) is calculated as the 'Total overnight patients bed' divided by the 'Number of overnight stays'.

PROBLEM STATEMENT

It is been claimed that patients in South western sydney have an Average Length of stay(ALOS) of 4.5 days. We are going to test this hypothesis. After loading the data we will handle the missing values then we will remove the outliers present. We are going to check the normality by Q-Q plot. After doing the hypothesis tests by different approaches we will infer whether the claim is statistically significant or not. We are going to achieve this at 0.05 significance level.

DATA

1) Loading and subsetting data

Data is taken from Australian institute of health and welfare. Data is loaded with read_excel() function from readxl package. We are storing the data in df variable. We subset the dataframe by considering only observations of South Western Sydney for ALOS.

Hide

```
library(readxl)
df <- read_excel("average-length-of-stay-multilevel-data.xlsx", skip = 12)
df <- subset(df, df$`Local Hospital Network (LHN)`== "South Western Sydney")
df <- df[, "Average length of stay (days)"]
```

2) Handling missing value

We can see below that 'NP' and '-' are unacceptable values hence we need to remove them. We eliminate observations with NP value by subsetting it. For further analysis we want to convert our dataframe to numeric hence we cast the dataframe to numeric by as.numeric(). After casting '-' values become NA so we eliminate those by subsetting the dataframe with the help of is.na(). After handling them completely we check the count of NA values in our dataframe which is displayed to 0.

Hide

```
levels(factor(df$`Average length of stay (days)`))
```

```
[1] "-"      "1.30"  "1.50"  "1.60"  "1.70"  "1.80"  "1.90"  "10"    "10.10" "10.20" "10.30" "10.40" "10.50" "10.90" "11.30"
[16] "11.60" "11.70" "11.80" "11.90" "2"      "2.10"  "2.20"  "2.30"  "2.40"  "2.50"  "2.60"  "2.70"  "2.80"  "2.90"  "3"
[31] "3.10"  "3.20"  "3.30"  "3.40"  "3.50"  "3.60"  "3.70"  "3.80"  "3.90"  "4"      "4.10"  "4.20"  "4.30"  "4.40"  "4.50"
[46] "4.60"  "4.70"  "4.80"  "4.90"  "5"      "5.10"  "5.20"  "5.30"  "5.40"  "5.50"  "5.60"  "5.70"  "5.80"  "5.90"  "6.10"  "6.80"
[61] "6.90"  "7"      "7.10"  "7.20"  "7.30"  "7.40"  "7.50"  "7.60"  "7.80"  "7.90"  "8"      "8.10"  "8.30"  "8.40"  "8.50"
[76] "8.60"  "8.70"  "8.80"  "8.90"  "9"      "9.10"  "9.20"  "9.30"  "9.40"  "9.50"  "9.60"  "9.70"  "9.80"  "9.90"  "NP"
```

Hide

```
df <- df[df$`Average length of stay (days)`!="NP",]
df$`Average length of stay (days)` <- as.numeric(df$`Average length of stay (days)`)
```

NAs introduced by coercion

Hide

```
df <- df[!is.na(df$`Average length of stay (days)`),]
cat("\nAfter handling : ")
```

After handling :

Hide

```
sum(is.na(df$`Average length of stay (days)`))
```

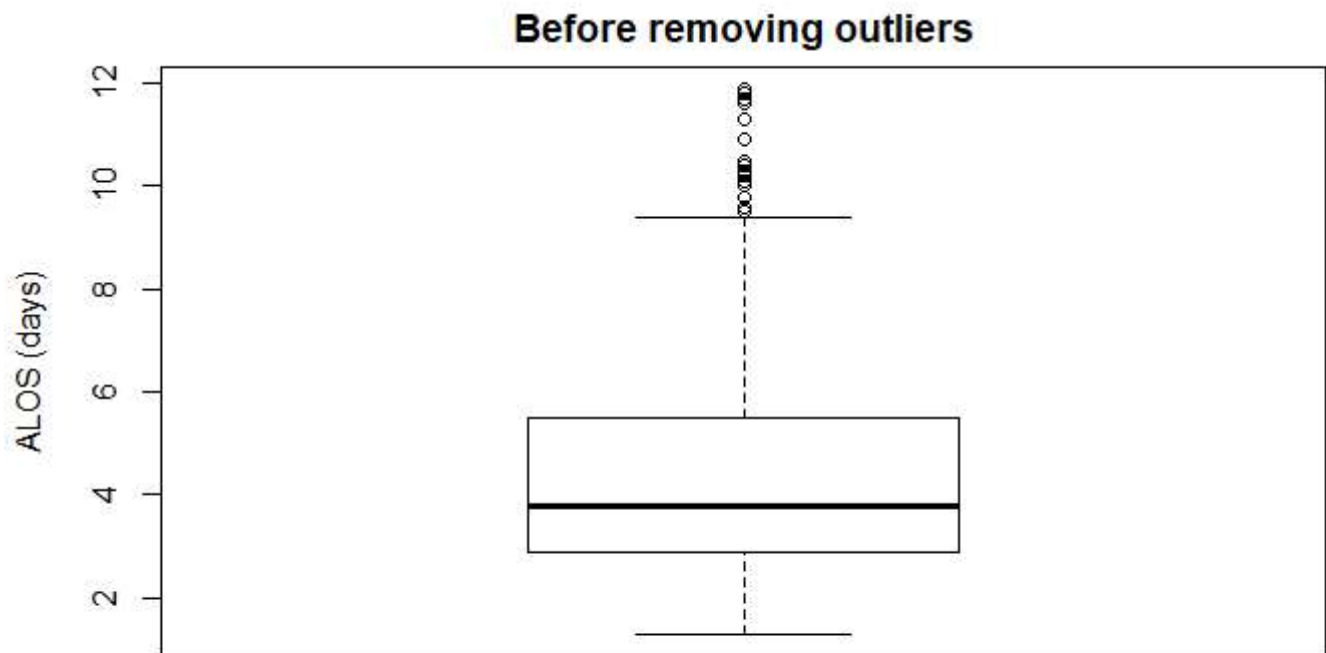
```
[1] 0
```

3) Removing Outliers

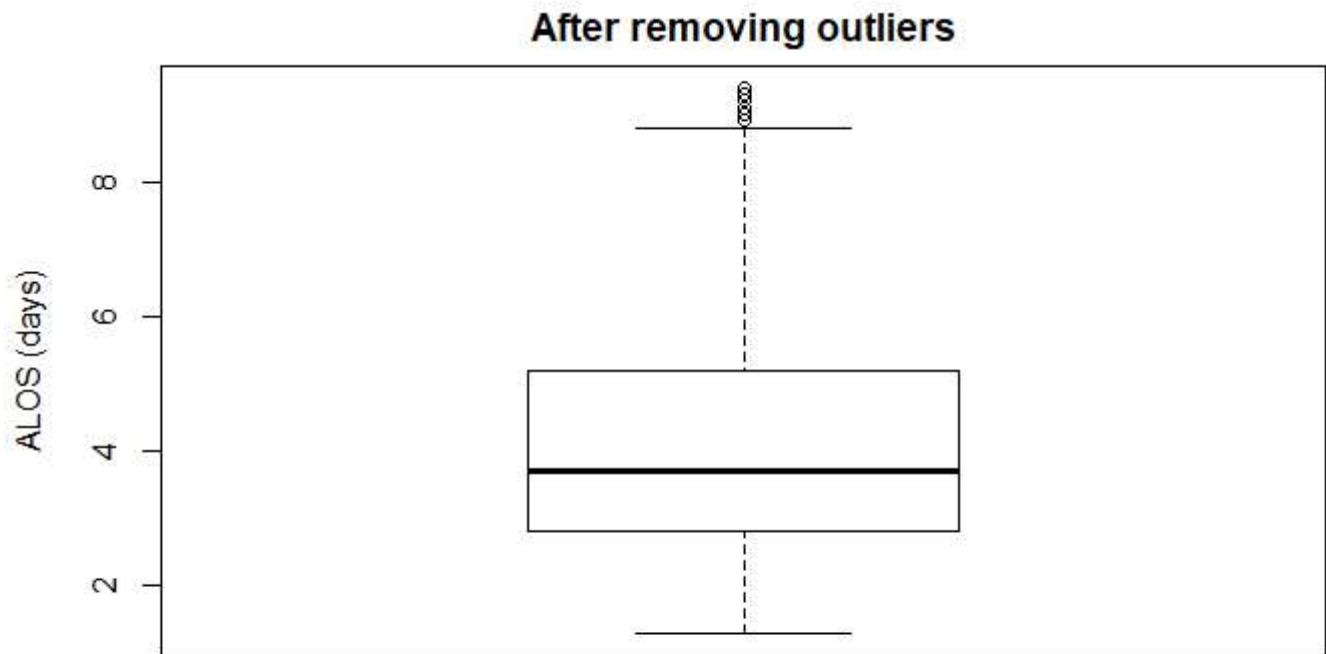
We can visualize the summary of data by boxplot. After plotting we see that there are outliers present in our dataset which might influence the hypothesis testing to give wrong results. Hence we need to eliminate them. The efficient method to eliminate them is by removing the values which fall beyond $(Q3 + IQR \times 1.5)$ and $(Q1 - IQR \times 1.5)$. From the first boxplot it could be inferred that outliers are present after ALOS = 8, hence we need to remove the values falling beyond the upper limit. Therefore we check whether values fall beyond $(IQR \times 1.5 + Q3)$, if yes then we remove them.

Hide

```
boxplot(df$`Average length of stay (days)`, main = "Before removing outliers", ylab = "ALOS (days)")
```

[Hide](#)

```
iqr = IQR(df$`Average length of stay (days)`)
q3 <- quantile(df$`Average length of stay (days)`, prob=0.75, na.rm = TRUE)
limit_max <- q3 + 1.5*iqr
df <- df[df$`Average length of stay (days)`<=limit_max,]
boxplot(df$`Average length of stay (days)`, main = "After removing outliers", ylab = "ALOS (days)")
```



DESCRIPTIVE STATISTICS AND VISUALIZATION

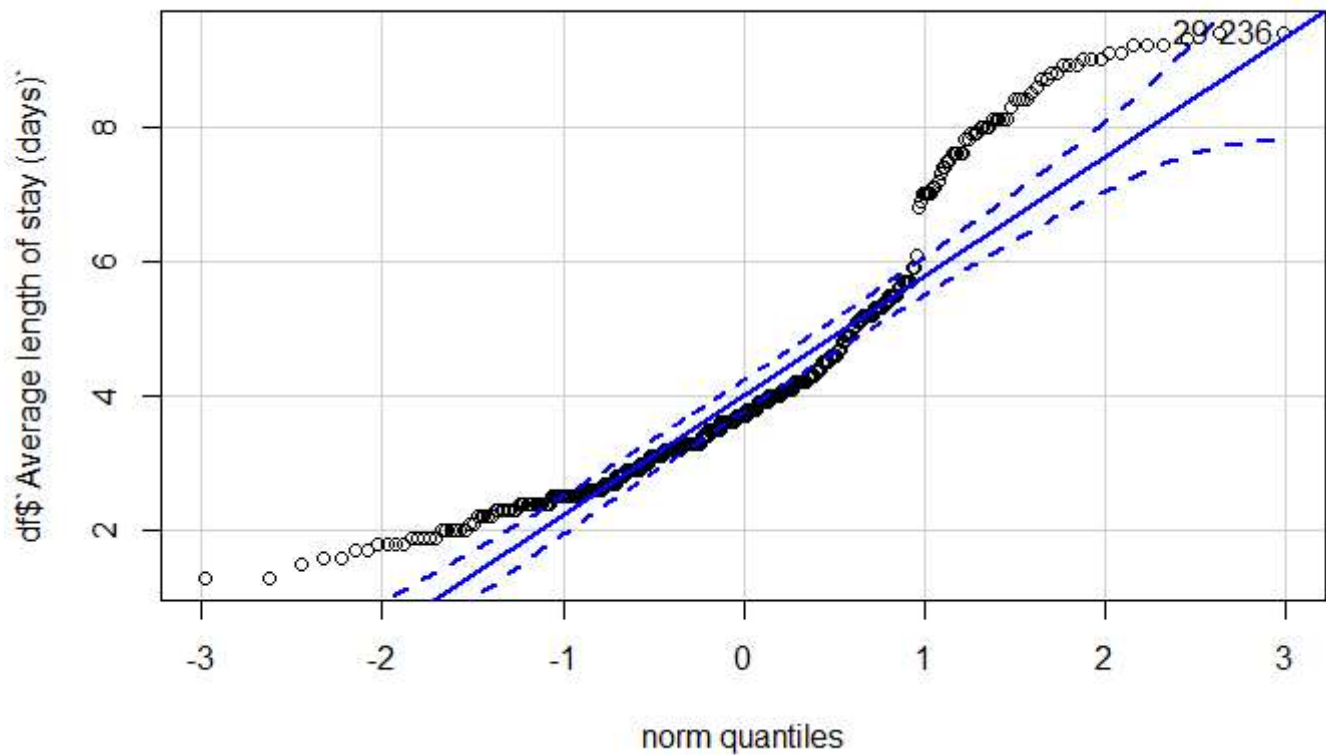
1) Analyzing normality by Q-Q Plot

When data is less than 30 we need to check the normality of dataset. Hence we plot Q-Q plot. If all points falls on the diagonal then the data is said to be normal. If the curve is S shaped then the data is not normally distributed. From the plot below we can say that data is not normally distributed. But Central Limit theorem tells us that if the dataset is large i.e $n > 30$ then we can assume the data is normally distributed irrespective of the the underlying populations distribution. Hence we can assume our data tends to act normally distributed. Therefore we can perform T-test.

[Hide](#)

```
library(car)
qqPlot(df$`Average length of stay (days)`, dist = "norm")
```

```
[1] 29 236
```



2) Descriptive Statistics

Descriptive statistics is about the attributes which describes our data for example how our data is spread, what is its average etc. Mean is calculated by `mean()`. Standard deviation is calculated by `sd()`. `length()` gives the how many values does our dataset contain. Descriptive statistics can also be found by `summary()`.

[Hide](#)

```
mean <- mean(df$`Average length of stay (days)`)
sd <- sd(df$`Average length of stay (days)`)
n <- length(df$`Average length of stay (days)`)
summary(df$`Average length of stay (days)`)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.300	2.800	3.700	4.289	5.200	9.400

[Hide](#)

```
cat("Mean : ",mean)
```

```
Mean : 4.288571
```

[Hide](#)

```
cat("\nStandard deviation : ",sd)
```

Standard deviation : 2.003975

HYPOTHESIS TESTING

As it is claimed that ALOS of patients from South Western Sydney is 4.5, our null and alternative hypothesis will be as following.

$$H_0 : \mu = 4.5$$

$$H_A : \mu \neq 4.5$$

Hence we will be testing the above hypothesis with one-sampled two tailed t-test. Level significance is 0.05.

1) Testing by critical value :

Here we are going to test the hypothesis by critical value approach.

[Hide](#)

```
t_stat <- (mean-4.5)/(sd/sqrt(n))
t_crit_lower <- qt(0.05/2, n-1, lower.tail = TRUE)
t_crit_upper <- qt(0.05/2, n-1, lower.tail = FALSE)
cat("T-statistics : ",t_stat)
```

T-statistics : -1.97381

[Hide](#)

```
cat("\nlower critical value : ",t_crit_lower)
```

lower critical value : -1.966785

[Hide](#)

```
cat("\nupper critical value : ",t_crit_upper)
```

upper critical value : 1.966785

We can see that t-statistics does not fall in the range of lower critical value and upper critical value. T-statistics falls in the rejection region. Therefore we can conclude that we fail to accept null hypothesis by CI approach.

2) Testing by p-value

Here we are going to test the hypothesis by p-value approach.

if $p < \alpha$, Reject H_0

if $p > \alpha$, Fail to reject H_0

[Hide](#)

```
p_of_tstatistics <- 2*pt(q=t_stat, n-1, lower.tail = TRUE)
cat("P-value : ",p_of_tstatistics)
```

P-value : 0.04919164

P-value is 0.04919 which is less than the area of significance i.e. 0.05, hence we fail to accept the null hypothesis by p-value approach.

3) Testing by CI

Here we are going to test the hypothesis by CI approach.

If the 95% CI does not capture H_0 , reject H_0

If the 95% CI capture H_0 , Fail to reject H_0

Hide

```
CIupper <- mean + abs(t_crit_upper)*(sd/sqrt(n))
CIlower <- mean - abs(t_crit_upper)*(sd/sqrt(n))
cat("CI lower bound : ",CIlower)
```

CI lower bound : 4.077895

Hide

```
cat("\nCI upper bound : ",CIupper)
```

CI upper bound : 4.499247

Hide

```
t.test(df$`Average length of stay (days)`, conf.level = 0.95, mu = 4.5, alternative = "two.sided")
```

One Sample t-test

```
data: df$`Average length of stay (days)`
t = -1.9738, df = 349, p-value = 0.04919
alternative hypothesis: true mean is not equal to 4.5
95 percent confidence interval:
 4.077895 4.499247
sample estimates:
mean of x
 4.288571
```

CI range [4.077, 4.499] does not include the mean of null hypothesis which is 4.5, hence we fail to accept null hypothesis by CI approach too.

DISCUSSION

A two-tailed, one-sample t-test was used to determine if the ALOS of patients from South Western Sydney was significantly different from the previously ALOS of 4.5. The 0.05 level of significance was used. The sample's mean ALOS was $M = 4.2885$, $SD = 2.003$. The results of the one-sample t-test found the mean ALOS to be statistically significantly lower than the previous claim, $t(349) = -1.9738$, $p = 0.04919$, 95% CI $[4.0778, 4.4992]$.

Question 2:

INTRODUCTION

We are provided with the data that has scores of the students before attending tutorial and after attending tutorial.

PROBLEM STATEMENT

Studies show that tutorials improve the understanding of students which leads to increase in their scores. Hence we need to prove this claim.

DATA

1) Loading and subsetting data

We load the data by read.csv function. As we need only two columns from the dataset, we subset it.

[Hide](#)

```
df <- read.csv("Assignment4b.csv")  
df <- df[,c(6,7)]
```

2) Handling missing value

If there are any missing values we remove the observation or replace them with some statistics. In the following code we see that there are no NA values in the dataset, Hence no need to remove anything.

[Hide](#)

```
cat(sum(is.na(df$Score.before.tutorial)))
```

```
0
```

[Hide](#)

```
cat(sum(is.na(df$Score.after.tutorial)))
```

```
0
```

DESCRIPTIVE STATISTICS AND VISUALIZATION

1) Descriptive statistics

Descriptive statistics is all about describing the dataset for example its average, how the data is spread etc. We need to find the difference between both the columns for further testing. Hence we perform the subtraction and store it in new column.

Hide

```
df$difference <- df$Score.after.tutorial - df$Score.before.tutorial
head(df, 2)
```

	Score.before.tutorial <int>	Score.after.tutorial <int>	difference <int>
1	50	42	-8
2	13	38	25
2 rows			

descriptive statistics for the difference column is :

Hide

```
diff_mean <- mean(df$difference)
diff_sd <- sd(df$difference)
n = length(df$difference)
summary(df$difference)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-14.00   0.00    0.00   5.35  14.00   31.00
```

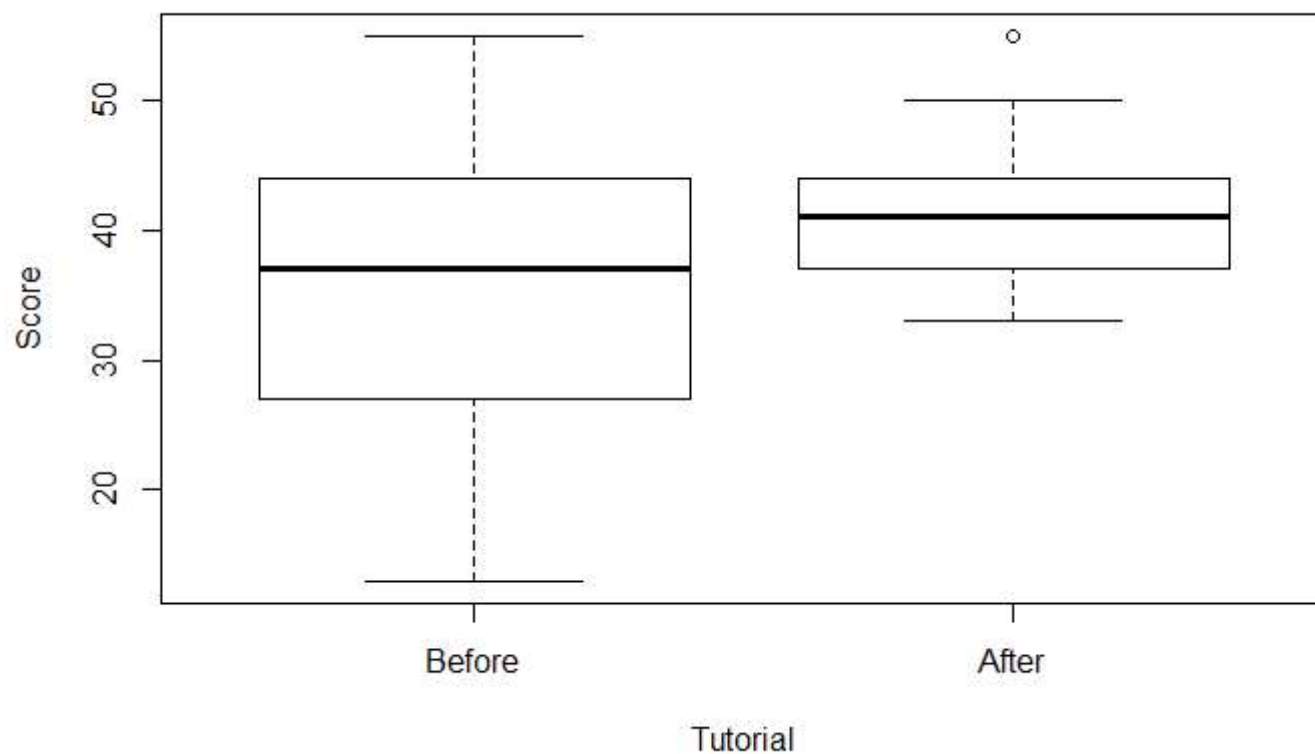
2) Visualization

a) Box plot:

Box plot summarizes the data visually. From the boxplot we can see that mean of the scores has increased after students attend tutorial.

Hide

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

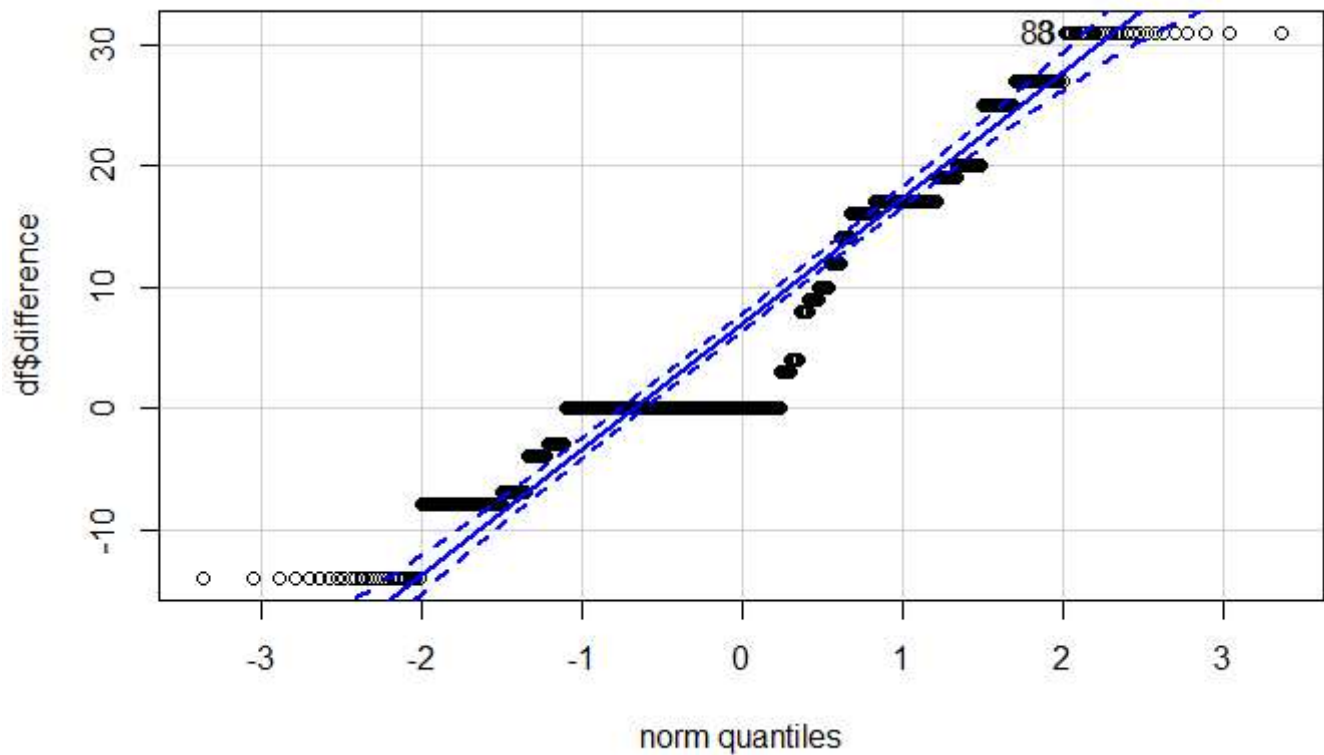
**b) Q-Q plot :**

Q-Q plot shows that most of the data points lie on the diagonal. Sample size is 1290 which is greater than 30 hence we can assume it as normally distributed.

[Hide](#)

```
library(car)
qqPlot(df$difference, distribution = "norm")
```

```
[1] 8 83
```



HYPOTHESIS TESTING

Studies show that tutorial is important to improve the scores which means after attending tutorial, score of the students increase. Hence the difference between the score after attending tutorial and before attending tutorial is greater than or equal to zero which becomes our null hypothesis.

$$H_0 : \mu_a - \mu_b \geq 0$$

$$H_A : \mu_a - \mu_b < 0$$

We are going to test the above hypothesis by paired sample t-test which will be lower tailed. Area of significance is 0.05.

1) Testing by critical value

Testing the Hypothesis by critical value.

Hide

```
t_statistics <- (diff_mean - 0)/(diff_sd/sqrt(n))
t_critical <- qt(0.05, n-1, lower.tail = TRUE)
cat("T-statistics : ",t_statistics)
```

```
T-statistics : 19.14415
```

Hide

```
cat("\nT-critical : ",t_critical)
```

```
T-critical : -1.646037
```

We can see that t-statistics fall beyond the t-critical. It does not fall in rejection region. Hence by this approach we fail to reject null hypothesis.

2) Testing by CI

Here we will be testing the hypothesis by CI approach.

If the 95% CI does not capture mean of difference, reject H_0

If the 95% CI capture mean of difference, Fail to reject H_0

[Hide](#)

```
CIupper <- diff_mean + abs(tcrit)*(diff_sd/sqrt(n))
CIlower <- diff_mean - abs(tcrit)*(diff_sd/sqrt(n))
cat("CI lower : ",CIlower)
```

```
CI lower : 4.890355
```

[Hide](#)

```
cat("\nCI upper : ",CIupper)
```

```
CI upper : 5.81042
```

Range of CI [4.89, 5.81] captures the original mean i.e. 5.35. Hence we fail to reject null hypothesis by this approach. We cannot find any statistical significance.

3) Testing by p-value

Testing hypothesis by p-value is as follows.

if $p < \alpha$, Reject H_0

if $p > \alpha$, Fail to reject H_0

[Hide](#)

```
p_value <- pt(q=t_statistics, n-1, lower.tail = TRUE)
cat("P value : ",p_value)
```

```
P value : 1
```

As p-value is 1 we fail to reject the null hypothesis.

DISCUSSION

A paired-samples t-test was used to test for a significant reduced mean difference between scores before and after attending tutorial. The mean difference was found to be 5.3503 (SD = 10.0379). Visual inspection of the Q-Q plot of the difference scores suggested that the data were approximately normally distributed. The paired-samples t-test did not find a statistically significant mean difference in reducing the scores before and after attending tutorial, $t(df=1289)=19.1441$, $p=1$, 95% [4.903, 5.8104]. Scores were found to be significantly increased after attending tutorial.

REFERENCES

- Applied Analytics Module 7 notes
- (R Documentation and manuals | R Documentation, 2020)
- (t.test function | R Documentation, 2020)
- (Admitted patients - Australian Institute of Health and Welfare, 2020)