Question 1:

Introduction

Admitted patients are patients who undergo a public or private hospital's formal admission process to receive treatment and/or care. The details of the admitted patients are recorded which include their No of overnight stays, Type of the hospital, Name of the hospital, etc.

The problem is to check whether the claim, South Western Sydney hospitals have an average of length of stay (ALOS) of 4.5 days is true or not.

Statistics for the South Western Sydney hospitals is calculated. The results of the statistics are then interpreted.

Problem Statement

The goal is to understand that South Western Sydney hospitals have an average of length of stay (ALOS) of 4.5 days is statistically significant or not.

The problem is established by assuming a null hypothesis that South Western Sydney hospitals have an average of length of stay (ALOS) of 4.5 days.

T-tests are a statistical hypothesis test which is conducted to test if the hypothesis is plausible.

Data

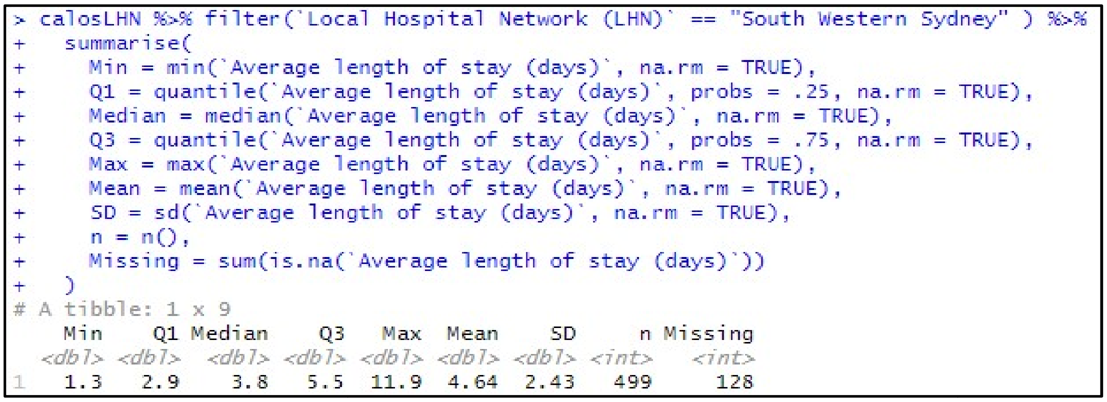
Data preprocessing is extremely important because it allows improving the quality of the raw experimental data. The primary aim of preprocessing is to eliminate those small data contributions associated with the experimental error.

The avgstay dataset and necessary packages are imported into the studio for analysis. The dataset has 13 columns and 30,021 entries.

The average length of stay (days) attribute includes NP values (i.e. Reported values which didn't meet the criteria).
The NP values include criteria where patients reported deaths or transfers to another hospital. Therefore, Replacing NP values with any constant value would be illogical. The NP values could be assumed as outliers. These outliers are thus left untouched for analysis.

The attributes of interest are selected to calculate the statistical information.
The 'average length of stay (days)' and 'Local Hospital Network (LHN)' are the attributes which are selected to perform the statistics.
'average length of stay (days)' is calculated as the number of bed days for overnight stays divided by the number of overnight stays.
'Local Hospital Network (LHN)' is the hospital network for different regions.

Descriptive Statistics



```
> calosLHN %>% filter(`Local Hospital Network (LHN)` == "South Western Sydney" ) %>%
+   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)`))
+   )
# A tibble: 1 x 9
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1   1.3   2.9    3.8   5.5  11.9  4.64  2.43   499     128
```

Fig 1

The average length of stay median and mean for South Western Sydney hospitals are 3.8 and 4.64 respectively. The values are far from each other; therefore, the data is not normally distributed. There are 128 missing average length of stay values.

Visualization

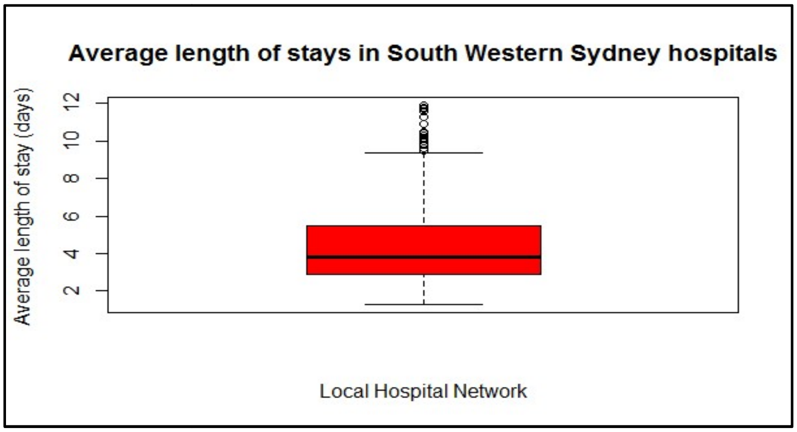


Fig 2

As we can see in Fig 2, the plotted points are the outliers. The median value is close to 4.

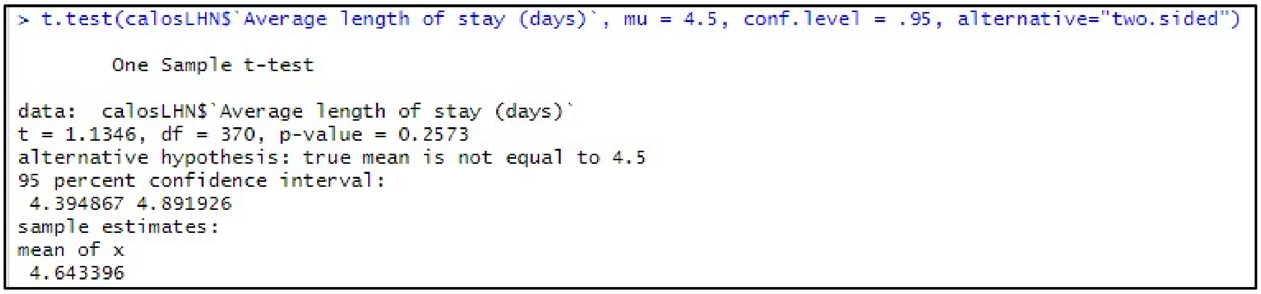
Hypothesis Testing

According to CLT, When the sample size we use is large, typically defined as n>30, the sampling distribution of the mean is approximately normal, regardless of the variable’s underlying population distribution.

Null Hypothesis is an assumption on which statistical T-test is computed to find out if the plausible hypothesis is true or not.

Assuming null hypothesis:

H0:- South Western Sydney hospitals have an average of length of stay (ALOS) of 4.5 days



```
> t.test(calosLHN$`Average length of stay (days)`, mu = 4.5, conf.level = .95, alternative="two.sided")

        One Sample t-test

data:  calosLHN$`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
```

Fig 3

As we can see in Fig 3, the One Sample t-test is conducted with mu = 4.5 and confidence level as 95 i.e. (Significance level = 0.05) with a two sided as an alternative.
We found out that the mean of average length of stay for South Western Sydney hospitals is 4.64. The results of the one-sample t-test found the mean average length of stay to be not statistically significant, t (370) = 1.13, p > 0.05, 95% CI [4.39, 4.89].

We failed to reject H0 due to the variation of average length of stays and the outliers (possibly the farther values as well as the NA values).

Question 2:

Introduction

The details of students such as their profession, school, gender, scores before and after the tutorial, etc are recorded.

The problem is to understand the effectiveness of tutorials in students' performance.

Statistics is conducted to find the relationship between scores before tutorial and scores after tutorial. Depending on the statistics, the t-test is calculated for further analysis which is then concluded by the interpretation of results.

Problem Statement

The goal is to determine whether tutorials are effective in improving students' performance.

The problem is established by assuming a null hypothesis that the population mean differences between 'score before tutorial' and 'score after tutorial' are equal to 0.

A paired t-test is conducted to check for a statistically significant mean change or difference in dependent samples.

Data

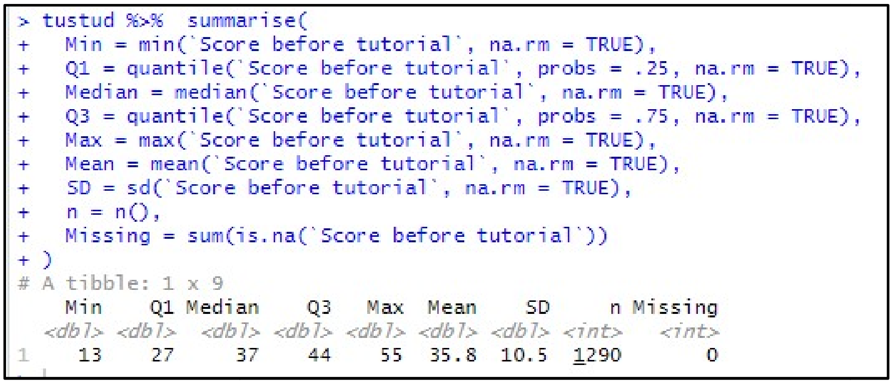
Data preprocessing is extremely important because it allows improving the quality of the raw experimental data. The primary aim of preprocessing is to eliminate those small data contributions associated with the experimental error.

The Assignment 4b-3 dataset and necessary packages are imported into the studio for analysis. The dataset has 7 columns and 1,290 entries.

The is.na function was used to find the null values and then the sums of those values were calculated.
The dataset didn't have any missing values present in it.
Moreover, the data was already cleaned with appropriate datatype for each feature.

The attributes of interest are selected to calculate the statistical information.
The 'Score before tutorial' and 'Score after tutorial' are the attributes which are selected to perform the statistics.
'Score before tutorial' are the scores before the student took the tutorials.
'Score after tutorial' are the scores after the student took the tutorials.

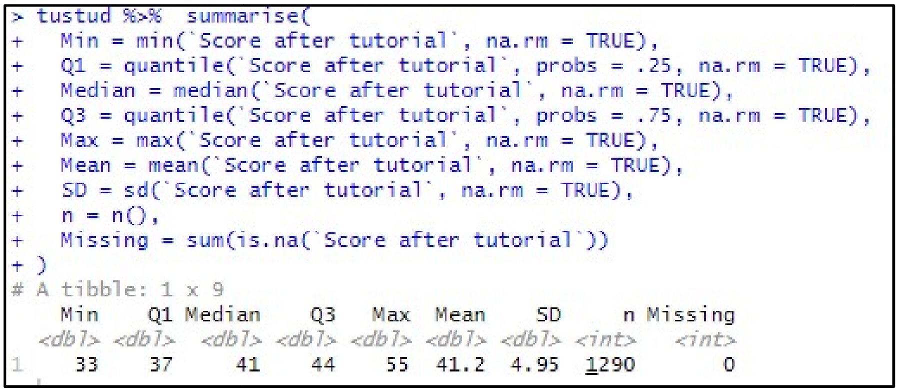
Descriptive Statistics



```
> tustud %>%  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`))
+ )
# A tibble: 1 x 9
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1    13    27     37    44    55  35.8  10.5  1290       0
```

Fig 4

The median and mean for Score before tutorial are 37 and 35.8 respectively. The values are far from each other; therefore, the data is not normally distributed for Score before tutorial.



```
> tustud %>%  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`))
+ )
# A tibble: 1 x 9
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1    33    37     41    44    55  41.2  4.95  1290       0
```

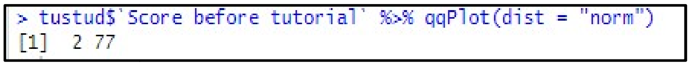
Fig 5

The median and mean for Score after tutorial are 41 and 41.2 respectively. The values are close to each other; therefore, the data is fairly balanced or Symmetric on both sides for Score after tutorial.

As we can see in both the figures 4 and 5, there is significant mean difference between both the variables. Score before tutorial has mean = 35.8 and score after tutorial has mean = 41.2.

When we measure the same sample twice, the measurements are said to be “paired” or “dependent”. Here, score before tutorial and score after tutorial are taken from the same sample; therefore a paired sample t-test should be undertaken.

Visualization



```
> tustud$`Score before tutorial` %>% qqPlot(dist = "norm")
[1]  2 77
```

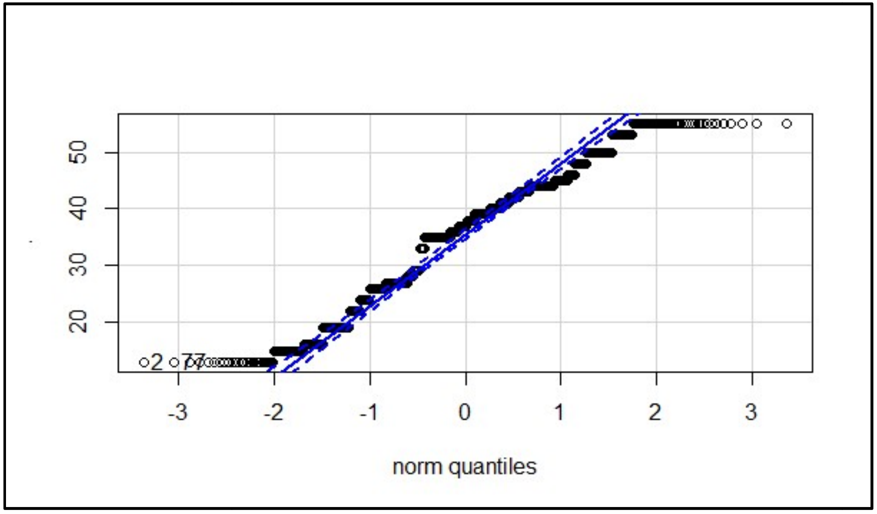
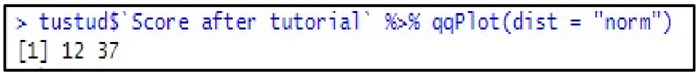


Fig 6

In the Fig 6, we can see the QQ plot of Score before tutorial, there are some points deviating from the blue dashed line. The blue dashed line is the confidence interval for the data. Therefore, the Score before tutorial is not normally distributed.



```
> tustud$`Score after tutorial` %>% qqPlot(dist = "norm")
[1] 12 37
```

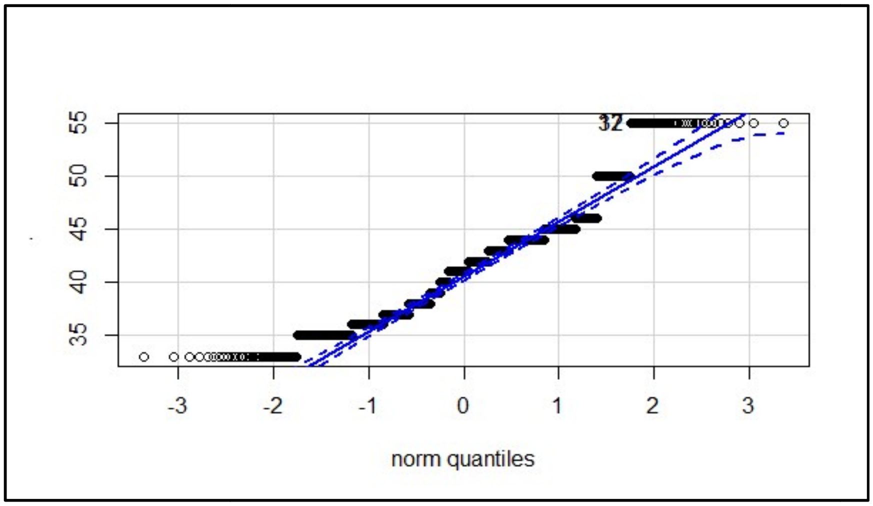


Fig 7

As we can see in the Fig 7, the QQ plot is shown for the Score after tutorial, majority of points are deviating from the blue dashed line. The blue dashed line is the confidence interval for the data. Therefore, the Score after tutorial is not normally distributed.

We can see points falling outside the tails of the distribution for both the variables.

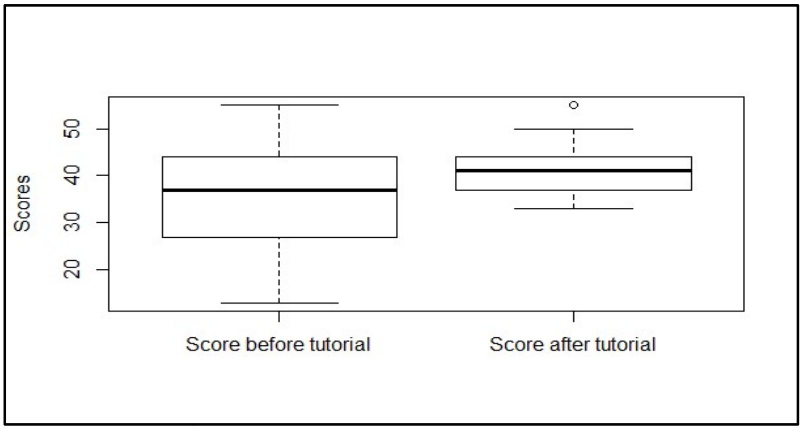


Fig 8

In the Fig 8, we can there is a mean difference between Score before tutorial and Score after tutorial. The Score before tutorial seems to have variation of scores whereas; Score after tutorial have more similar scores than the other way round.

Hypothesis Testing

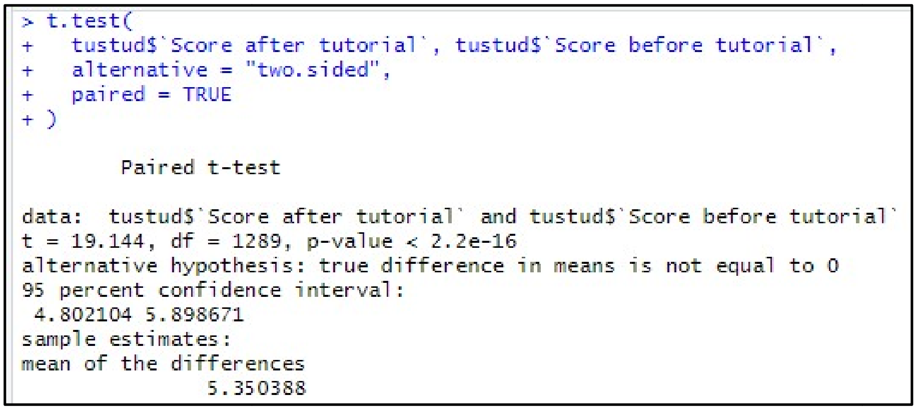
According to CLT, we know that the sampling distribution of a mean will be approximately normally distributed, regardless of the underlying population distribution when the sample size is large (i.e. n>30).

This means that we can proceed with the two-sample t-test if the normality assumption is violated when the sample sizes in each group are greater than 30.

In this example, the sample sizes are much greater than 30 in both the variables, so we can effectively ignore the issue with normality for the Score before tutorial and Score after tutorial.

Assuming null hypothesis:

H0:- The populations mean differences between ‘Score before tutorial’ and ‘Score after tutorial’ are equal to 0.



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

        Paired t-test

data:  tustud$`Score after tutorial` and tustud$`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
```

Fig 9

A paired-samples t-test was used to test for a significant mean difference between scores before and after tutorial. From the Fig 9, we can say that the paired-samples t-test found a statistically significant mean difference between scores before and after tutorial, t (1289) = 19.14, $p < 0.05$, 95% [4.80, 5.90]. Scores were found to be statistically effective after students undertook tutorials. This means that the students' performance improved after taking the tutorials.

References

Science direct. Data Pre-processing [online]. Available at
< https://www.sciencedirect.com/topics/engineering/data-preprocessing >
[Accessed 10 May 2020]

Astral theory. Appspot. MATH1324_Module _05 [online]. Available at
< https://astral-theory-157510.appspot.com/secured/MATH1324_Module_05.html >
[Accessed 1 June 2020]

Astral theory. Appspot. MATH1324_Module _07 [online]. Available at
< https://astral-theory-157510.appspot.com/secured/MATH1324_Module_07.html >
[Accessed 1 June 2020]