

Name: Nikhil H

INTRODUCTION

The data consists of the Details from the Australian Institute of Health and Welfare.

Admitted patients are patients who undergo a public or private hospital's formal admission process to receive treatment and/or care. The types of care provided include surgical care, medical care, intensive care, newborn care, rehabilitation care, palliative care, and mental health care.

The ALOS measure is the average length of stay in hospital. The average is calculated as the number of bed days for overnight stays divided by the number of overnight stays and is reported for selected conditions and procedures.

The ALOS data set consists of average length of Stay values for Different Peer group and it also contains the other attributes.

PROBLEM STATEMENT

The aim is to investigate and understand if there is any statistical significant difference in the average length of stay (ALOS) between large and medium hospitals which might make patients choose one over the other.

DATA PREPARATION


The dataset is collected from the Australian Institute of Health and Welfare website.

Import the data in R studio skip the first 12 rows which contains the description of the data.

The important variables in Average length of stay dataset are the Peer group and Alos(Average length of stay).

The Peer group variable consists of categorical variables so factor the values and the class() of the attribute should be changed to ordered factor

The data type of Alos variable is character as there are NP and – values in the records. Change the data type of the attribute to numeric which converts the NP and – values to NA.



Filter the records which contain Peer group (Large hospitals, Medium hospitals) and store them in a new data frame called `alos_var`

DATA PREPARATION

➤ Missing values:

The NP and – values will be changed to NA.

As mentioned in the description of the dataset :

NP – Reported data did not meet the criteria to calculate this indicator

- – There were no patients reported for this indicator in this time period

As the Description of the NP and – stated as above we will not include these records in the analysis as it might affect the results. So we will remove the NA values ➤ Outliers:

Plot a box plot to check for the outliers separately in both the peer groups

Remove the upper and lower filter by using the formula

Upper outlier $> Q3 + [IQR * 1.5]$, Lower Outlier $< Q1 - [IQR * 1.5]$

DESCRIPTIVE STATISTICS AND VISUALISATION

➤ Summary of alos attribute grouped by Peer group before removing the outliers:

```
alos_var %>% group_by(alos_var$`Peer group`)%>%  
  summarise(Min = min(alos_var$`Average length of stay (days)`,na.rm = TRUE),  
    Q1 = quantile(alos_var$`Average length of stay (days)`,probs = .25,na.rm =TRUE),  
    Median = median(alos_var$`Average length of stay (days)`,na.rm = TRUE),  
    Q3 = quantile(alos_var$`Average length of stay (days)`,probs = 0.75,na.rm = TRUE),  
    Max = max(alos_var$`Average length of stay (days)`,na.rm = TRUE),  
    Mean = mean(alos_var$Average length of stay (days)`,na.rm = TRUE),
```

DESCRIPTIVE STATISTICS AND

```
SD = sd(alos_var$`Average length of stay (days)`,na.rm = TRUE), n = n(),
```

```
# A tibble: 2 x 10  
  `alos_var$`Peer group`  Min    Q1 Median    Q3    Max  Mean    SD    n Missing  
  <ord>    <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>  
1 Large hospitals      1.2    2.5    3.5     5   12.6  3.99  1.98   5692    1281  
2 Medium hospitals      1     2.4    3.4    4.5   13.2  3.71  1.85   3877    1695
```

```
Missing = sum(is.na(alos_var$`Average length of stay (days)`))
```

VISUALISATION

➤ Missing values :

According to the summary statistics both there are 2976 missing values

As mentioned in the Data Preparation step we will remove the missing values in both the Peer group using the below code snippet:

```
alos_var <- alos_var [! (is.na ( alos_var$`Average length of stay (days)`)),]
```

DESCRIPTIVE STATISTICS AND

Verify the summary statistics after removing the missing values. The number of missing values should be zero as displayed below:

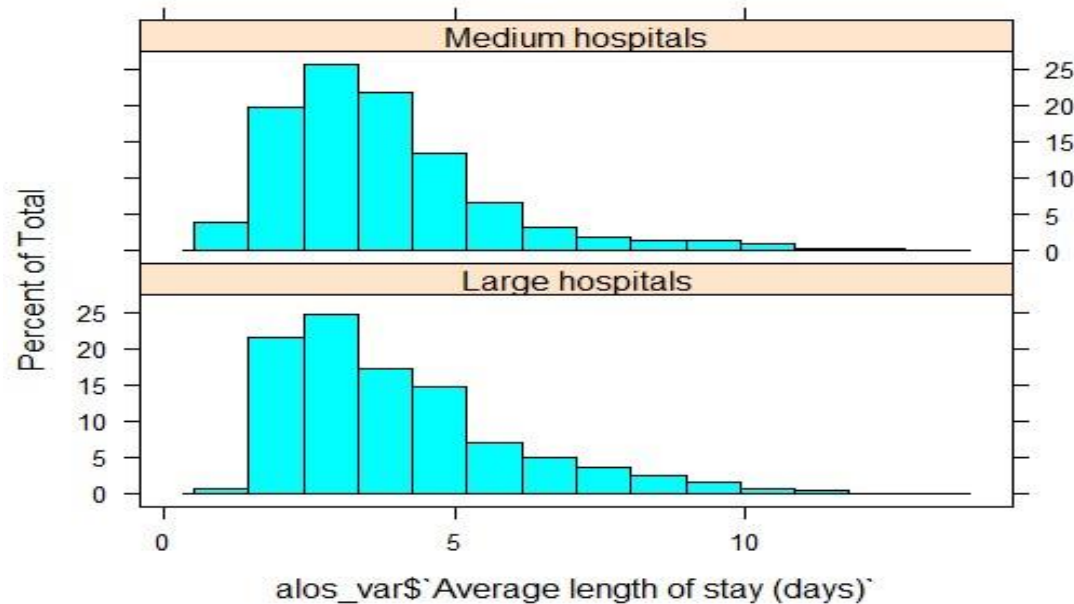
```
# A tibble: 2 x 10
  `alos_var$`Peer group`  Min    Q1 Median    Q3    Max  Mean    SD    n Missing
  <ord>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>    <int>
1 Large hospitals      1.2    2.5    3.5     5   12.6   3.99   1.98   4411      0
2 Medium hospitals      1     2.4    3.4    4.5   13.2   3.71   1.85   2182      0
```

VISUALISATION

Plot the Histogram to check the skewness of the data using the below code:

```
library(lattice) library(ggplot2) alos_var %>% histogram(~alos_var$`Average length of stay (days)` |
alos_var$`Peer group`, data = ., layout=c(1,2))
```


DESCRIPTIVE STATISTICS AND



DESCRIPTION:

The data is not normally distributed in both the Peer groups as it positively Skewed.

VISUALISATION

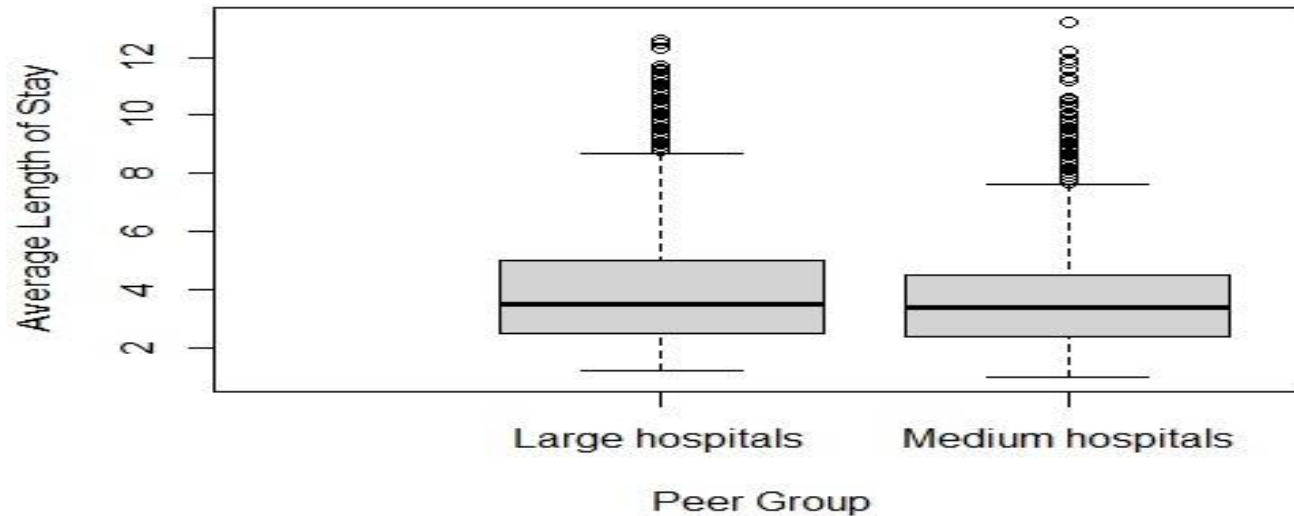
➤ Boxplot of Average Length of stay with

respect to the Peer Groups: Plot the boxplot using the below code snippet:

DESCRIPTIVE STATISTICS AND

```
alos_var %>% boxplot ( alos_var$`Average length of stay (days)` ~ alos_var$`Peer group`, data = ., xlim =  
c( 0,2.5 ),ylab = 'Average Length of Stay', main = 'Boxplot of Alos goruped by Peer Group', xlab = 'Peer  
Group')
```

Boxplot of Alos goruped by Peer Group



Description:

The Boxplot displays the Quartiles, mean and the outliers in each of the Peer Groups. We need to remove these outliers as these are rare cases where the average length of stay is greater than 7.6, and hence these outliers might affect our analysis.

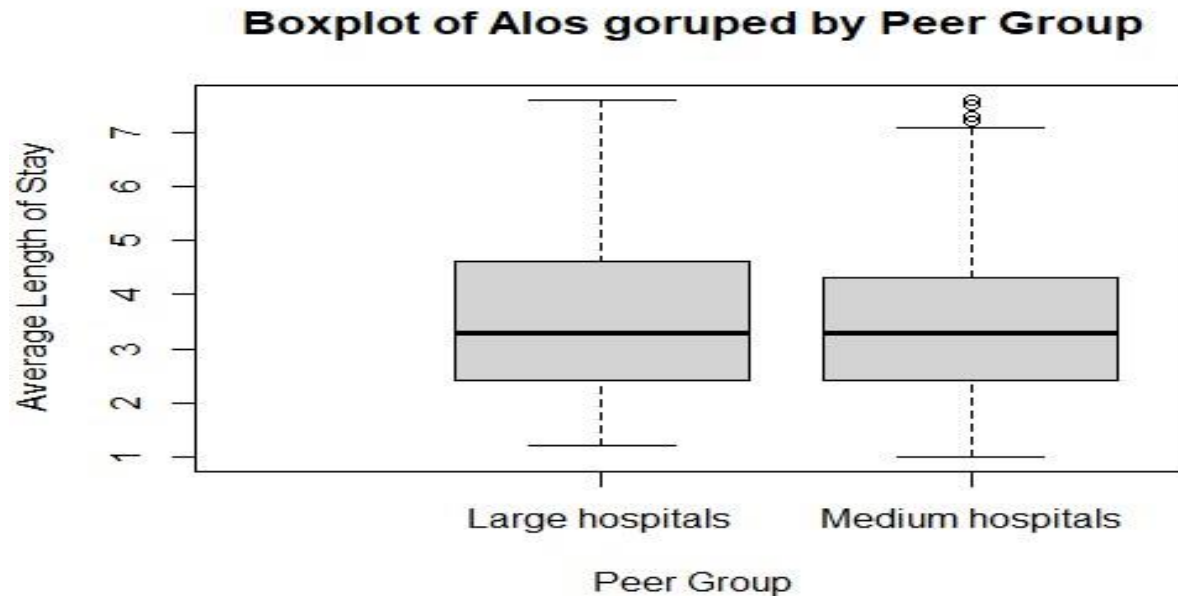
VISULAISATION

➤ Boxplot after removing the outliers:

DESCRIPTIVE STATISTICS AND

```
## plot the boxplot with the clean data
```

```
alos_filter %>% boxplot($`Average length of stay (days)` ~ $`Peer group`, data=., xlim = c(0,2.5), ylab  
= 'Average Length of Stay', main = 'Boxplot of Alos goruped by Peer Group', xlab = 'Peer Group' )
```



Description:

The figure displays a boxplot after removing the outliers in each of the Peer Group.

VISUALISATION

➤ Outliers:

We will remove the outliers in each of the peer group using the below formula:

DESCRIPTIVE STATISTICS AND

Upper outlier $> Q3 + [IQR * 1.5]$, Lower Outlier $< Q1 - [IQR * 1.5]$

Below is the code snippet for removing the outliers:

```
## get the outlier values
```

```
boxplot <- alos_var %>% boxplot(alos_var$`Average length of stay (days)`~alos_var$`Peer group`, data = ., plot = FALSE)
```

```
## filter matrix to get the outliers
```

```
Filt_mat <- data.frame(group = boxplot$group, outliers = boxplot$out)
```

```
Filt_mat$group <- Filt_mat$group %>% factor(levels=c(1,2), labels = c("Large hospitals","Medium hospitals"))
```

```
## remove the outliers
```

```
alos_filter <- alos_var %>% filter(!(alos_var$`Peer group` %in% Filt_mat$group)| !(alos_var$`Average length of stay (days)` %in% Filt_mat$outliers))
```

VISUALISATION

Comparison of the summary of data frame before and after removing the Outliers:

DESCRIPTIVE STATISTICS AND

Summary statistics before removing the outliers:

```
# A tibble: 2 x 10
  `alos_var$`Peer group`  Min    Q1 Median    Q3    Max    Mean    SD    n Missing
  <ord>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <int>
1 Large hospitals      1.2    2.5    3.5    5    12.6    3.99    1.98    4411      0
2 Medium hospitals      1    2.4    3.4    4.5    13.2    3.71    1.85    2182      0
```

Summary statistics after removing the outliers:

```
# A tibble: 2 x 10
  `Peer group`  Min    Q1 Median    Q3    Max    Mean    SD    n Missing
  <ord>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <int>
1 Large hospitals      1.2    2.4    3.3    4.6    7.6    3.63    1.49    4118      0
2 Medium hospitals      1    2.4    3.3    4.3    7.6    3.42    1.38    2073      0
```

Comparison : As there were no outliers in the lower region the minimum value remains the same. Whereas the Maximum value is changed for both the peer group as the outliers were removed from the Upper region. So the Alos(Average Length of Stay) values which were greater than 7.6 are removed as they were outliers.

We select a two-sample t-test because we want to compare mean ALOS(Average Length Of Stay) of two independent groups. But before conducting the two-sample t-test we need to check the normality and variance homogeneity assumption.

Null Hypothesis:

The Null hypothesis states that , there is no statistical significant difference between the Average length of stay for admitted patients at Large hospitals and medium hospitals.

H0: M1(Mean ALOS Large Hospitals) = M2(Mean ALOS Medium Hospitals)

Alternative Hypothesis

The Alternative hypothesis states that , there is a statistical significant difference between the Average length of stay for admitted patients at Large hospitals and medium hospitals.

HA: M1(Mean ALOS Large Hospitals) != M2(Mean ALOS Medium Hospitals)

HYPOTHESIS TESTING

HYPOTHESIS TESTING

➤ Assumptions:

Comparing two independent population means with unknown population variance.

Population data are normally distributed or large sample used ($n > 30$ for both groups).

Population homogeneity of variance.

➤ Decision Rules:

Reject H_0 :

If -value $p < 0.05$ (significance level)

If CI of the mean difference does not capture $H_0 : M_1 - M_2 = 0$

Otherwise, fail to reject .

QQ plot to check for normality of the data:

Create a separate dataframe for each Large and Medium hospitals using the below code snippet:

```
alos_large <- alos_filter %>% filter(alos_filter$`Peer group` == 'Large hospitals')
```

```
alos_medium <- alos_filter %>% filter(alos_filter$`Peer group` == 'Medium hospitals')
```

Plot qq plot for each of the Peer groups :

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

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

Description of QQ plot:

The blue dashed lines of QQ plot correspond to 95% CI for the normal quantiles.

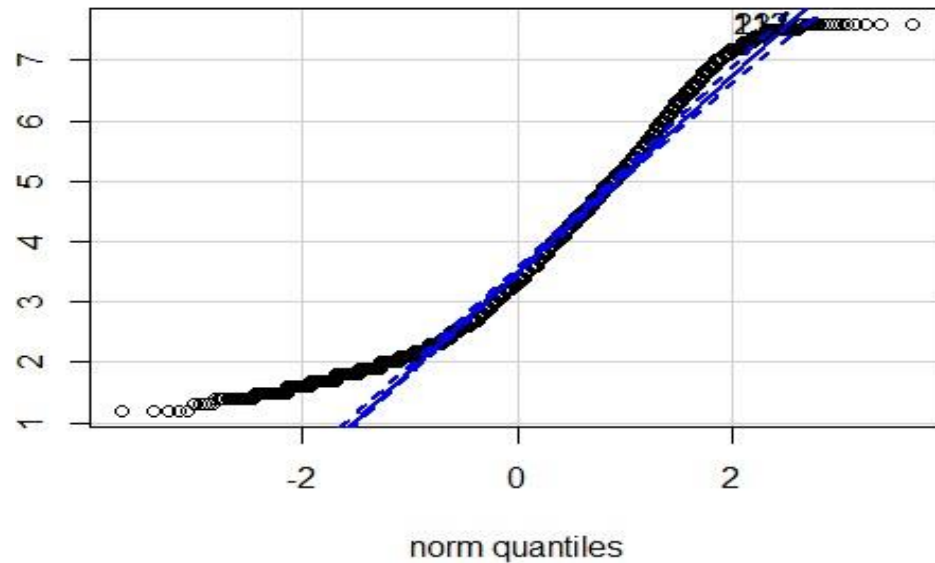
HYPOTHESIS TESTING

The data points should fall inside the blue lines for the data to be normal.

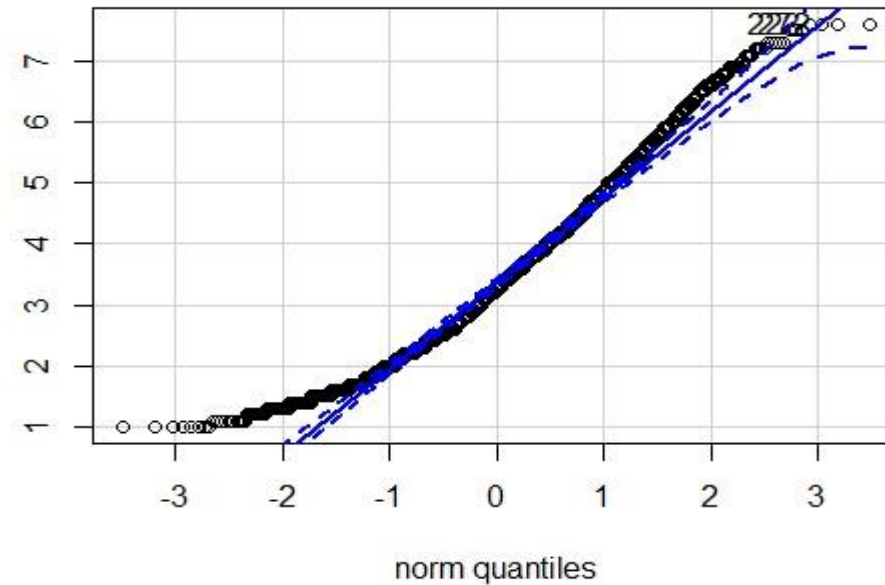
HYPOTHESIS TESTING

QQ plot to check for normality of the data:

QQ plot for Large hospitals



QQ plot for Medium hospitals



Conclusion: We can conclude that the data is not normally distributed for both Large hospitals and Medium hospitals as the data points are falling outside the blue dashed line. As the sample size of

HYPOTHESIS TESTING

both the Large hospitals($n=4118$) and Medium hospitals($n=2073$) is greater than 30, we can assume that the data is normally distributed based on the CLT(Central Limit Theorem).

➤ Two sample t-test:

To test the homogeneity of Variance using Levene's test:

H_0 : Variance of Large hospitals = Variance of Medium hospitals

H_A : Variance of Large hospitals \neq Variance of Medium hospitals

Apply the levene test using the below code snippet:

```
leveneTest(alos_filter$`Average length of stay (days)` ~ alos_filter$`Peer group`,
```

```
data = data.frame(alos_filter$`Average length of stay (days)`))
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group   1   14.25 0.0001616 ***
      6189
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: The output of Levene test has $p < 0.05$ so we Reject H_0 .

HYPOTHESIS TESTING

➤ Welch Test:

As the variance of the two samples is not equal we can use welch test to check the two sample t-test of samples with unequal variance. Use the below code snippet to test the Welch test: `t.test(`

```
alos_filter$`Average length of stay (days)`~alos_filter$`Peer group`,  
data = data.frame(alos_filter$`Average length of stay (days)`),  
var.equal = FALSE, alternative = "two.sided" )
```

 Output of Welch

Test:

```
welch Two Sample t-test  
data: alos_filter$`Average length of stay (days)` by alos_filter$`Peer group`  
t = 5.5554, df = 4450.2, p-value = 2.931e-08  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 0.1373519 0.2871626  
sample estimates:  
mean in group Large hospitals mean in group Medium hospitals  
      3.632807                3.420550
```

➤ Conclusion:

We assumed normality as , $n_1, n_2 > 30$.

Levene test, variances are not homogeneous

Estimated difference between means: $3.632807 - 3.420550 = 0.212257$

95% CI of difference between means $[0.1373519 \ 0.2871626]$ p-value =
 2.93×10^{-8}

Decision: Reject H_0

HYPOTHESIS TESTING

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

Discussion

The results of this investigation found a statistical significance difference between the Average length of stay of admitted patients at Large Hospitals and Average length of stay of admitted patients at Medium Hospitals of 0.212257, 95% CI for the difference in means [0.1373519 0.2871626]

The average length of stay of patients admitted in Medium hospitals is less compared to Large hospitals . So the patients would prefer to be admitted in Medium hospitals rather than Large hospitals to undergo the treatment.