Harvinder Singh Sethi

Introduction

Our dataset is about patients who undergo a public or private hospital's formal admission process to receive treatment.

We will seek to understand if there is any significant difference in the average length of stay (ALOS) between large and medium hospitals using some basic statistics which might make patients choose one over the other.

Problem statement

We would be applying some statistical tests and plot them on a graph to see that's if there is any significant difference in the average length of stay between large and medium hospitals.

My assumption for Null Hypothesis in this case is :- Means are Equal.

Data

- I Got the data from Australian Institute of Health and Welfare

- I uploaded the xlsx file into R Studio using appropriate library readxl

- I created a separate data frame of Peer group and the Average length of stay (days), renamed the columns and converted the NP values to 0 indeed changing their datatype also.

- Filtered out the large and medium hospitals peers and attached average length of stay into 2 different data frames where the average length of stay > 0, so as to make more sense of data I dropped those rows containing NP Values.

Data (Codes)

```r
library(readxl)
library(dplyr)
library(magrittr)
alos<- read_excel("Assignment-02/average-length-of-stay-multilevel-data (1).xlsx",
                  col_types = c("text", "text", "text",
                                "text", "text", "text", "text", "text",
                                "skip", "text", "skip", "text", "skip",
                                "text", "skip", "text", "skip", "text",
                                "skip"), skip = 12)

View(alos)

df <- data.frame(alos$`Peer group`, alos$`Average length of stay (days)`)

names(df)[1] <- "peer"                  #Renamed the column1 to peer
names(df)[2] <- "avg"                   #Renamed the column2 to avg
names(df)
View(df)

class(df$avg)                           #Type of column avg comes out to be "factor"

df$avg = as.numeric(as.character(df$avg))  #Converted the avg type to numeric INDEED. as. numeric() returns a vector of the levels of your factor and not the original values.
df$avg[is.na(df$avg)] = 0               #Checked for the NA values and replaced it with 0.

alos_large = filter(df,df$peer == 'Large hospitals' & df$avg > 0 )
alos_medium = filter(df,df$peer == 'Medium hospitals' & df$avg > 0)
```

Descriptive Statistics

```
alos_medium %>% summarise(Min = min(avg,na.rm = TRUE), Max = max(avg,na.rm = TRUE),
                          Q1 = quantile(avg, probs = .25,na.rm = TRUE),
                          Q3 = quantile(avg,probs = .75,na.rm = TRUE),
                          IQR = Q3-Q1,
                          Mean = mean(avg, na.rm = TRUE),
                          Median = median(avg, na.rm = TRUE),
                          SD = sd(avg, na.rm = TRUE))
Min  Max  Q1  Q3 IQR     Mean Median       SD
  1 13.2 2.4 4.5 2.1 3.706049    3.4 1.85253

alos_large %>% summarise(Min = min(avg,na.rm = TRUE), Max = max(avg,na.rm = TRUE),
                         Q1 = quantile(avg, probs = .25,na.rm = TRUE),
                         Q3 = quantile(avg,probs = .75,na.rm = TRUE),
                         IQR = Q3-Q1,
                         Mean = mean(avg, na.rm = TRUE),
                         Median = median(avg, na.rm = TRUE),
                         SD = sd(avg, na.rm = TRUE))
Min  Max  Q1 Q3 IQR     Mean Median       SD
1.2 12.6 2.5  5 2.5 3.986874    3.5 1.978679
```

- Compare the two groups base on the descript stats their means and the sample size.
- We can easily see that the mean for large and medium hospitals and the sample sizes are different.

Visualization

- Plotted the two histograms for average length of stay for large and medium hospitals to show the distribution.

- Settled upon a decision that the plot is Right-Skewed and not normally distributed. As we can clearly see from the data that both of them are independent columns, we therefore decided to go for welch test base on plot distribution.

hist(alos_large$avg)
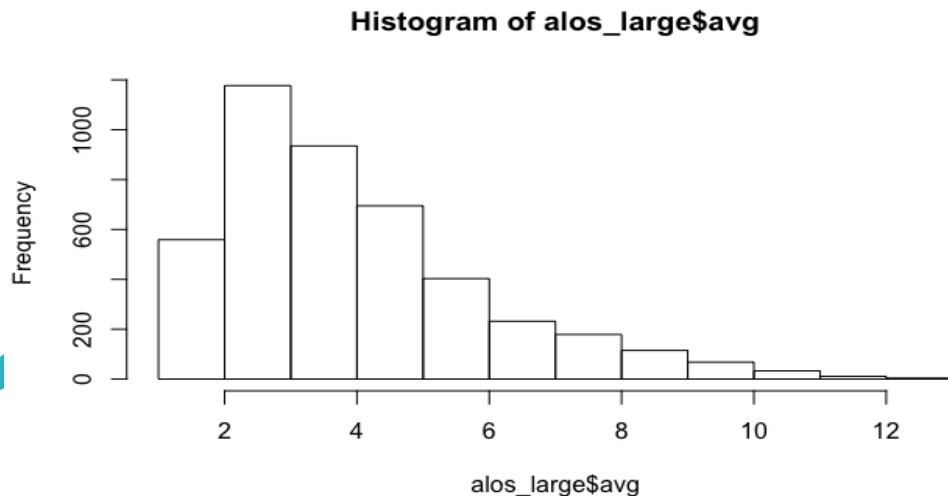
hist(alos_medium$avg)
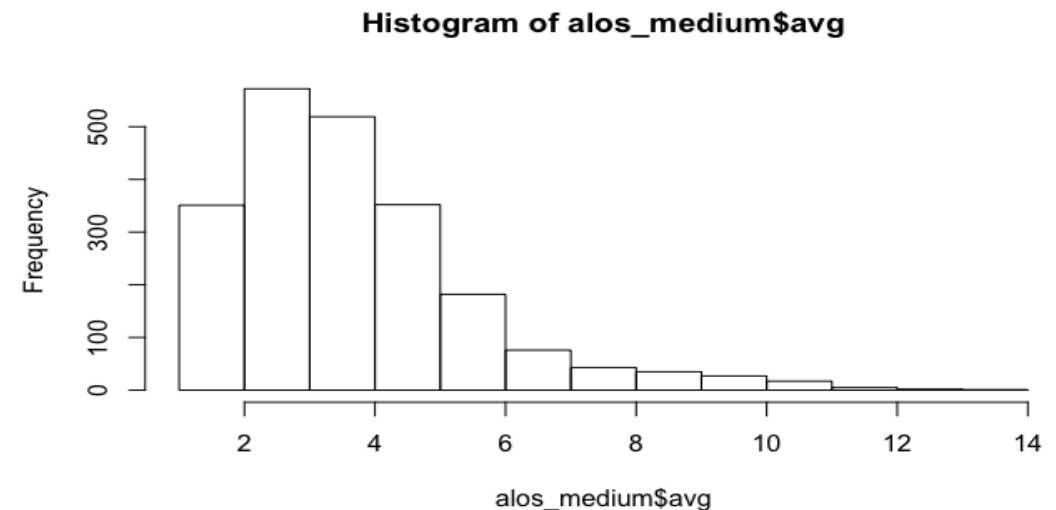
Welch test

- As we visualized the the data was not paired and there was a significant difference between their mean values, we decided to perform the Welch Test.

- t.test() is used for testing and we have passed the average length of stay for both large and medium hospitals.

- From the Descriptive Statistics we can see the difference between means of average length of stay of large and medium hospitals.

- The default values for Significance level is set as 0.05 and Confidence Interval is set at 95%.

```
#We will be using the Welch test because we can see the mean is different. And he homogeneity of variance doesn't hold true.
t.test(alos_large$avg, alos_medium$avg ,var.equal = FALSE, paired = FALSE, alternative = "two.sided")
```

Determine factor to reject or fail to reject H0

- If P-Value < 0.05
- 95% CI means that there's 95% chance our mean will fall between the the upper bound and lower bound of the mean.
- If these conditions don't meet , fail to reject

Result of the rest

```
> t.test(alos_large$avg, alos_medium$avg ,var.equal = FALSE, paired = FALSE, alternative = "two.sided")

        Welch Two Sample t-test

data:  alos_large$avg and alos_medium$avg
t = 5.6615, df = 4611, p-value = 1.592e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1835797 0.3780687
sample estimates:
mean of x mean of y
 3.986874  3.706049
```

What's your decision after the test results

- P value = 1.592e-08

- Observation show that there **IS A STATISCALLY SIGNIFICANCE DIFFERENCE** for the average length of stay between Large and Medium hospitals

- since **p-value** is **less than 0.05**, we reject the null hypothesis and there is a difference between the means and Alternative hypothesis becomes true.

Discussion

- We can Actually get the Info about the average length of stay for large and medium hospitals basics on which people can decide in future which one is more Renowned and where they can easily get seats.

- There are restriction due to the type of data as this can only work with numerical values.

- If we could work on the combined categorical and numerical data that would be more useful as currently we can only work with the numerical values indeed.

Reference

- https://www.aihw.gov.au/reports-data/myhospitals/sectors/admitted-patients
- https://www.listendata.com/2015/05/r-converting-factor-to-integer.html

Thank You