Project Title: Data analysis using inferential statistics

Author name: Samiul Azam

Overview:

In this project, I conduct statistical analysis on the well-known **ToothGrowth** data. Initially, I provide basic statistical measures of the data and simple exploratory analysis. Lastly, 95% T confidence interval has been calculated to infer the population mean (mean difference of tooth length) from estimated sample mean. However, analysis shows that we need to increase the sample size to get appropriate confidence on population mean.

Highlights:

- Summary of the data (in a Table) with some exploratory statistics, such as min, max, mean, median, and few quantiles. Also include histogram and scatter plots for important variables.
- Show 95% T confidence interval to compare two groups (OJ and VC).
- List all the key assumptions relate to T confidence interval.
- Provide interpretation of 95% T confidence interval.
- Provide a statistical conclusion.
- Provide the complete R code for this analysis (at the end of this report).

Basic summary of the data:

- The dataset ToothGrowth contains 60 rows and 3 columns (3 variables: len, supp, dose)
- Summary of the data is as follows:

Variable	Type	Exploratory statistics
len	Number	Min. 1st Qu. Median Mean 3rd Qu. Max. 4.20 13.08 19.25 18.81 25.28 33.90
supp	Number	Min. 1st Qu. Median Mean 3rd Qu. Max. 0.500 0.500 1.000 1.167 2.000 2.000
dose	Categorical	OJ VC (OJ = orange juice, VC = Vitamin C) 30 30

- Figure 1 shows the histogram of len variable.
- Figure 2 shows the scatter plot of point (*len,dose*) where red dots are the VC and black dots are the OJ. From the plot, we see that there are 3 levels of dose (0.5, 1.0 and 2.0), and OJ shows higher *len* values than the VC in most of the cases.

Key assumptions:

- Assuming the variable *len* is an iid Gaussian data.
- Assuming t distribution in confidence interval. T distribution is widely used, and it becomes standard normal distribution as more and more samples are being used.
- Assuming OJ and VC as two independent groups.
- Assuming constant variance across the groups.

T Confidence interval to compare tooth growth by delivery method (supp):

- The 95% T confidence interval of the mean difference (OJ - VC) between the *len* of two groups (OJ and VC) is
  -0.167 to 7.567.
- The sample mean difference is 3.7 (falls within the interval).
- **Interpretation:** The interpretation of the T confidence interval is that there are 95% chance being the population mean (mu) within the interval -0.167 to 7.567.
- **Conclusion:** Here, the positive region is much larger than the negative and zero region. However, the interval still contains the zero whereas our target is to find an narrower interval without the zero (one delivery method increase the tooth size more significantly than the other). Therefore, obviously, we need to increase the sample size (n) to narrow down the confidence interval.

Appendix

Figure 1: Histogram of *len* variable of ToothGrowth data.

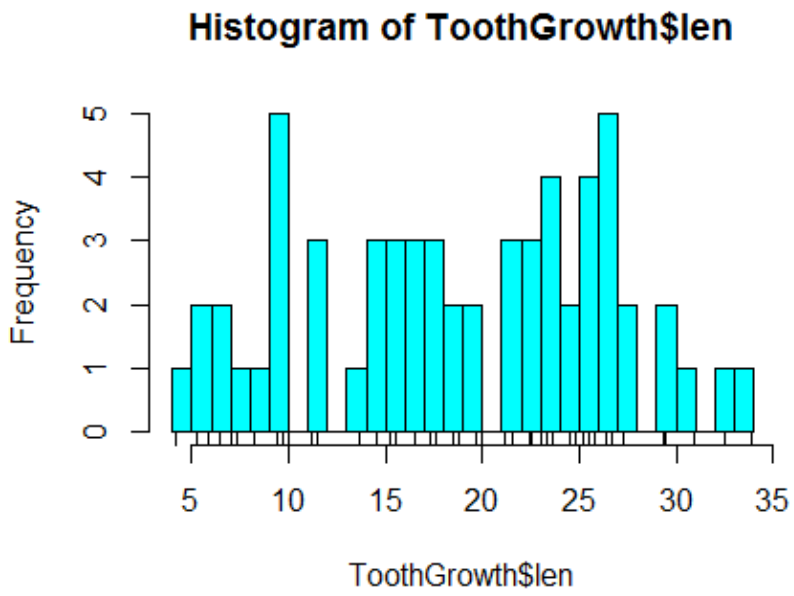
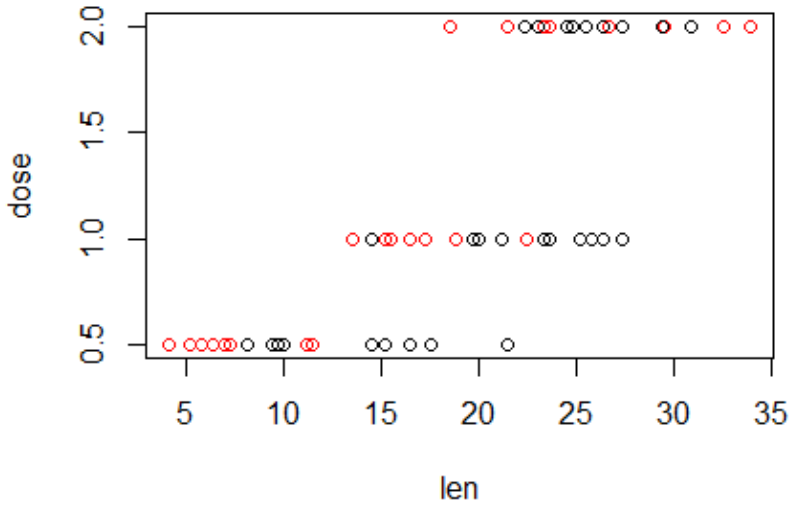


Figure 2: Scatter plot of *len* and *dose* variable for two groups OJ (black) and VC (red).



R code

```
# Import data
data(ToothGrowth)

# Description of the data
help(ToothGrowth)

#Summary of the data
dim(ToothGrowth)
summary(ToothGrowth$len)
summary(ToothGrowth$dose) # Unit mg/day
summary(ToothGrowth$supp)

# Histogram of the len variable
hist(ToothGrowth$len, col = "cyan", breaks = 30)
rug(ToothGrowth$len)

# Scatter plot of len and dose variable (for two groups VC and OJ)
with(ToothGrowth, plot(len, dose, col = supp))

# Splitting into groups
vc_data = ToothGrowth[1:30,c(1,3)]
oj_data = ToothGrowth[31:60,c(1,3)]
g1 = vc_data[,1]
g2 = oj_data[,1]

# 95% T confidence interval
t.test(g2,g1, paired = FALSE, var.equal = TRUE)$conf
```