# ADS 511 - Statistical Inference Methods With Applications

## Part 1

- In this part, we are asked to apply Central Limit Theorem (CLT) to the exponential distribution in R and illustrate yielding a normal distribution from sample means. For this part, we are given Exponential lambda value, which is 0.3 and the sample size which is 50. We will create samples from an exponential distribution using "rexp(sample size , lambda)" where the number of samples for illustration is 1000.
  - Population mean = Projected mean = 1 / lambda
  - Population variance = (1 / lambda)**2
  - Projected variance = (1 / lambda)**2 / sample size

```
rateparam=0.3
samplesize=50
numsample=1000

populationmean= 1 / rateparam # Population mean
populationmean
```

```
## [1] 3.333333
```

```
populationVar=(1 / rateparam)**2 # Population variance
populationVar
```

```
## [1] 11.11111
```

```
#----------------------

projectedmean= 1 / rateparam # Projected mean
projectedmean
```

```
## [1] 3.333333
```

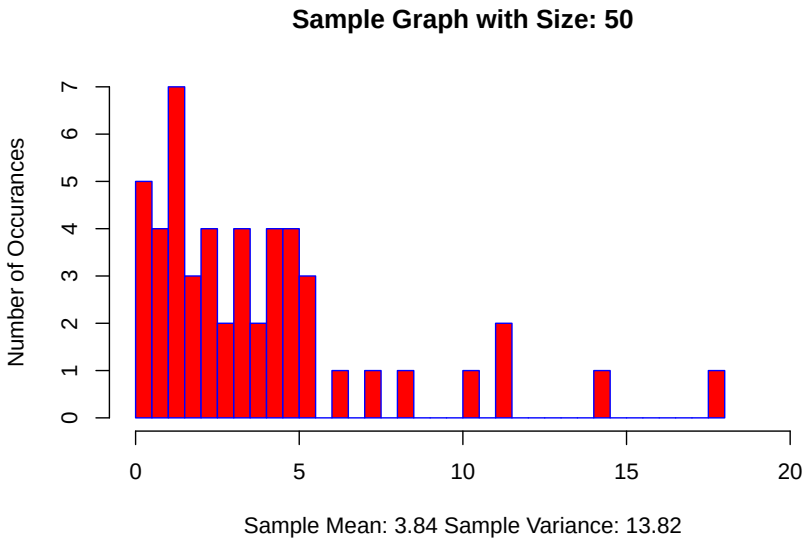
```
projectedVar=populationVar / samplesize # Projected variance
projectedVar
```

```
## [1] 0.2222222
```

## Question 1 & Question 2 & Question 3

- Here, we will take 1000 samples whose size is 50 and calculate the sample mean after CLT along with the sample variance after CLT and compare them to the projected mean and variance. In the figure below, the graph of a sample whose size is 50, can be seen.

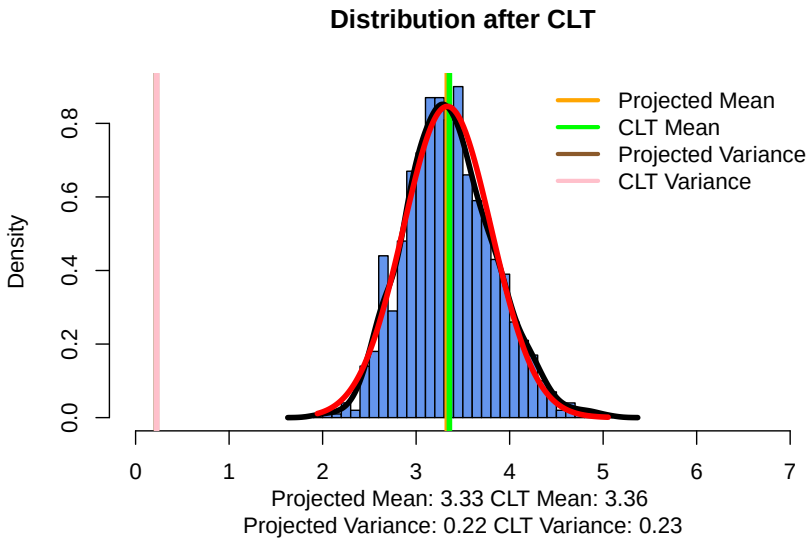
```
z = rexp(samplesize,rateparam)

hist(z,main = paste("Sample Graph with Size:",samplesize),
     xlim = c(0, 20),
     br = 30,
     border = "blue",
     col = "red",
     xlab = paste("Sample Mean:",round(mean(z), digits = 2),
                  "Sample Variance:",round(var(z), digits = 2)),
     ylab = paste("Number of Occurances"),
     col.main ="black",
     col.lab="black")
```




- Now, 1000 samples will be taken to compare the resulting distribution's mean and variance to the projected mean and variance. In the figure below, the results can be seen.

```
mean_taken=c()
std=c()

for (i in 1:numsample){

  mean_taken[i] = mean(rexp(samplesize,rateparam))
  std[i] = ((mean_taken[i] - populationmean)/(sqrt(populationVar/samplesize)))

}

CLTmean = mean(mean_taken) # mean after CLT
```

```r
CLTVariance =var(mean_taken) # Variance after CLT

hist(mean_taken,main = "Distribution after CLT",
     freq=FALSE,
     br = 30,
     xlim=c(0,7),
     border = "black",
     col = "cornflowerblue",
     xlab = paste("Projected Mean:",round(projectedmean, digits = 2),
                  "CLT Mean:",round(CLTmean, digits = 2),
                  "\nProjected Variance:",round(projectedVar, digits = 2),
                  "CLT Variance:",round(CLTVariance, digits = 2)),
     col.main ="black",
     col.lab="black")

abline(v=projectedmean,col="orange", lwd=4)
abline(v=CLTmean,col="green", lwd=4)
abline(v=projectedVar,col="tan4", lwd=4)
abline(v=CLTVariance,col="pink", lwd=4)

xfit<-seq(min(mean_taken),max(mean_taken),length=160)
yfit<-dnorm(xfit,mean=1/rateparam,sd=(1/rateparam)/sqrt(samplesize))

lines(density(mean_taken), lwd=4, col="black")
lines(xfit, yfit, col="red", lwd=4)

legend("topright",
       c("Projected Mean", "CLT Mean","Projected Variance","CLT Variance"),
       lty=c(1, 1),
       col=c("orange","green","tan4","pink"),
       bty = "n",
       lwd = c(3,3))
```

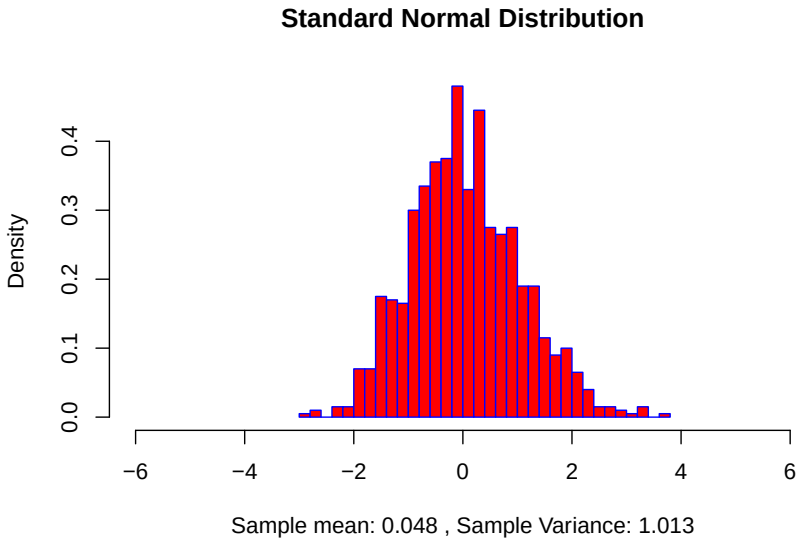


- After CLT, the mean approached the projected mean while the variance approached the projected variance. In the beginning, the distribution of the "one sample" was non-normal. However, after applying CLT, the distribution's shape turned out to be normal distribution shape. If the number of samples increases (if we repeat an experiment independently), sample's mean and variance will get closer to the projected mean and variance (law of large numbers). At the same time, the error rate will decrease.

### Question 4

- Here, the standard normal distribution will be constructed.

```
#Drawing Standard Normal Distribution graph
hist(std,
     freq=FALSE,
     main=paste("Standard Normal Distribution"),
     xlim=c(-6,6),
     br=30,
     border="blue",
     col="red",
     xlab=paste("Sample mean:",round(mean(std),digits=3),
     ", Sample Variance:", round(var(std),digits=3)))
```




- As seen in the figure above, the mean of the standardized samples is close to 0 and the variance is close to 1.

## Part 2

- In Part 2, we will use "ToothGrowth" dataset. This dataset is imported by default in R by running "ToothGrowth". It displays the result of the effects of 2 supplements at different dosage levels on tooth length in 60 guinea pigs. Here, supplement orange juice is represented by "OJ" while supplement vitamin C is represented by "VC"

## Basic Statistical Info about the Data

- In this part, basic statistical information related to the data will be provided.

### Data

- An outlook of this dataset and its structure can be seen below.

```
head(ToothGrowth,6)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

### Types of Variables

- According to "str" function, in the data set, there are 60 rows and 3 columns.
  - "len" column stores the tooth lengths in mm and it is numeric.
  - "supp" column stores the supplement types and it is factor.
  - "dose" column stores the dosage levels of the supplements in mg and it is numeric.

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

### Summary

- The output of "summary" function is shown below. Here, since "supp" column is factor, the result is the total number of each item in the column. Just before visualizing the variables, it can be said that the distribution of "len" column is skewed since the "len" column's mean and median values are not equal to each other. (medianLen != meanLen) However, since the difference between mean and median is not very high, the shape of the distribution is not too skewed.

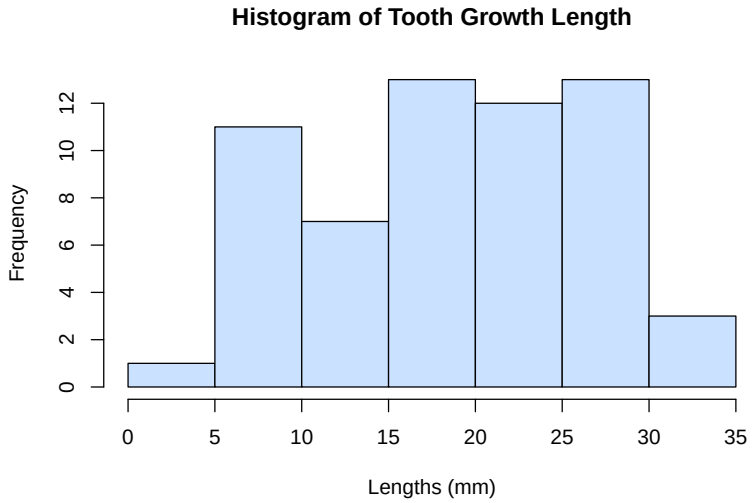
```
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

### Visualizing the Data

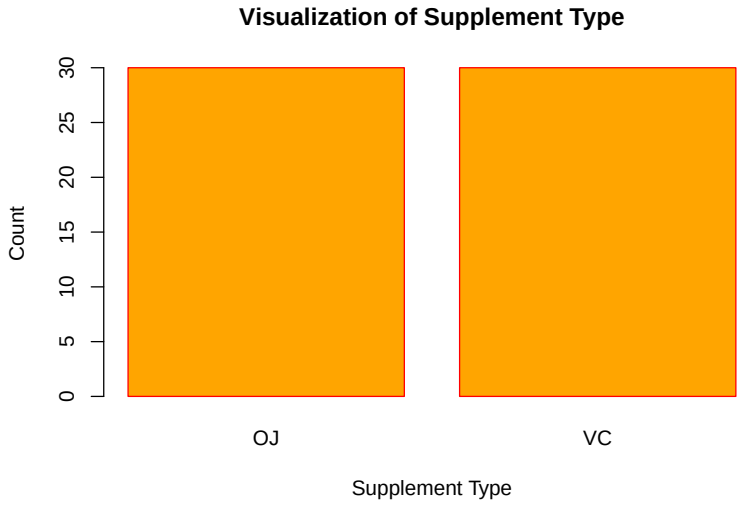
- As seen below, "len" column is visualized. It can be seen that it is skewed a bit.

```
hist(ToothGrowth$len, main = "Histogram of Tooth Growth Length",
     xlab = "Lengths (mm)", prob=FALSE, col = "lightsteelblue1")
```



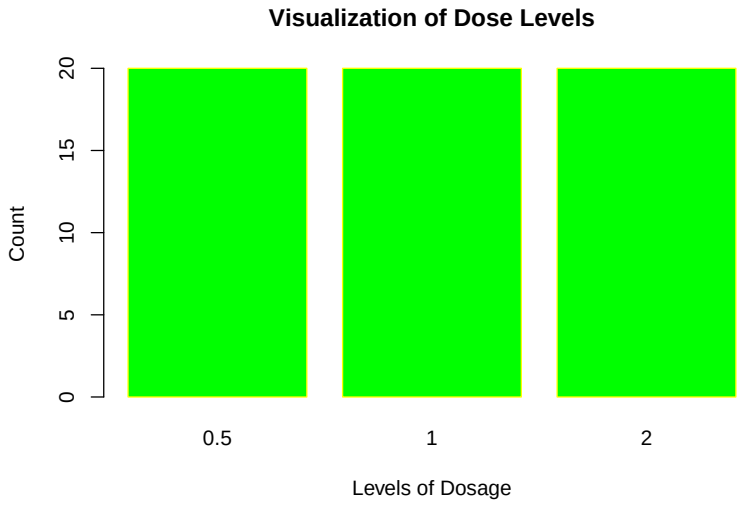

- The categorical variable which is "supp" column is visualized. According to the figure below, supplement OJ is used 30 times and supplement VC is used 30 times.

```
supplement <- table(ToothGrowth$supp)
barplot(supplement,
        main = "Visualization of Supplement Type",
        xlab = "Supplement Type", ylab = "Count",
        border="red", col = "orange")
```



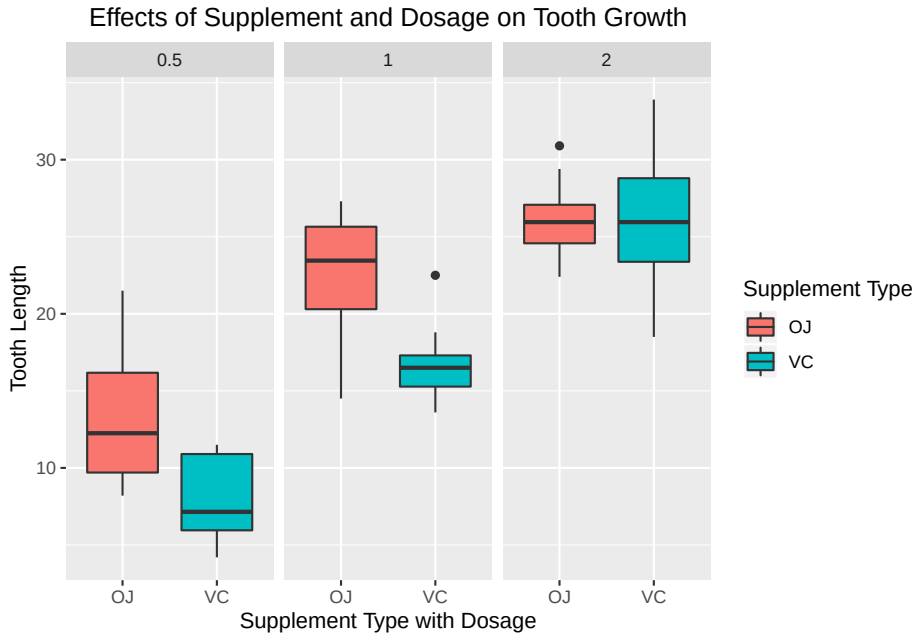
- After visualizing "dose" column as seen below, it can be stated that there are only 3 levels in the set. Previously, it was concluded that the column was numeric. However, since there is no other value than 0.5, 1 and 2, the column can be converted to factor. This conversion will be made in Anova part.

```
dose <- table(ToothGrowth$dose)
barplot(dose,
        main = "Visualization of Dose Levels",
        xlab = "Levels of Dosage", ylab = "Count",
        border="yellow", col = "green")
```




- In addition, a boxplot is created to see the effect of supplement type and dose on the length of tooth growth.

```
library(ggplot2)
ggplot(ToothGrowth, aes(x = supp, y = len, fill = supp))+
  geom_boxplot()+
  facet_grid(. ~ dose)+
  scale_x_discrete("Supplement Type with Dosage")+
  scale_y_continuous("Tooth Length")+
  labs(title = "Effects of Supplement and Dosage on Tooth Growth")+
  labs(fill = "Supplement Type")+
  theme(plot.title = element_text(hjust = 0.5))
```




- According to the figure above, when the amount of supplement dose is increased, it can be claimed that the tooth growth increases, too.
- Also, at the 0.5 mg and 1 mg dose level, supplement OJ is more effective than supplement VC. On the other hand, at the 2 mg dose level, it can be stated that supplement type has no effect on tooth growth.
- According to the figure above, it can be seen that some boxes' median lines don't overlap. So, it is stated that there is likely to be a difference between these groups.
- There are some box plots whose median lines are outside of the other boxes. Then, there is a possibility that there is a significant difference between these groups.
- By comparing the box lengths, we can gain an insight about how the data is dispersed between each sample. If the box length is longer, the data is more dispersed.
- There are some outliers that are located outside the whiskers as seen in the figure.
- By considering the place of the median line, it is possible to make comments on the skewness. If the median line is not in the middle of the box, it can be stated that the data don't appear to be symmetric. Most probably, the most asymmetry data is the data with supplement VC at 0.5 mg dosage. Most probably, the most symmetric data is the data with supplement VC at 2 mg dosage.

### Additional Descriptive Statistics

- In order to provide more information about the data set, "stat.desc" function can be used. As seen from the results, there is no missing value in the columns below. In addition to this, this function enables us to examine other essential information related to data set such as min, max, range, sum, median, mean, SE of the mean, 95% CI of the mean, var, standard deviation, etc.

```
options(warn=-1) # To stop getting errors.
library(pastecs)
stat.desc(ToothGrowth[,c(1,3)])
```

```
##                      len        dose
## nbr.val        60.0000000 60.00000000
## nbr.null        0.0000000  0.00000000
## nbr.na          0.0000000  0.00000000
## min             4.2000000  0.50000000
## max            33.9000000  2.00000000
## range          29.7000000  1.50000000
## sum          1128.8000000 70.00000000
## median         19.2500000  1.00000000
## mean           18.8133333  1.16666667
## SE.mean         0.9875223  0.08118705
## CI.mean.0.95    1.9760276  0.16245491
## var            58.5120226  0.39548023
## std.dev         7.6493152  0.62887219
## coef.var        0.4065901  0.53903330
```

## Constructing and Applying Hypothesis Tests

- After examining the data, especially after visualizing it, it was claimed that there was a positive relationship among some columns at some dosage levels. Is it really so? To be able to understand if there is a significant change in the tooth length by considering some factors, hypothesis tests will be applied. Under this title, there are 7 different hypothesis tests.
- Before starting, there are a few assumptions:
  - The variables are independent and identically distributed.
  - For each supplement at different dose levels, variances of tooth growth length are different.
  - The tooth growth data set is normally distributed.
  - These 60 pigs represent the whole guinea pig population so that in the end, the results can be generalized. The pigs are selected randomly.

For each test, independent and unequal variance T-Test will be applied and 0.95 (95%) will be accepted as the confidence level.

### Hypothesis 1

- H0: Using different supplements for tooth growth length doesn't make a difference. (length after supplement OJ = length after supplement VC)
- H1: The supplement OJ is more effective than VC in tooth growth. (length after supplement OJ > length after supplement VC)

```
#Splitting the data set
OJ = ToothGrowth$len[ToothGrowth$supp == 'OJ']
VC = ToothGrowth$len[ToothGrowth$supp == 'VC']

#The test is one-sided
table=t.test(OJ, VC, alternative = "greater",
             paired = FALSE, var.equal = FALSE, conf.level = 0.95)

paste("P-Value:",table$p.value)
```

```
## [1] "P-Value: 0.030317253940467"
```

- Since P value is less than 0.05, H0 is rejected. So, the supplement OJ is more effective with a high probability.

### Hypothesis 2

- H0: Increasing of the dose amount from 0.5mg to 2mg has no effect on tooth growth (length at dose 0.5 = length at dose 2)
- H1: Increasing of the dose amount from 0.5mg to 2mg has a positive effect on tooth growth (length at dose 0.5 < length at dose 2)

```
#Splitting the data set
dosehalf = ToothGrowth$len[ToothGrowth$dose == 0.5]
dose2 = ToothGrowth$len[ToothGrowth$dose == 2]

#The test is one-sided
table=t.test(dose2,dosehalf, alternative = "greater",
             paired = FALSE, var.equal = FALSE, conf.level = 0.95)

paste("P-Value:",round(table$p.value,16))
```

```
## [1] "P-Value: 2.2e-14"
```

- Since P value is less than 0.05, H0 is rejected. So, increasing of the dose amount from 0.5mg to 2mg has a positive effect, likely.

### Hypothesis 3

- H0: Increasing of the dose amount from 0.5mg to 1mg has no effect on tooth growth (length at dose 0.5 = length at dose 1)
- H1: Increasing of the dose amount from 0.5mg to 1mg has a positive effect on tooth growth (length at dose 0.5 < length at dose 1)

```
#Splitting the data set
dosehalf = ToothGrowth$len[ToothGrowth$dose == 0.5]
dose1 = ToothGrowth$len[ToothGrowth$dose == 1]

#The test is one-sided
table=t.test(dose1,dosehalf, alternative = "greater",
             paired = FALSE, var.equal = FALSE, conf.level = 0.95)

paste("P-Value:",round(table$p.value,8))
```

```
## [1] "P-Value: 6e-08"
```

- Since P value is less than 0.05, H0 is rejected. So, increasing of the dose amount from 0.5mg to 1mg has a positive effect, likely.

### Hypothesis 4

- H0: Increasing of the dose amount from 1mg to 2mg has no effect on tooth growth (length at dose 1 = length at dose 2)
- H1: Increasing of the dose amount from 1mg to 2mg has a positive effect on tooth growth (length at dose 1 < length at dose 2)

```
#Splitting the data set
dose1 = ToothGrowth$len[ToothGrowth$dose == 1]
dose2 = ToothGrowth$len[ToothGrowth$dose == 2]

#The test is one-sided
table=t.test(dose2,dose1, alternative = "greater",
             paired = FALSE, var.equal = FALSE, conf.level = 0.95)

paste("P-Value:",round(table$p.value,7))
```

```
## [1] "P-Value: 9.5e-06"
```

- Since P value is less than 0.05, H0 is rejected. So, increasing of the dose amount from 1mg to 2mg has a positive effect, likely.

### Hypothesis 5

- H0: At the amount of dose 0.5mg, the supplement type has no effect on the lenght of tooth growth (length for supplement OJ = length for supplement VC)
- H1: At the amount of dose 0.5mg, VC is less effective than OJ on tooth growth (length for supplement VC < length for supplement OJ)

```
#Splitting the data set
VC = as.vector(ToothGrowth$len[ToothGrowth$dose == 0.5 & ToothGrowth$supp == 'VC'])
OJ = as.vector(ToothGrowth$len[ToothGrowth$dose == 0.5 & ToothGrowth$supp == 'OJ'])

#The test is one-sided
table=t.test(VC,OJ, alternative = "less",
             paired = FALSE, var.equal = FALSE, conf.level = 0.95)

paste("P-Value:",round(table$p.value,7))
```

```
## [1] "P-Value: 0.0031793"
```

- Since P value is less than 0.05, H0 is rejected. So, at the amount of dose 0.5mg, VC is less effective, most probably.

### Hypothesis 6

- H0: At the amount of dose 1mg, the supplement type has no effect on the lenght of tooth growth (length for supplement OJ = length for supplement VC)
- H1: At the amount of dose 1mg, VC is less effective than OJ on tooth growth (length for supplement VC < length for supplement OJ)

```r
#Splitting the data set
VC = as.vector(ToothGrowth$len[ToothGrowth$dose == 1 & ToothGrowth$supp == 'VC'])
OJ = as.vector(ToothGrowth$len[ToothGrowth$dose == 1 & ToothGrowth$supp == 'OJ'])

#The test is one-sided
table=t.test(VC,OJ, alternative = "less",
             paired = FALSE, var.equal = FALSE, conf.level = 0.95)

paste("P-Value:",round(table$p.value,7))
```

```
## [1] "P-Value: 0.0005192"
```

- Since P value is less than 0.05, H0 is rejected. So, at the amount of dose 1mg, VC is less effective than OJ, likely.

### Hypothesis 7

- H0: At the amount of dose 2mg, the supplement type has no effect on the lenght of tooth growth (length for supplement OJ = length for supplement VC)
- H1: At the amount of dose 2mg, the supplement type has effects on tooth growth (length after supplement OJ != length after supplement VC)

```r
#Splitting the data set
VC = as.vector(ToothGrowth$len[ToothGrowth$dose == 2 & ToothGrowth$supp == 'VC'])
OJ = as.vector(ToothGrowth$len[ToothGrowth$dose == 2 & ToothGrowth$supp == 'OJ'])

#The test is two-sided
table=t.test(OJ,VC, alternative = "two.sided",
             paired = FALSE, var.equal = FALSE, conf.level = 0.95)

paste("P-Value:",round(table$p.value,7))
```

```
## [1] "P-Value: 0.9638516"
```

- Since P value is greater than 0.05, We can't reject H0. So, at the amount of dose 2mg, the supplement type has no effect, likely

## One Way Anova

- Anova can be used to check if there are any significant differences between the means of three or more independent groups. In our case, there are 3 different dose levels, which are 0.5mg, 1mg, 2mg in the data set as mentioned before. To gain better solutions, "dose" column, which is the independent variable, needs to be converted from numeric to factor. "Low" represents 0.5mg, "Medium" represents 1mg and "High" represents 2mg. There are a few assumptions before starting as followings:
  - The data set is normally distributed.
  - 3 levels in "dose" column are independent.
  - "len" column and "dose" column are dependent and their variances are equal to each other.
  - These 60 pigs represent the whole guinea pig population so that in the end, the results can be generalized. The pigs are selected randomly.

- Also, it is necessary to make a hypothesis test. After Anova, according to the p-value, we will either fail to reject the null hypothesis or reject it. Here, 0.95 (95%) will be accepted as the confidence level.
  - H0: There is no relationship between the level of supplement dose and tooth growth
  - H1: There is a significant relationship between the level of supplement dose and tooth growth

```
#Calculating critical f value
paste("Critical F value: ", qf(0.95,2,57))
```

```
## [1] "Critical F value:  3.15884271926064"
```

```
#Converting the dose values into categorical values.
vec = c(0.5,1,2)
ToothGrowth$dose <- factor(x = ToothGrowth$dose,
                           levels = sort(unique(vec)),
                           labels = c( "Low", "Medium", "High"))

#Applying Anova
summary(aov(len ~ dose, data = ToothGrowth))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## dose         2   2426    1213   67.42 9.53e-16 ***
## Residuals   57   1026      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- According to the results above, critical f value is less than F value which is generated by "aov" function. In addition to this, since p value $< 0.05$, H0 is rejected at alpha=0.05. It can be concluded that the dosage affects the tooth length, with a high probability.

## Conclusion

- By using "ToothGrowth" data set, basic statistical information has been examined, hypothesis tests have been made and then, Anova has been conducted.
- In the light of these, it can be said that supplement dosage level has a positive effect on tooth growth length with a high probability.
- At the level of dose 0.5 and 1mg, supplement OJ has more positive effects (more effective) on the tooth growth length than supplement VC. However, supplement types have no effect on the tooth growth length at the level of dose 2mg, most likely.