Basic Inferential Data Analysis

Oluwadare, Margaret

8/27/2020

DATA DESCRIPTION

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid). We're going to analyze the ToothGrowth data in the R datasets package by performing the following task: 1. Load the ToothGrowth data and perform some basic exploratory data analyses 2. Provide a basic summary of the data. 3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supplement and dose by using Only the techniques from class, (even if there's other approaches worth considering) 4. State your conclusions and the assumptions needed for your conclusions.

We start the program by loading neccesary library and the ToothGrowth Data to investigate its structure

```r
library(knitr)
library(ggplot2)
library(dplyr)
library(datasets)
library(gridExtra)
data(ToothGrowth)
attach(ToothGrowth)
```

PRE-PROCESSING AND EXPLORATORY DATA ANALYSIS

The following code will give us a brief on the nature of data we are dealing with.

```r
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 ...
```

```r
head(ToothGrowth)
```

```
##    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
```

```r
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
```

```r
unique(ToothGrowth$len)
```

```
##  [1]  4.2 11.5  7.3  5.8  6.4 10.0 11.2  5.2  7.0 16.5 15.2 17.3 22.5 13.6 14.5
## [16] 18.8 15.5 23.6 18.5 33.9 25.5 26.4 32.5 26.7 21.5 23.3 29.5 17.6  9.7  8.2
## [31]  9.4 19.7 20.0 25.2 25.8 21.2 27.3 22.4 24.5 24.8 30.9 29.4 23.0
```

```r
unique(ToothGrowth$supp)
```

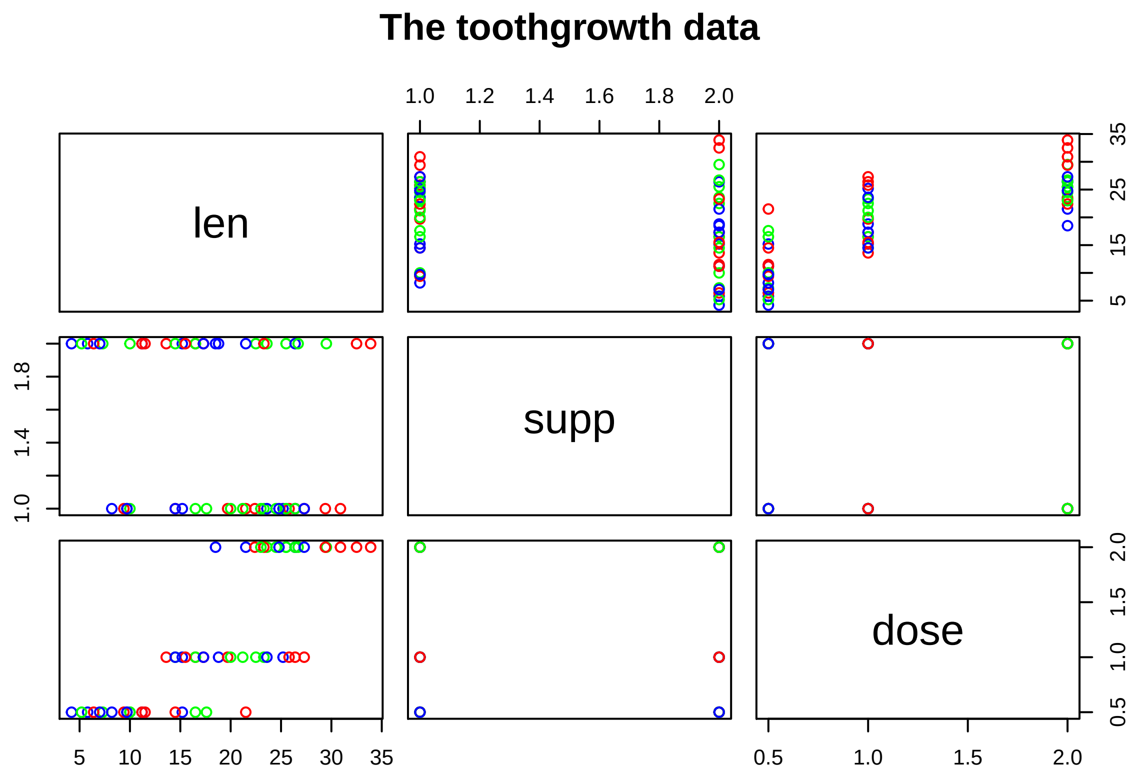
```
## [1] VC OJ
## Levels: OJ VC
```

```r
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

From the above, our data is a data frame of 60 observations and 3 varaibles vis: `len`(length), `supp` (supplements) and `dose`(dose level administered). We also notice that `len` and `dose` are number class whereas `supp` is a factor variable with two levels: `OJ` (Orange juice) and `VC` (vitamin C or ascorbic acid). The summary statistics indicates that the minimum and maximum tooth `length` is 4.20 and 33.90 respectively with an average of 18.81, minimum and maximum `dose` level is 0.5 and 2.0 respectively with an average of 1.167, `suppliment` have a minimum level of 30 for both `OJ` and `VC`. We will plot a scater plot to visualize our data.

```r
plot(ToothGrowth, main = "The toothgrowth data", col = c("blue","red", "green"))
```



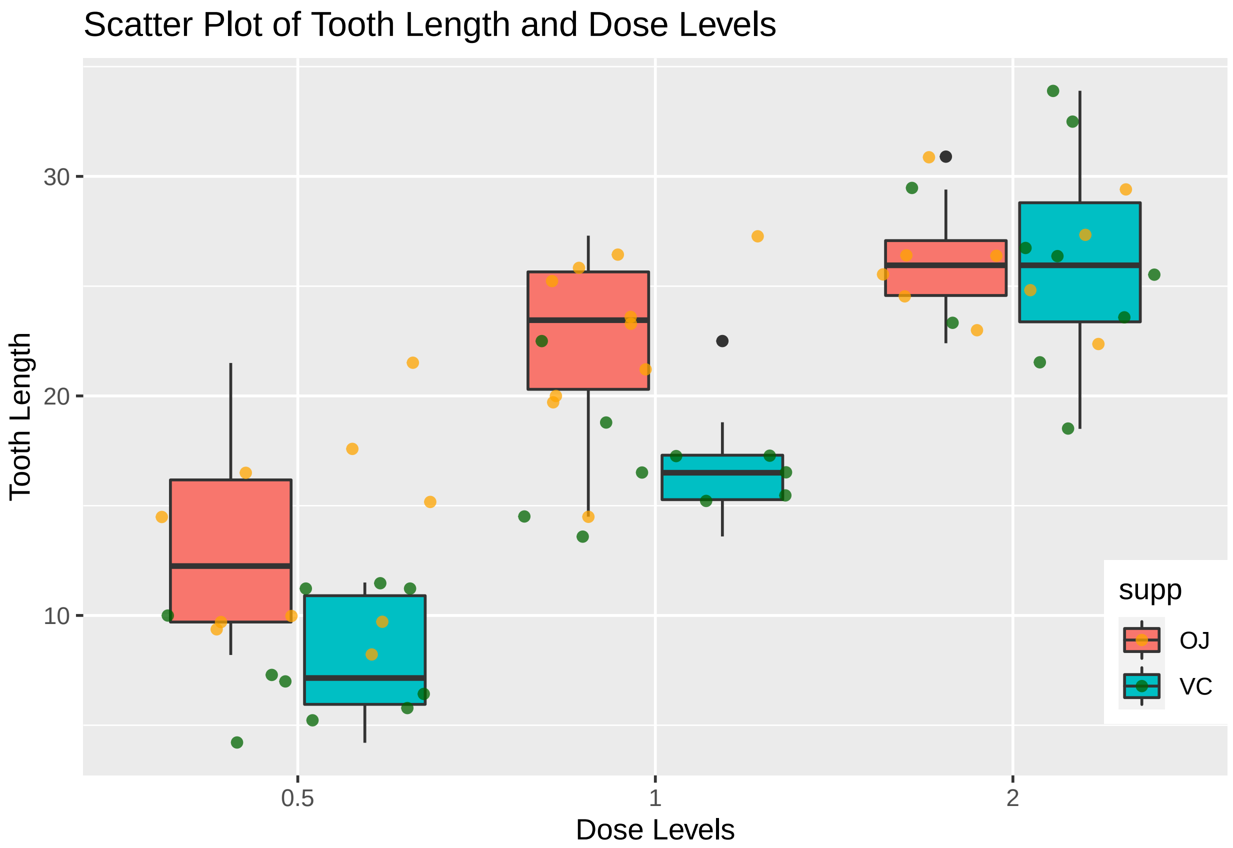
It seems that variable `dose` is rather a factor then a numeric value as seen in its unique value entries.

Therefore, it will be converted into a factor variable.

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

We will produce a scatter box plot to depict tooth length by dose and type of adminstration

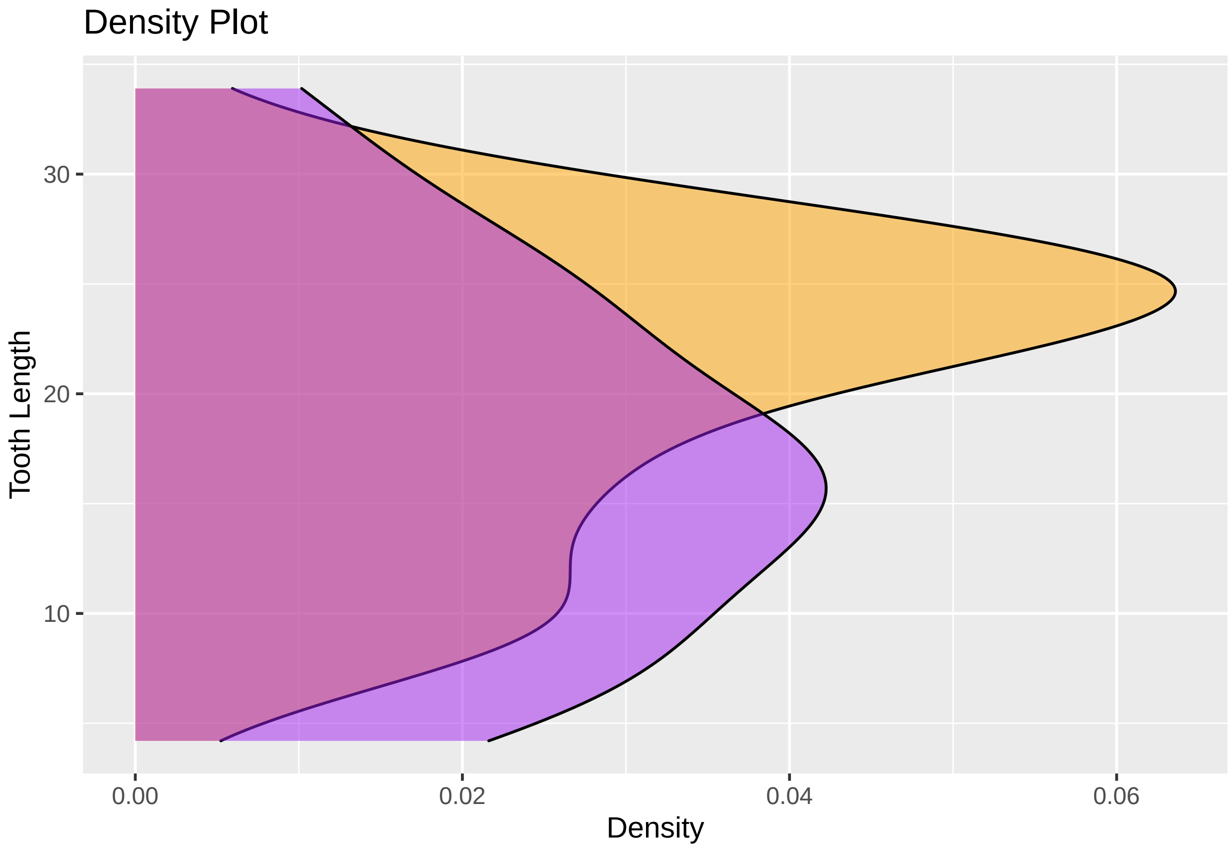
```
set.seed(123)
boxcat <- ggplot(ToothGrowth, aes(dose, len)) +
  geom_boxplot(aes(fill = supp)) +
  geom_jitter(alpha = I(3/4), aes(color = supp)) +
  scale_color_manual(values = c("orange","darkgreen")) +
  theme(legend.position = c(1,0.3), legend.justification = c(1,1)) +
  labs(title = "Scatter Plot of Tooth Length and Dose Levels",x = "Dose Levels", y = "Tooth Length")

boxcat
```



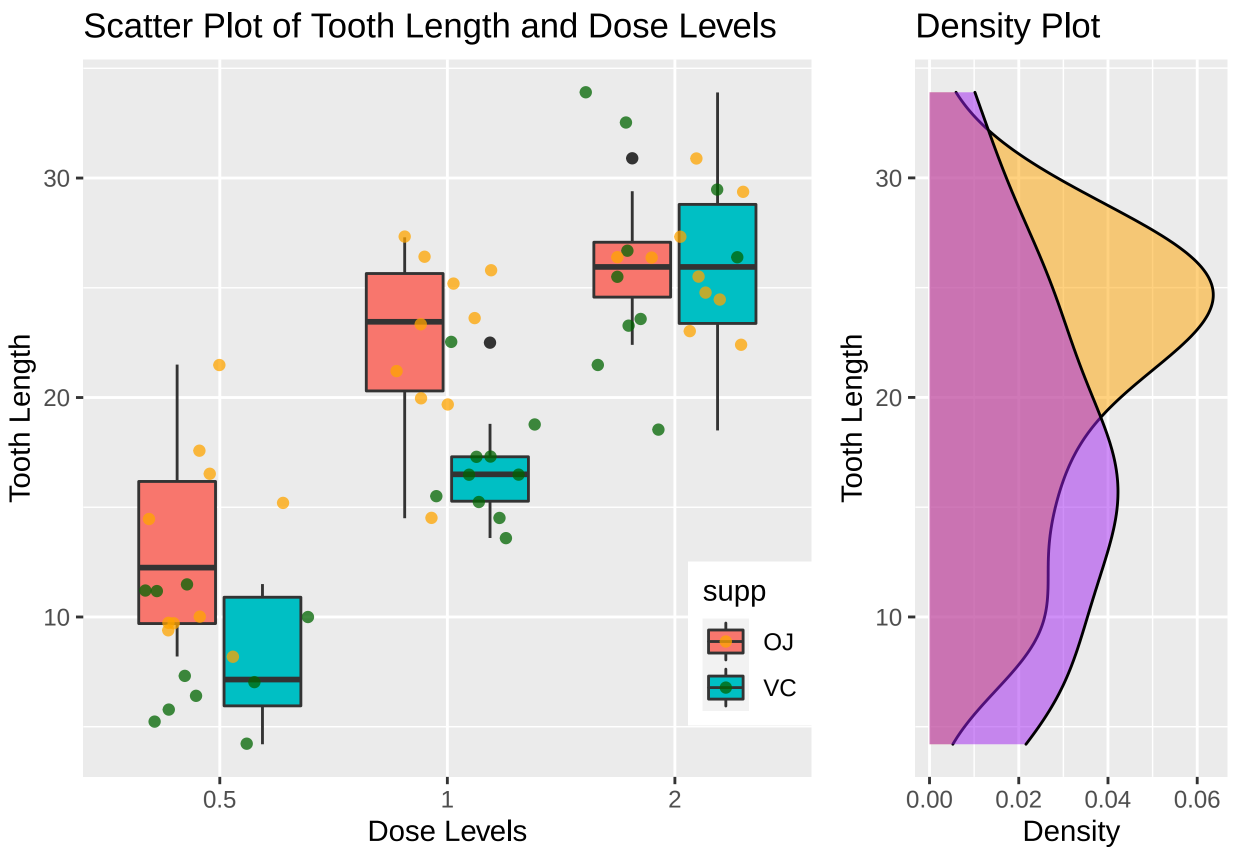
The box plots seem to show, increasing the `dosage` increases the tooth growth. Orange juice is more effective than ascorbic acid for tooth growth when the dosage is 0.5 to 1.0 milligrams per day. Both types of supplements are equally as effective when the dosage is 2.0 milligrams per day. To get a clearer picture a density plot for comparison between `Tooth Lengths` with respect to `Dose Levels` and `supplement` is performed in the following codes below.

```
# Plotting Marginal Density of `Tooth Lengths`.

plt <- ggplot(ToothGrowth,aes(len,fill = supp)) +
  geom_density(alpha = .5) +
  coord_flip() +
  scale_fill_manual(values = c("orange","purple")) +
```

```r
  theme(legend.position = "none") +
  labs(title = "Density Plot", y = "Density", x = "Tooth Length")

plt
```



```r
grid.arrange(boxcat, plt, ncol = 2, nrow = 1, widths = c(4, 2))
```



FURTHER ANALYSIS:

We will compute the mean and varaince by application method (supp). from our results OJ have a mean of 20.66 and a varaince of 43.63. Ascorbic acid have a mean of 16.96 and a varinace of 68.32.

```
appmthd <- split(ToothGrowth$len, ToothGrowth$supp)
sapply(appmthd, mean)
```

```
##       OJ       VC
## 20.66333 16.96333
```

```
# Varaince of supplement
sapply(appmthd, var)
```

```
##       OJ       VC
## 43.63344 68.32723
```

We will perform similar analysis for dose level.

```
dozmean <- split(ToothGrowth$len, ToothGrowth$dose)
sapply(dozmean, mean)
```

```
##    0.5      1      2
## 10.605 19.735 26.100
```

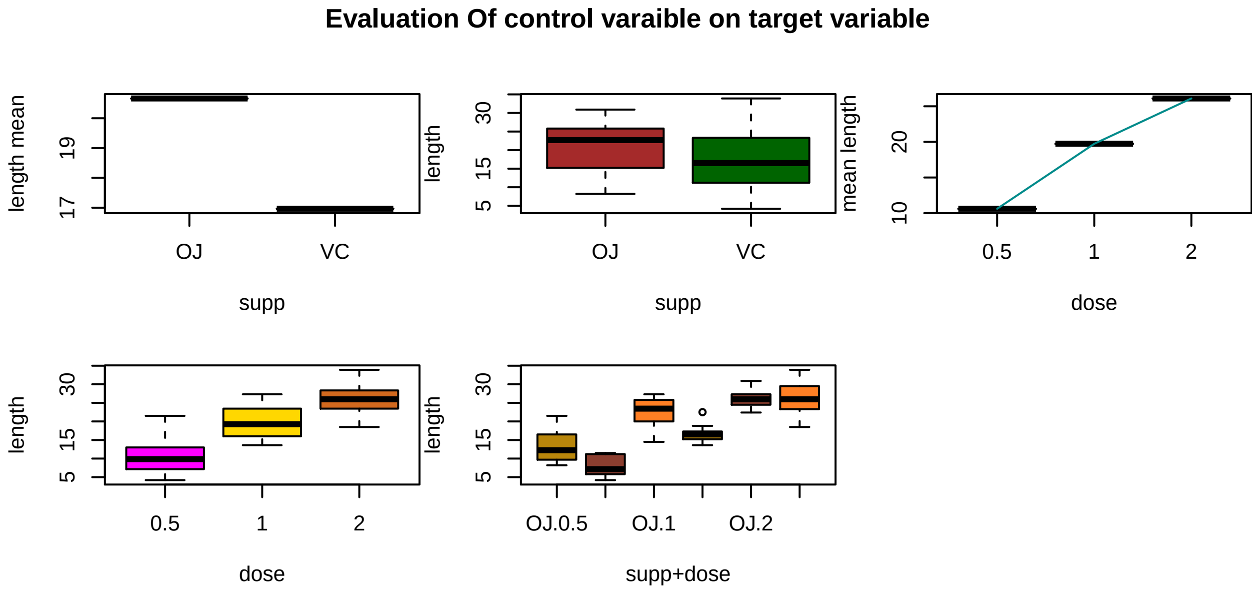
```
# Variance for each dose
sapply(dozmean, var)
```

```
##      0.5        1        2
## 20.24787 19.49608 14.24421
```

We will perform a plot of the control varaible as a realtion to the target varaible to aid us in formulating our test of hyothesis for this analysis. The following plots produces plots for the measure of the following relationship:

1. Plot of tooth length (`len`) against supplement (`supp`)
2. box plot of `len` vs `supp`
3. line plot of `len` vs `dose`
4. box plot of `len` vs `dose`
5. box plot of `len` vs `dose` and `supp` interaction effect.

```r
par(mfrow = c(3,3), mar = c(4,4,2,0), oma = c(0,0,2,0))
plot(aggregate(len~supp,ToothGrowth,mean), ylab = "length mean", col = c("blue","red"))
boxplot(len~supp,ToothGrowth,xlab = "supp", ylab = "length", col = c("brown", "darkgreen"))
plot(aggregate(len~dose,ToothGrowth,mean), pch = 19, ylab = "mean length")
lines(aggregate(len~dose,ToothGrowth,mean), col = c("cyan4","red", "blue"))
boxplot(len~dose,ToothGrowth,xlab = "dose", ylab = "length", col = c("magenta","gold", "chocolate") )
boxplot(len~supp+dose,ToothGrowth,xlab = "supp+dose", ylab = "length", col = c("darkgoldenrod","coral4"
title(main = "Evaluation Of control varaible on target variable",outer = T)
```



HYPOTHESIS TESTING USING CONFIDENCE INTERVAL:

In this work we are going to evaluate the individual effect of control variables `supp` and `dose` on the target variable `len`, as well as their interaction effect. Assuming that a higher tooth length `len` value indicates a higher impact and a higher measure of `dose` indicates a higher dose, a first evaluation of the last plot above yields the following hypotheses : 1. `Supp(OJ)` has a higher impact on the target variable `len`. 2. Higher measure of control variable `dose` effect on target variable `len`. 3. The combined effect of the control variables `supp` and `dose`, shows that `OJ` has higher impact on target variable `len` for `dose` measures of 0.5 and 1. 4. For combined impact of control variables `supp` and `dose`, `OJ` and `VC` have same impact on target variable `len` for `dose` at measure 2mg.

Hypothesis #1:

`Supp`(`OJ`) has a higher impact on the target variable `len`.

```
HYP1 = t.test(len~supp, paired=F, var.equal=F,data=ToothGrowth)
print(HYP1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

```
HYP1$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

```
HYP1$p.value
```

```
## [1] 0.06063451
```

From our result, we notice that the `p-Value = 0.0606` is greater than `=0.05` (for confidence interval of 95%). The `confidence interval =(-0.171, 7.571)` for the difference of the means of each group spans 0, hence null hypothesis is Failed to Reject. Hence hypothesis one cannot be rejected implying that Orange juice has higher impact on tooth length.

Hypothesis #2:

Higher measure of control variable `dose` effect on target variable `len`.

For this hypothesis, we will consider it in three ways: A. That `dose` measure of 1.0mg has a higher impact on tooth length than `dose` measure of 0.5mg. B. That `dose` measure of 2.0mg has a higher impact on tooth length than `dose` measure of 1.0mg C. That `dose` measure of 2.0mg has a higher impact on tooth length than `dose` measure of 0.5mg

2A: `dose` measure of 1.0mg has a higher impact on tooth length than `dose` measure of 0.5mg——————————————————-

```
HYP2A <- t.test(len~dose,paired=F,var.equal=F,data=ToothGrowth[ToothGrowth$dose%in%c(0.5,1),])
print(HYP2A)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```
HYP2A$conf.int
```

```
## [1] -11.983781  -6.276219
## attr(,"conf.level")
## [1] 0.95
```

```
HYP2A$p.value
```

```
## [1] 1.268301e-07
```

From our result, we notice that the `p-Value = 1.268301e-07` is less than `=0.05` (for confidence interval of 95%). The `confidence interval =(-11.983781  -6.276219)` for the difference of the means of each measure level does not spans 0, hence null hypothesis is Rejected. Hence hypothesis 2A is failed to reject implying that dose level of 1.0mg does not really have higher impact on tooth length than dose level 0.5mg.

#2B: dose measure of 2.0mg has a higher impact on tooth length than dose measure of 1.0mg ————————————————-

```
HYP2B <- t.test(len~dose,paired=F,var.equal=F,data=ToothGrowth[ToothGrowth$dose%in%c(1,2),])

print(HYP2B)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

```
HYP2B$p.value
```

```
## [1] 1.90643e-05
```

```
HYP2B$conf.int
```

```
## [1] -8.996481 -3.733519
## attr(,"conf.level")
## [1] 0.95
```

From our result, we notice that the `p-Value = 1.90643e-05` is less than `=0.05` (for confidence interval of 95%). The `confidence interval =(-8.996481 -3.733519)` for the difference of the means of each measure level does not spans 0, hence null hypothesis is Rejected. Hence hypothesis 2B is failed to reject implying that dose level of 2.0mg does not really have higher impact on tooth length than dose level 1.0mg.

2C: dose measure of 0.5mg has a higher impact on tooth length than dose measure of 2.0mg ———————————————-

```
HYP2C <- t.test(len~dose,paired=F,var.equal=F,data=ToothGrowth[ToothGrowth$dose%in%c(0.5,2),])

print(HYP2C)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

```
HYP2C$p.value
```

```
## [1] 4.397525e-14
```

```
HYP2C$conf.int
```

```
## [1] -18.15617 -12.83383
## attr(,"conf.level")
## [1] 0.95
```

From our result, we notice that the `p-Value = 4.397525e-14` is less than `=0.05` (for confidence interval of 95%). The `confidence interval =(-18.15617 -12.83383)` for the difference of the means of each measure level does not spans 0, hence null hypothesis is Rejected. Hence hypothesis 2C is failed to reject implying that dose level of 2.0mg does not really have higher impact on tooth length than dose level 1.0mg.

By way of conclusion from the above two analysis Hypothesis 2 is Failed to Reject.

Hypothesis #3:

In this hypothesis we are looking at: The combined effect of the control variables `supp` and `dose`, to show that `OJ` has higher impact on target variable `len` for `dose` measures of 0.5, 1 and 2. The following code will prepare the data for further analysis.

```
Dose0.5 <- subset(ToothGrowth, dose %in% c(0.5))
Dose1.0 <- subset(ToothGrowth, dose %in% c(1.0))
Dose2.0 <- subset(ToothGrowth, dose %in% c(2.0))
```

The `Null hypothesis` is: There is no correlation between the `Delivery Method` and `Tooth Length` for the given `Dose Level`. And we will consisder it in three different level

3A: "OJ" has higher impact for dose 0.5

```
HYP3a <- t.test(len~supp,paired = F,var.equal = F, data = Dose0.5)
print(HYP3a)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
HYP3a$p.value
```

```
## [1] 0.006358607
```

```
HYP3a$conf.int
```

```
## [1] 1.719057 8.780943
## attr(,"conf.level")
## [1] 0.95
```

From the result above `p-Value =0.006358607` is less than =0.05 (for confidence interval of 95%). The `confidence interval = (1.719057 8.780943)` for the difference of the means of the `supp` and `dose = 0.5` does not span 0, hence null hypothesis is Rejected, hence hypothesis 3a is Failed to Reject.

3B: "OJ" has higher impact for dose 1.0mg

```
HYP3b <- t.test(len~supp,paired = F,var.equal = F, data = Dose1.0)
print(HYP3a)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
HYP3b$p.value
```

```
## [1] 0.001038376
```

```
HYP3b$conf.int
```

```
## [1] 2.802148 9.057852
## attr(,"conf.level")
## [1] 0.95
```

From the result above `p-Value =0.001038376` is less than =0.05 (for confidence interval of 95%), `confidence interval = (2.802148 9.057852)` for the difference of the means the `supp` and `dose=1.0mg` does not span 0, hence null hypothesis is Rejected, implying that hypothesis 3b is Failed to Reject.

3C: "OJ" has higher impact for dose 2.0mg

```
HYP3c <- t.test(len~supp, paired = F,var.equal = F, data = Dose2.0)
print(HYP3c)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

```
HYP3c$p.value
```

```
## [1] 0.9638516
```

```
HYP3c$conf.int
```

```
## [1] -3.79807  3.63807
## attr(,"conf.level")
## [1] 0.95
```

From the result above `p-Value =0.9638516` is greater than =0.05 (for confidence interval of 95%), `confidence interval = (-3.79807  3.63807)` for the difference of the means the `supp` and `dose=2.0mg` does span 0, hence null hypothesis does not failed to Rejected, implying that hypothesis 3b is accepted.

By way of conclusion for Hypothesis 3 is Failed to Reject for lower doses of `supp` and does not fail to reject for higher `dose`.

Hypothesis #4:

In this hypothesis we are looking at: The combined effect of the control variables `supp` and `dose`, to show that `OJ and VC` has higher impact on target variable `len` for `dose` measures of 2.0mg. The following code will produce the result of our analysis.

```
HYP4 <- t.test(len~supp, paired = F,var.equal = F, data = ToothGrowth[ToothGrowth$dose == 2,])
print(HYP4)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

```
HYP4$p.value
```

```
## [1] 0.9638516
```

```
HYP4$conf.int
```

```
## [1] -3.79807  3.63807
## attr(,"conf.level")
## [1] 0.95
```

From the result above `p-Value =0.9638516` is greater than =0.05 (for confidence interval of 95%), `confidence interval = (-3.79807  3.63807)` for the difference of the means the `supp` and `dose=2.0mg` does span 0, hence null hypothesis does not failed to Rejected, implying that hypothesis 4 is accepted.

It will be observed that the result for hypothesis four is similar to the result of hypothesis 3c. Hence we will say that `supp`types and `dose at higher levels` does not differ in effect towards tooth `length`.

CONCLUSIONS:

Based on our analysis the following conclusions are arrived at:

1. Increase in Supplement `Dose Levels` leads to overall increase in `Tooth Length`.

2. `Supplement types` has no overall significant impact on `Tooth Length`, but for `0.5` and `1.0 Dose levels`. `Orange Juice` increases `Tooth Length` more faster compared to `Ascorbic Acid/ Vitamin`, but for `2.0mg Dose Level` there is no significant difference in the increase of `Tooth Length` by both `Supplement`.

3. For combined impact of control variables, there is significant difference on target variable `Tooth length` for different values of `supplement` for `dose 0.5 and 1`. There is no significant difference for different values of `supplement` for `dose 2`.

ASSUMPTIONS NEEDED FOR THE CONCLUSIONS:

1. Data provided is independently distributed. Members of the sample population, i.e. the 60 guinea pigs, are representative of the entire population of guinea pigs. This assumption allows us to generalize the results.

2. The experiment was done with random assignment of guinea pigs to different Supplement `Dose Level` categories and Supplement `Delivery Methods` to take care of noise that might affect the outcome.

3. For the `t-tests`, the variances are assumed to be different for the two groups being compared. This assumption is less stronger than the case in which the variances are assumed to be equal.
   4.Higher value of “length” indicates a higher impact of the `supplement`.
4. Higher value of “dose” indicates increased dosages administration.
5. Data follows T distribution as the observations are limited.