Statistical Inference Project - Part 2

Overview:

The ToothGrowth data has 60 observations that capture length of the tooth *(len)*, supply methods *(Orange Juice- OJ or Vitamin C - VC)*, and dosage *(0.5,1,2 mg)*. The objective of this report is to do exploratory analysis on this data set and derive conclusions on the probably correlation or impact of supply methods or dosage of each method over the length of the tooth.

Loading and Exploratory analysis :

load data ToothGrowth and create a data table for ease of subsetting etc. Taking a summary of the data we observe that there are two types in supp and 4 levels for dose and against each there are about 10 observations each for length of tooth as seen in the table output below summary details.

```
data(ToothGrowth)

dt <- data.table(ToothGrowth)
dt[which(dt[, len > (max(len)-5)]), ] # Look at top 5 length records
```

```
##     len supp dose
## 1: 33.9   VC    2
## 2: 32.5   VC    2
## 3: 29.5   VC    2
## 4: 30.9   OJ    2
## 5: 29.4   OJ    2
```

```
summary(dt)
```

```
##       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
```

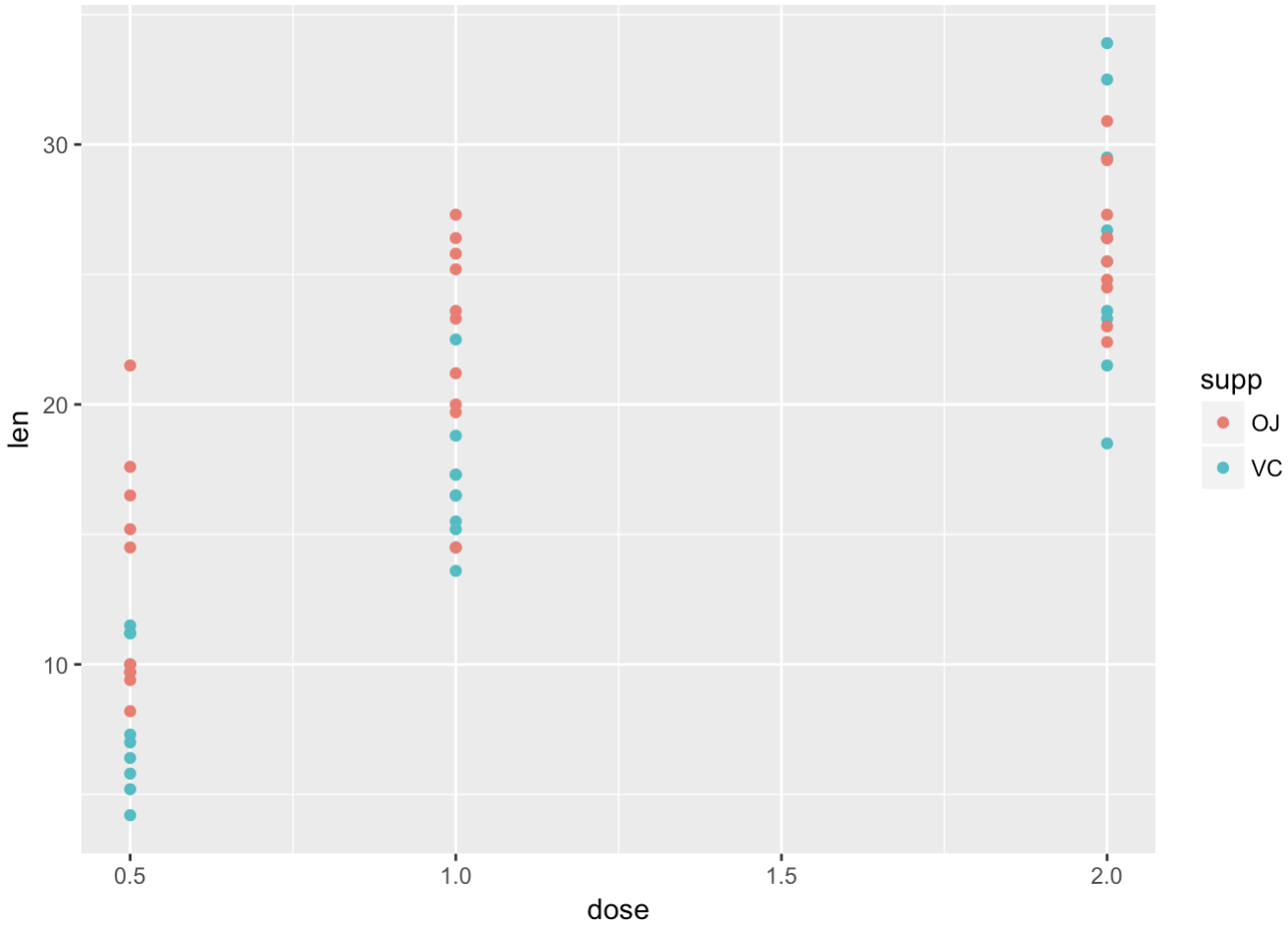
```
table(dt$supp, dt$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

Cursory analysis:

Let us also plot length of tooth using dose as a predictor by supply methods.

```
with(dt, qplot(dose,len, colour=supp) )
```




```
ag <- aggregate(dt$len, by = list(dt$supp), mean) # aggregate mean grouped by supp
ag2 <- aggregate(dt$len, by = list(dt$dose,dt$supp), mean) # aggregate mean grouped by dose and supp.

print(ag)
```

```
##   Group.1        x
## 1      OJ 20.66333
## 2      VC 16.96333
```

```
print(ag2)
```

```
##   Group.1 Group.2     x
## 1     0.5      OJ 13.23
## 2     1.0      OJ 22.70
## 3     2.0      OJ 26.06
## 4     0.5      VC  7.98
## 5     1.0      VC 16.77
## 6     2.0      VC 26.14
```

From the figure above, it is evident that

1. At lower doses OJ seems to have more impact on the length compared to VC but at 2 mg doese, VC overtakes by having to contribute for maximum length in teh given dataset. While this cannot be a conclusion, given the information / data this is the behaviour it is observed.
2. There is no strong correlation of relationship/correlation between dose and length given the supp type.

Let us evaluate the predictability of the variables *doese vs supp* over the outcome (length) by taking mean grouped by different predictors.
Difference between means of length when grouped by supp or when grouped by *dose, supp*. At a high glance, there is not much that we can conclude. The objective was to see if the computed mean is same for both the groups to see which variable has more influence on the predictability on length. Hence, it is valuable to perform hypothesis testing using T distribution tests on the dataset grouped by either supp or by dose and supp.

T Distribution Analysis to evaluate Hypothesis Testing, Confidence Interval and p-Value:

First we evaluate the relation between two groups of len by supp methods running a *t.test*.

```
t.test(dt$len[dt$supp=="OJ"], dt$len[dt$supp=="VC"], paired = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  dt$len[dt$supp == "OJ"] and dt$len[dt$supp == "VC"]
## 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 of x mean of y
##  20.66333  16.96333
```

Note that p-Value is greater than the level of significance (.05). Which means that mean of both the groups are not same but more than that given that the confidence interval includes 0 and pValue > significance level, it is **not statistical significant and hence we fail to reject *null* hypothesis**, which in our case means that both means are close and same and have not very significant influence on the the length, given the data set..

We then run t tests for various combination of two groups of doses over the len outcome variable. In all the t-tests, we consistently find that pValue is 0 (rounded) and confidence interval does not include the mean. Hence, it is **statistically significant and hence reject the null hypothesis**, which in our case means that the dosage have considerable impact on the length of tooth, given the data set.

```
t.test(dt$len[dt$dose== 0.5], dt$len[dt$dose== 1], paired = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  dt$len[dt$dose == 0.5] and dt$len[dt$dose == 1]
## 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 of x mean of y
##    10.605    19.735
```

```
t.test(dt$len[dt$dose== 1], dt$len[dt$dose== 2], paired = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  dt$len[dt$dose == 1] and dt$len[dt$dose == 2]
## 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 of x mean of y
##    19.735    26.100
```

```
t.test(dt$len[dt$dose== 0.5], dt$len[dt$dose== 2], paired = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  dt$len[dt$dose == 0.5] and dt$len[dt$dose == 2]
## 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 of x mean of y
##    10.605    26.100
```

Conclusion:

From the statistical significance for different sets of variables, it is evident that the mean for both the groups are different when grouped by supp or dose, but statistical significance shows marked difference when run toward *dose*. Hence, it is safer to conclude **(subject to the key assumptions listed below)** that *dose* have more impact on the length of the tooth compared to sheer influence of *supp OJ versus VC*.

Assumptions: considered here to draw conclusion are:

1. Though the dataset has 60 observations which is very small to decide the influence of predictors on the outcome, it is assumed to be a good representaton of the population.
2. Popullation will not be too skewed due to external influences on which there will be no control.