# Statistical Inference - Course Project - Part II

***Woldetsadick Selam Getachew***

***Thursday, October 23, 2014***

---

# Part I: Simulation Exercises

The following analysis was made using the R statistic software version R version 3.1.1 (2014-07-10) on a computer using Windows OS version x86_64-w64-mingw32/x64 (64-bit) and completed on Thu, Oct 23 2014, 9:50:04 PM.

## 1. Explanatory Data Analysis

For the second part of the project, the ToothGrowth dataset in the R datasets package is used to make some basic inferential data analysis.
The dataset monitors the length of teeth (mm) 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).

```
#Loading data set
library(datasets)
data(ToothGrowth)
data <- ToothGrowth
# Some Observations on the Data
name <- names(data)
nas <- sum(is.na(data))
Dim <- dim(data)
data$dose <- ifelse(data$dose == .5, "LOW", ifelse(data$dose == 1, "MID", "HIGH"))
for(i in 1:3){if(i!=1) {data[,i] <- as.factor(data[,i])} else {data[,i] <- as.numeric(data[,i])}}
```

First some basic observations of the dataset. After loading **ToothGrowth** dataset, it is renamed **data** for the sake of convenience. The names of the variables in the dataset are `len, supp, dose` corresponding to the respectively to the length of teeth (in mm) in each guinea pig, the delivery methods (or supplement type) and the dose levels of vitamin C.
Observe that the number of missing values in the dataset is exactly equal to `0` and that dimensions of said dataset are `60, 3`. It is to be remembered that three dose levels of Vitamin are coded by level in mg in the dataset. For this analysis however the preference is to make the `dose` variable more descriptive by substituting levels .5, 1 and 2 by LOW, MID and HIGH respectively, corresponding to more descriptive levels of dose.
This process might change the dose variable a factor variable. However to make sure that it is indeed the case, class numeric was assigned to `len` variable while `supp` and the new `dose` variables are assigned class factor, meaning that the latter two variables are categorical variables.

```
#Short Summary of Data
library(psych)
a <- summary(data$supp); b <- summary(data$dose); c <- describe(data$len)
```

Observe that `supp` variable takes values `OJ, VC` respectively for Orange Juice and Ascorbic Acid and exactly `30, 30` times for each value. Similarly, `dose` variable takes values `HIGH, LOW, MID` respectively and exactly `20, 20, 20` times for each value.
A more detailed summary for variable `len` is given by the function `describe()` of `psych package`. The mean of the variable is `18.8133` while its median is `19.25`, the near equality of these statistics show that the mass of observations with values higher and lower to the mean are equal. The variance of the variable is `1` while its maximum and minimum values are `33.9` and `4.2` respectively, for a range of `29.7`. The skewness of `-0.1425` even if slightly negative indicating a slightly longer lower tail basically assumes a symmetrical distribution. However with a kurtosis value of `-1.0425` the distribution is clearly platykurtic, corresponding to thin tails, indicating that distribution is less clustered around the mean value compared to the normal distribution.

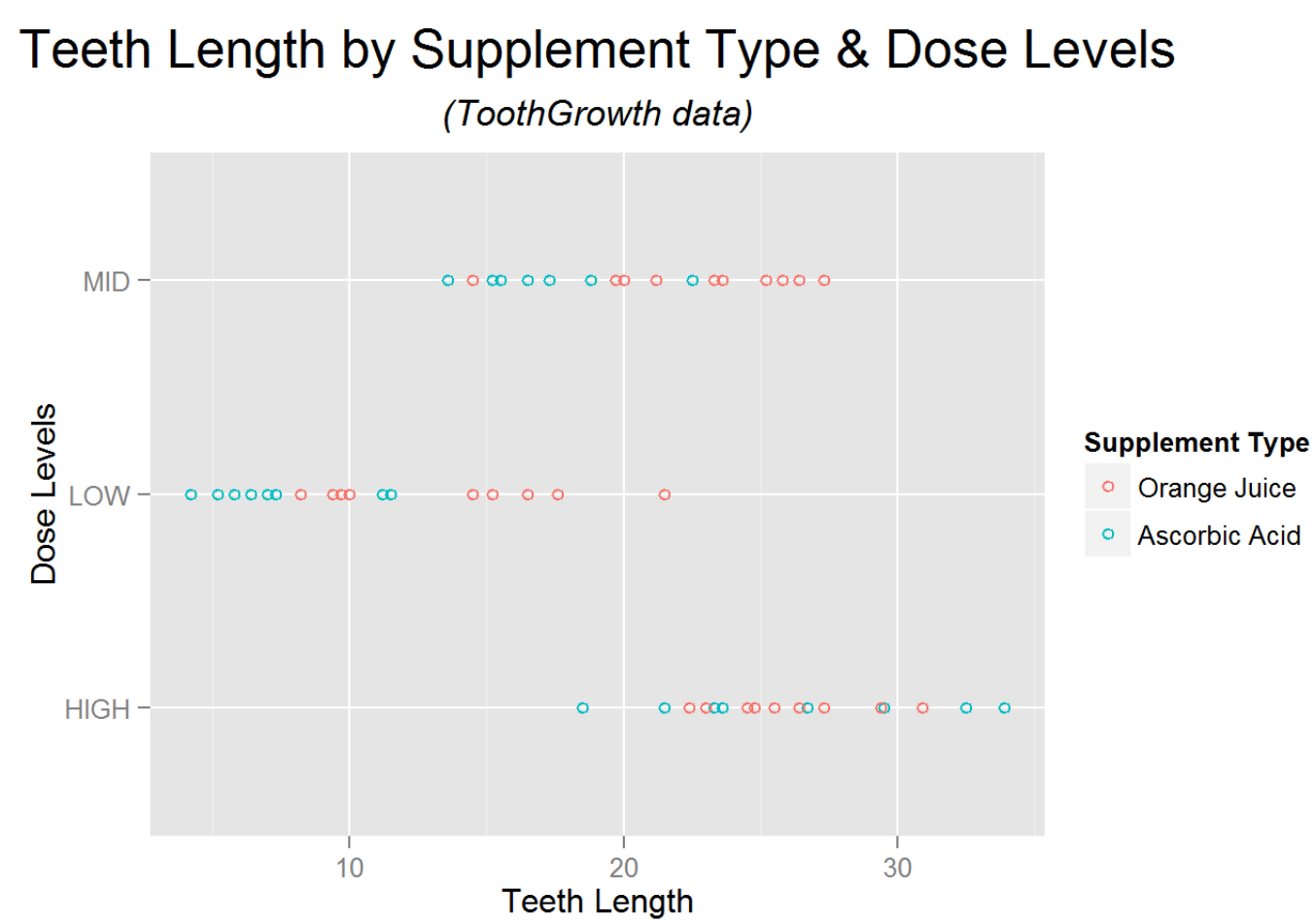
# 2. Summary of the Data

In sections one above, one can see that the data is evenly spitted between groups within `supp` variable and within `dose` variables. But not only should the data be evenly spitted between groups looking at just one categorical variable but also the combination of the two.

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

```
##
##      HIGH LOW MID
##   OJ   10  10  10
##   VC   10  10  10
```

The table up shows how many observations we have for all combinations of values of the two categorical variables, each combinations has 20 observations. The data is hence perfectly balanced.



First observe that the data doesn't seem to have outliers, data points being mostly grouped. As we can imagine higher dose of Vitamin C seem to stimulate teeth growth, but differences in teeth growth along supplement type line is more difficult to assess. In average it seems however that Orange Juice does better in low and medium levels, while ascorbic acid does better in higher levels.

```
##                     Tooth Length [ 4.2,13.6) [13.6,19.7) [19.7,25.5) [25.5,33.9]
## Supplement Type Dose
## OJ              HIGH                      0           0           4           6
##                 LOW                       5           4           1           0
##                 MID                       0           1           6           3
## VC              HIGH                      0           1           3           6
##                 LOW                      10           0           0           0
##                 MID                       0           9           1           0
```

One can also cut length in quartile groups and study a 3-ways contingency table to investigate further the above assumptions. The table up is such a table. Observe that no matter the dosage levels, it seems like supplement Orange Juice does systematically better.
Let's directly look at differences in mean of Tooth length between the different groups, along `dose` first, `supp` second and both categorical variables in the end.

```
tapply(data$len, list(data$dose), mean);tapply(data$len, list(data$supp), mean);tapply(data$len, l
ist(data$supp, data$dose), mean)
```

```
##  HIGH   LOW   MID
## 26.10 10.61 19.73
```

```
##    OJ    VC
## 20.66 16.96
```

```
##     HIGH   LOW   MID
## OJ 26.06 13.23 22.70
## VC 26.14  7.98 16.77
```

Indeed it seems like in average of teeth length, higher doses seem to perform better than than lower doses, and Orange Juice seem to perform better than Ascorbic Acid when considering only one of the categorical variables. However, when considering the interaction between the two variables, Orange Juice perform much better in lower and medium doses, while at higher doses the difference in performance between the two types of delivery is very small.

# 3. T - test, Confidence Intervals and Hypothesis Tests

All conclusion hereinafter are made at 95% confidence interval. Now one can test the differences in mean between the different groups using the t-test.
For variable supp this is not a problem as it has only two categories that we need to compare, for variable dose however one need to create 3 subset of the original data, each using only two categories of the three dose variable categories. Then 3 T-Tests is conducted, one for each combination of categories.

```
d <- t.test(len ~ supp, data = data)
data1 <- subset(data, dose=="LOW"|dose=="MID"); data2 <- subset(data, dose=="LOW"|dose=="HIGH"); d
ata3 <- subset(data, dose=="MID"|dose=="HIGH")
e <- t.test(len ~ dose, data = data1); f <- t.test(len ~ dose, data = data2); g <- t.test(len ~ do
se, data = data3)
```

The null hypothesis of the t-test is that the mean between the two groups are equal, the alternative being that difference is not equal to 0. In the case of the group OJ mean against group VC mean in teeth length, the t-test gives a p-value of `0.0606`, and hence one cannot reject the hypothesis that difference between the two means are equal to zero. And so difference in mean teeth length observed in data could be accidental.
For the difference between low dose group mean and medium dose group mean, the p-value for the t-test is ≈ `0`, between low dose group and high dose group the p-value for the t-test is ≈ `0` and finally between medium dose group and high dose group the p-value for the t-test is ≈ `0`. Hence one can reject the null hypothesis that difference between the two means are equal to zero, meaning difference in mean teeth length observed in data could not be accidental, vitamin C dose levels have incidence on teeth length.
Following, t-tests to test the differences between the different groups created by the interaction of the two categorical variable can be done. The easiest way to test difference of means in groups created by interactions of these two variables is to simply cut the data along the dosage variable, then test difference of mean between groups formed by supplement type.This is what is done in the code chunk below.

```
dosevals <- c("LOW","MID","HIGH")
pvals <- sapply(dosevals, function(vals){
         subd <- subset(data, dose == vals)
         h <- t.test(len ~ supp, data = subd)
         round(h$p.value, digits=4)})
```

Within the low dosage group, the difference in means of teeth length between the groups using Ascorbic Acid and Orange Juice as supplement are found to be statistically significant as the p-value of the t -test equal to `0.0064` can permit to reject the null hypothesis. Similarly, within the medium dosage group, the difference in means of teeth length between the groups using Ascorbic Acid and Orange Juice as supplement are found to be statistically significant as the p-value of the t -test is equal to `0.001`. On the contrary, within the high dosage group the t-test gives a p-value of `0.9639` cannot permit to reject the null hypothesis, and hence difference in mean observed between the two groups within the high dosage group can be due to chance.

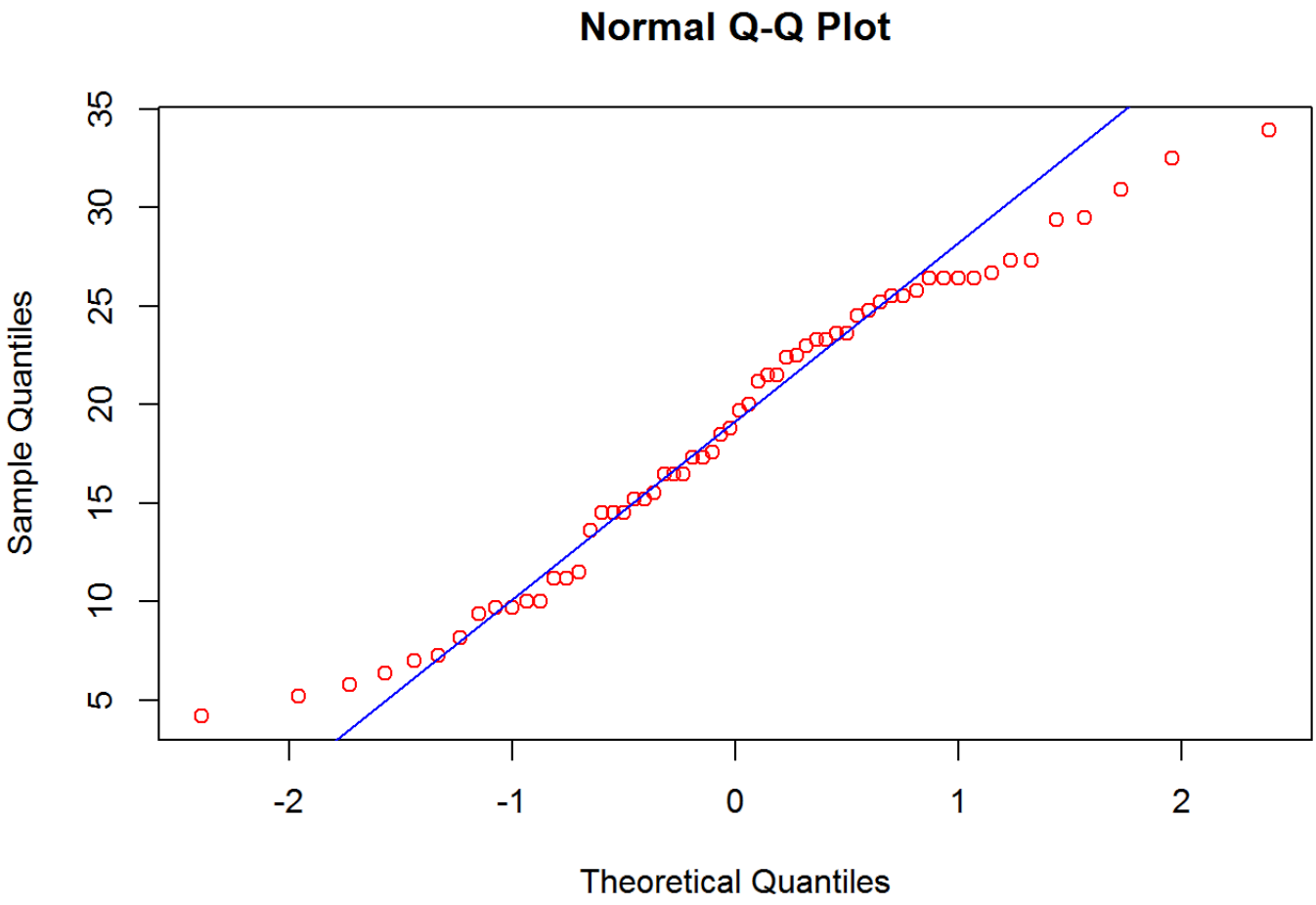
# 4. conclusions and Assumptions

In **conclusion** one can say that the supplement type Orange Juice or Ascorbic Acid by itself has no effect on the growth of teeth. On the contrary the dosage levels have a statistically significant impact on teeth growth, in fact the higher the dosage, the higher the teeth growth.
On the other side, while within the high dosage group the supplement type has no significant impact on teeth growth, it seems like that for low and medium dosages, using Orange Juice as supplement has more effect than Ascorbic Acid.
However, some assumptions were made when applying t-test.

- Randomized assignment of the animals to the different sub-groups controlling effectively for effect of counfonders
- Equal variance across pairwise groups being compared (Homogeneity).
- Normality of the outcome variable distribution (specially Skewness)
- Outliers
- Sample representativeness

# 5. Annexes

While assumptions 1 & 5 can be estimated to be met because of a good study design. One can check for outliers in scatterplot above, and there are none. The `len` as seen above present no skewness and the QQplot below shows that it follows roughly a normal distribution.One can study homogeneity with Bartlett test.




```
i <- bartlett.test(len ~ interaction(supp, dose), data = data)
```

The null hypothesis of this test being that all K groups variances are equal against each other (Talking about groups along the two categorical variables here). The p value of the test is equal to `0.2261` which doesn't permit to reject equality of variance at 95% confidence interval.