# Coursera Johns Hopkins -- Statistical Inference Project Report

## Part 2 : ToothGrowth data Analysis in the R datasets package.

### Understanding the Data

The aim of experiment that generated the `TooGrowth` data set was to see if the bioassay on odontoblast cell size was a reliable test of Vitamin C intake, so the question the original authors were trying to answer was "Is this a reliable test?".

The gathered data consists of measurements of the mean length of the odontoblast cells harvested from the incisor teeth of a population of 60 guinea pigs, fed a diet with one of 6 Vitamin C supplement regimes. The Vitamin C was administered either in the form of **Orange Juice (OJ)** or chemically pure Vitamin C in **aqueous solution (VC)**. Each animal received the same daily **dosage of Vitamin C (either 0.5, 1.0 or 2.0 milligrams)** consistently.

Since each combination of supplement type and dosage was given to 10 randomly selected animals, 60 animals were required for the study. After 42 days of experimentation, the animals were euthanized, their incisor teeth were harvested and subject to analysis to determine the mean length (in microns) of the odontoblast cells.

Now, we'll be trying to answer that question : **Is the odontoblast cell length a reliable test for Vitamin C intake** ?

### Hypothesis Tests (HT)

Let's set some hypotheses before exploring the data

$H_0$ : The odontoblast cell length is not a reliable test for Vitamin C intake

$H_a$ : The odontoblast cell length is a reliable test for Vitamin C intake

### Exploratory Data Analysis (EDA)

```
str(ToothGrowth)
```

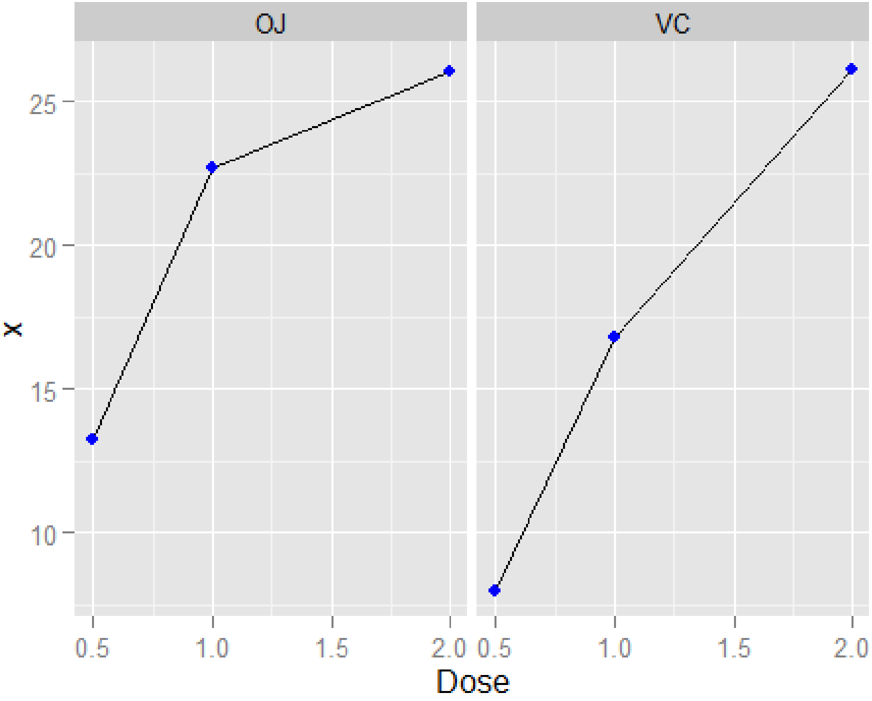
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
summary(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
```

Now, let's plot data according to supp and dose

```
agg <- aggregate(ToothGrowth$len,list(Supp = ToothGrowth$supp,Dose = ToothGro
wth$dose),mean)
```



Looks like the data plotting go in favor of the alternative Hypothesis. In specific, for the two supports of Vitamin C, the two plots show that the cell lenghts evolve according to the Dose.

## Statistical Inference

We'll consider groups based on the support and doses. The groups according to the experiment protocol are independent, thus not paired.

As the size of groups is small, we'll rely on t-tests to draw statistical inferences. We'll assume that the **significance level is : 0.05**.

Consider following data sets

```
library(datasets)
OJ20 <- subset(ToothGrowth, supp == "OJ" & dose == 2.0)
OJ10 <- subset(ToothGrowth, supp == "OJ" & dose == 1.0)
VC20 <- subset(ToothGrowth, supp == "VC" & dose == 2.0)
VC10 <- subset(ToothGrowth, supp == "VC" & dose == 1.0)
```

We'll do 2 separate t-tests, on OJ observations and on VC observations.

### *OJ data*

```
t.test(OJ20$len , OJ10$len, mu=0, paired = FALSE, alternative = "greater")

## 
##  Welch Two Sample t-test
## 
## data:  OJ20$len and OJ10$len
## t = 2.2478, df = 15.842, p-value = 0.0196
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7486236       Inf
## sample estimates:
## mean of x mean of y 
##     26.06     22.70
```

We can notice here that the `p-value = 0.0196` for the hypothesis  that the 2 groups are similar  is very low, thus the null Hypothesis could be rejected for this sub case.

### *VC data*

```
t.test(VC20$len , VC10$len, mu=0, paired = FALSE, alternative = "greater")

## 
##  Welch Two Sample t-test
## 
## data:  VC20$len and VC10$len
## t = 5.4698, df = 13.6, p-value = 4.578e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.346525      Inf
## sample estimates:
## mean of x mean of y 
##     26.14     16.77
```

We can notice here that the `p-value = 0.00004578` for the hypothesis for the hypothesis that the 2 groups are similar  is extremely low, thus the null Hypothesis could be rejected for this sub case.

We can conclude from the 2 t-tests above that **the $H_0$ hypotheis could be rejected, thus The odontoblast cell length can be considered as a reliable test for Vitamin C intake**.