# ToothGrowth Analysis

## Synopsis

In this project, we take a look at the ToothGrowth data available in base R, which is an experiment conducted on the tooth of Guinea Pigs. Our goal is to conduct an analysis of this data using hypothesis testing and/or confidence intervals. In particular, we are interested in the impact of Vitamin C delivery method and on the impact of Vitamin C dosage in the growth of the Guinea Pig's teeth.

### Loading & understanding the Data

According to the documentation, this dataset shows the results of an experiment on 60 guinea pigs run in 1952 by C. I. Bliss. During the experiment, each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice (OJ) or ascorbic acid (a form of vitamin C and coded as VC). The response is the length of odontoblasts (cells responsible for tooth growth). Let's run several basic requests in order to better understand the results.

```
data(ToothGrowth)
colnames(ToothGrowth) <- c("length", "supplement", "dose")
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ length    : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supplement: 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)
```

```
##      length      supplement      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  
```

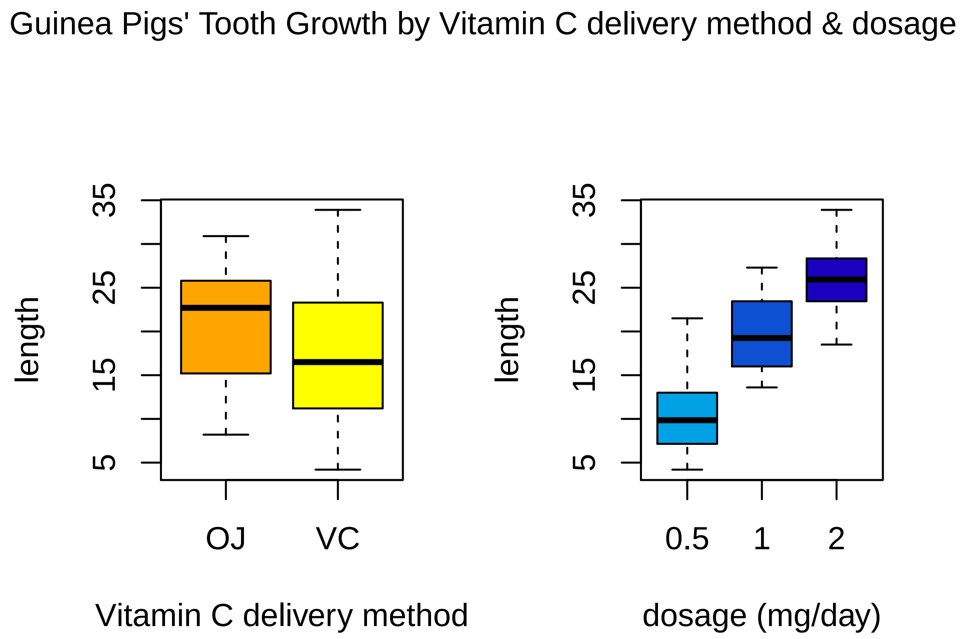
The mean of the length of odontoblasts is 18.81.

The supplement variable is a factor variable with 2 levels "OJ" (Orange Juice) and "VC" (Ascorbic Acid). Both levels have 30 occurences each.

Similarly, the dose variable is a numeric variable with only 3 possible values : 0.5, 1, 2. Each value has 20 occurences (10 under OJ factor and 10 under VC factor).

Let's draw two boxplots :

```
par(mfrow=c(1,2), oma = c(0,0,1,0))
boxplot(length~supplement, data = ToothGrowth, col = c("orange", "yellow"), xlab = "Vitamin C delivery
boxplot(length~dose, data = ToothGrowth, col = c("#00A1E5", "#0D50D2", "#1B00BF"), xlab = "dosage (mg/da
mtext("Guinea Pigs' Tooth Growth by Vitamin C delivery method & dosage", outer = TRUE)
```



From the 2 boxplots, we can observe that the Orange Juice delivery method seems to have a stronger impact on tooth growth on average. The ascorbic acid has a much lower median, however, it shows a wider range of values.

Unsurprisingly, the dosage boxplot seems to tells us that a stronger dosage of Vitamin C is associated with a faster tooth growth.

## Laying a null hypothesis and run permutations

Now that we have a good understanding of the ToothGrowth data, let's formulate the hypothesis that the Vitamin C delivery method has no impact on the tooth growth. If our hypothesis is true, the labels "OJ" and "VC" would be irrelevant.

Let's analyze the distribution of the observations in these 2 categories to see if they are similar. To do so, we use the permutation method. Let's sample 10 000 times the supplement labels and recalculate an overall mean.

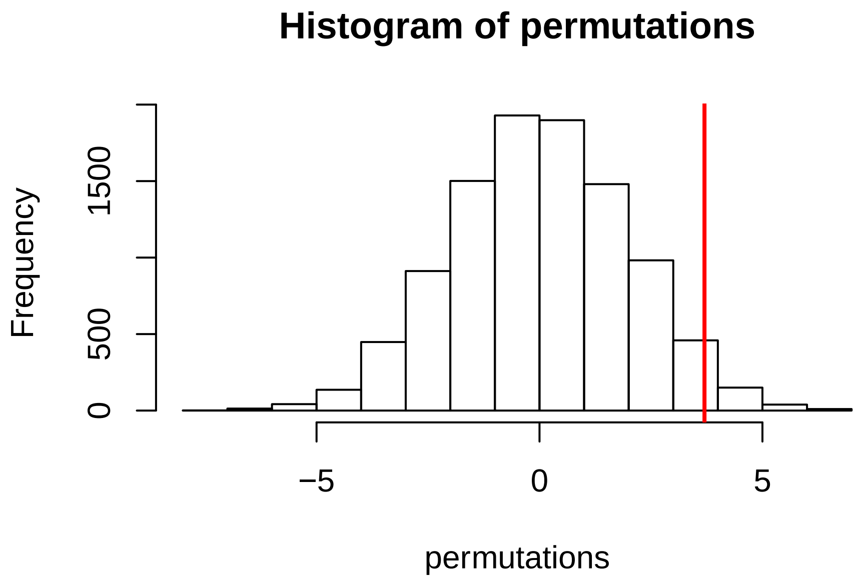
```
set.seed(2019)
y <- ToothGrowth$length
group <- as.character(ToothGrowth$supplement)
testStat <- function (w,g) mean(w[g=="OJ"])-mean(w[g=="VC"])
observedStat <- testStat(y, group)
permutations <- sapply(1:10000, function(i) testStat(y, sample(group)))
observedStat
```

```
## [1] 3.7
```

```
mean(permutations > observedStat)
```

```
## [1] 0.0286
```

```
hist(permutations)
abline(v = observedStat, col = "red", lwd = 2)
```



Only 2.9% of the permutations had an higher difference average than our original dataset. Since the p-value is so low, we reject our null hypothesis for any reasonable level of alpha.

Let's run another analysis on dosage. This time our null hypothesis is that the mean of measures with dosage = 1 is equal to the mean of the measures with dosage = 2.
Let' subset the initial data in two series d1 (dosage = 1) and d2 (dosage =2) and calculate a confidence interval between the mean of both series.

```
d1 <- subset(ToothGrowth, ToothGrowth$dose == 1)
d2 <- subset(ToothGrowth, ToothGrowth$dose == 2)
t.test(d2$length,d1$length,paired = FALSE, var.equal = TRUE)$conf
```

```
## [1] 3.735613 8.994387
## attr(,"conf.level")
## [1] 0.95
```

The result shows confidence at 95% that the difference in means of the two series is located in the [3.74 - 8.99] interval. Our null hypothesis is rejected and we can formulate an alternate hypothesis that the mean of the d2 series is superior of the mean of d1 series.

## Results

In this project, we ran an exploratory analysis on the ToothGrowth data. Our first impressions using boxplots were that delivery method and dosage both had an impact on the tooth growth.
To make sure of it, we stated two null hypothesis : 1. delivery method has no impact on tooth growth
2. dosage has no impact on tooth growth
Both null hypothesis were rejected using the permutation method (for the first hypothesis) and a t.test (for the 2nd hypothesis).