# Part 2: Basic Inferential Data Analysis

Alexander Kuznetsov

05/27/2018

## Overview

The purpose of this project is to analyze ToothGrowth data collected during experiments on Guinea Pigs. In these experiments animals were given different doses of vitamin C and orange juice as supplements to study their effect on the growth of teeth.

```r
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=65),tidy=TRUE)
```

## Exploratory Data Analysis

"ToothGrowth" dataset is stored in variable 'data'. First, functions such as 'dim', 'summary' and 'str' are to be used. Functions 'head' and 'tail' will be used to get an understanding of how actual dataset looks like.

```r
data <- ToothGrowth
head(data)
```

```
##    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
tail(data)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

```r
dim(data)
```

```
## [1] 60  3
```

```r
summary(data)
```

```
##       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
```

```
str(data)
```

```
## '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 of Data

Basic exploratory analysis indicates that dataset has 3 columns for length of teeth, type of supplement and dose of that supplement given to subjects. There are 60 rows with observations. Length and dose are stored as numeric values, while supplement (column 'supp') column is factor variable with two levels: "VC" and "OJ". "VC" stands for Vitamin C and "OJ" - for orange juice. For convenience and to make visualization more informative, these variables can be written explicitly using full names.

```
data$supp <- ifelse(data$supp == "VC", "Vitamin C", "Orange Juice")
```

Length values have quite significant variability ranging from 4.2 to 33.9 with mean 18.81. Ggplot2 is great way to visualize observations of this dataset. Data will be plotted using boxplot with whiskers where dose and supplement are used as factor variables.

```
library(ggplot2)
ggplot(data, aes(x = factor(dose), y = len, fill = factor(supp))) +
    geom_boxplot(position = position_dodge(1)) + labs(title = "Tooth Length",
    x = "Dose (mg/day)", y = "Length", fill = "Supplement")
```
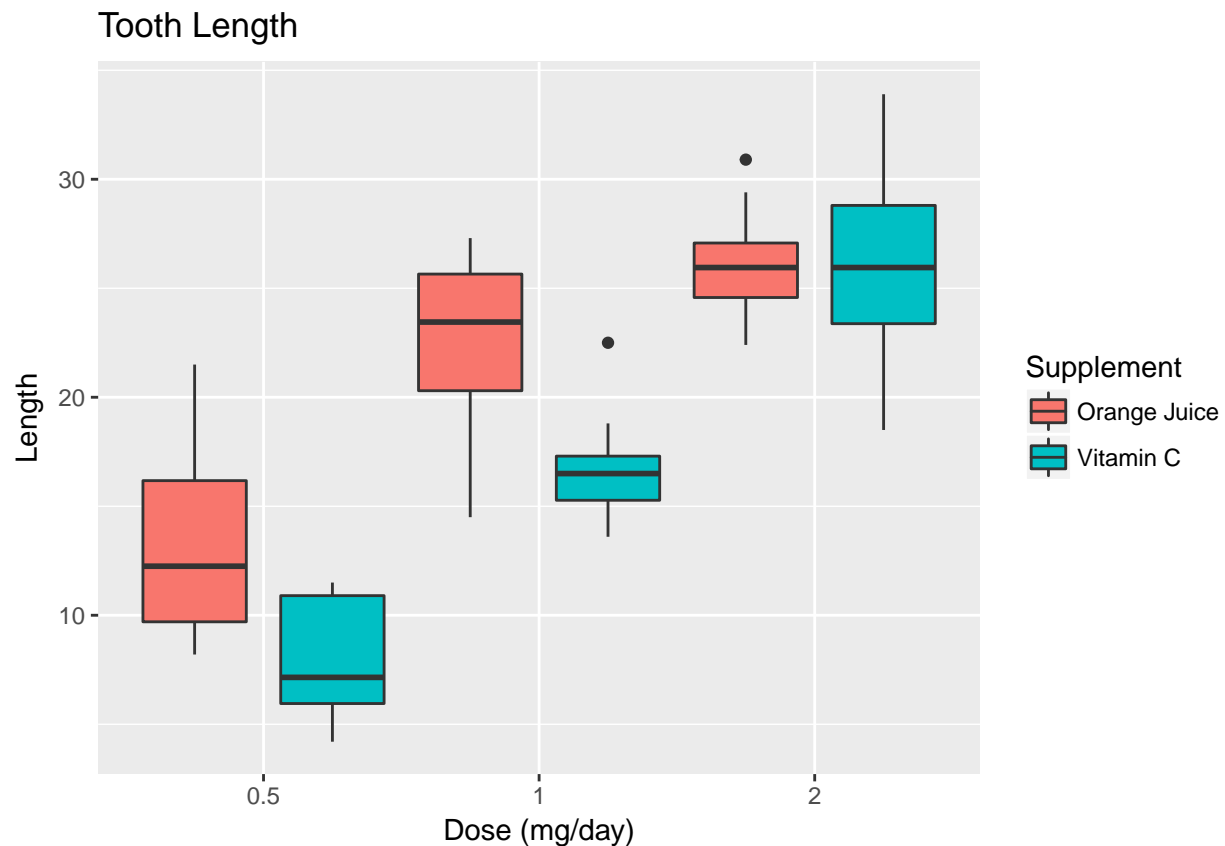


There is clear correlation between length of teeth and dose, as higher doses of both supplements result in

longer teeth length. At lower doses of 0.5 mg/day and 1 mg/day orange juice demonstrates more significant effect on the length of teeth with higher averages than in case of vitamin C. However at 2 mg/day dose both supplements provide on average the same teeth length.

## Hypothesis Tests

Following code is to test significance of differences in observations for each dose of supplement applied to subjects, using t-statistics. As null hypothesis, we assume that there are no differences in average observations for each dose of each supplement. Alternative hypothesis states that differences in means for supplements are statistically significant.

```
ttest05 <- t.test(len ~ supp, data = subset(data, dose == 0.5))
ttest1 <- t.test(len ~ supp, data = subset(data, dose == 1))
ttest2 <- t.test(len ~ supp, data = subset(data, dose == 2))
```

First, assume that at dose 0.5 mg/day, both supplements result in observations which are not statistically different, e.g. there is no difference in using vitamin C or orange juice.

```
ttest05
```

```
##
##  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 Orange Juice    mean in group Vitamin C
##                     13.23                       7.98
```

However, as can be seen from the output of the 't.test', p-value is very small, indicating that probability of rejecting null hypothesis by mistake is quite low. Therefore, null hypothesis is rejected in favor of alternative hypothesis for a dose equal to 0.5 mg/day. At this dose, orange juice has more significant impact on the teeth growth than vitamin C. One can arrive to similar conclusions for dose = 1 mg/day:

```
ttest1
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group Orange Juice    mean in group Vitamin C
##                     22.70                      16.77
```

T-test for 2 mg/day dose returns very high p-value:

```
ttest2
```

```
##
##  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 Orange Juice    mean in group Vitamin C
##                     26.06                      26.14
```

In this case null hypothesis cannot be rejected. Thus, both supplements result in the same effect on teeth growth on average when applied at 3 mg/day dose.


## Conclusions

"ToothGrowth" dataset records 60 observations collected by conducting tests with Guinea Pigs where animals were given two supplements: vitamin C and orange juice to study their effect on the tooth growth. Given relatively small number of observations, Student's t statistics was used to analyze the data. As a null hypothesis we assumed that on average both supplements resulted in similar effect on tooth growth at each dose. Alternative hypothesis stated that effects are different. For doses 0.5 mg/day and 1 mg/day, null hypothesis could be rejected, indicating that differences in the effect of each supplements were quite significant. For dose 2 mg/day null hypothesis could not be rejected in favor of alternative one. For this dose, effects from both supplements were similar.