# Basic Inferential Data Analysis on ToothGrowth Data

Benedict Neo

16/12/2020

## Overview

I will be taking the ToothGrowth data and do a basic eda on the data, then I will use t.test to perform hypothesis testing for the effectiveness of the supplement types on tooth growth length under the respective dose levels.

## Load data

```
data("ToothGrowth")
dt <- ToothGrowth
```

## Basic exploratory Data Analyses

```
str(dt)
```

```
## '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 ...
```

The data has 60 rows and 3 columns

```
unique(dt$dose)
```

```
## [1] 0.5 1.0 2.0
```

Since there are three unique values for dose, I will convert it to a factor type.

```
dt$dose <- factor(dt$dose)
str(dt)
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

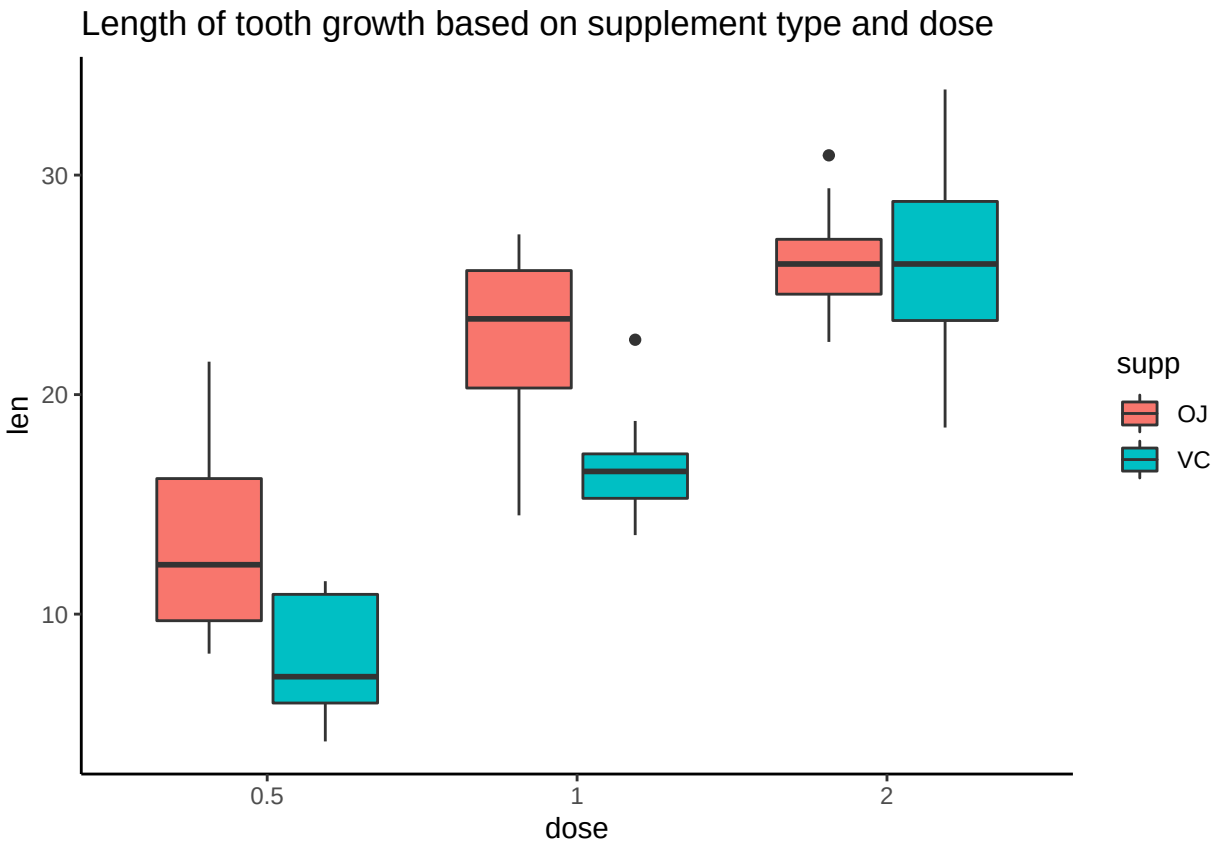
Using summary function to look at statistics of variables

```
summary(dt)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

Plotting box plot based on supplement type and dose

```
library(ggplot2)
ggplot(dt, aes(x = dose, y = len, fill = supp)) +
    geom_boxplot() +
    ggtitle("Length of tooth growth based on supplement type and dose") +
    theme_classic()
```




From the box plot, we can observe that the dosage for 0.5 and 1.0 has bigger differences in length as compared to the dosage for 2.0 mg/day. We can see a trend as the dose increases, the tooth length increases as well. Moreover, from the plot alone, we can tell supplement OJ is more effective for doses 0.5 and 1.0, where as for dose 2, there's not much difference between them.

We can look at the mean for the length of tooth growth for each dose and supp.

load libraries

```
library(dplyr)
library(tidyr)
```

```
dt %>%
    group_by(supp, dose) %>%
    summarize(mean = mean(len), .groups='drop') %>%
    spread(supp, mean) %>%
    mutate(diff = abs(VC - OJ))
```

```
## # A tibble: 3 x 4
##   dose     OJ    VC  diff
##   <fct> <dbl> <dbl> <dbl>
## 1 0.5    13.2  7.98  5.25
## 2 1      22.7 16.8   5.93
## 3 2      26.1 26.1   0.08
```

Observing the mean, we can see 2mg/day doses has very small differences as compared to dose of 0.5 and 2. This means it's harder to compare the effectiveness between OJ and VC for dose 2.

To formally test the effectiveness between the two supplement types, we will use t.test to find the p-values and confidence intervals to perform hypothesis testing.

## t.test Hypothesis Testing

- Our null hypothesis would is there is no difference between using OJ and VC
- Our alternative hypothesis is there is a difference between sign OJ and VC
- alpha rate is set at 0.05 as standard

Even though dose 2 has the smallest differences, I will still be performing hypothesis testing for all three dose types.

First we filter the data based on dose. This makes it easier and cleaner for t.test.

```
dose_half <- filter(dt, dose == 0.5)
dose_one <- filter(dt, dose == 1)
dose_two <- filter(dt, dose == 2)
```

#### t-test for 0.5 mg/day dose

```
t.test(len ~ supp, dose_half)
```

```
##
##  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 OJ mean in group VC
##            13.23             7.98
```

**t-test for 1 mg/day dose**

```
t.test(len ~ supp, dose_one)
```

```
##
##  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 OJ mean in group VC
##            22.70            16.77
```

**t-test for 2 mg/day dose**

```
t.test(len ~ supp, dose_two)
```

```
##
##  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 OJ mean in group VC
##            26.06            26.14
```

Forming a table to summarize the t.tests results

```
dose <- c(0.5, 1.0, 2.0)
p_value <- c(0.0064, 0.0010, 0.9639)
conf.int <- c("1.72, 8.78", "2.80, 9.06", "-3.80, 3.64")
decision <- c("Reject null", "Reject null", "Do not reject null")
data.frame(dose, conf.int, p_value, decision)
```

```
##   dose    conf.int p_value           decision
## 1  0.5  1.72, 8.78  0.0064        Reject null
## 2  1.0  2.80, 9.06  0.0010        Reject null
## 3  2.0 -3.80, 3.64  0.9639 Do not reject null
```

As expected, the p-values for dose 0.5 and 1.0 will be very small because of the big differences in mean between them.

Thus, for dose 0.5 and 1.0, since p-values are smaller than 0.5, we reject the null hypotheses that the supplement types don't have a difference on tooth growth. But for dose 2.0 mg/day, we can reject the null as the p-value is greater than 0.5.

## Conclusion

The central assumption for the results is that the sample is representative of the population, and the variables are IID random variables.

For the t.test, two assumptions are made,

1. The data isn't paired, meaning they're independent
2. The variance are different.

With that, in reviewing the t.test, supplement type OC are more effective than VC for doses less than 1.0. But for dose at 2.0 mg/day, there is no difference between the supplement types.