# Basic Inferential Analysis on ToothGrowth Data

Guimiao Zhang

## Overview

This project does a preliminary exploration on ToothGrowth data and provides pairwise T-tests along with 95% Confidence Interval (CI) tests to compare tooth growth by supplement and dose.

## Loading Data & Getting a Basic Data Idea

```
library(datasets); attach(ToothGrowth) # load
str(ToothGrowth) # basic format
with(ToothGrowth, table(supp, dose)) # idea of possible independent variables
```

```
## '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 ...
##     dose
## supp 0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

The dataset is a data frame with 60 observations on 3 variables: tooth length (`len`, numeric), supplement type (`supp`, factor: ascorbic acid, a form of vitamin C `VC` or orange juice `OJ`), and dose in milligrams/day (`dose`, numeric, 0.5, 1, and 2 mg/day).

## Exploratory Data Analysis & Data Summary

```
library(data.table)
dat <- data.table(ToothGrowth) # put into data.table for easy summary
summaries <- dat[, .(obs = .N, mean = mean(len), min = min(len), qt25 = quantile(len, 0.25),
                     median = median(len), qt75 = quantile(len, 0.75),
                     max = max(len), std = sd(len)), by = .(supp, dose)]
# plotting
library(ggplot2); library(gridExtra)
box <- ggplot(ToothGrowth, aes(factor(dose), len)) + ylim(4, 34) + # base
    geom_boxplot(aes(fill = supp), alpha = .8) + # boxplot
    scale_fill_manual(name = 'supplement', values = c('yellow', 'purple')) + # col adjust
    labs(title = 'Boxplot of tooth length', x = 'dose (mg/day)', y = 'tooth length') +
    theme(legend.position = c(1, 0.4), legend.justification = c(1, 1),
          legend.background = element_rect(fill = NA),
          plot.title = element_text(hjust = 0.5))
scatter <- ggplot(summaries, aes(x = factor(dose), y = mean, color = supp, group = supp)) +
    ylim(4, 34) + geom_point(size = 3, alpha = .8) + geom_line() +
    scale_color_manual(name = 'supplement', values = c('yellow', 'purple')) +
    labs(title = 'Scatterplot of the mean tooth length',
```

```
        x = 'dose (mg/day)', y = 'mean tooth length') +
    theme(legend.position = c(1, 0.4), legend.justification = c(1, 1),
          legend.background = element_rect(fill = NA),
          plot.title = element_text(hjust = 0.5))
grid.arrange(box, scatter, ncol = 2, nrow = 1)
```

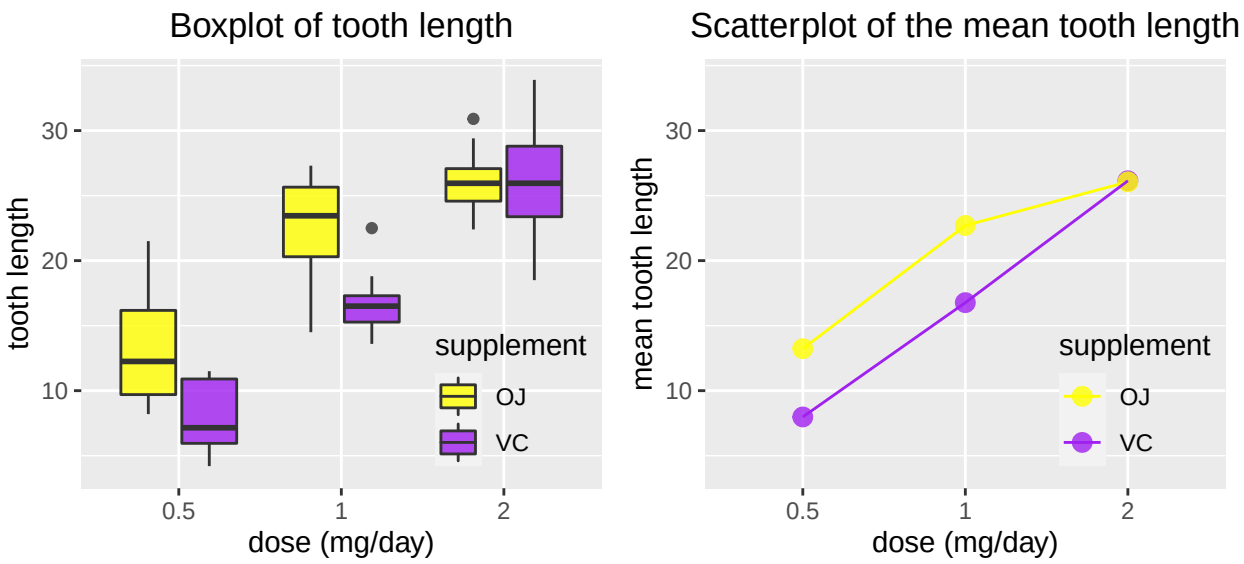


Figure 1: Boxplot and scatterplot for tooth length

The boxplot in figure 1 indicates that there is not significant difference for the means of tooth length VC and OJ controlling the `dose`, whereas the scatterplot suggests an increasing trend as the dose level increases controlling the `supp`. The roughly parallel lines in the scatter plot indicate no significant interaction between `dose` and `supp`.

Table 1: Summary of the tooth length by dose and supplement

| supp | dose | obs | mean | min | qt25 | median | qt75 | max | std |
|---|---|---|---|---|---|---|---|---|---|
| VC | 0.5 | 10 | 7.98 | 4.2 | 5.95 | 7.15 | 10.90 | 11.5 | 2.75 |
| VC | 1.0 | 10 | 16.77 | 13.6 | 15.28 | 16.50 | 17.30 | 22.5 | 2.52 |
| VC | 2.0 | 10 | 26.14 | 18.5 | 23.38 | 25.95 | 28.80 | 33.9 | 4.80 |
| OJ | 0.5 | 10 | 13.23 | 8.2 | 9.70 | 12.25 | 16.18 | 21.5 | 4.46 |
| OJ | 1.0 | 10 | 22.70 | 14.5 | 20.30 | 23.45 | 25.65 | 27.3 | 3.91 |
| OJ | 2.0 | 10 | 26.06 | 22.4 | 24.58 | 25.95 | 27.08 | 30.9 | 2.66 |

## Hypothesis Testing

From the exploratory data analysis above, we assume no significant interaction between `dose` and `supp`. We are going to explore the marginal effects of `supp` and `dose`. As we can only use the method mentioned in the class, we are using pairwise T-test for means to explore the effects of supplement and dose on tooth length. A better approach would be ANOVA test.

**Supplement effect:** $H_{null}:\ \mu_{OJ} = \mu_{VC}$ $H_{alternative}:\ \mu_{OJ} \neq \mu_{VC}$

```
result <- t.test(len ~ supp, ToothGrowth)
c('p-value' = result$p.value, '95 CI lower' = result$conf.int[1],
  '95 CI upper' = result$conf.int[2])
```

```
##     p-value 95 CI lower 95 CI upper
##  0.06063451 -0.17101562  7.57101562
```

$P - value = 0.06063 > \alpha = 0.05$. Fail to reject $H_{null}$ at 0.05 level of significance. Also, the 95% CI contains **0** indicating not enough evidence to claim $H_{alternative}$.

**Dose effect:** $H_{null}:\ \mu_{dose_{0.5}} = \mu_{dose_{1.0}} = \mu_{dose_{2.0}}$ $H_{alternative}:$ **at least one mean differs**

```
result1 <- t.test(len ~ dose, subset(ToothGrowth, dose < 2))
dose.5_vs_1 <- c('p-value' = result1$p.value, '95 CI lower' = result1$conf.int[1],
                 '95 CI upper' = result1$conf.int[2])
result2 <- t.test(len ~ dose, subset(ToothGrowth, dose > .5))
dose1_vs_2 <- c('p-value' = result2$p.value, '95 CI lower' = result2$conf.int[1],
                '95 CI upper' = result2$conf.int[2])
rbind(dose.5_vs_1, dose1_vs_2)
```

```
##                  p-value 95 CI lower 95 CI upper
## dose.5_vs_1 1.268301e-07  -11.983781   -6.276219
## dose1_vs_2  1.906430e-05   -8.996481   -3.733519
```

```
rm(list = ls(all.names = TRUE))
```

At dose level 0.5 and 1.0, $P - value < 0.001 < \alpha = 0.05$. Reject $H_{null}:\ \mu_{dose_{0.5}} = \mu_{dose_{1.0}}$ and claim $H_{alternative}:\ \mu_{dose_{0.5}} \neq \mu_{dose_{1.0}}$ at 0.05 level of significance. Also, the 95% CI is below **0** indicating enough evidence to say $\mu_{dose_{0.5}} < \mu_{dose_{1.0}}$.

At dose level 1.0 and 2.0, $P - value < 0.001 < \alpha = 0.05$. Reject $H_{null}:\ \mu_{dose_{1.0}} = \mu_{dose_{2.0}}$ and claim $H_{alternative}:\ \mu_{dose_{1.0}} \neq \mu_{dose_{2.0}}$ at 0.05 level of significance. Also, the 95% CI is below **0** indicating enough evidence to say $\mu_{dose_{1.0}} < \mu_{dose_{2.0}}$.

As from previous 2 comparisons, we can claim $H_{alternative}:$ at least one mean differs, and even claim $\mu_{dose_{0.5}} < \mu_{dose_{1.0}} < \mu_{dose_{2.0}}$, without comparing `dose` 0.5 with `dose` 2.0.

## Conclusions & Assumptions Needed

To sum up:

- there is no significant difference for the means of tooth length using different supplements `OJ` and `VC`;
- there is a significant increasing trend for the means of tooth length as the dose increases.

To achieve these conclusions, we need to assume:

- all observation subjects, the 60 guinea pigs, are random chosen from the population and independent with others;
- randomly assign one of the 6 combination of treatments (`supp` x `dose`) to 10 randomly chosen pigs from the 60;
- the population are approximately normally distributed with equal variances;
- no significant interaction between `dose` and `supp`.