Effects of vitamin C on tooth growth in guinea pigs

Ricardo Leon

February 23, 2020

Overview

We will going to perform a basic analysis of the tooth growth for guinea pigs database in R. After the study we conclude that the greater the dose, the greater the tooth growth and the most effective delivery method is orange juice (OJ) except by the dose of 2 mg/day that it was not possible to determine between both methods with 95% of confidence.

Exploratory data analysis

Importing the data and checking its structure, we know that it has three columns with two of them numerical (len and dose) and one as factor (supp). From the documentation, we know that dose has only three values so it can be converted to type factor to make easier plotting the data.

```
library(dplyr)
library(ggplot2)
set.seed(42)

data("ToothGrowth")
str(ToothGrowth)
```

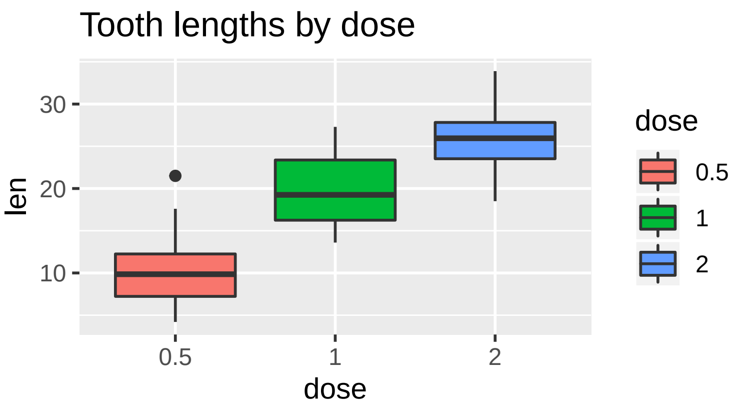
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
tooth_data <- ToothGrowth %>% mutate(dose = factor(dose))
```

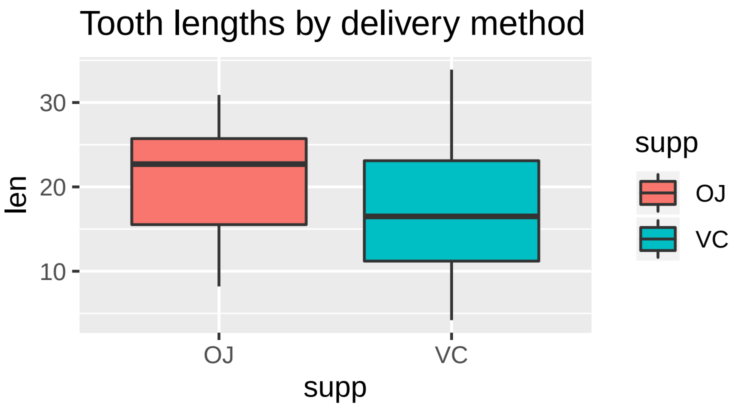
Below there are three plots to visualize data behavior. For each we can state the following hypothesis:

```
H1. The most effective delivery method is orange juice.
H2. The greater the dose, the greater the tooth growth.
```

```
ggplot(tooth_data, aes(x=dose, y=len, group=dose)) +
    geom_boxplot(aes(fill=dose)) +
    ggtitle('Tooth lengths by dose')
```



```r
ggplot(tooth_data, aes(x=supp, y=len, group=supp)) +
    geom_boxplot(aes(fill=supp)) +
    ggtitle('Tooth lengths by delivery method')
```



Basic summary of the data.

Using the dplyr library, we calculate the details of the data grouped by delivery method and dose.

```r
tooth_data %>% group_by(supp) %>%
    summarise(min = min(len), median = median(len), max = max(len), mean = mean(len), std = sd(len))
```

```
## # A tibble: 2 x 6
##   supp    min median   max  mean   std
##   <fct> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 OJ      8.2   22.7  30.9  20.7  6.61
## 2 VC      4.2   16.5  33.9  17.0  8.27
```

```r
tooth_data %>% group_by(dose) %>%
    summarise(min = min(len), median = median(len), max = max(len), mean = mean(len), std = sd(len))
```

```
## # A tibble: 3 x 6
##   dose    min median   max  mean   std
##   <fct> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 0.5     4.2   9.85  21.5  10.6  4.50
## 2 1      13.6  19.2   27.3  19.7  4.42
## 3 2      18.5  26.0   33.9  26.1  3.77
```

Confidence intervals

All tests are going to be executed under the following assumptions:

```
* Data is not paired.
* Variance for different dataset is considered not equal.
* Alternative hypothesis is always trying to prove that the mean of the first dataset is greater
  than the second one.
```

To test the hyphotesis H1, tooth data was filtered by delivery method and a t test was applied:

```r
filter_by_supp <- function(data, supp_value) data %>% filter(supp == supp_value) %>% select(len)
len_by_oj <- filter_by_supp(tooth_data, 'OJ')
len_by_vc <- filter_by_supp(tooth_data, 'VC')
t.test(len_by_oj, len_by_vc, paired = FALSE, var.equal = FALSE, alternative = "g")$conf
```

```
## [1] 0.4682687       Inf
## attr(,"conf.level")
## [1] 0.95
```

The same logic was applied when studying effect by dose (H2). First, look at the confidence interval comparing doses 0.5 and 1:

```r
filter_by_dose <- function(data, dose_value) data %>% filter(dose == dose_value) %>% select(len)
len_by_dose_0.5 <- filter_by_dose(tooth_data, '0.5')
len_by_dose_1.0 <- filter_by_dose(tooth_data, '1')
t.test(len_by_dose_1.0, len_by_dose_0.5, paired = FALSE, var.equal = FALSE, alternative = "g")$conf
```

```
## [1] 6.753323      Inf
## attr(,"conf.level")
## [1] 0.95
```

And then comparing doses 1 and 2:

```r
len_by_dose_2.0 <- filter_by_dose(tooth_data, '2')
t.test(len_by_dose_2.0, len_by_dose_1.0, paired = FALSE, var.equal = FALSE, alternative = "g")$conf
```

```
## [1] 4.17387     Inf
## attr(,"conf.level")
## [1] 0.95
```

Analysis Assumptions

```
* Data is not paired.
* Variance for different dataset is considered not equal.
* Alternative hypothesis is always trying to prove that the mean of the first dataset is
  greater than the second one.
```

Conclusions:

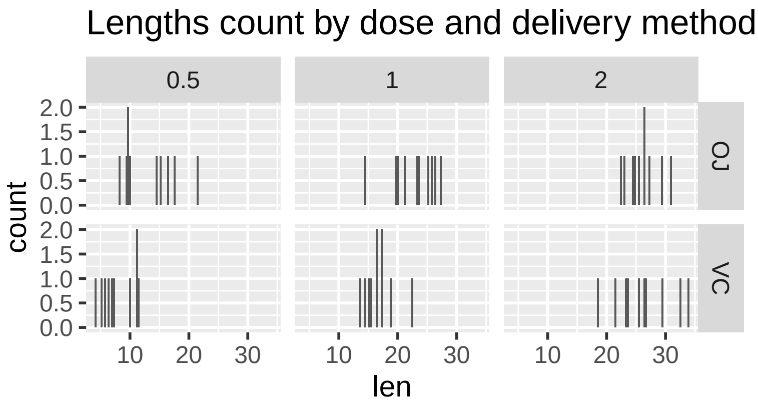
```
1. Dividing the dataset by delivery method we get a confidence interval greater than
   0.4682687 so we confirm than orange juice is (in average) better method than vitamic C.
2. Comparing doses, from 0.5 to 1 and 1 to 2 we got intervals greater than 6.753323 and
   4.17387 so we reach the conclusion that the greater the dose, the greater the tooth growth
3. Both hypothesis are true.
```

Appendix

If we look more in depth for H2, we will find that the truth assertion is not neccesary true when subseting by delivery method is considered.

Looking at the plot, we can get an intuition about this:

```
ggplot(tooth_data, aes(x=len)) +
    geom_bar() +
    facet_grid(supp ~ dose) +
    ggtitle('Lengths count by dose and delivery method')
```



It looks like Orange Juice is not the best option when the dose is high. Repeating the test separating by delivery method we get the following intervals:

```
t_test_by_dose <- function(data, dose_value) {
    oj_data <- data %>% filter(dose == dose_value, supp == 'OJ') %>% select(len)
    vc_data <- data %>% filter(dose == dose_value, supp == 'VC') %>% select(len)
    t.test(oj_data, vc_data, paired = FALSE, var.equal = FALSE, alternative = "g")$conf
}

t_test_by_dose(tooth_data, '0.5')
```

```
## [1] 2.34604      Inf
## attr(,"conf.level")
## [1] 0.95
```

```
t_test_by_dose(tooth_data, '1')
```

```
## [1] 3.356158      Inf
## attr(,"conf.level")
## [1] 0.95
```

```
t_test_by_dose(tooth_data, '2')
```

```
## [1] -3.1335      Inf
## attr(,"conf.level")
## [1] 0.95
```

Thus, for doses 0.5 and 1, orange juice is better (positive confidence interval) but for dose 2 we cannot reject null hypothesis with 95% confidence level.