Analyzing Tooth Growth

Konstantinos Patelis

26/01/2021

Overview - Exploratory Analysis

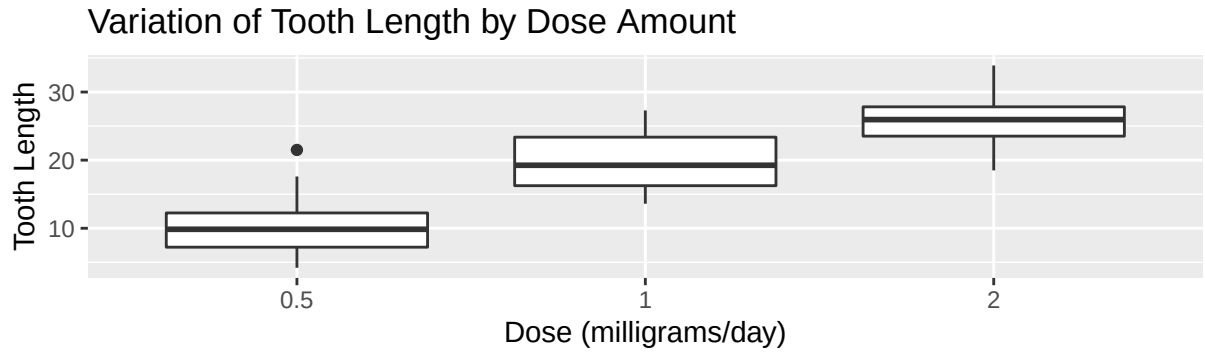
For this report we will have a look at the ToothGrowth dataset from the datasets package. Let's look at some summary statistics for the data and explore the data.

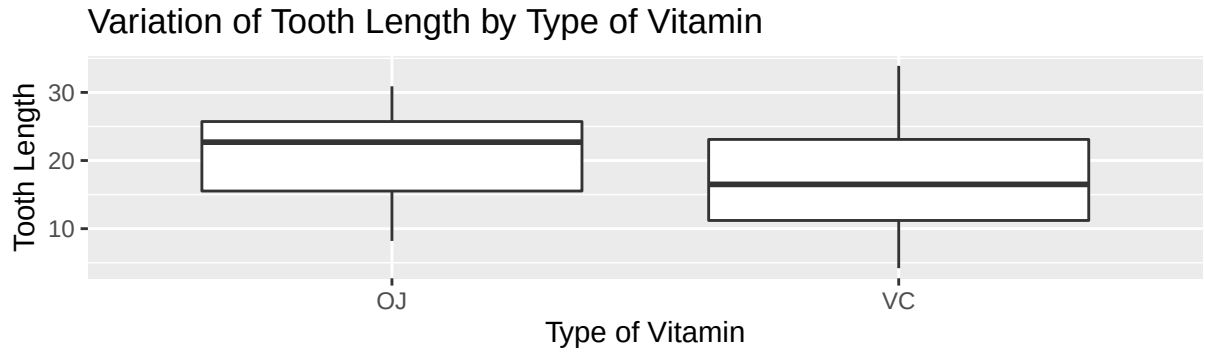
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
##       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
```

```
## # A tibble: 6 x 3
## # Groups:   supp, dose [6]
##   supp   dose     n
##   <fct> <dbl> <int>
## 1 OJ      0.5    10
## 2 OJ      1      10
## 3 OJ      2      10
## 4 VC      0.5    10
## 5 VC      1      10
## 6 VC      2      10
```

The dataset contains sixty observations with three variables, two of which are numeric and one is a factor variable with two levels. The dose variable takes on three distinct values, so it might be more suited to be coded as a factor. Furthermore, there's an equal number of observations across all combinations of the supp and dose variables. The dose variable measures the dose of a vitamin provided in milligrams per day to the guinea pigs, while supp encodes the type of vitamin provided. The final variable, len, measures tooth length of the guinea pigs.





Looking at the above figures, tooth length clearly seems to vary by the amount of vitamin provided to the guinea pig. Similarly, average tooth length is larger after receiving the OJ treatment. We can perform a t-test to conclude with more certainty whether the type of supplement and the dosage amount have an effect on tooth length.

Hypothesis testing

We will perform separate t-tests to check whether there is a difference in average means of tooth length between: - different supplements - 0.5 vs 1 milligram dosage per day - 1 vs 2 milligram dosage per day - 0.5 vs 2 milligram dosage per day

```
##
##  Welch Two Sample t-test
##
## data:  data %>% filter(supp == "OJ") %>% pull(len) and data %>% filter(supp == "VC") %>% pull(len)
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

```
##
##  Welch Two Sample t-test
##
## data:  data %>% filter(dose == 0.5) %>% pull(len) and data %>% filter(dose == 1) %>% pull(len)
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

```
##
##  Welch Two Sample t-test
##
## data:  data %>% filter(dose == 1) %>% pull(len) and data %>% filter(dose == 2) %>% pull(len)
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

```
##
##  Welch Two Sample t-test
##
## data:  data %>% filter(dose == 0.5) %>% pull(len) and data %>% filter(dose == 2) %>% pull(len)
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

We can see that there is not enough evidence to reject the hypothesis that means are different in the case of different supplements provided on the 95% confidence level; we would reject on a 90% confidence level. On the other hand, looking at the different t-tests for varying dosages, it is clear that the means are different, even on 99% confidence level.

Appendix

```
knitr::opts_chunk$set(
    echo = FALSE,
    eval = TRUE
)
library(tidyverse)
library(patchwork)

data <- datasets::ToothGrowth

str(data)
summary(data)
data %>% group_by(supp, dose) %>% count()

p1 <- ggplot(data, aes(x = factor(dose), y = len)) +
        geom_boxplot() +
        labs(
          title = "Variation of Tooth Length by Dose Amount",
          x = "Dose (milligrams/day)",
          y = "Tooth Length"
```

```
        )

p2 <- ggplot(data, aes(x = supp, y = len)) +
        geom_boxplot() +
        labs(
          title = "Variation of Tooth Length by Type of Vitamin",
          x = "Type of Vitamin",
          y = "Tooth Length"
        )

p1 / p2

t.test(data %>% filter(supp == "OJ") %>% pull(len),
       data %>% filter(supp == "VC") %>% pull(len))

t.test(data %>% filter(dose == 0.5) %>% pull(len),
       data %>% filter(dose == 1) %>% pull(len))

t.test(data %>% filter(dose == 1) %>% pull(len),
       data %>% filter(dose == 2) %>% pull(len))

t.test(data %>% filter(dose == 0.5) %>% pull(len),
       data %>% filter(dose == 2) %>% pull(len))
```