Part 2: Basic Inferential Data Analysis

Simone

2/22/2020

Overview

The goal of this project is to explore the ToothGrowth dataset and compare tooth length by supplement type and dose.

Load the dependencies and the dataset

Below is the code to load the libraries required for the analysis and the data.

```
# Dependencies
library(datasets)
library(ggplot2)

# Load the dataset
data(ToothGrowth)
```

Let's check the dataset structure.

```
# Dataset structure
dim(ToothGrowth)
```

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

```
head(ToothGrowth)
```

```
##    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
```

```
sum(!complete.cases(ToothGrowth))
```

```
## [1] 0
```

We can see that the dataset has three columns:

- **len** *numeric* Tooth length
- **supp** *factor* Supplement type ("VC" = ascorbic acid, a fom of vitamin C or "OJ" = orange juice)
- **dose** *numeric* Dose in milligrams/day

Moreover there are 60 observations in total and no NA occurrences.

Summary of the data

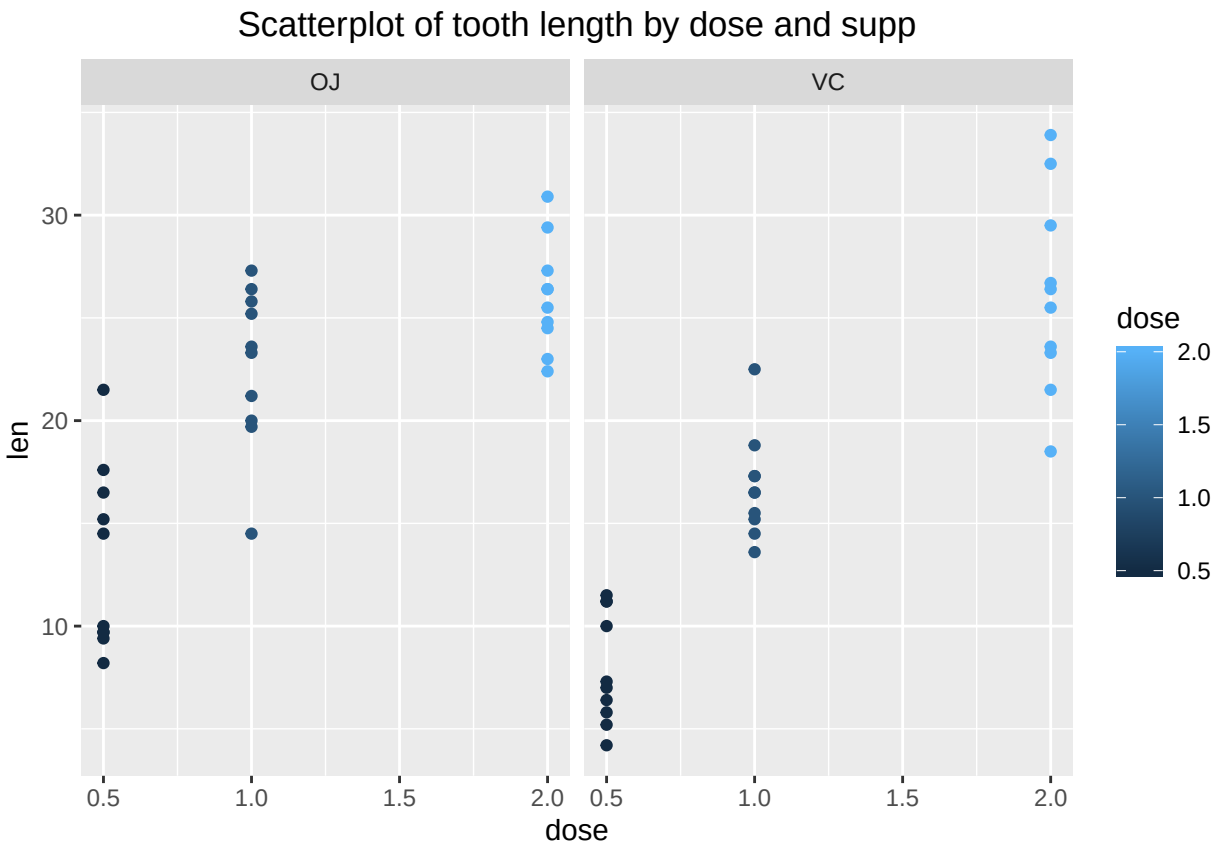
Below are standard ways to summarize the data. We can see that the number of observations are evenly split between all the supplement types and dose categories (10 for each combination of supplement type and dose).

```
# Basic summary
with(ToothGrowth, table(supp, dose))
```

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

Furthermore, we can also plot the tooth length vs dose by supplement type, to start to understand the relationships between the different variables.

```
# Plot of tooth length vs dose by supplement type
ggplot(ToothGrowth, aes(x = dose, y = len, col = dose)) + geom_point() + facet_grid(. ~ supp) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Scatterplot of tooth length by dose and supp")
```

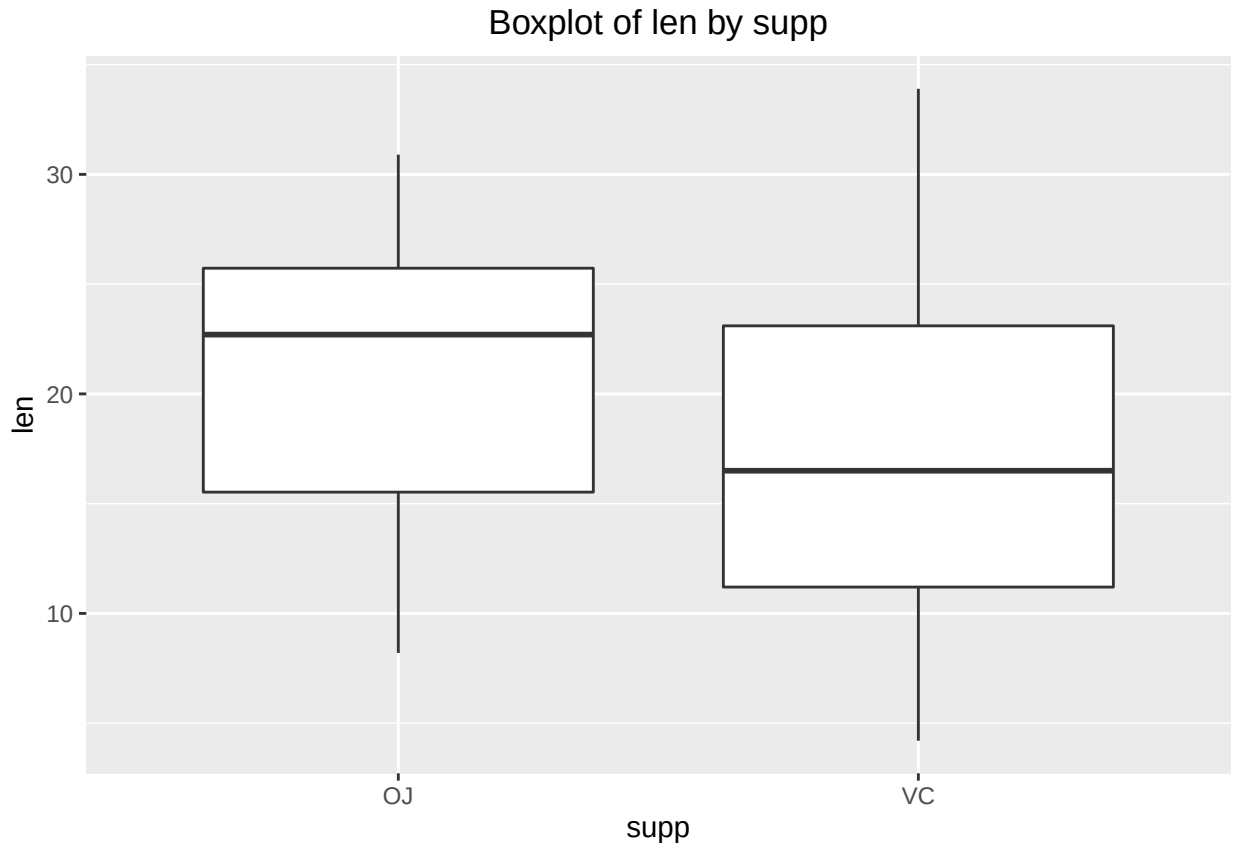


Qualitatively, we can see that the tooth length is increasing with the dose, while the relationship with the supplement type is not so clear.

Explore the tooth growth by supplement type and dose.

To explore the relationship between tooth length (tooth growth) and supplement type, first of all a boxplot of the 2 variables is built.

```
# Plot of tooth length by supplement type
print(ggplot(ToothGrowth, aes(x = supp, y = len)) + geom_boxplot() +
        theme(plot.title = element_text(hjust = 0.5)) +
        labs(title = "Boxplot of len by supp", x = "supp"))
```



Then a t-test on the averages of tooth length for both values of the supplement type is made, assuming as null hypothesis that the averages are equal.

```
# Subsetting the dataset by supplement type
subset_supp_1 <- subset(ToothGrowth, supp == "VC")$len
subset_supp_2 <- subset(ToothGrowth, supp == "OJ")$len

## t-test on the averages of tooth length for the two different supplement types (function t.test)
with(ToothGrowth, t.test(subset_supp_2, subset_supp_1, paired = FALSE, var.equal = FALSE))
```

```
##
##  Welch Two Sample t-test
##
## data:  subset_supp_2 and subset_supp_1
## 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
```

```
## t-test on the averages of tooth length for the two different supplement types (manual calculation)
s2_supp_1 <- var(subset_supp_1)
s2_supp_2 <- var(subset_supp_2)
n_supp_1 <- length(subset_supp_1)
n_supp_2 <- length(subset_supp_2)
df_supp <- (s2_supp_1/n_supp_1 + s2_supp_2/n_supp_2)^2/((s2_supp_1/n_supp_1)^2/(n_supp_1-1) +
                                                         (s2_supp_2/n_supp_2)^2/(n_supp_2-1))
m_supp_1 <- mean(subset_supp_1)
m_supp_2 <- mean(subset_supp_2)
CI <- m_supp_2 - m_supp_1 + c(-1,1) * qt(0.975, df_supp) *
  sqrt(s2_supp_1/n_supp_1 + s2_supp_2/n_supp_2)
p_value <- 2 * (1 - pt((m_supp_2 - m_supp_1) /
                       sqrt(s2_supp_1/n_supp_1 + s2_supp_2/n_supp_2), df_supp))
print(paste("The confidence interval of the difference between the averages is [",
            round(CI[1], 2), ",", round(CI[2], 2), "]"))
```

```
## [1] "The confidence interval of the difference between the averages is [ -0.17 , 7.57 ]"
```

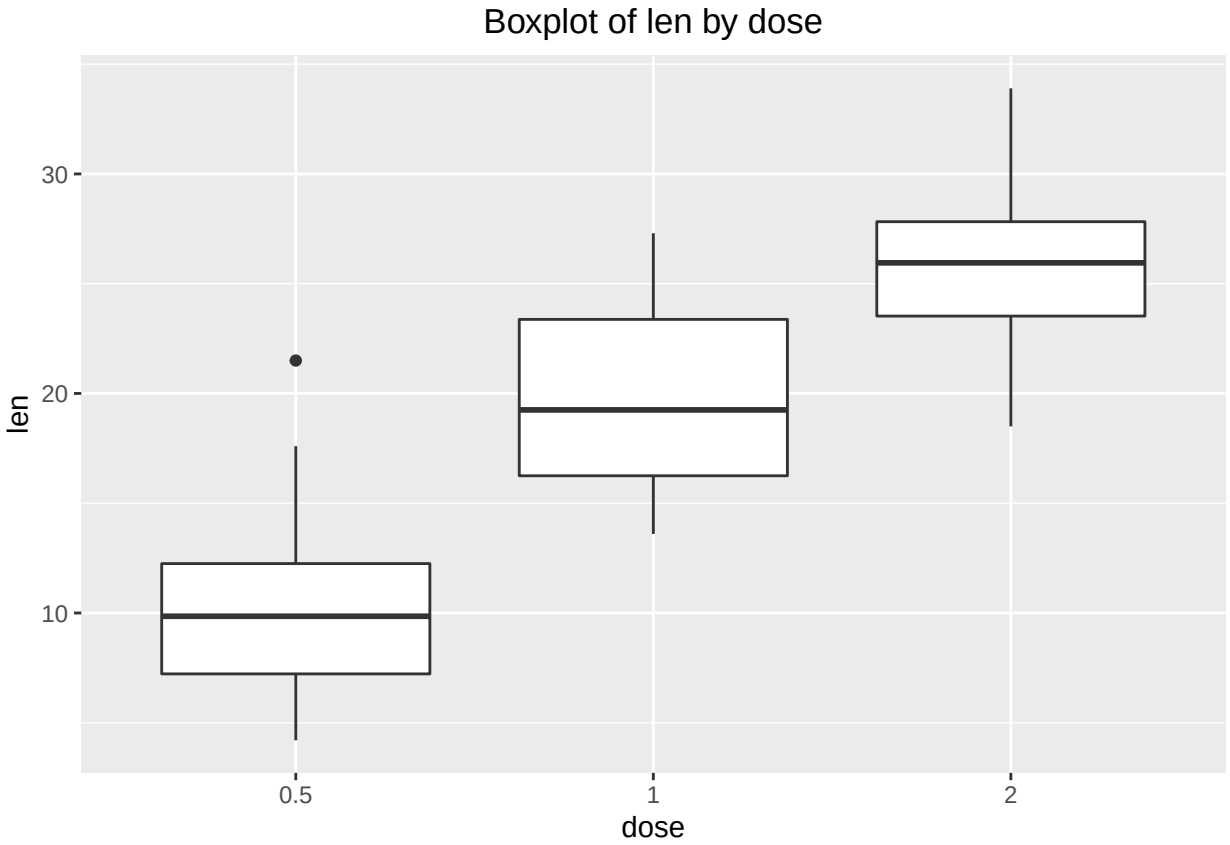
```
print(paste("The p-value is", round(p_value, 3)))
```

```
## [1] "The p-value is 0.061"
```

Based on the results, since the p-value is greated that 0.05 (or equivalently, the confidence interval includes 0), we cannot conclude with a confidence level of 95% that the 2 means are different, and therefore we accept the null hypothesis (in other words, that tooth length does not depend on the supplement type). We must observe, though, that the p-value is very close to 0.05, so probably more data is needed to reach a final conclusion.

Similarly, to explore the relationship between tooth length (tooth growth) and dose, first of all a boxplot of the 2 variables is built.

```
# Plot of tooth length by dose
ggplot(ToothGrowth, aes(x = factor(dose), y = len)) + geom_boxplot() +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Boxplot of len by dose", x = "dose")
```



Then a t-test on the averages of tooth length for the two extreme values of dose (0.5 and 2.0) is made, again assuming as null hypothesis that the averages are equal.

```
# Subsetting the dataset by dose
subset_dose_1 <- subset(ToothGrowth, dose == 0.5)$len
subset_dose_2 <- subset(ToothGrowth, dose == 2.0)$len

## t-test on the averages of tooth length for the two extreme doses (function t.test)
with(ToothGrowth, t.test(subset_dose_2, subset_dose_1, paired = FALSE, var.equal = FALSE))
```

```
##
##  Welch Two Sample t-test
##
## data:  subset_dose_2 and subset_dose_1
## 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:
##  12.83383 18.15617
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

```
## t-test on the averages of tooth length for the two extreme doses (manual calculation)
s2_dose_1 <- var(subset_dose_1)
s2_dose_2 <- var(subset_dose_2)
n_dose_1 <- length(subset_dose_1)
```

```
n_dose_2 <- length(subset_dose_2)
df_dose <- (s2_dose_1/n_dose_1 + s2_dose_2/n_dose_2)^2/((s2_dose_1/n_dose_1)^2/(n_dose_1-1) +
                                                        (s2_dose_2/n_dose_2)^2/(n_dose_2-1))
m_dose_1 <- mean(subset_dose_1)
m_dose_2 <- mean(subset_dose_2)
CI <- m_dose_2 - m_dose_1 + c(-1,1) * qt(0.975, df_dose) *
  sqrt(s2_dose_1/n_dose_1 + s2_dose_2/n_dose_2)
p_value <- 2 * (1 - pt((m_dose_2 - m_dose_1) /
                         sqrt(s2_dose_1/n_dose_1 + s2_dose_2/n_dose_2), df_dose))
print(paste("The confidence interval of the difference between the averages is [",
            round(CI[1], 2), ",", round(CI[2], 2), "]"))
```

```
## [1] "The confidence interval of the difference between the averages is [ 12.83 , 18.16 ]"
```

```
print(paste("The p-value is", round(p_value, 3)))
```

```
## [1] "The p-value is 0"
```

Based on the results, since the p-value is much smaller than 0.05 (or equivalently, the confidence interval does not include 0), we can conclude with a confidence level of 95% that the 2 means are different, and therefore we reject the null hypothesis (in other words, we can say that tooth length depends on the dose).

Conclusions and assumptions

In summary, we concluded that the tooth length depends on the dose level of vitaminc C, in particular it's different when the dose level is 2.0 and 0.5 (the two extreme values), but does not depend on the supplement type (ascorbic acid or orange juice).

Below are some of the key assumptions we had to make in order to reach this conclusion:

- we assumed that the underlying data is normal, because we used the t-test to compare the averages (however we didn't assume that the variances are equal).
- we assumed that randomization was done effectively and the samples are representative.
- we assumed that 95% for the confidence interval is a good value, and that both the Type I and Type II errors risks are acceptable.
- finally, concerning the dose, we assumed that only the 2 extreme values are of interest, as we didn't check if the dose effect is significant between 0.5 and 1.0 and between 1.0 and 2.0.