Basic inferential Data Analysis, Statistical Inference Course Project Week 4

Jose Manuel Coello

NOTE: My native language is not English, so I apologize for any grammatical errors.

Basic Inferential and Data Analysis

In this part of the project we work with the ToothGrowth dataset, this dataset contains information about the length of "odontoblasts" (cells responsible for tooth growth) in 60 guinea pigs, for more informations we can type **help(ToothGrowth)** in the R console.

```
library(ggplot2)
library(dplyr)
library(tidyr)

# Load the ToothGrowth data and perform some basic exploratory data analyses
tooth_growth <- ToothGrowth
tooth_growth <- transform(tooth_growth, dose = factor(dose))

head(tooth_growth)
```

```
##    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
```

Dataset description:

1. `len`: response variable, numeric Tooth length
2. `supp`: Supplement type, orange juice (OJ), or ascorbic acid (a form of vitamin C and coded as VC)
3. `dose`: dose numeric Dose in milligrams/day

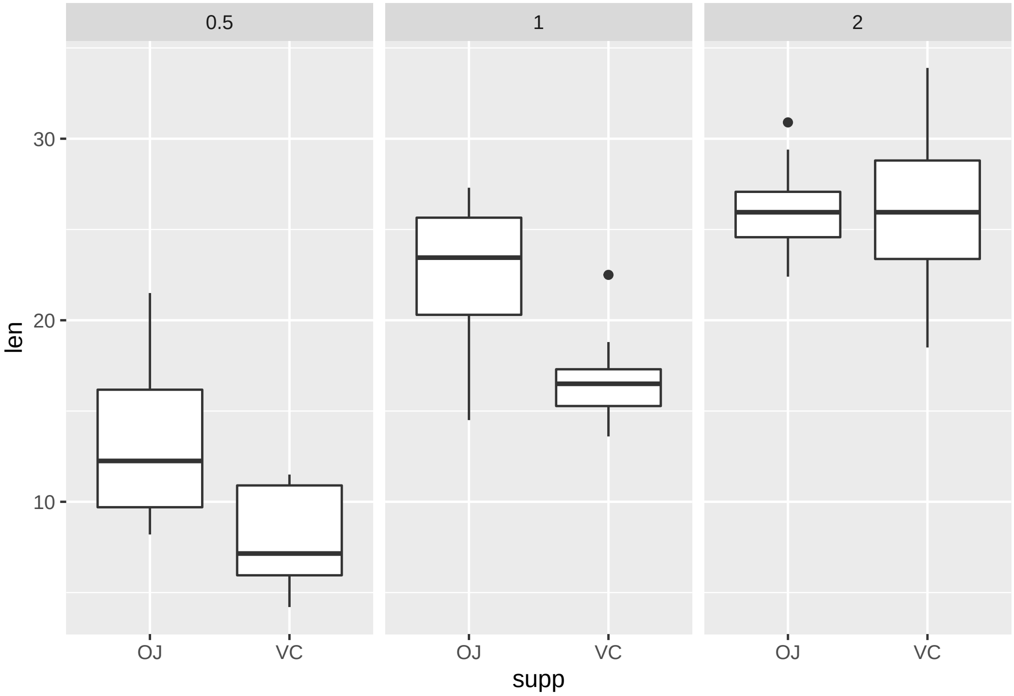
Summary of the data

We proceed to see the average length of teeth by supplements and dose.

```
# basic exploratory
ToothGrowth %>%
  count(supp, dose, wt = mean(len)) %>%
  spread(key = supp, value = n)
```

```
##   dose    OJ    VC
## 1  0.5 13.23  7.98
## 2  1.0 22.70 16.77
## 3  2.0 26.06 26.14
```

```
tooth_growth %>%
  ggplot(aes(x = supp, y = len)) +
  geom_boxplot() +
  facet_wrap(~ dose)
```



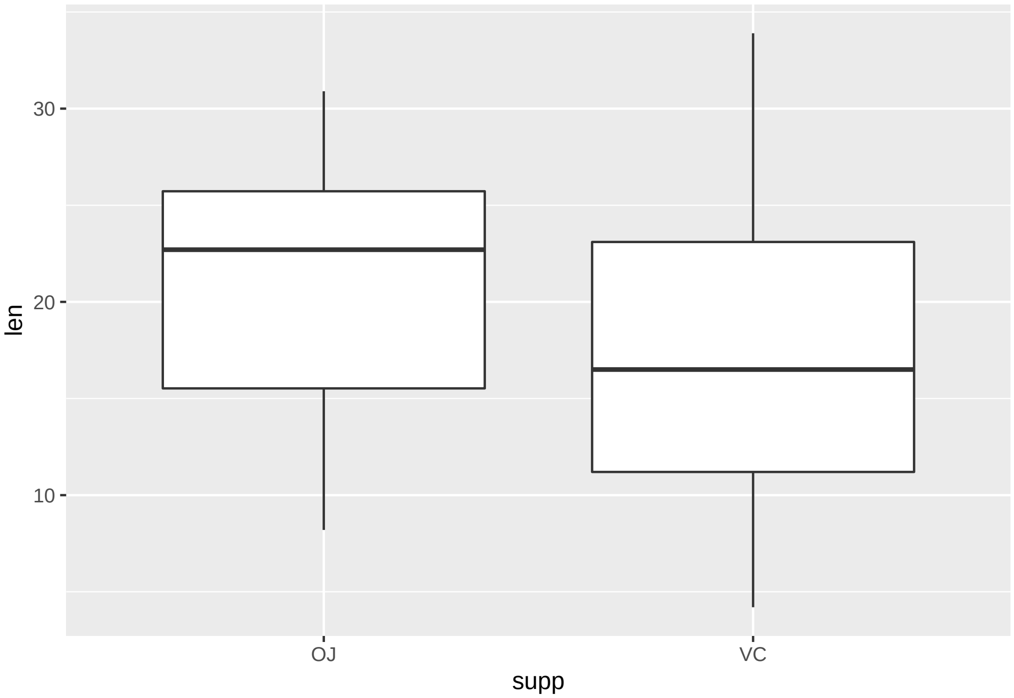
We can see that the average length of teeth varies between doses 0.5 and 1, with the average length of teeth for orange juice being greater than the average length of teeth for ascorbic acid. When the dose is 2 milligrams per day, the average length of the teeth is very similar.

Confidence intervals and hypothesis tests to compare tooth growth by supp

```
ToothGrowth %>%
  count(supp, wt = mean(len))
```

```
##   supp        n
## 1   OJ 20.66333
## 2   VC 16.96333
```

```
tooth_growth %>%
  ggplot(aes(x = supp, y = len)) +
  geom_boxplot()
```



A difference can be observed between the average length of teeth by orange juice and ascorbic acid supplements. I set a hypothesis test where:

$H_0 : \mu_{oj} = \mu_{vc}$ $H_A : \mu_{oj} \neq \mu_{vc}$

with $\alpha = 0.05$

```
len_oj <- tooth_growth[tooth_growth$supp == 'OJ', 'len']
len_vc <- tooth_growth[tooth_growth$supp == 'VC', 'len']

# hypothesis tests
my_ht <- t.test(x = len_oj, y = len_vc, alternative = "two.sided", mu = 0)
my_ht
```

```
##
##  Welch Two Sample t-test
##
## data:  len_oj and len_vc
## 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
```

p-value is 0.06 it's very close to the alpha value 0.05, nevertheless p-value is higher than alpha, so we say that it's likely to observe the data even if the null hypothesis is true, and hence do not reject H_0.

The confidence interval (-0.1710156, 7.5710156) agrees with the hypothesis test.

I conclude that these data don't indeed provide convincing evidence that there is a difference between the average length of teeth by orange juice and average length of teeth for ascorbic acid.