Statistical Inference Course Project Part 2

Masha Maziuk

Overview

In this project we are going to analyze the ToothGrowth data in the R datasets package.

```
#Loading the dataset
data(ToothGrowth)
```

Basic exploratory data analyses

```
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 ...
```

```
summary(ToothGrowth)
```

```
##       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
```

```
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
```

```
table(ToothGrowth$supp)
```

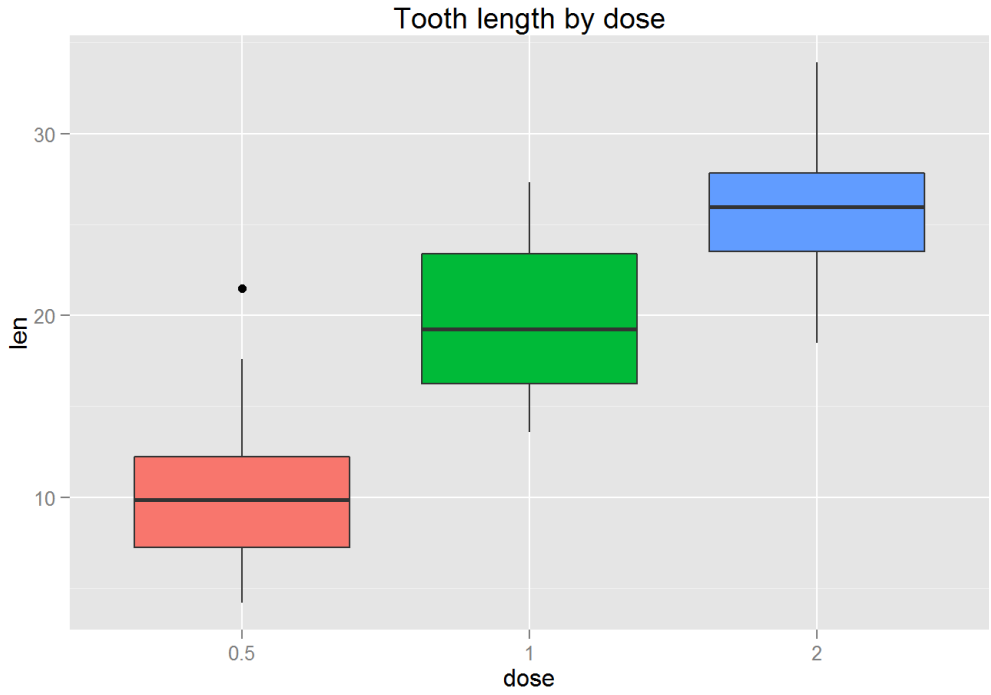
```
##
## OJ VC
## 30 30
```

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
table(ToothGrowth$dose)
```

```
##
## 0.5   1   2
##  20  20  20
```

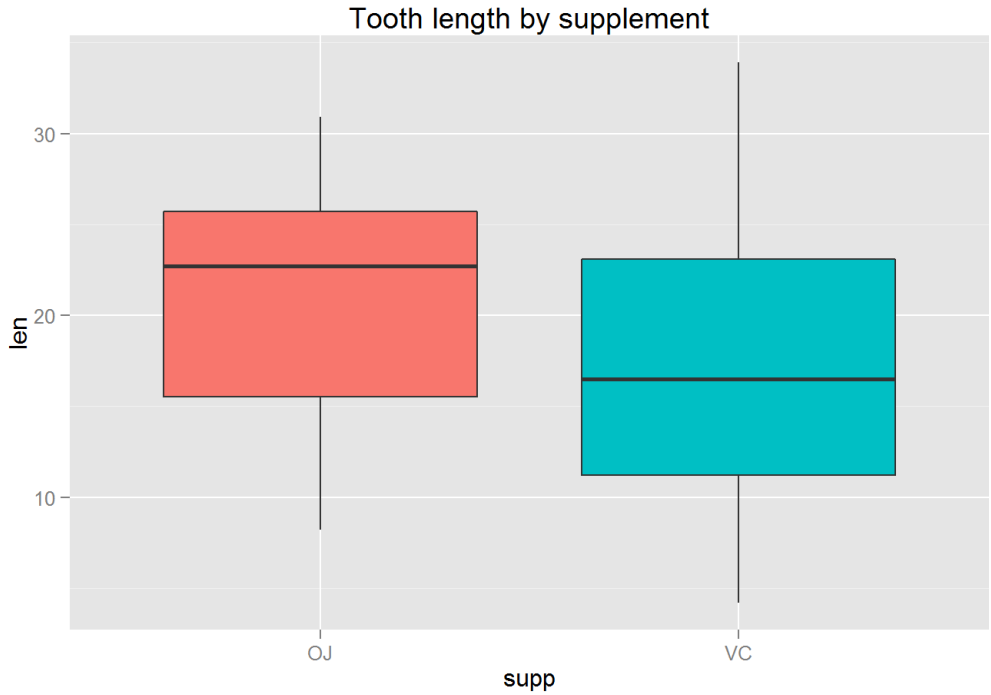
It's easy to see the difference between factor levels using boxplot. Let's create plots for both `supp` and `dose` variables, and see how the tooth length varies by different supp and dose levels.

```
library(ggplot2)
plot1 <- ggplot(ToothGrowth, aes(x = dose, y = len, fill = dose)) +
  geom_boxplot() +
  theme(legend.position="none") +
  labs(title = "Tooth length by dose")
plot1
```



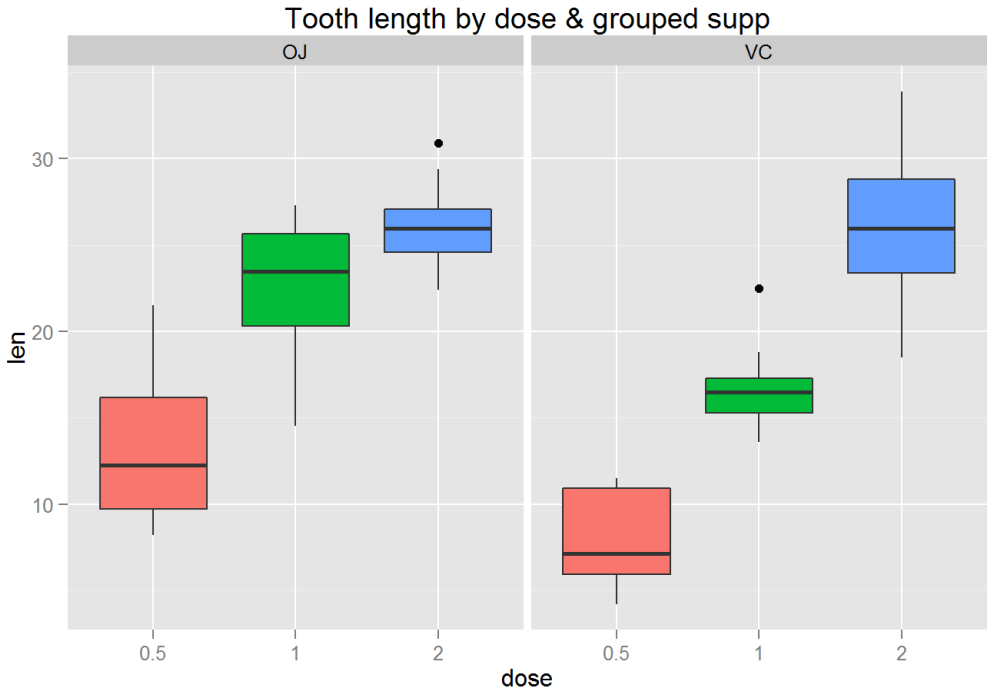
As we can see from the plot above, tooth length varies quite strong by dose. This variable seems to have a strong effect on the tooth length. There is a positive correlation between these variables.

```
plot2 <- ggplot(ToothGrowth, aes(x = supp, y = len, fill = supp)) +
  geom_boxplot() +
  theme(legend.position="none") +
  labs(title = "Tooth length by supplement")
plot2
```



The supplement seems to affect the tooth length, too. But this difference is less easy to distinguish, as interquantile ranges of two supplement types largely overlap.

```
plot3 <- ggplot(ToothGrowth, aes(x = dose, y = len, fill = dose)) +
  geom_boxplot() +
  theme(legend.position="none") +
  facet_grid(.~supp) +
  labs(title = "Tooth length by dose & grouped supp")
plot3
```



From the plot above we can see that for “VC” supplement the tooth length varies more by dose, than for “OJ” supplement.

Confidence Intervals and Hypothesis Testing

Since our sample size is quite small, we’ll use the t-distribution for our hypothesis tests. We need to conduct hypothesis tests for groups of upaired and independent observations.

1. Tooth growth by supp

Our `H0` hypothesis is that the difference in means of two groups is equal to 0. Alternative `Ha` hypothesis is that the difference in means of two groups is not equal to 0.

```
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

The `p-value` is `0.06`, and the confidence interval `[-0.1710; 7.5710]` contains zero. Thus, we fail to reject the null hypothesis, as our data do not provide convincing evidence in favor of the alternative hypothesis.

2. Tooth growth by dose

As there are three dose levels we need to apply 3 hypothesis tests for all the combinations of tooth length with three dose levels.

Each time our `H0` hypothesis will be that the difference in means of two groups is equal to 0. While the alternative `Ha` hypothesis will be that the difference in means of two groups is not equal to 0.

```
dose2dose1 <- t.test(ToothGrowth$len[ToothGrowth$dose == 2], ToothGrowth$len[ToothGrowth$dose == 1])

dose2dose05 <- t.test(ToothGrowth$len[ToothGrowth$dose == 2], ToothGrowth$len[ToothGrowth$dose == 0.5])

dose1dose05 <- t.test(ToothGrowth$len[ToothGrowth$dose == 1], ToothGrowth$len[ToothGrowth$dose == 0.5])
```

The test results for `dose = 2` and `dose = 1`:

```
dose2dose1$p.value
```

```
## [1] 1.90643e-05
```

```
dose2dose1$conf.int
```

```
## [1] 3.733519 8.996481
## attr(,"conf.level")
## [1] 0.95
```

The test results for `dose = 2` and `dose = 0.5`:

```
dose2dose05$p.value
```

```
## [1] 4.397525e-14
```

```
dose2dose05$conf.int
```

```
## [1] 12.83383 18.15617
## attr(,"conf.level")
## [1] 0.95
```

The test results for `dose = 1` and `dose = 0.5`:

```
dose1dose05$p.value
```

```
## [1] 1.268301e-07
```

```
dose1dose05$conf.int
```

```
## [1]  6.276219 11.983781
## attr(,"conf.level")
## [1] 0.95
```

In all three tests the p-value is smaller than our significance level `0.05`. Thus, we reject the null hypothesis H0 in all three cases. The data provide convincing evidence that the tooth length of guinea pigs is different for each dose level.

The interpretation of the first result will be the following: We are 95% confident that the average tooth length of guinea pigs who received a `2mg` dose of vitamin C is on average `3.733519` to `8.996481` millimeters higher than those who received a `1mg` dose.

Conclusions and assumptions

- Conclusion 1: The tooth lengths for different supplement types are not significantly different.
- Conclusion 2: The tooth lengths for different doses are significantly different.

For our conclusions we assume that the observations are independent (a random sampling was used and the sample is less than 10% of the population).