Analysis of the Tooth Growth Data Set

Dominik Sudwischer

6 September 2017

Introduction

The 'ToothGrowth' data set contains data about guinea pig tooth growth under influence of two different supplements in different doses. We will find out whether there is statistically significant evidence that higher doses of supplements are correlated with longer teeth and whether there is a signifcant difference between the two supplements used. The significance level we will use for our analysis will be 5%.

Loading the Data Set

Let us start by loading the data set which can be found in the package 'datasets'. It contains data about the length of teeth of guinea pigs in relation to two different supplements fed to the guinea pigs in different doses.

```
library(datasets)
library(ggplot2)
df = ToothGrowth
```

Exploring the Data Set

```
str(df)
```

```
## '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 ...
```

The data set comprises 60 observations of tooth lengths ('len') and the dose of the respective supplement ('supp') that was fed to the guinea pig. It turns out that there are actually two different supplements, 'VC' and 'OJ', the former of which is vitamin C and the latter of which is orange juice. The dose has three different levels, .5, 1 and 2 mg respectively. We can verify this claim by looking at the distinct values per column:

```
unique(df[, 'supp'])
```

```
## [1] VC OJ
## Levels: OJ VC
```

```
unique(df[, 'dose'])
```

```
## [1] 0.5 1.0 2.0
```

Let us have a quick glance on the summary of the data set:

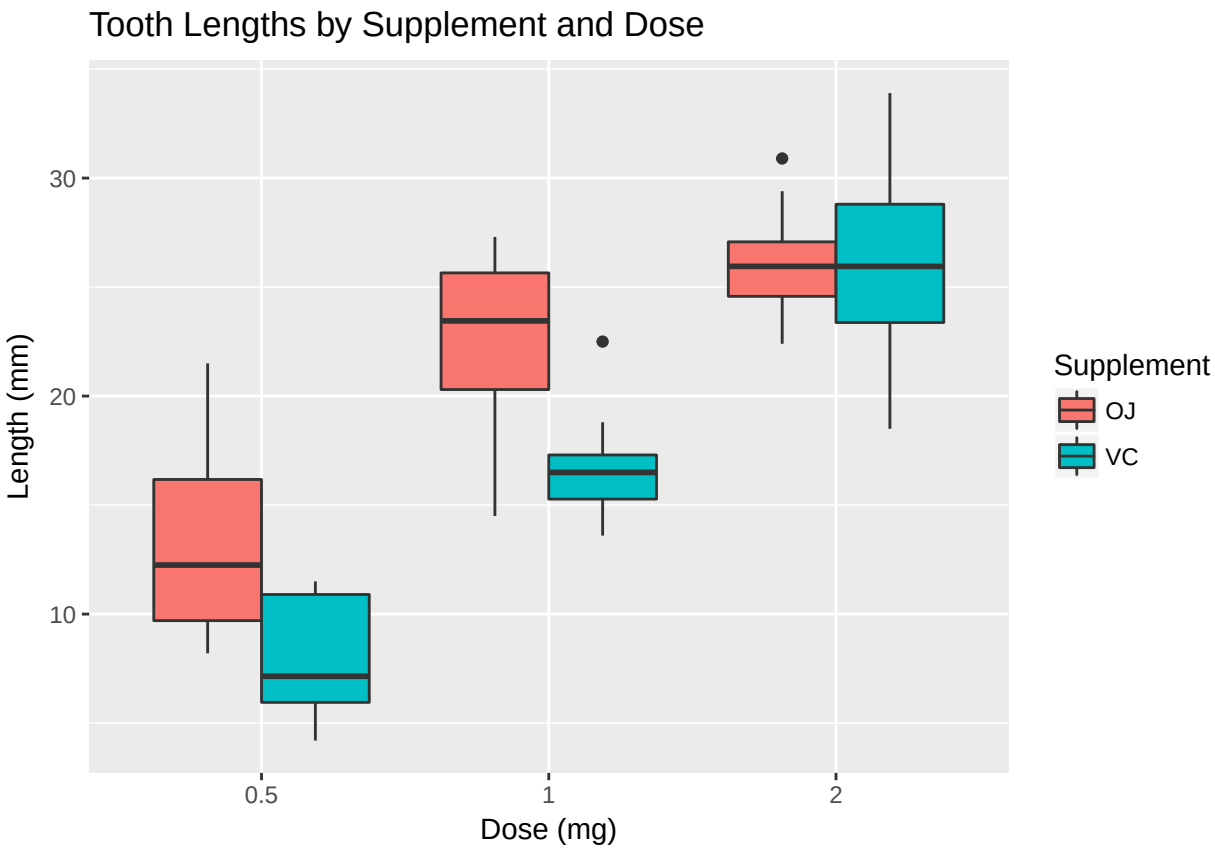
```
summary(df)
```

```
##       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
```

To get a basic understanding of the data, we will use a box plot. The two levels of the variable 'supp' combined with the three levels of the variable 'dose' leaves us with 6 combinations in total, so we create 6 boxes.

```
ggplot(df, aes(x=factor(dose), y=len, fill=supp)) +
  labs(title="Tooth Lengths by Supplement and Dose", x="Dose (mg)", y = "Length (mm)",
       fill = 'Supplement') + geom_boxplot()
```



The box plot indicates that there might be a correlation between the length of teeth and the dose of a supplement. Higher doses seem to be related to longer teeth. A slight difference in tooth lengths between both supplements for lower doses is visible from the plot. We can easily look at the underlying data of the box plot.

```
by(data = df$len, INDICES = df[, c("dose", "supp")], FUN = summary)
```

```
## dose: 0.5
## supp: OJ
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20    9.70   12.25   13.23   16.18   21.50
## ------------------------------------------------------------
## dose: 1
## supp: OJ
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.50   20.30   23.45   22.70   25.65   27.30
## ------------------------------------------------------------
## dose: 2
```

```
## supp: OJ
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.40   24.57   25.95   26.06   27.07   30.90
## ------------------------------------------------------------
## dose: 0.5
## supp: VC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20    5.95    7.15    7.98   10.90   11.50
## ------------------------------------------------------------
## dose: 1
## supp: VC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   15.28   16.50   16.77   17.30   22.50
## ------------------------------------------------------------
## dose: 2
## supp: VC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.50   23.38   25.95   26.14   28.80   33.90
```

Questions and Assumptions

After looking at the plot and the grouped summaries, we will consider the following to questions:

- Is the mean tooth length for guinea pigs fed with a higher dose of any supplement higher than the mean tooth length for such fed with a lower dose?
- Is there a difference in mean tooth length between the groups treated with different supplements?

Before formulating the corresponding hypotheses, let us note our assumptions for the upcoming analysis. First, we assume that the observations are independent from each other. This means in particular that the data is not assumed to be paired. Since there is no guarantee that the population variance among different groups is equal, we will not make this assumption either. However, in order to perform a t-test, we will need to assume that tooth growth of guinea pigs follows a normal distribution. This will ensure that our sample means follow a t-distribution and enable us to calculate meaningful t-test statistics.

Hypothesis Testing

We will continue with our analysis. First, we will compare the mean tooth length of the part of the group with only .5 mg doses and the mean tooth length of the part with doses of 2 mg. Our test will be a one-sided t-test which analyzes whether the mean of the low-dose group is strictly less than the mean of the high-dose group, which is our null hypothesis in this case.

```
t.test(len ~ dose, data = df[df$dose %in% c(0.5, 2),], paired = FALSE, var.equal = FALSE,
       alternative = "less")$p.value
```

```
## [1] 2.198762e-14
```

The p-value of 2.2e-14 is less than the the significance level of .05, so we can safely reject the null hypothesis 'Both true means are equal' in favor of the alternative hypothesis 'the true mean tooth length in the low-dose group is less than the true mean tooth length of the high-dose group'.

Now that we have seen that there is a correlation between the supplement dose and the tooth growth, we will have a look on the effectiveness of the different supplies. Since we initially considered that the difference in means for the two supplements might by different among the levels of the doses, we should do three t-tests for this case. The corresponding null hypothesis for each $x \in \{.5, 1.0, 2.0\}$ is: 'The true mean tooth length is equal for a dose of x mg for both supplements'.

```
t.test(len ~ supp, data = df[df$dose == .5,], paired = FALSE,
       var.equal = FALSE)$p.value
```

```
## [1] 0.006358607
```

```
t.test(len ~ supp, data = df[df$dose == 1.0,], paired = FALSE,
       var.equal = FALSE)$p.value
```

```
## [1] 0.001038376
```

```
t.test(len ~ supp, data = df[df$dose == 2.0,], paired = FALSE,
       var.equal = FALSE)$p.value
```

```
## [1] 0.9638516
```

The result of the t-tests shows that we can reject the null hypothesis for both $x = .5$ and $x = 1.0$. However, it cannot be rejected for $x = 2.0$.

Conclusion

According to our analysis, we have established the following results:

- The mean tooth length differs significantly between the parts of the sample that were fed high doses and those that were fed low doses of either supplement.
- The mean tooth length differs significantly between the parts of the sample that were fed with low or medium doses vitamin C and those that were fed with low or medium doses of orange juice, respectively. For doses of 2 mg, there is no statistically significant evidence that the true mean tooth length differs between the groups for each of the two supplements.

That being said, we have to keep in mind that we cannot just claim that there is a causal relationship between the variables and the tooth length without knowing that the data was observed during a specifically designed experiment that ensures all other factors are equal. If the data is indeed from a scientific study designed for this purpose, we can conclude that higher doses of either supplement lead to an increase in tooth length. For doses of .5 or 1 mg, there is a significant difference in effectiveness between the two sorts of supplements.