Inferential data analysis on the ToothGrowth dataset

Bert Lijnen

10 August 2017

Summary

In this report we will investigate whether the administration of orange juice and vitamin C at different doses will result in an increase in the tooth growth of guinee pigs. By means of hypothesis testing we will verify whether the results observed in our sample can be generalized to the population of guinee pigs.

The dataset: ToothGrowth

We load the datset and have a look at the variables and the values they can take.

```r
library(datasets)
data(ToothGrowth)
View(ToothGrowth)
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 ...
```

```r
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
```

```r
unique(ToothGrowth$dose)
```

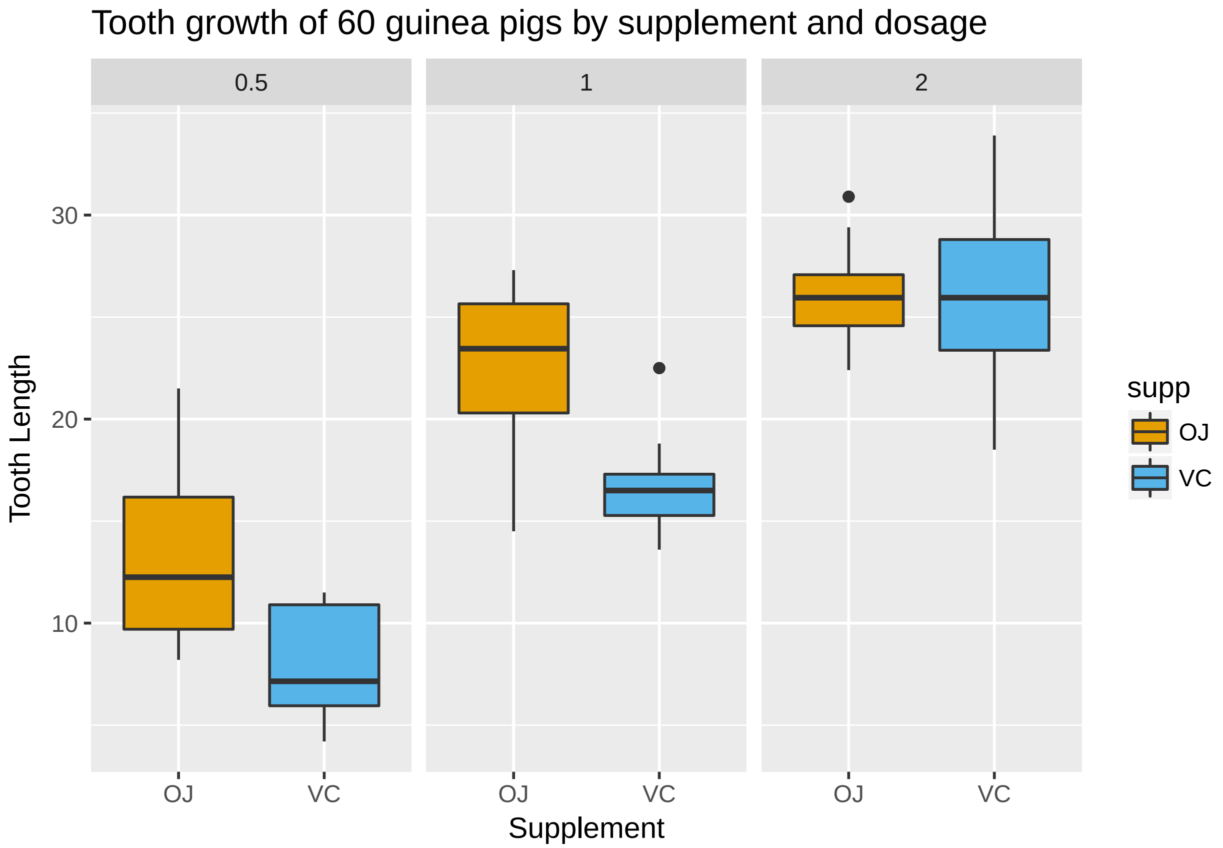
```
## [1] 0.5 1.0 2.0
```

The dataset consists of 60 observations and 3 variables. The variables are:
* len (numeric) - the tooth length
* supp (factor with 2 levels) - Supplement type (vitamin C and orange juice)
* dose (numeric) - dose in mg/day (0.5, 1.0 or 2.0)

Exploratory data analysis

First we make a plot that captures as much of the information as available in the dataset

```r
library(ggplot2)
ggplot(ToothGrowth, aes(x=supp, y=len))+
        geom_boxplot(aes(fill=supp))+
        facet_grid(.~dose)+
        labs(title="Tooth growth of 60 guinea pigs by supplement and dosage",
             x="Supplement", y="Tooth Length")+
        scale_fill_discrete(name="supplement")+
        scale_fill_manual(values=c("#E69F00", "#56B4E9"))
```



The boxplots clearly indicate that the tooth length increases as the dosage increases and this for both types of supplements. When using Vitamin C as a supplement, the higher the dosage the higher the thooth growth. However, when the dosage is low, orange juice seems to result in longer teeth. When using the highest dosage (2mg) there does not appear to be a significant difference between the two supplements.

Let's now give a numeric overview of the boxplots. We calculate the most important descriptive measures for each combination (supplement * dose) and store the results in a data frame.

```
library(dplyr)
groups_tooth <- group_by(ToothGrowth, supp, dose)
overview<-summarise(groups_tooth, count= n(), mean=mean(len),median=median(len),
                    std = sd(len))
overview <- as.data.frame(overview)
knitr::kable(overview)
```

supp	dose	count	mean	median	std
OJ	0.5	10	13.23	12.25	4.459708
OJ	1.0	10	22.70	23.45	3.910953
OJ	2.0	10	26.06	25.95	2.655058
VC	0.5	10	7.98	7.15	2.746634
VC	1.0	10	16.77	16.50	2.515309
VC	2.0	10	26.14	25.95	4.797731

We can see that 2mg/day of vitamin C result in the highest tooth growth (but also has the highest standard deviation), closely followed by 2mg/day of orange juice. 0.5mg/day of vitamin C has the smallest effect on tooth growth among guinea pig. In order to verify whether these differences are significant, we need to run some hypothesis tests.

Hypothesis testing

We will compare the average tooth length between the different conditions. First, we subset the data in three groups: one group per dosage. Next, we run three tests including the supplements and using the subsets as data. This allows us to compare the means in the group "orange juice" with the means in the group "vitamin c" at three different doses.

Subsetting the data:

```
dose05<-ToothGrowth[ToothGrowth$dose==0.5,]
dose10<-ToothGrowth[ToothGrowth$dose==1.0,]
dose20<-ToothGrowth[ToothGrowth$dose==2.0,]
```

Running the t-tests

```
t05<-t.test(len~supp, var.equal=FALSE, data=dose05)
t10<-t.test(len~supp, var.equal=FALSE, data=dose10)
t20<-t.test(len~supp, var.equal=FALSE, data=dose20)
```

We summarize the results in a data frame

```
options(digits=3)
t_table<-data.frame("t.statistic"=c(t05$statistic, t10$statistic, t20$statistic),
                    "df"=c(t05$parameter, t10$parameter, t20$parameter),
                    "p.value"=c(t05$p.value, t10$p.value, t20$p.value),
                    "lower_bound"=c(t05$conf.int[1], t10$conf.int[1],
                                    t20$conf.int[1]),
                    "upper_bound"=c(t05$conf.int[2], t10$conf.int[2],
                                    t20$conf.int[2]),
                    "mean_OJ"=c(t05$estimate[1], t10$estimate[1],
                                    t20$estimate[1]),
                    "mean_VC"=c(t05$estimate[2], t10$estimate[2],
                                    t20$estimate[2]))
row.names(t_table)<-c("OJ/VC (0.5mg)", "OJ/VC (1.0mg)","OJ/VC (2.0mg)")
knitr::kable(t_table)
```

	t.statistic	df	p.value	lower_bound	upper_bound	mean_OJ	mean_VC
OJ/VC (0.5mg)	3.170	15.0	0.006	1.72	8.78	13.2	7.98
OJ/VC (1.0mg)	4.033	15.4	0.001	2.80	9.06	22.7	16.77
OJ/VC (2.0mg)	-0.046	14.0	0.964	-3.80	3.64	26.1	26.14

Conclusions and assumptions

1. Our analysis shows that the administration of orange juice yields a higher effect on the tooth growth of guinea pigs than the administration of vitamin C, at least when administrating it a dose of 0.5 or 1.0 mg/day.

2. The average tooth length is significantly different when administrating orange juice at a dose of 0.5mg/day and 1.0 mg/day. The p-values for both conditions are significantly lower than 0.05.

3. On the other hand, when administrating a dose of 2.0mg/day, the differences between the two supplements are not significant and therefore we cannot generalize these results to the overall population of guinee pigs. The p-value is higher than 0.05 (p=0.96) and the confidence interval comprises 0.

4. In the t-test command we have used var.equal=FALSE since the variances between the groups cannot be assumed equal as suggested by both the boxplots and the standard deviation values in the summary table.

5. We assume that the sample of 60 pigs is representative for the general population of pigs in order to make inferences about the entire population.

6. We assume that the pigs have been randomly assigned to the different experimental conditions.