Statistical Inference - Course project - Part 2

Christoph Wagner

03/15/2020

1.) Introduction

This document contains my solutions for part 2 of the final course project of Coursera course "Statistical Inference" (Specialization "Data Mining"). The document has been created using R Markdown and the knitr framework. Part 2 of the course project deals with the analysis of the "ToothGrowth" dataset, which is part of the R datasets package. The analysis comprises a basic summary of the data and hypothesis tests to compare the effects of different supplement types and doses of vitamin C on tooth growth. All experiments below have been executed in R (IDE RStudio). Graphics have been plotted using the ggplot2 package.

2.) The ToothGrowth dataset

The ToothGrowth dataset contains 60 observations of odontoblasts (cells which produce tooth growth) of guinea pigs. The animals received vitamin C in three different doses and by one of two different delivery methods.

The columns of the dataset contain the following variables:

- len: Length of the odontoblasts
- supp: Delivery method of vitamin C, either VC (Asorbic acid) or OJ (orange juice)
- dose: Dose of vitamin C in mg/day

3.) Basic summary of the data

First of all, we inspect the dataset using the standard summary functionality of R.

```
data("ToothGrowth")
data<-ToothGrowth
str(data) #Determine data types and inspect content
```

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

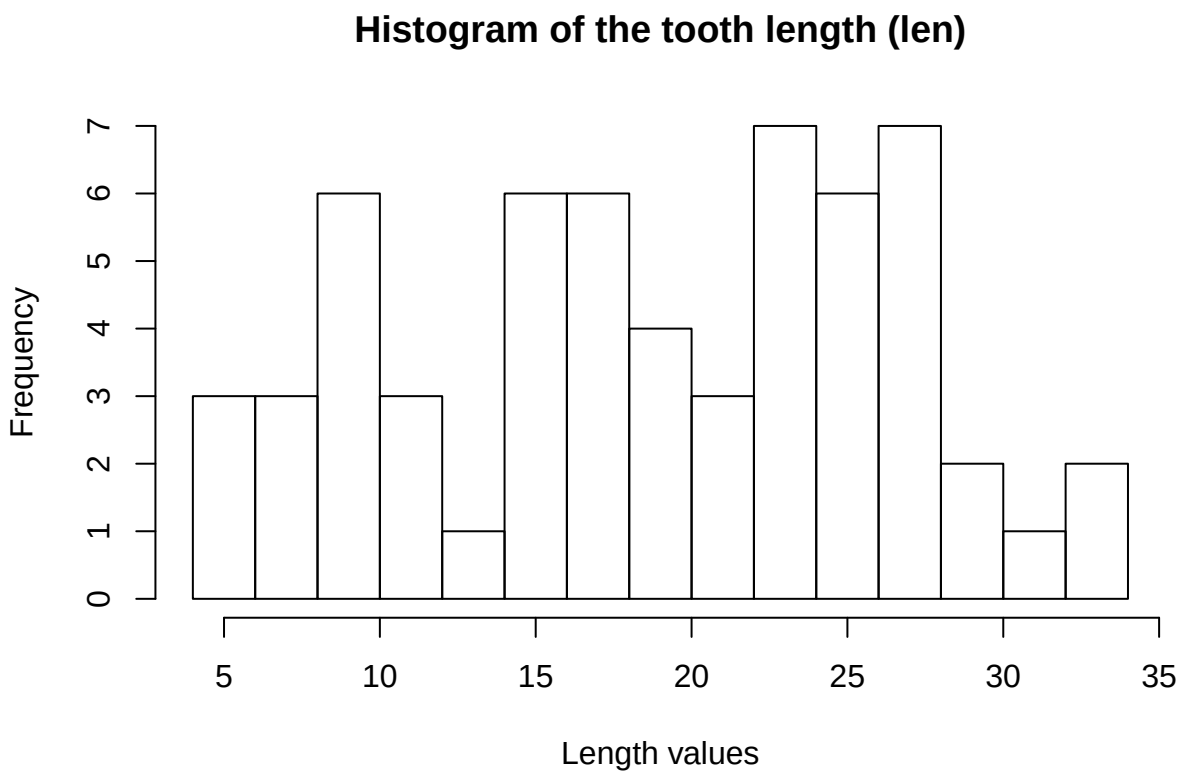
```
summary(data) #Display basic statistics about the dataset
```

```
##       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
```

We can visualize a histogram over the len values of all observations using the hist function. Of course, this doesn't provide much information because the measured odontoblasts were created under different conditions.

```
hist(data$len, breaks=15, main="Histogram of the tooth length (len)",
     xlab="Length values")
```



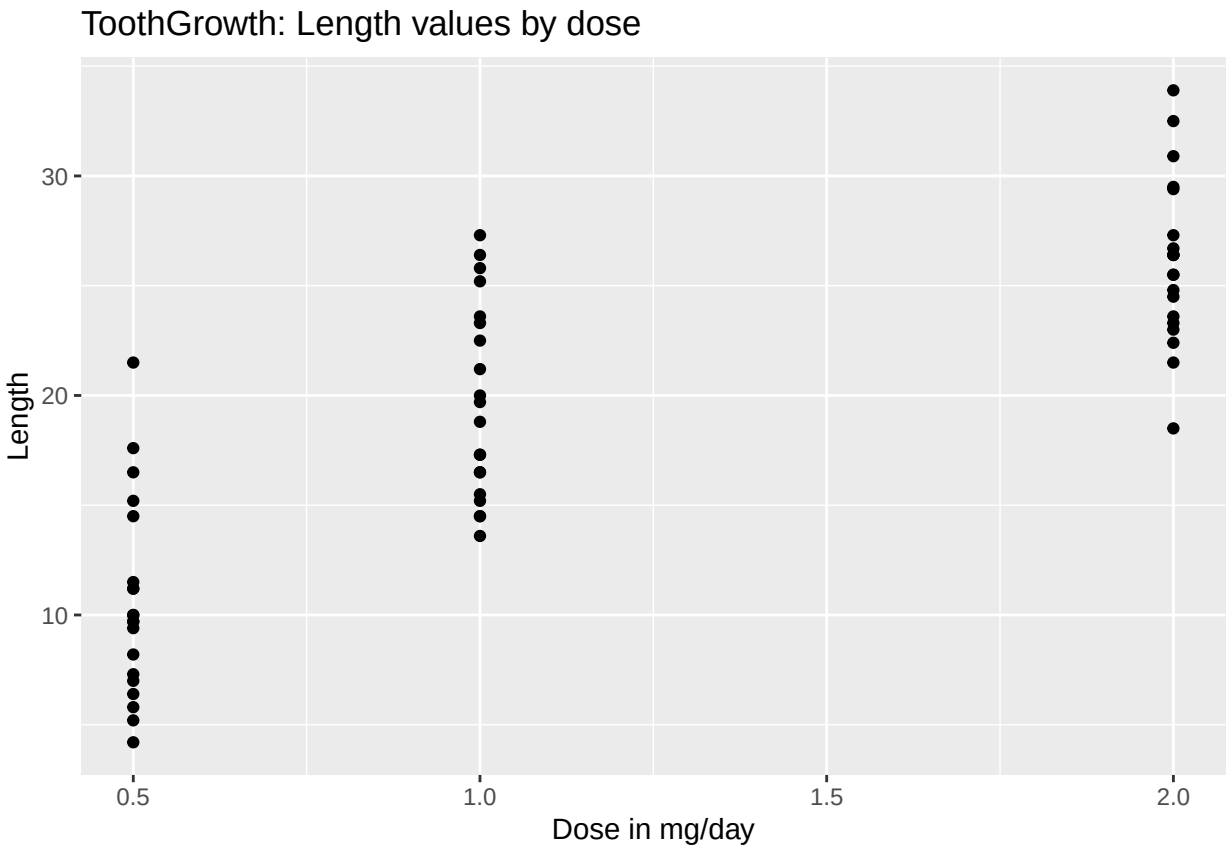
Until now, the data don't seem to follow the normal distribution.

If we plot the len values by dose, the graph indicates that len grows with higher dose value.

```
library("ggplot2")
```

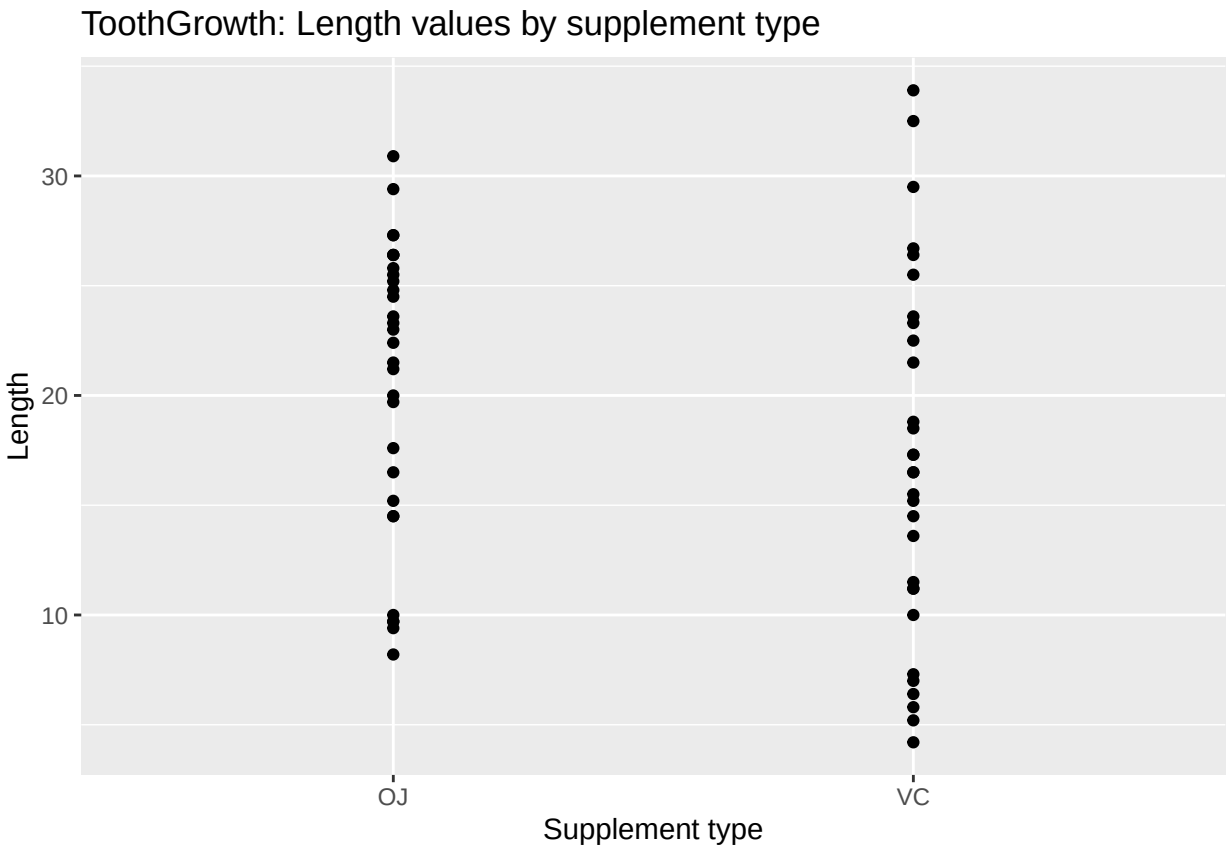
```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
ggplot(data, aes(x=dose, y=len))+geom_point()+
  ggtitle("ToothGrowth: Length values by dose")+xlab("Dose in mg/day")+ylab("Length")
```



In contrast, there is no indication that the growth differs for different applications methods of the vitamin. But they seem to spread more for supplement method "Asorbic Acid".

```
ggplot(data, aes(x=supp, y=len))+geom_point()+
  ggtitle("ToothGrowth: Length values by supplement type")+
  xlab("Supplement type")+ylab("Length")
```



All in all, at least there is indication that the len data differ depending on dose level. It is not obvious that they also differ by supplement method.

4.) Hypothesis tests: Tooth growth by supp and dose

In this chapter, we want to test the assumption that tooth growth differs significantly for different dose levels and supplement methods.

4.1) Tooth growth and supplement method

Null hypothesis (H_0): Len values of the tooth in average DON'T differ if vitamin C is applied via different supplement methods.

Alternative hypothesis (H_a): Len values in average differ if vitamin C is applied via different supplement methods.

Assumptions: The len values of the two different supplement methods do not have the same variance (e.g. it could be possible that individuals have a different response to the application method). Additionally, the individuals are different and so, this is an unpaired test. Significance level 5 %.

```
data_OJ<-data[data[,2]=="OJ",1]
data_VC<-data[data[,2]=="VC",1]
t_1<-t.test(data_OJ, data_VC, paired=FALSE, var.equal=FALSE,
            alternative="two.sided", conf.level=0.95)
t_1
```

```
##
##  Welch Two Sample t-test
##
## data:  data_OJ and data_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
```

Confidence interval: The difference of the average tooth length lies within [-0.1710156;7.5710156] with 95 % probability.

Conclusion: The mean values indicate that for supplement method "Asorbic Acid", the tooth growth is higher. But as the confidence intervals contains value 0, we cannot reject the assumption of equal average tooth length on a significance level of 5 %.

4.2) Tooth growth and dose level

Null hypothesis (H_0): Len values of the tooth in average ARE NOT higher if vitamin C is applied in dose 1.0 instead of 0.5 mg/day.

Alternative hypothesis (H_a): Len values in average ARE HIGHER if vitamin C is applied in 1.0 instead of 0.5 mg/day

Assumptions: As in 4.1) we assume differing variances. Test is unpaired. Significance level 5 %.

```
data_05<-data[data[,3]==0.5,1]
data_10<-data[data[,3]==1.0,1]
t_2<-t.test(data_05, data_10, paired=FALSE, var.equal=FALSE,
            alternative="less", conf.level=0.95)
t_2
```

```
##
##  Welch Two Sample t-test
##
## data:  data_05 and data_10
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Confidence interval: With 95 % probability, tooth of pigs which get dose 0.5 mg/day are smaller than pigs which ghet 1.0 mg/day and the difference lies in the interval $[-\infty;-6.7533227]$. The small p value indicates that it is very improbable to get the tested data samples under the condition that the null hypothesis is true.

Conclusion: With a significance level of 95 % we can reject the null hypothesis. If vitamin C is applied in 1.0 mg/day instaed of 0.5 mg/day, tooth length is bigger.