

Statistical Inference Course Project - Part 2

Willem Hogewoning

8-2-2021

Statistical Inference Course Project

Part 2: Basic Inferential Data Analysis Instructions

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
4. State your conclusions and the assumptions needed for your conclusions.

First load the data and take a look

```
library(datasets)
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 ...
```

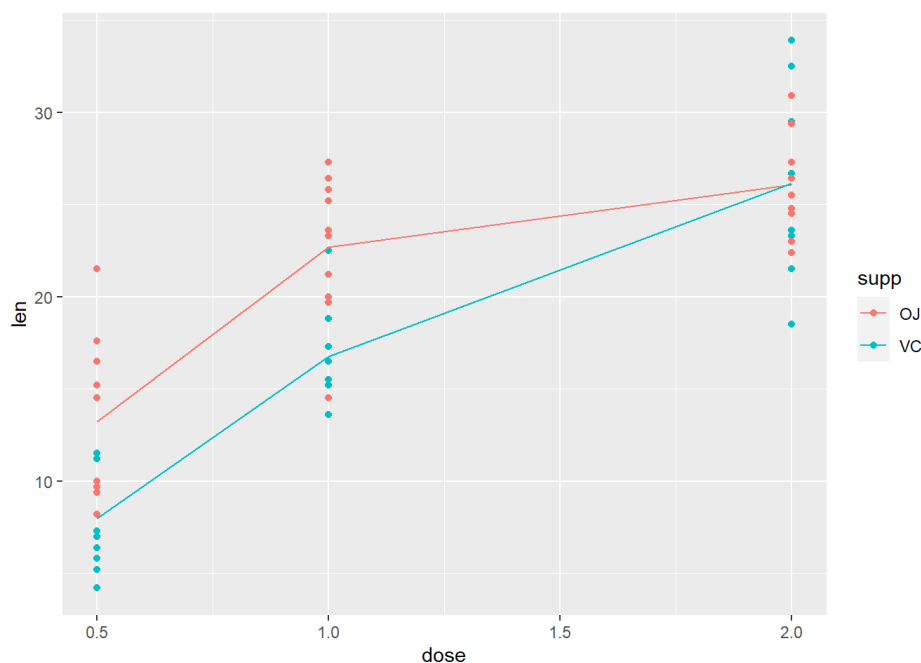
```
summary(ToothGrowth)
```

```
##      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
```

The data consists out of 60 observations, divided into two groups of 30.

Lets see the data in a visual

```
library(ggplot2)
datamean <- aggregate(len~.,data=ToothGrowth,mean)
plot <- ggplot(ToothGrowth,aes(x=dose,y=len))
plot <- plot + geom_point(aes(group=supp,colour=supp))
plot <- plot + geom_line(data=datamean,aes(group=supp,colour=supp))
plot
```



In the plot we can see that as the dosage increases, the average length also increases. However for the dosis of 2.0 the average doesn't show a difference anymore.

Comparison

Lets see the comparison between the 3 different dosses that are in the data set

```
data_OJ <- ToothGrowth[31:60,]
data_VC <- ToothGrowth[1:30,]
len_difference <- data_OJ$len - data_VC$len
data <- data.frame(len_difference, data_OJ$dose)
names(data) <- c("len_difference", "dose")
```

For the dosis of 0.5

```
dose_0.5 <- subset(data, dose == 0.5)
t.test(dose_0.5)
```

```
##
## One Sample t-test
##
## data: dose_0.5
## t = 2.8295, df = 19, p-value = 0.01071
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.7483243 5.0016757
## sample estimates:
## mean of x
##      2.875
```

Has a positive mean in the confidence interval, with a p-value of 0.01071, meaning both methods are different.

For the dosis of 1.0

```
dose_1.0 <- subset(data, dose == 1.0)
t.test(dose_1.0)
```

```
##
## One Sample t-test
##
## data: dose_1.0
## t = 3.3779, df = 19, p-value = 0.003158
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.318017 5.611983
## sample estimates:
## mean of x
##      3.465
```

Has a positive mean in the confidence interval, with a p-value of 0.003158, meaning both methods are different.

For the dosis of 2.0

```
dose_2.0 <- subset(data, dose == 2.0)
t.test(dose_2.0)
```

```
##
## One Sample t-test
##
## data: dose_2.0
## t = 1.0162, df = 19, p-value = 0.3223
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -1.01732  2.93732
## sample estimates:
## mean of x
##      0.96
```

Has a positive mean in the confidence interval, but with a p-value of 0.3223, meaning we cannot make any conclusions