Statistical Inference Course Project Part 2

Madhur

July 25, 2015

Overview

Analyze the ToothGrowth data in the R datasets package.

Load the ToothGrowth data and perform some basic exploratory data analyses Provide a basic summary of the data. 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) State your conclusions and the assumptions needed for your conclusions.

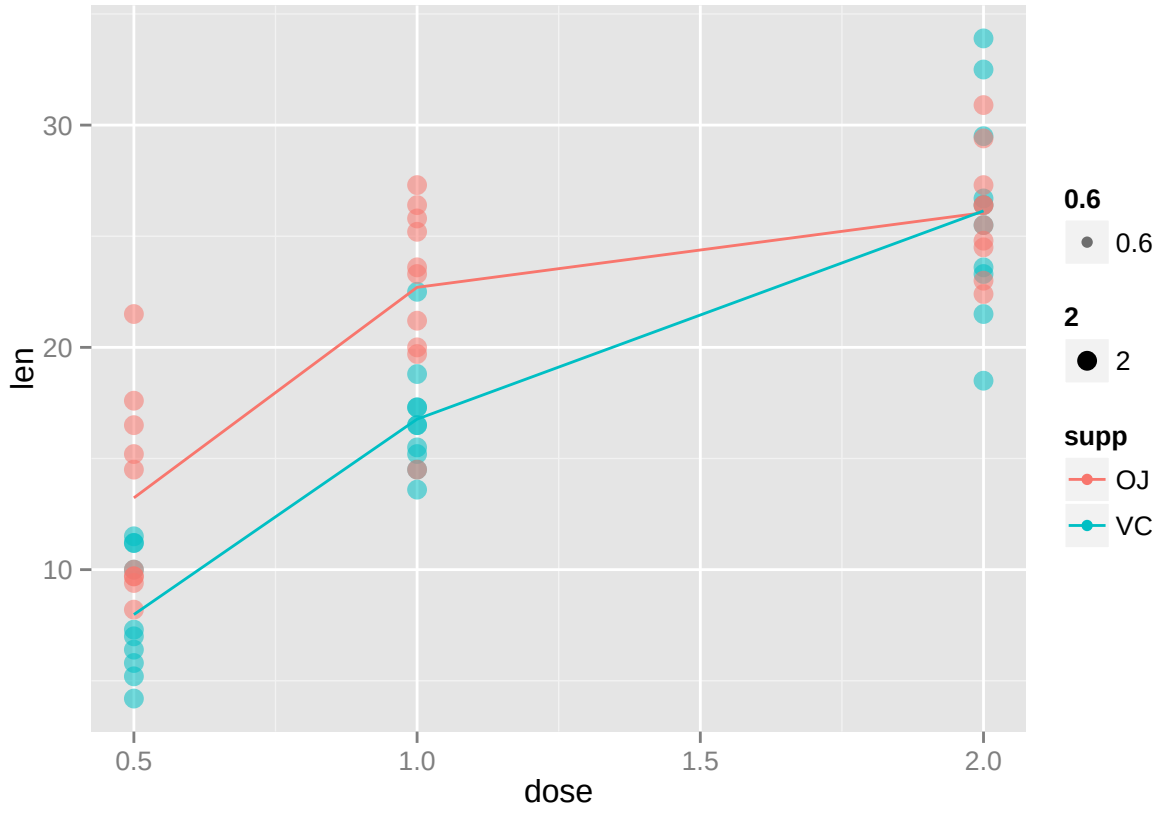
```
library(datasets)
str(ToothGrowth)
```

Load Data

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

```
# Calculate the mean length of the tooth for every supp and dose
avg <- aggregate(len~.,data=ToothGrowth,mean)

# plot the scatter points and the avg lines for the average length
library(ggplot2)
g <- ggplot(ToothGrowth,aes(x=dose,y=len))
g <- g + geom_point(aes(group=supp,colour=supp,size=2,alpha=0.6))
g <- g + geom_line(data=avg,aes(group=supp,colour=supp))
print(g)
```



Data Analysis

The figure above shows that as the dosage increases, the average tooth length increases. Lets try to create a box plot for the same data to see if we can find more.

Basic Summary of Data The summary for the data collection is:

```
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
```

```
print(paste("Number of rows in the data are ",nrow(ToothGrowth)))
```

```
## [1] "Number of rows in the data are  60"
```

```
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
```

Essentially, we have two different methods of delivery (orange juice and ascorbic acid) and we are measuring the tooth length with respect to different doses of the two different methods of delivery. Also, the first 30 rows are for VC supp while the next 30 are for OJ sub.

Tooth Growth Comparison Before doing hypothesis and confidence interval analysis, lets transform the data:

```
datOJ <- ToothGrowth[31:60,]
datVC <- ToothGrowth[1:30,]
length_comp <- datOJ$len - datVC$len
comp_data <- data.frame(length_comp, datOJ$dose)
names(comp_data) <- c("diff_length", "dose")
```

Now, we have a mechanism to compare the difference in tooth length for different modes of delivery namely, OJ and VC and if diff_length is positive, it means OJ is more effective

```
n <- subset(comp_data, dose == 0.5)
t.test(n)
```

When dose == 0.5

```
##
##  One Sample t-test
##
## data:  n
## 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
```

Positive mean and confidence interval meaning that OJ is more effective at dose == 0.5 and the p-value is 0.01547 which means that the two methods are different.

```
n <- subset(comp_data, dose == 1.0)
t.test(n)
```

When dose == 1.0

```
##
##  One Sample t-test
##
## data:  n
## 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
```

Positive mean and confidence interval meaning that OJ is more effective at dose == 1.0 and the p-value is 0.03 which means that the two methods are different

```
n <- subset(comp_data, dose == 2.0)
t.test(n)
```

When dose == 2.0

```
##
##  One Sample t-test
##
## data:  n
## 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
```

Positive mean is low in number and confidence interval is both positive and negative meaning that the two methods have similar effectiveness at dose == 2.0 and the p-value is 0.3 (closer to 0.5) which means that the two methods are approaching parity.

Conclusions

1. From the box plot and the scatter plot, it is clear that as the dose increases, the effectiveness of the both the methods increases.
2. From the hypothesis and confidence interval analysis, for lower dose values (0.5 and 1.0), OJ is an effective method while for higher dose value (2.0), both methods achieve parity.

An assumption that we have made is that we have used t-model for such a small set of data.