Statistical Inference Project - Part 2

Srotaswati Panda

01/10/2019

This is a project for the Coursera Data Science class on Statistical Inference

Part 1: Basic Inferential Data Analysis

Overview:

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.

Setup Environment and Load Data

Load libraries and set global options

```
library(knitr)
opts_chunk$set(echo = TRUE,message=FALSE,warning=FALSE,tidy=TRUE)
library(ggplot2)
library(datasets)
library(xtable)
```

1. Load ToothGrowth dataset and provide basic summary of the data

Display the first few rows of the data and show the unique values of Dose and Supplement Delivery Method

```
data(ToothGrowth)
print(kable(head(ToothGrowth)))
```

len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5
5.8	VC	0.5
6.4	VC	0.5
10.0	VC	0.5

```
print(kable(table(ToothGrowth$supp,ToothGrowth$dose)))
```

	0.5	1	2
OJ	10	10	10
VC	10	10	10

```
sum(!complete.cases(ToothGrowth))
```

```
## [1] 0
```

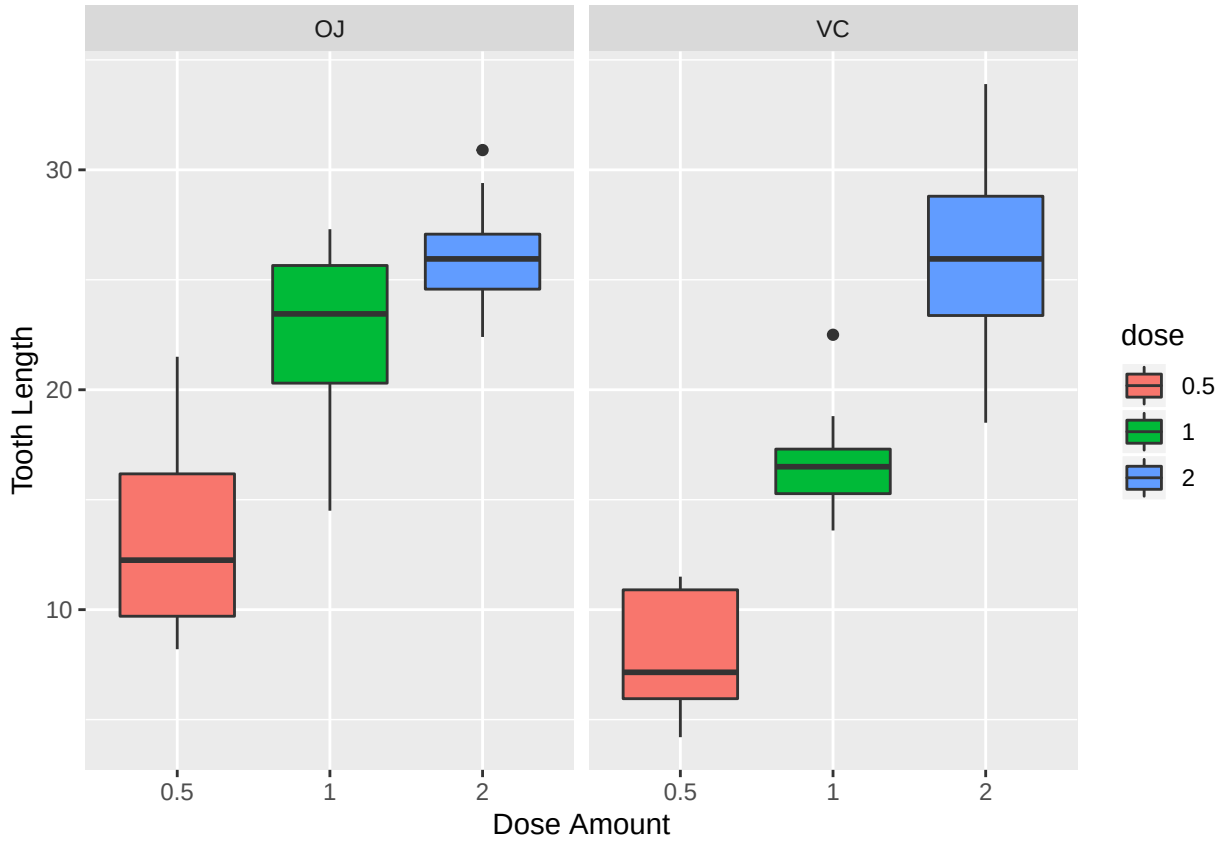
```
summary(ToothGrowth$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.07   19.25   18.81   25.27   33.90
```

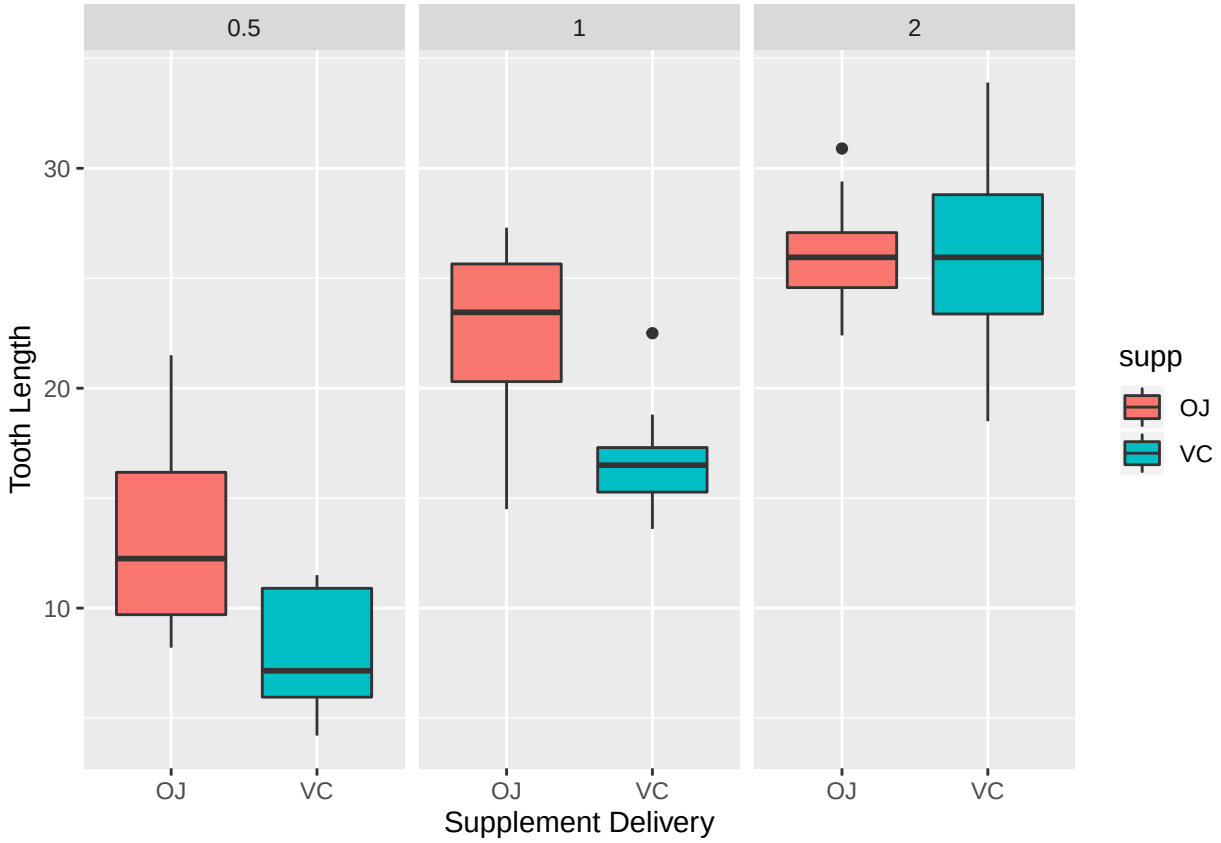
2. Perform basic exploratory analyses

Boxplots would show the range of Tooth Lengths vis a vis Dose Amount and Supplement Delivery Method. Tooth Lengths seem to be positively correlated with both Dose and Delivery Method. Further evidence can be established by running t-tests.

```
ToothGrowth$dose<-as.factor(ToothGrowth$dose)
g<-ggplot(data=ToothGrowth,aes(x=dose,y=len))
g<-g+geom_boxplot(aes(fill=dose))
g<-g+labs(x="Dose Amount",y="Tooth Length")
g<-g+facet_grid(.~supp)
g
```




```
g<-ggplot(data=ToothGrowth,aes(x=supp,y=len))
g<-g+geom_boxplot(aes(fill=supp))
g<-g+labs(x="Supplement Delivery",y="Tooth Length")
g<-g+facet_grid(.~dose)
g
```



3. Hypothesis testing

```
t.test(len~supp,data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##         20.66333         16.96333
```

At the significance level $\alpha = 0.5$, the p-Value is higher at 0.06, we fail to reject the null hypothesis H_o: difference in means for the groups 'OJ' and 'VC' is equal to 0. Also, the confidence interval includes 0, so the test is not significant.

```
sub<-subset(ToothGrowth,ToothGrowth$dose %in% c(0.5,1.0))
t.test(len~dose,data = sub)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```
sub<-subset(ToothGrowth,ToothGrowth$dose %in% c(1.0,2.0))
t.test(len~dose,data = sub)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

```
sub<-subset(ToothGrowth,ToothGrowth$dose %in% c(2.0,0.5))
t.test(len~dose,data = sub)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

At the significance level $\alpha = 0.5$, the p-Values of all the above tests is very close to 0 which implies that the null hypothesis H_o will be rejected in favor of the alternative hypothesis H_a that the difference in means for the three groups of Dosage Amounts in not equal to 0. Also, the confidence intervals do not contain 0.

4. Conclusions

It is not very clear if the Tooth Lengths are independent of Delivery Methods. However from the t-tests and the confidence intervals, Tooth Lengths increase with the increase in Dosage Amounts, which is also evident from the boxplots.