Data Science: Statistical Inference - Project Part 2 (Basic Inferential Data Analysis)

Jagannatha Reddy

July 26, 2016

Problem Description

In this project you will analyze the ToothGrowth data in the R datasets package. You should

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.

Synopsis

In this project we will analyze the ToothGrowth data avaiable as part of the **datasets** package of R. ToothGrowth data has data to track the Effect of Vitamin C on Tooth Growth in Guinea Pigs

1) Load the ToothGrowth data and perform some basic exploratory data analyses

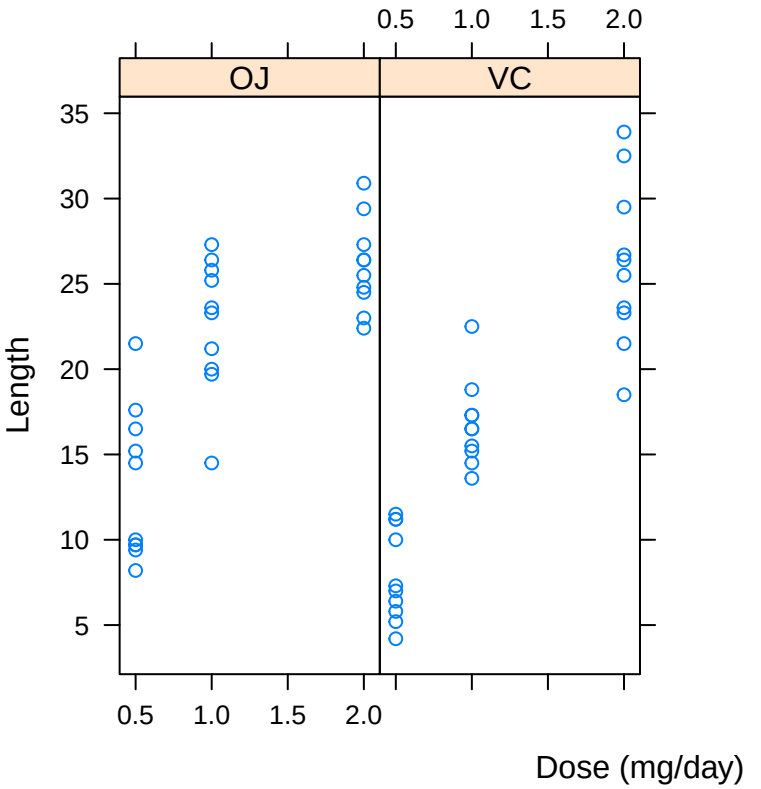
```
echo=TRUE
cache=TRUE
# load the required libraries
library(datasets)
library(lattice)
head(ToothGrowth)
```

```
##    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
```

```
dim(ToothGrowth)
```

```
## [1] 60  3
```

```
# do a basic plot
xyplot(len~dose|supp, data=ToothGrowth, layout=c(4,1), xlab="Dose (mg/day)", ylab="Length")
```



You can observe that there are totally 60 observations in the TotalGrowth dataset with each observation having 3 columns to represent Tooth length (len), Supplement type (supp), and Dose in milligrams/day (dose)

2) Provide a basic summary of the data.

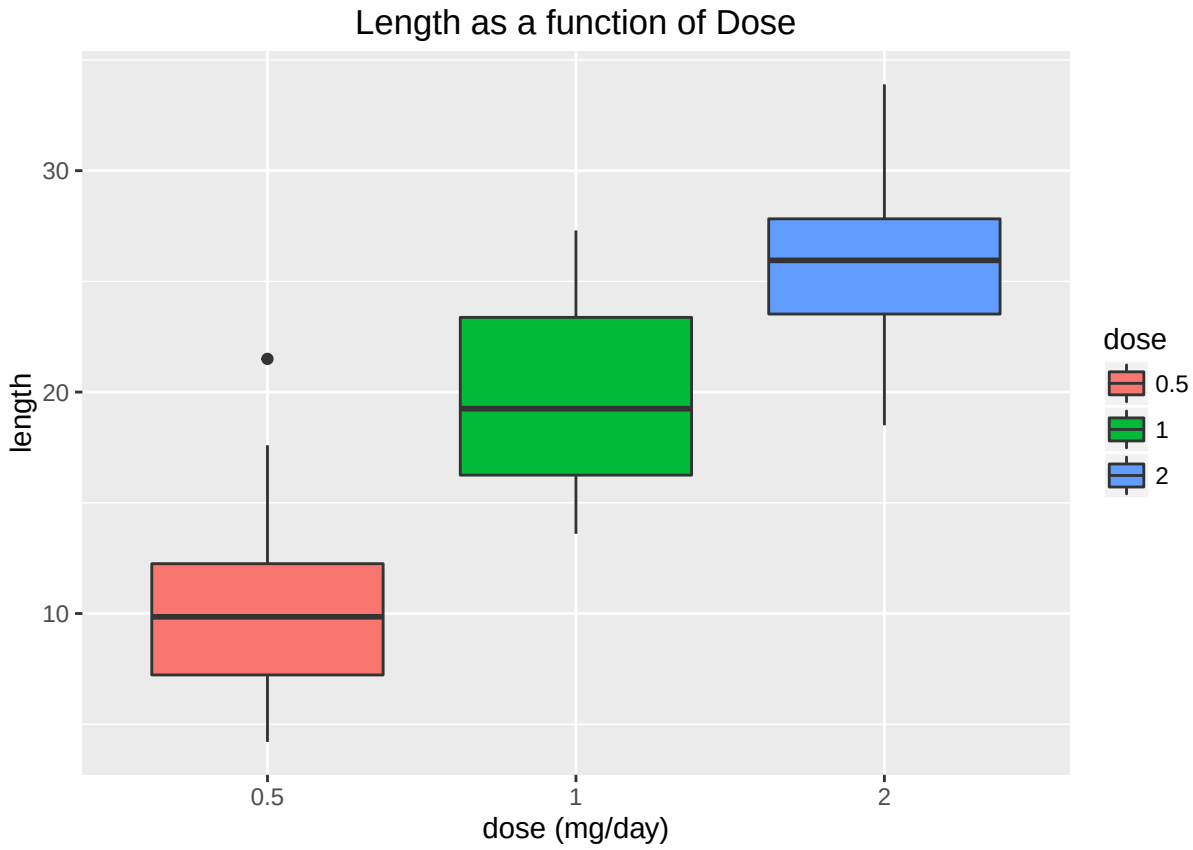
```
library(ggplot2)
summary(ToothGrowth)
```

```
##       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
```

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose) # make the dose value as factor
summary(ToothGrowth) # print the summary of data
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

```
g<-ggplot(aes(y=len,x=dose), data=ToothGrowth)+geom_boxplot(aes(fill=dose))
g+labs(x="dose (mg/day)", y="length", title="Length as a function of Dose")
```




```
mean(ToothGrowth[ToothGrowth$supp=='OJ',]$len)
```

```
## [1] 20.66333
```

```
mean(ToothGrowth[ToothGrowth$supp=='VC',]$len)
```

```
## [1] 16.96333
```

You can observe that length consistently increases with the dose value and also geinea pigs that received the dose through Orange Juice (OJ) has better growth compared to ascorbic acid (VC).

Some additional insights about the data where we summarize based on Supplement type & Dose combinations as we have only 10 observations for each of these combinations

```
for(s in c("VC", "OJ")) {
  for(d in c(0.5, 1.0, 2.0)) {
    combData <- ToothGrowth[ToothGrowth$supp==s&ToothGrowth$dose==d,]$len
      print(paste("Summary for supp", s, "& dose", d), quote=FALSE)
      print(summary(combData), quote=FALSE)
      print(paste("Variance for supp", s, "& dose", d, "is", round(var(combData),2)), quote=FALSE)
      print("", quote=FALSE)
  }
}
```

```
## [1] Summary for supp VC & dose 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20    5.95    7.15    7.98   10.90   11.50
## [1] Variance for supp VC & dose 0.5 is 7.54
## [1]
## [1] Summary for supp VC & dose 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   15.27   16.50   16.77   17.30   22.50
## [1] Variance for supp VC & dose 1 is 6.33
## [1]
## [1] Summary for supp VC & dose 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.50   23.38   25.95   26.14   28.80   33.90
## [1] Variance for supp VC & dose 2 is 23.02
## [1]
## [1] Summary for supp OJ & dose 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20    9.70   12.25   13.23   16.18   21.50
## [1] Variance for supp OJ & dose 0.5 is 19.89
## [1]
## [1] Summary for supp OJ & dose 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.50   20.30   23.45   22.70   25.65   27.30
## [1] Variance for supp OJ & dose 1 is 15.3
## [1]
## [1] Summary for supp OJ & dose 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.40   24.58   25.95   26.06   27.08   30.90
## [1] Variance for supp OJ & dose 2 is 7.05
## [1]
```

3) Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose

As seen the previous section geinea pigs that received the dose through Orange Juice (OJ) has better growth compared to ascorbic acid (VC). In this section we will illustrate this through t.test method. We also investigate how different dose values play a role in the growth of the tooth

```
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
```

```r
t.test(len~dose, data=ToothGrowth[ToothGrowth$dose==0.5|ToothGrowth$dose==2.0,])
```

```
##
##  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
```

```r
t.test(len~dose, data=ToothGrowth[ToothGrowth$dose==0.5|ToothGrowth$dose==1.0,])
```

```
##
##  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
```

```r
t.test(len~dose, data=ToothGrowth[ToothGrowth$dose==1.0|ToothGrowth$dose==2.0,])
```

```
##
##  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
```

You can observe that length consistently increases with the dose value and also geinea pigs that received the dose through Orange Juice (OJ) has better growth compared to ascorbic acid (VC)

3) State your conclusions and the assumptions needed for your conclusions

Based on the analysis of the ToothGrowth data, the results are summarized below:

1. Increase in the dose value consistently increases the growth of the tooth irrespective of the Supplement

2. Supplement administered through Orange Juice (OJ) has better growth compared to gascorbic acid (VC)

These conclusions are based on the limited dataset of 10 observations for each of the combination (Supplement & Dose). Though the results are consistent the data used for analysis doesn't appear to be statistically significant. Also looking at the high variance for each of the Supplement type & Dose combinations it is clear that data available for this analysis is not statistically significant