Statistical Inference - Course Project Part 2

Anjali Singh

1 October 2017

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.

Load the libraries and Data

We will first set up the environment and load up the library needed to run the project

```
library(ggplot2)
# Load ToothGrowth data
data("ToothGrowth")
set.seed(111)
# Display a summary and head of the data
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
```

Supp is a factor variable, with 30 obversation under "QJ" and 30 under "VC. Additionally we also see tooth length ranges from from 4.20 to 33.90, with mean being is 18.8133333 and a standard deviation being 7.6493152.

```
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
```

```
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 ...
```

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

```
#Dose has only three unique values, its better we convert it to factor variable for analysis
```

Summary of Data

There are total 60 observations, 3 Variables in the Dataset We now convert the dose from numeric to factor

```
# convert variable dose from numeric to factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(ToothGrowth)
```

```
##       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
```

Analysis

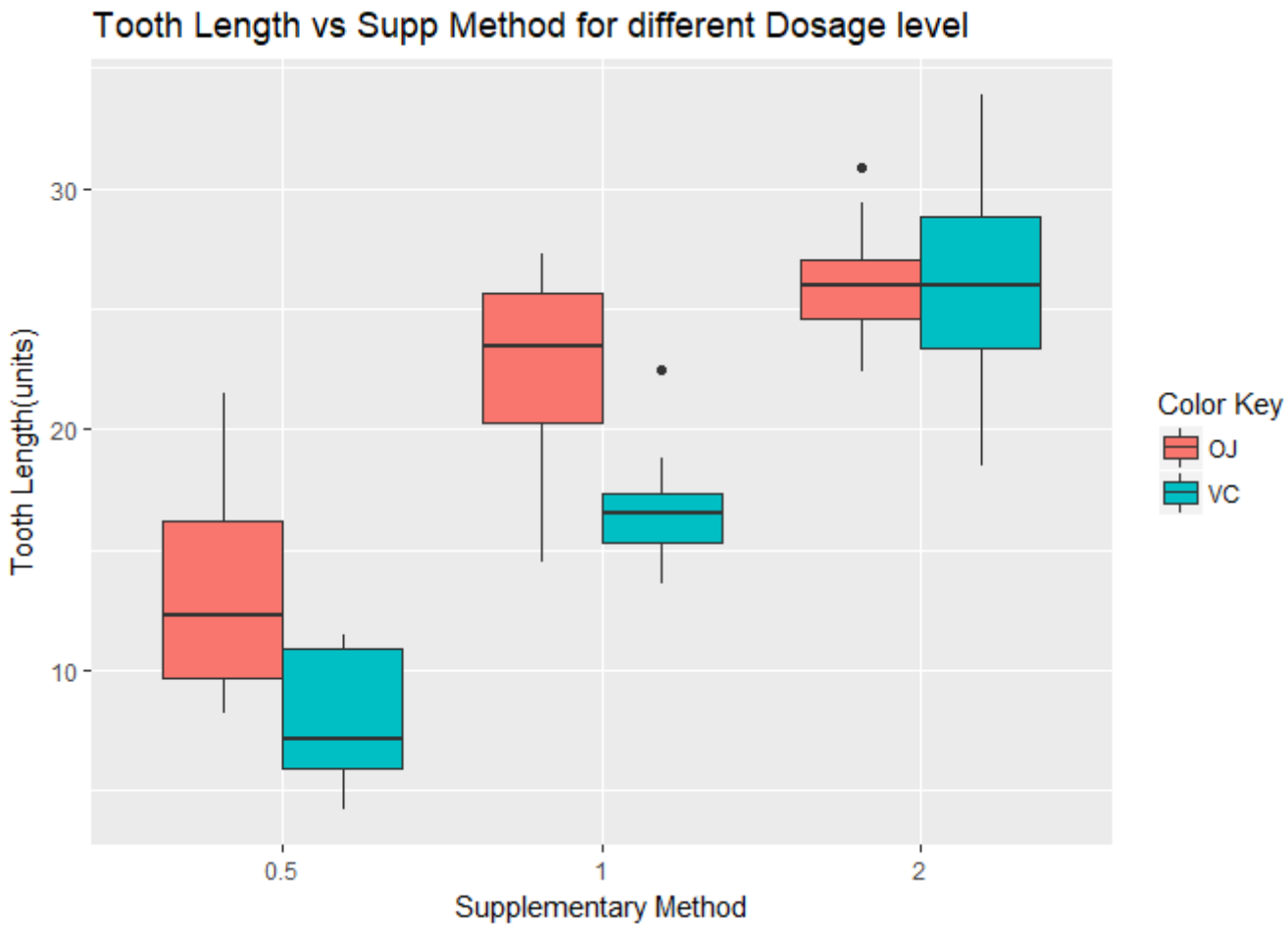
We will try and breakdown between different dose level and supplement level

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

```
##
##       OJ VC
##   0.5 10 10
##   1   10 10
##   2   10 10
```

Plotting Tooth growth for both dose and supplement

```
ggplot(aes(x=dose, y=len), data=ToothGrowth) + geom_boxplot(aes(fill=supp)) + xlab("Supplementary Method") + ylab("Tooth Length(units)") + ggtitle("Tooth Length vs Supp Method for different Dosage level")+labs(fill = "Color Key")
```



There appears to be a +ve relationship between tooth length and the dose.Also it appears as though “OJ” might be more effective in increasing tooth length than “VC”.

Now we will compare tooth growth by supplement using a t-test

Hypothesis:Supplement methods have no impact on tooth growth

H0: Both group have the same mean.

HA: Means are different.

```
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
```

The p-value of this test is 0.06. Since the p-value is greater than 0.05 and the confidence interval of the test contains zero, we can safely reject the null hypothesis. Therefore, Supplement menthod have no impact on tooth growth

Now we'll compare tooth growth by dose, looking at the different pairs of dose values.

By Dose Level: Dose has three factors 0.5,1.0,2.0, we have perform three hypothesis test to come to some conclusion

Hypothesis 1: Higher doses cause less tooth growth.

H0: Mean of level 2.0 is smaller or equal than level 0.5.

HA: Mean of level 2.0 is greater than level 0.5

```
# run t-test using dose amounts 0.5 and 2.0
tsub<-subset(ToothGrowth, ToothGrowth$dose %in% c(0.5,2.0))
t.test(len~dose,data=tsub)
```

```
##
##  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
```

Hypothesis 2 :

H0: Mean of level 1.0 is smaller or equal than level 0.5.

HA: Mean of level 1.0 is greater than level 0.5.

```
# run t-test using dose amounts 0.5 and 1.0
tsub<-subset(ToothGrowth, ToothGrowth$dose %in% c(0.5,1.0))
t.test(len~dose,data=tsub)
```

```
##
##  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
```

Hypothesis 3:

H0: Mean of level 2.0 is smaller or equal than level 1.0.

HA: Mean of level 2.0 is greater than level 1.0.

```
# run t-test using dose amounts 0.5 and 1.0
tsub<-subset(ToothGrowth, ToothGrowth$dose %in% c(1.0,2.0))
t.test(len~dose,data=tsub)
```

```
##
##  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
```

The p-value of each test above is essentially zero and the confidence interval of each test does not cross over zero.Therefore, we can reject the null hypothesis and safely say that the average tooth length increases with an inceasing dose.

Overall Conclusion

We can conclude that

1. Suppliment method has no impact of tooth growth.
2. As we increase the dose, we see growth in the tooth length.

Assumption

1. The sample is representative of the population
2. The distribution of the sample means follows the Central Limit Theorem