# Statistical Inference Project Part 2

Jamison Stevens

6/10/2020

## Exploratory Data Analysis

This program starts by loading the ToothGrowth data from datasets then looking at the data contained in ToothGrowth. The data is factored by dose, and boxplots are made for len with respect to supp and len with respect to dose. The purpose of this is to gain a general understanding of the code before performing an in-depth analysis or manipulating the data.

```
library(datasets)
data("ToothGrowth")

# Perform some basic exploratory data analyses
head(ToothGrowth)
```

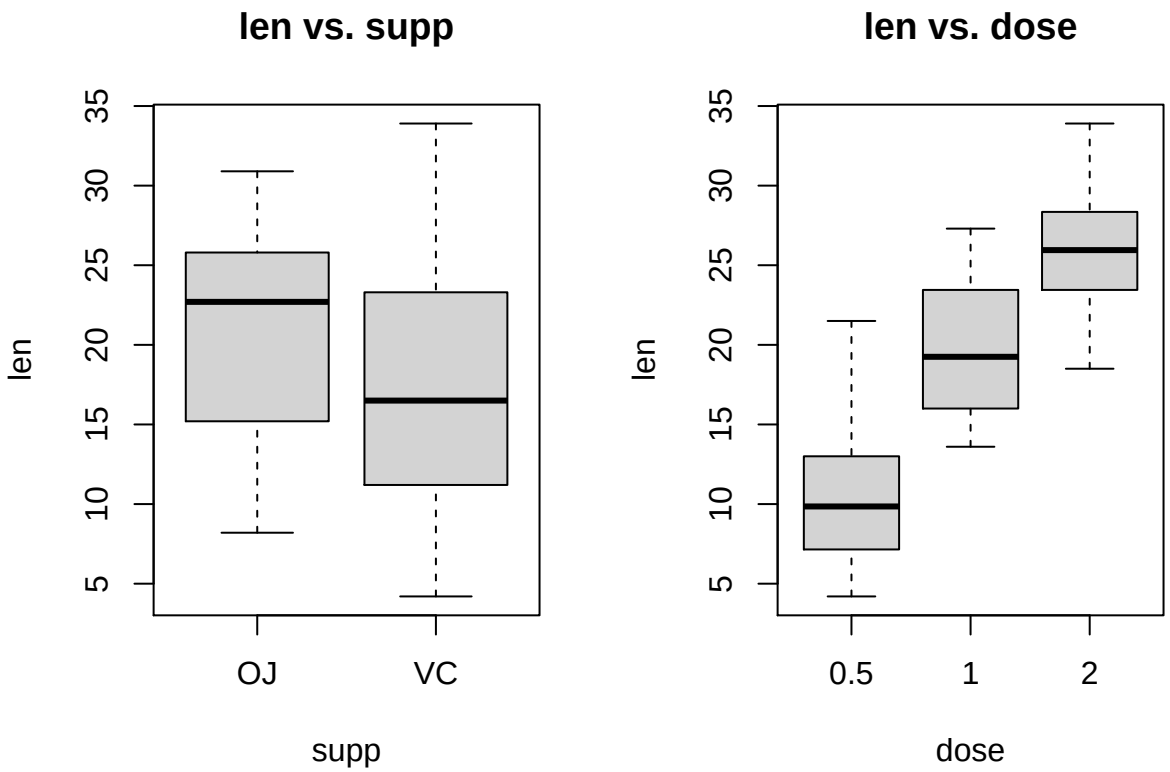
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
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 ...
```

```
# Factor Data by Dose
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

# Create Boxlots for len with respect to supp and len with respect to dose
par(mfrow = c(1,2))
boxplot(ToothGrowth$len~ToothGrowth$supp, main = "len vs. supp", xlab = "supp", ylab = "len")
boxplot(ToothGrowth$len~ToothGrowth$dose, main = "len vs. dose", xlab = "dose", ylab = "len")
```




The boxplot of len vs. supp shows that the average value for len is lower for VC than OJ, but the range of values for len is greater for VC than OJ. The boxplot of len vs. dose shows the average, maximum, and minimum values for len increase as dose increases.

## Summary of Data

This code provides a summary of the ToothGrowth data now that dose has been factored.

```
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
```

## Comparisons

This code uses a Welch Two Sample T-Test to compare the effect of different values of supp and dose on len. Since this test is for two samples, the supp data can be used directly in this test. However, the dose data has three different values, so only two values from dose can be compared at a time.

```
print(t.test(len ~ supp, 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
```

```
print(t.test(ToothGrowth[ToothGrowth$dose == 1.0, "len"],
  ToothGrowth[ToothGrowth$dose == 0.5, "len"]))
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$dose == 1, "len"] and ToothGrowth[ToothGrowth$dose == 0.5, "len"]
## 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:
##   6.276219 11.983781
## sample estimates:
## mean of x mean of y
##    19.735    10.605
```

```
print(t.test(ToothGrowth[ToothGrowth$dose == 2.0, "len"],
  ToothGrowth[ToothGrowth$dose == 0.5, "len"]))
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$dose == 2, "len"] and ToothGrowth[ToothGrowth$dose == 0.5, "len"]
## 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:
##  12.83383 18.15617
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

```
print(t.test(ToothGrowth[ToothGrowth$dose == 2.0, "len"],
  ToothGrowth[ToothGrowth$dose == 1.0, "len"]))
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$dose == 2, "len"] and ToothGrowth[ToothGrowth$dose == 1, "len"]
## 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:
##  3.733519 8.996481
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

These tests show that the p-value for the comparison of data from supp is greater than 0.05, and the p-values for the three comparisons of data from dose are less than 0.05.

## Conclusions and Assumptions

Since the p-value for the comparison of data from supp is greater than 0.05, the null hypothesis fails to be rejected. This means that the type of supplement doesn't impact tooth growth. One the other hand, the p-values for the three comparisons of data from dose are less than 0.05, the null hypotheses is rejected. This means that any increase in dose results in greater tooth growth.

There were several assumptions needed to come to these conclusions. One assumption is that the data is representative of the entire population. The data must also follow a normal distribution and be of a reasonably large sample size. The creation of this sample must also include random selection and random assignment to supplement types and dosage levels. We assume that measurements are not paired and variances are not equal.