Statistical Inference Course Project

Part 2: Basic Inferential Data Analysis Instruction

Data Reading and Summary

We must begin by loading the necessary libraries for the analysis

```
library(datasets)
library(ggplot2)
```

The ToothGrowth Dataset is then assigned to the variable *df*, we also perform a basic summary of this dataet:

```
data(ToothGrowth)
df <- ToothGrowth
summary(df)
```

```
##       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
```

```
head(df)
```

```
##    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
```

```
unique(df$len)
```

```
##  [1]  4.2 11.5  7.3  5.8  6.4 10.0 11.2  5.2  7.0 16.5 15.2 17.3 22.5 13.6 14.5
## [16] 18.8 15.5 23.6 18.5 33.9 25.5 26.4 32.5 26.7 21.5 23.3 29.5 17.6  9.7  8.2
## [31]  9.4 19.7 20.0 25.2 25.8 21.2 27.3 22.4 24.5 24.8 30.9 29.4 23.0
```

```
unique(df$supp)
```

```
## [1] VC OJ
## Levels: OJ VC
```

```
unique(df$dose)
```

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

```
by(df$len, INDICES = list(df$supp, df$dose), summary)
```

```
## : OJ
## : 0.5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20    9.70   12.25   13.23   16.18   21.50
## ------------------------------------------------------------
## : VC
## : 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20    5.95    7.15    7.98   10.90   11.50
## ------------------------------------------------------------
## : OJ
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.50   20.30   23.45   22.70   25.65   27.30
## ------------------------------------------------------------
## : VC
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   15.28   16.50   16.77   17.30   22.50
## ------------------------------------------------------------
## : OJ
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.40   24.57   25.95   26.06   27.07   30.90
## ------------------------------------------------------------
## : VC
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.50   23.38   25.95   26.14   28.80   33.90
```

Dataset Description

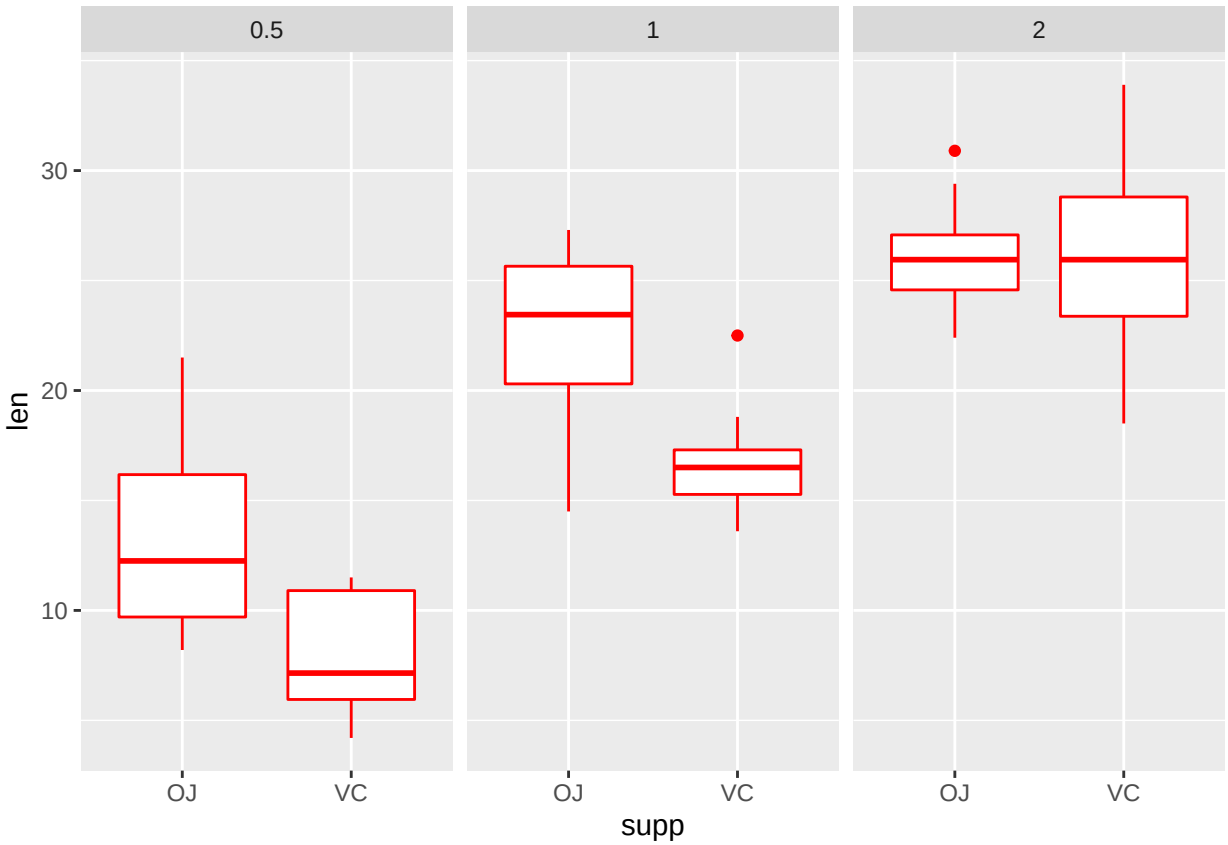
The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

In this project we are asked to use confidence intervals amd/or hypothesis tests to comapre tooth growth by *sup* (suplement type) and *dose* (dose levels of said vitamin).

Exploratory Data Analysis

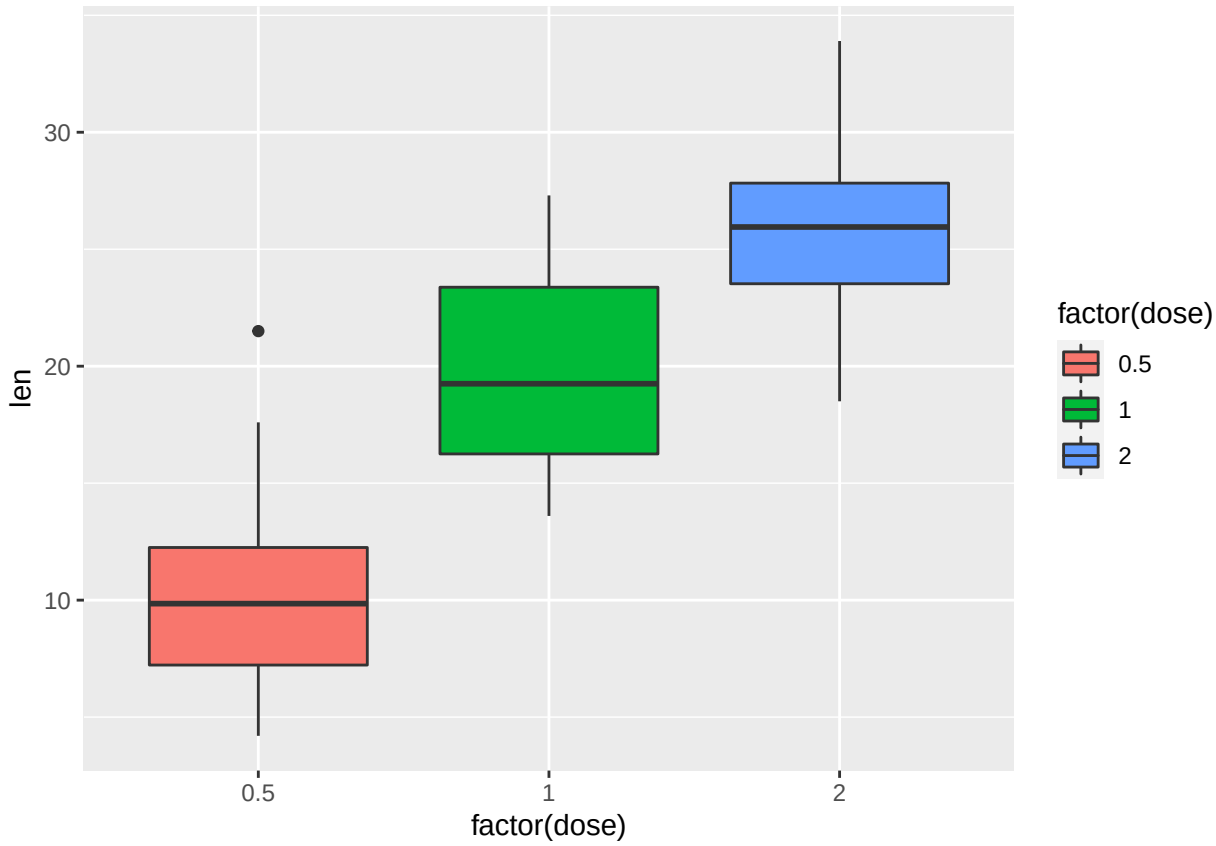
We start the EDA by performing a basic exploratory data analysis to see if ther is a visible pattern on the growth of the tooth based on the type of vitamin and its dosage:

```
g <- ggplot(df, aes(supp,len))
g + geom_boxplot(color = "red") +
  facet_grid(.~dose)
```



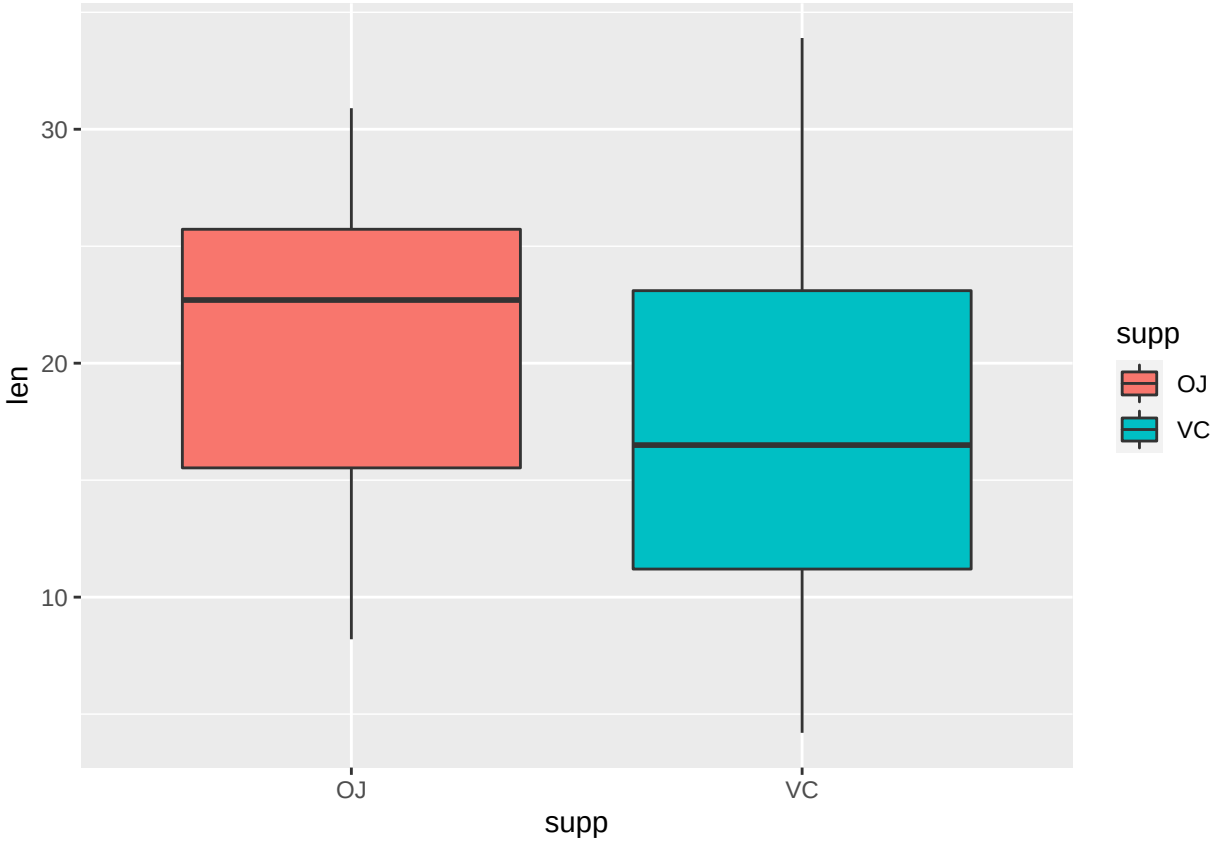
Based on the graph, it can be con be concluded that a bigger dosage results in longer teeth in both types of supplements, and that in both 0.5 and 1 mg/day orange juice leads to better results, while vitamin C is more effective with 2 mg/day. It would then be beneficial to explore both variables in isolation:

```
g <- ggplot(df, aes(x = factor(dose), y = len))
g + geom_boxplot(aes(fill = factor(dose)))
```



In this graphic we can prove that a bigger dose leads to longer teeth

```
g <- ggplot(df, aes(x = supp, y = len))
g + geom_boxplot(aes(fill = supp))
```



In this case, the results are not clear, while Orange Juice seems to be better than Vitamin C, it only is so by a small margin. It would be better to proceed with statistical methods to probe these hypotheses. ## Hypothesis testing to compare tooth length by Vitamin Supplement and Dosage

```
t.test(len ~ supp, paired = FALSE, val.equal = F, data = df)
```

```
##
##  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 between group OJ and group VC 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
```

We have a p-value of 0.06, meaning we cannot reject the null hypothesis that supplement types ## Analysis for correlation between dose level and tooth growth:

```
d1 <- subset(df, dose %in% c(0.5, 1.0))
d2 <- subset(df, dose %in% c(0.5, 2.0))
d3 <- subset(df, dose %in% c(1.0, 2.0))
t.test(len ~ dose, paired = F, var.equal = F, data = d1)
```

```
##
##  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 between group 0.5 and group 1 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
```

```
t.test(len ~ dose, paired = F, var.equal = F, data = d2)
```

```
##
##  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 between group 0.5 and group 2 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
```

```
t.test(len ~ dose, paired = F, var.equal = F, data = d3)
```

```
##
##  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 between group 1 and group 2 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
```

Analysis for correlation between dose level and tooth growth for each dose level:

```
d4 <- subset(df, dose == 0.5)
d5 <- subset(df, dose == 1.0)
d6 <- subset(df, dose == 2.0)
t.test( len ~ supp, paired = F, var.equal = F,  data = d4)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
t.test( len ~ supp, paired = F, var.equal = F,  data = d5)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

```
t.test( len ~ supp, paired = F, var.equal = F,  data = d6)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

The confidence intervals for 0.5 mg and 1.0 mg allow the rejection of the null hypothesis and confirm that there is correlation between those dose levels and tooth length. The interval for 2.0 mg is not enough to reject the null hypothesis. ## Conclussions and Assumptions We made the following assumptions to reach our conclusions:

- The populations are independent
- A random population was used
- Similar guinea pigs made the population
- At least 60 guinea pigs would have to be used for each combination of the variables

We can then conclude that there is a significant correlation between tooth length and dose levels with both vitamin supplements. There is a significant difference between 0.5 and 1.0 mg, but not significant enough with 2.0 mg. The better delivery method is orange juice, above maximum dosage there is no further improvement.