Inferential Data Analysis of the ToothGrowth Dataset

Shenay

June 9, 2019

Synopsis

In this report, I will apply statistical inference tools taught during the course to a specific dataset called ToothGrowth. This data contains tooth length for 60 guinea pigs, after they are given three dosage levels of vitamin C (0.5, 1, and 2 milligrams per day) via two delivery methods (orange juice and ascorbic acid).

The key question to address is the effects of different dosage levels as well as the two supplement types. I will focus on comparing the differences in tooth growth for each dosage level, holding the supplement type constant. Additionally, I aim to analyze the changes in tooth growth for each supplement type, given that the amount of dosages are consistent.

Load and Explore the Data

```r
library(datasets)
library(moments)
library(pastecs)
library(ggplot2)

data <- datasets::ToothGrowth
str(data)  #no missing data, 60 by 3 dimension
```

```
## '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 ...
```

```r
head(data, 3)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
```

```r
tail(data, 3)
```

```
##     len supp dose
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

```r
summary(data)  #summary statistics
```

```
##       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  
```

```r
skewness(data$len)  #negatively skewed to the left
```

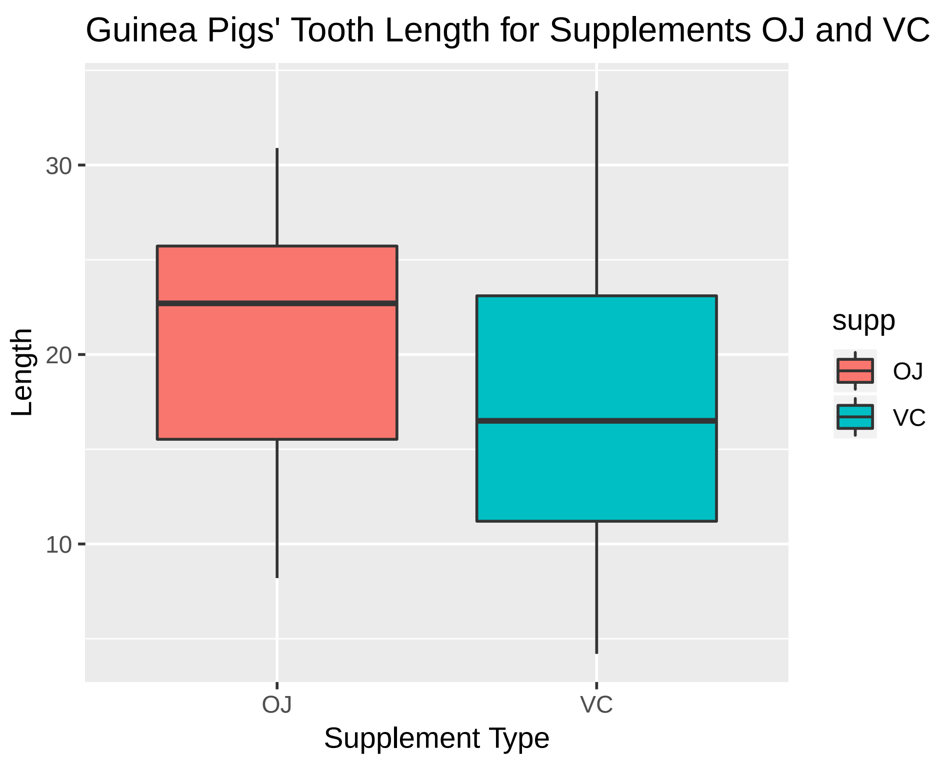
```
## [1] -0.1461768
```

```r
kurtosis(data$len)  #flatter tails since less than 3
```

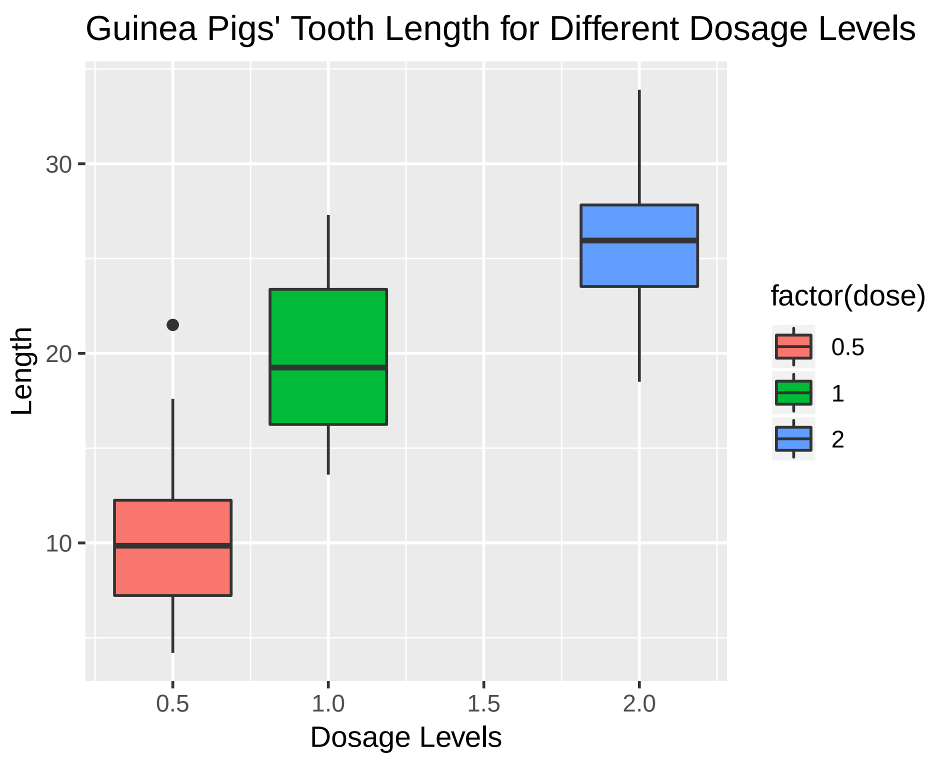
```
## [1] 2.024403
```

Visualize with Boxplots

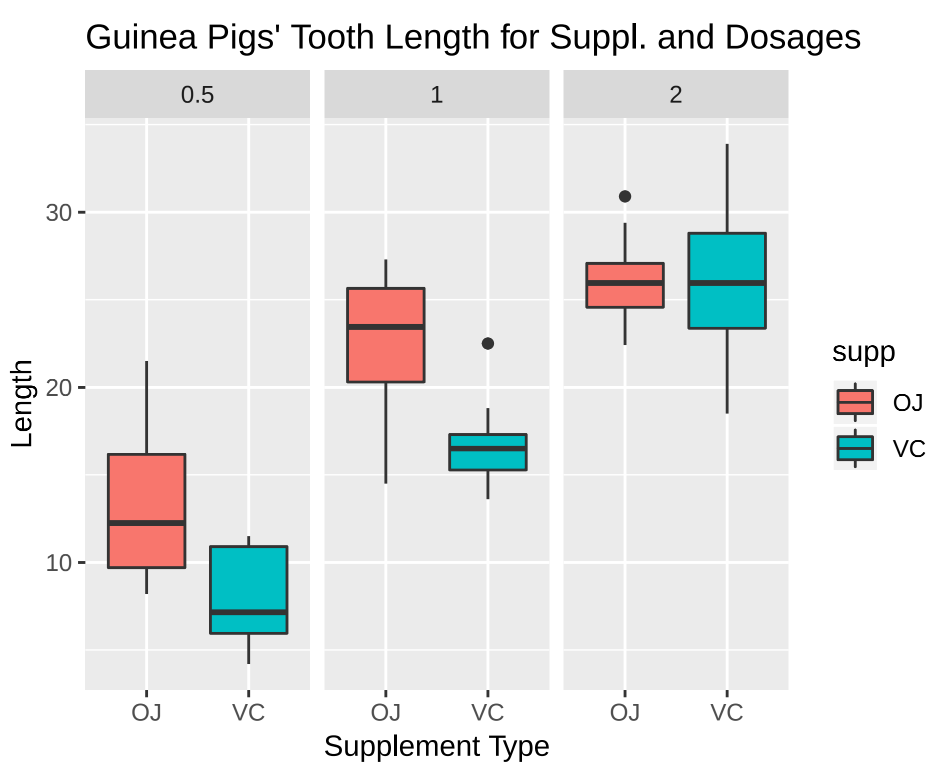
```r
ggplot(aes(x = supp, y = len), data = data) + geom_boxplot(aes(fill = supp)) +
           labs(title = "Guinea Pigs' Tooth Length for Supplements OJ and VC",
                x = "Supplement Type", y = "Length")
```




```r
ggplot(aes(x = dose, y = len), data = data) + geom_boxplot(aes(fill = factor(dose))) +
           labs(title = "Guinea Pigs' Tooth Length for Different Dosage Levels",
                x = "Dosage Levels", y = "Length")
```




```r
ggplot(aes(x = supp, y = len), data = data) + geom_boxplot(aes(fill = supp)) +
        facet_grid(.~ dose) +
        labs(title = "Guinea Pigs' Tooth Length for Suppl. and Dosages",
             x = "Supplement Type", y = "Length")
```



According to the three boxplots above, we can observe the following patterns:

- The type of supplement provided has a positive impact on guinea pigs' tooth length. It appears that the median length is higher via orange juice compared to that by ascorbic acid.
- The amount in dosage levels has a positive correlated relationship with guinea pigs' tooth length.
- After combining the effects of both dosage levels and supplement type, we can see that the median length is roughly similar by orange juice or ascorbic acid, provided that the dosage level is 2 mg/day. However, the range of values is much more spread out for vitamin c that is delivered by ascorbic acid.

Test Hypothesis

Effect from supplement type:

- Ho: For supplements OJ and VC, there are no differences in tooth growth

```
h1 <- t.test(len ~ supp, alt = "two.sided", mu = 0, paired = FALSE, data = data)
h1$p.value
```

```
## [1] 0.06063451
```

Effect from dosage level:

- Ho: Given dosage levels of 0.5, there are no differences in tooth growth for supplement OJ vs. VC

```
d0.5 <- subset(data[which(data$dose == "0.5"), ])

h2 <- t.test(len ~ supp, alt = "two.sided", mu = 0, paired = FALSE, data = d0.5)
h2$p.value
```

```
## [1] 0.006358607
```

Given a constant dosage level, effect from supplement type:

- Ho: Given dosage levels of 2.0, there are no differences in tooth growth for supplement OJ vs. VC

```
d2.0 <- subset(data[which(data$dose == "2"), ])

h3 <- t.test(len ~ supp, alt = "two.sided", mu = 0, paired = FALSE, data = d2.0)
h3$p.value
```

```
## [1] 0.9638516
```

Given supplement type OJ, effect from dosage levels:

- Ho: no difference in length for dosage levels of 2, given supplement is OJ

```
OJ_2.0 <- subset(data[which(data$dose == "2" & data$supp == "OJ"), ])

t.test(OJ_2.0$len, OJ_2.0$dose, alt = "two.sided", mu = 0, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ_2.0$len and OJ_2.0$dose
## t = 28.656, df = 9, p-value = 3.737e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  22.16069 25.95931
## sample estimates:
## mean of x mean of y
```

```
##     26.06      2.00
```

Given supplement type VC, effect from dosage levels:

- Ho: no difference in length for dosage levels of 2, given supplement is VC

```
VC_2.0 <- subset(data[which(data$dose == "2" & data$supp == "VC"), ])

t.test(VC_2.0$len, VC_2.0$dose, alt = "two.sided", mu = 0, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC_2.0$len and VC_2.0$dose
## t = 15.911, df = 9, p-value = 6.753e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  20.70791 27.57209
## sample estimates:
## mean of x mean of y
##     26.14      2.00
```

Conclusion and Assumptions

Based on our brief exploratory data analysis and hypothesis testings, we can state that there are differences in tooth growth for supplement OJ vs. VC, holding dosage levels constant. (p-values less than the significant level 0.05; 95% confidence intervals does not contain 0). Additionally, there are differences in tooth growth for different dosage levels, given that the supplement type is fixed.

Assumptions used: data observations are independent and identically distributed for different supplement types and dosage levels of 0.5, 1, and 2 mg/day. Further investigation using analysis of variance/ANOVA test may be more helpful since there are multiple independent and unrelated group factors.

Appendix

Other possible hypothesis:

- Ho: Given dosage levels of 0.5, there are no differences in tooth growth for supplement OJ vs. VC

```
d1.0 <- subset(data[which(data$dose == "1"), ])

t.test(len ~ supp, alt = "two.sided", mu = 0, paired = FALSE, data = d1.0)
```

```
##
##  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 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
```

- Ho: no difference in length for dose levels of 0.5 and 1.0, given supplement is OJ

```
OJ_0.5 <- subset(data[which(data$dose == "0.5" & data$supp == "OJ"), ])
OJ_1.0 <- subset(data[which(data$dose == "1" & data$supp == "OJ"), ])
```

```r
t.test(OJ_0.5$len, OJ_0.5$dose, alt = "two.sided", mu = 0, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ_0.5$len and OJ_0.5$dose
## t = 9.0266, df = 9, p-value = 8.335e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.539717 15.920283
## sample estimates:
## mean of x mean of y
##     13.23      0.50
```

```r
t.test(OJ_1.0$len, OJ_1.0$dose, alt = "two.sided", mu = 0, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ_1.0$len and OJ_1.0$dose
## t = 17.546, df = 9, p-value = 2.871e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  18.90227 24.49773
## sample estimates:
## mean of x mean of y
##      22.7       1.0
```

- Ho: no difference in length for dose levels of 0.5 and 1.0, given supplement is VC

```r
VC_0.5 <- subset(data[which(data$dose == "0.5" & data$supp == "VC"), ])
VC_1.0 <- subset(data[which(data$dose == "1" & data$supp == "VC"), ])

t.test(VC_0.5$len, VC_0.5$dose, alt = "two.sided", mu = 0, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC_0.5$len and VC_0.5$dose
## t = 8.6119, df = 9, p-value = 1.223e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.515176 9.444824
## sample estimates:
## mean of x mean of y
##      7.98      0.50
```

```r
t.test(VC_1.0$len, VC_1.0$dose, alt = "two.sided", mu = 0, paired = FALSE)$p.value
```

```
## [1] 9.805753e-09
```