Coursera Statistical Inference Part II - Tooth Growth Analysis

Charlie Astrada

10/28/2016

Overview

The purpose of this analysis is to explore the `ToothGrowth` dataset using statistical inference techniques learned in class. The `ToothGrowth` dataset shows the effect of vitamin c on tooth growth on 60 guinea pigs. Each animal received one of the following doses of vitamin c:

- 0.5 mg/day
- 1 mg/day
- 2 mg/day

The delivery methods, or supplement type, were one of:

- orange juice
- ascorbic acid (a form of vitamin c)

We will look at basic summary statistics of the data. We will then analyze the effect of tooth growth by the delivery method (supplement type) and dosage using hypothesis testing and confidence intervals. We will finalize this document with a brief summary conclusion of our findings.

Basic Summary of Data

We will load the libraries that we need to perform our analysis work.

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

# change dose to factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

The first summary we will perform is a simple `str`, which shows us the structure of the dataset.

```
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 ...
```

`str` confirms that we have 60 observations of 3 variables:

- len – variable containing the length of the guinea pig teeth

- supp – variable containing the supplement type, or delivery method, with values coded as “OJ” for orange juice and “VC” for ascorbic acid
- dose – variable containing the dosage given to the guinea pig, with the values coded as 0.5, 1, or 2

The second summary we will perform is a simple `summary`, which shows us quantiles and averages for each variable.

```
summary(ToothGrowth)
```

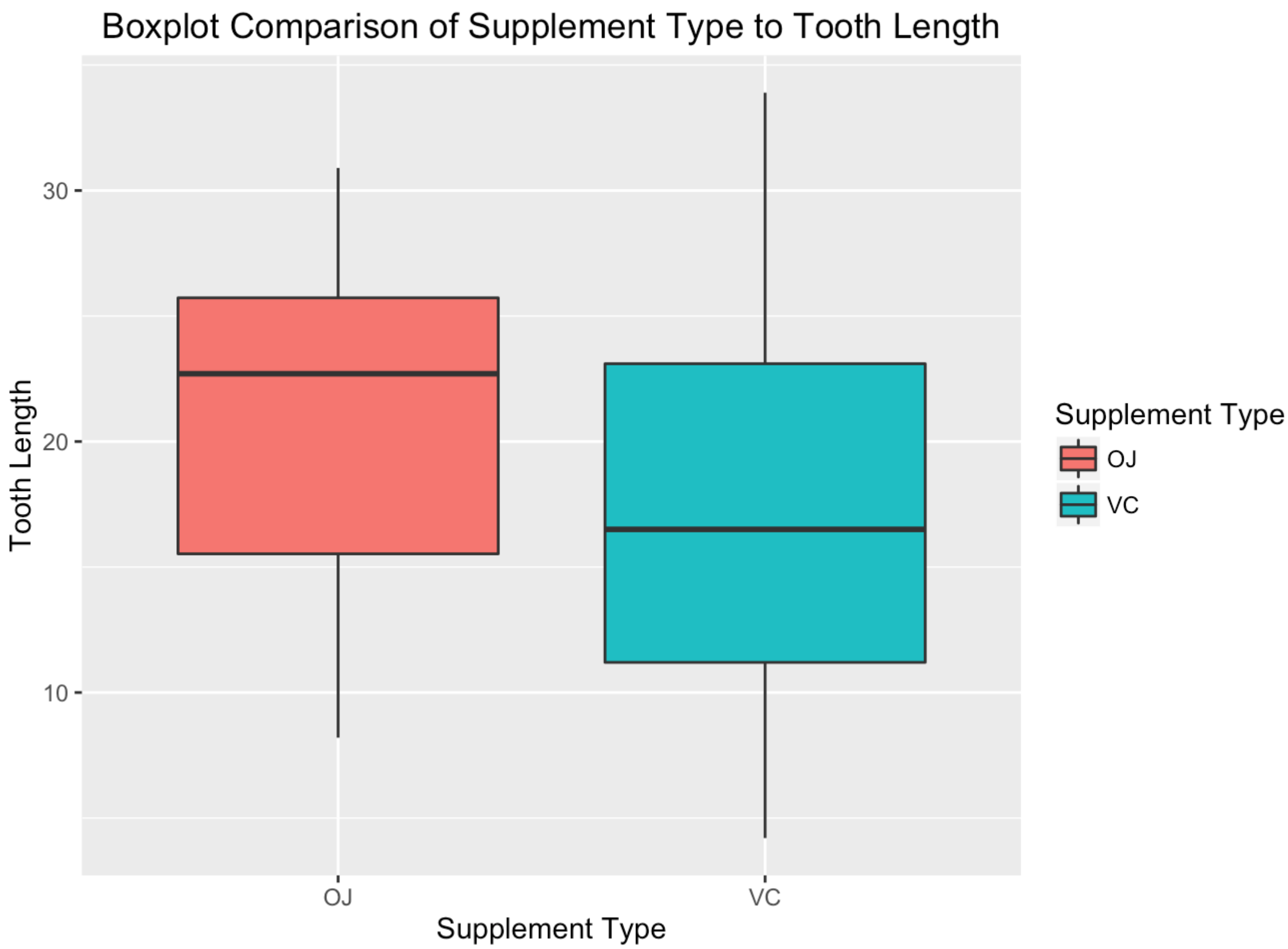
```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

`summary` shows us that the supplement type, `supp`, is split 50/50 between OJ and VC. Tooth length, `len`, has a range from 4.20 to 33.90 with a mean of 18.81. The vitamin c dosage, `dose`, shows that each dose was given to 20 guinea pigs in the experiment.

Comparison Analysis of Tooth Growth by Supp and Dose

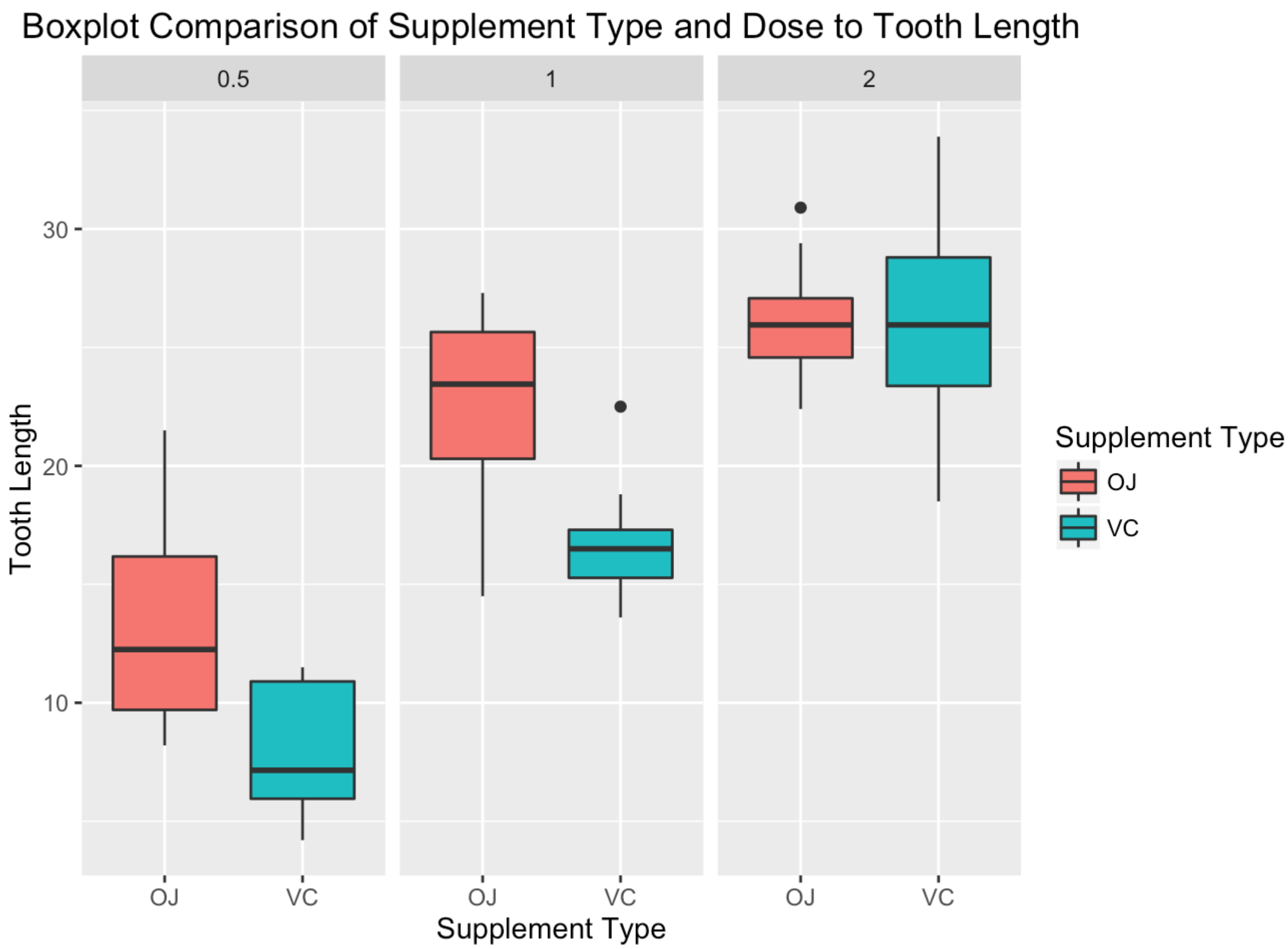
The exploratory graph below visually compares tooth length to both supplement type.

```
p.box <- ggplot(data = ToothGrowth, aes(x = supp, y = len))
p.box + geom_boxplot(aes(fill = supp)) +
        labs(title = "Boxplot Comparison of Supplement Type to Tooth Length", x = "Suppl
ement Type", y = "Tooth Length") +
        scale_fill_discrete(name = "Supplement Type")
```



Based on this initial boxplot, OJ has a higher average tooth length than VC. Given the range of values in each supplement type, it's clear that we need to drill down into the data. Let's take a look at the specific dosages relative to tooth length for each supplement type.

```
p.box.dose <- ggplot(data = ToothGrowth, aes(x = supp, y = len))
p.box.dose + geom_boxplot(aes(fill = supp)) +
        facet_wrap(~ dose) +
        labs(title = "Boxplot Comparison of Supplement Type and Dose to Tooth Length", x
 = "Supplement Type", y = "Tooth Length") +
        scale_fill_discrete(name = "Supplement Type")
```



Visually, it seems that there is a strong correllation between tooth growth and the supplement type, OJ, at doses of 0.5 and 1. At a dose of 2, however, it is not clear if OJ or VC has a stronger correllation to tooth growth.

In order to quantify what we see visually, and find out if our results are significant, we will run a T Test with confidence intervals for the different dosages. Our null hypothesis in all cases is that there is no difference in effect, at each dosage, of the two supplement types.

```
small.dose <- filter(ToothGrowth, dose == 0.5)
t.test(len ~ supp, var.equal = T, data = small.dose)
```

```
## 
##  Two Sample t-test
## 
## data:  len by supp
## t = 3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.770262 8.729738
## sample estimates:
## mean in group OJ mean in group VC 
##            13.23             7.98
```

At a dose of 0.5 mg/day, we get a t statistic of 3.1697 with a p-value of 0.005304, or a nearly a 0% chance of getting a t statistic that high by chance. Furthermore, the 95% confidence interval range goes from ~1.77 to ~8.73, which does not cross 0. We can safely reject the NULL hypothesis.

```
medium.dose <- filter(ToothGrowth, dose == 1)
t.test(len ~ supp, var.equal = T, data = medium.dose)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 18, p-value = 0.0007807
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.840692 9.019308
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

At a dose of 1 mg/day, we see a t statistic of 4.0328 with a p-value of 0.0007807, or a nearly 0% chance of getting a t statistic that high by chance. The 95% confidence interval range is from ~2.84 to ~9.02, which does not cross 0. We can safely reject the NULL hypothesis.

```
high.dose <- filter(ToothGrowth, dose == 2)
t.test(len ~ supp, var.equal = T, data = high.dose)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 18, p-value = 0.9637
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.722999  3.562999
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

At a dose of 2 mg/day, we see a t statistic of -0.046136 with a p-value of 0.9637, which is a very high likelihood that our t statistic can exist by chance. The 95% confidence interval range is from ~3.72 to ~3.56, which crosses 0. We cannot reject the NULL hypothesis at a dose of 2.

Finally, we will run a t test of supplement type compared to tooth length to see if we can find a statistically significant correllation of dosage, regardless of supplement type.

```
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
```

We get a t statistics of 1.9153 with a p-value of 0.06063, which is slightly above our cut-off alpha of 0.05. So, while there is strong evidence of a correllation between supplement type and tooth length, we cannot reject the NULL hypothesis outright.

Conclusion Summary

There is evidence of a correllation between supplement type and tooth length, although we did not reject the NULL hypothesis outright since the p-value of 0.06 is above our cutoff.

At doses of 0.5 mg/day and 1 mg/day, there is statistically significant evidence that the supplement type OJ is more strongly correllated than VC with regards to positive tooth length. At a dose of 2 mg/day, there is no significant evidence that OJ or VC is more strongly correllated to tooth length.