Investigation of ToothGrowth Dataset

Matthew Richards

October 2, 2016

Overview

In this report, we will be taking a look at the ToothGrowth dataset from R. Specifically, our goals are as follows:

- Explore the ToothGrowth dataset to get a general feeling for what it contains and summarize the data
- Compare tooth growth by *supp* and *dose* using confidence intervals and/or hypothesis tests
- State conclusions and assumptions needed for conclusions

Exploratory Data Analysis and Summary

Let's first load the data and take a quick look at the basic structure of the dataset using the *str()* command.

```
data("ToothGrowth")
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 ...
```

We can immediately see that *len* is a numeric vector that *supp* is a factor variable with 2 levels. The *dose* is a little less clear, it kind of looks like a factor but it's actually a numeric. And, of course, there are 60 observations. Let's look at how *supp* and *dose* are distributed

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

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

This clearly shows the distributions of factors: we know that our 60 subjects are divided in half with regard to *supp* and in thirds with regard to *dose*, so there are 6 distinct groups of 10 subjects each. Now that we've taken a look at the numbers, let's do a quick summary of the 2 different *supp* levels:

```
lapply(split(ToothGrowth$len,ToothGrowth$supp),summary)
```

```
## $OJ
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20   15.52   22.70   20.66   25.72   30.90
##
## $VC
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   11.20   16.50   16.96   23.10   33.90
```

Looks like "OJ" has larger lengths in almost every way, but that the spread of "VC" is larger. We can also do this for *dose* by converting it to a factor (Appendix, Text 1). Doing so, we see that as dose increases,

every statistic returned by *summary()* also increases. So at this point, we already have a pretty good idea of what's happening with regard to *supp* and *dose*. Let's see about making some quick graphs.

Histograms

We'll start by looking at the distribution of the data with some histograms for the different dosages. We'll color them by *supp* and split them by dose levels to give us 3 plots (see Appendix, Figure 1). Looking at these histograms, it's looks like the "OJ" might have higher lengths under the two lower dosages; it's less clear in the highest dose level, but the two distributions look pretty similar. Overall, there length of "OJ" subjects may very well be higher, but we'll have to look in another way to see more clearly. Tooth length also seems to have a positive correlation with dose, with higher doses corresponding to larger lengths.

Box and Whisker Plot

In order to look more closely at the *supp* variable, we can construct a box and whiskers plotch like the histograms, the box and whiskers plot (Appendix, Figure 2). Much like the histograms, this plot seems to show that "OJ" is correlated with larger tooth growth than "VC", but unlike the histograms it explicitly plot shows that "OJ" has a higher overall median. We also see that the two groups overlap quite a bit and the "VC" has more extreme values. We will likely want to look into this more rigorously later with a hypothesis test

Scatterplot

For our final exploratory plot, let's now look more closely at the correlation between dose and length using a scatter plot(Appendix, Figure 3). In addition to the data, we can plot a linear fit that shows what looks like a pretty clear positive correlation between length and dose, another relationship we can test later.

Summary of Data

Based upon our exploratory data analyses, we can summarize the data as follows:

- The "OJ" subjects have higher mean and median tooth growth than "VC" subjects, though "VC" subjects have a larger range.
- Subjects with higher dose levels have greater mean and median tooth growth than lower dose levels.

Note that at this point, we cannot speak to whether the differences we've seen are meaningful; for that, let's move on to some hypothesis testing

Hypothesis Testing

Now that we have a basic summary of the data, let's take a look at the effects of *supp* and *dose* on length. Using *?ToothGrowth*, we can see that the dataset contains 60 different subjects, so we can safely treat these observations as independent. We'll start with *supp*, which has only 2 levels. Earlier, we saw that "OJ" has a higher mean than does "VC", so let's conduct a one-sided T test where we hypothesize that $\mu_J > \mu_{VC}$. We'll test this by subsetting the data by *supp* and using the R *t.test()* command.

```
oj_growth <- subset(ToothGrowth,supp == 'OJ')$len
vc_growth <- subset(ToothGrowth,supp == 'VC')$len
t.test(oj_growth,vc_growth,alternative = "g")$conf
```

```
## [1] 0.4682687       Inf
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(oj_growth,vc_growth,alternative = "g")$p.value
```

```
## [1] 0.03031725
```

Our 95% confidence interval does not contain 0 and we have a p-value of about 0.03. Thus, we reject H_0 for $\alpha = 0.05$; it looks like the orange juice delivery method actually did result in larger tooth lengths.

Let us move on to *dose*, which is a bit more complicated because now we have 3 different groups rather than 2. We're going to need to test 3 different hypotheses then adjust either our p-values. Our hypotheses are:

1. $\mu_1 > \mu_{0.5}$
2. $\mu_2 > \mu_{0.5}$
3. $\mu_2 > \mu_1$

Let's go ahead and create the appropriate subsets, then run these tests, storing their p-values as we go along

```
dose0.5 <- subset(ToothGrowth,dose == 0.5)$len
dose1 <- subset(ToothGrowth,dose == 1)$len
dose2 <- subset(ToothGrowth,dose == 2)$len
dose_ps <- c(0,0,0)
dose_ps[1] <- t.test(dose1,dose0.5,alternative = "g")$p.value
dose_ps[2] <- t.test(dose2,dose0.5,alternative = "g")$p.value
dose_ps[3] <- t.test(dose2,dose1,alternative = "g")$p.value
dose_ps
```

```
## [1] 6.341504e-08 2.198762e-14 9.532148e-06
```

As we see from the output, our 3 raw p-values are each well below the threshold of $\alpha = 0.05$; however, because we have done multiple tests, we need to adjust. In this case, let's calculate adjusted p-values; we'll start with the more conservative "Bonferroni" method

```
p.adjust(dose_ps,method = "bonferroni")
```

```
## [1] 1.902451e-07 6.596287e-14 2.859644e-05
```

Our p-values are still all well below the threshold of $\alpha = 0.05$, even with this conservative adjustment, therefore there is no need to do a less conservative adjustment. Based upon these adjusted values, we can feel confident in rejecting H_0 for all 3 hypothesis tests.

Conclusions

Summing up our analyses, we explored the ToothGrowth dataset and conducted several hypothesis tests regarding *supp* (Delivery Method) and *dose*. Assuming all subjects were independent and that $\alpha = 0.05$ is an acceptable Type I Error Rate, our results were as follows:

- We rejected $H_0 : \mu_{OJ} = \mu_{VC}$ in favor of $H_a : \mu_{OJ} > \mu VC$. Our 95% confidence interval of [0.468, Inf] did not contain zero and our p-value was 0.03.
- We rejected three null hypotheses, summarized together as $H_0 : \mu_{0.5} = \mu_1 = \mu 2$ in favor of the alternate hypotheses, summarized together as $H_a : \mu_{0.5} < \mu_1 < \mu 2$. Because we conducted 3 hypothesis tests, we adjusted our p-values conservatively using the Bonferroni correction; these adjusted p-values of $(1.9x10^{-7}, 6.6x10^{-14}, 2.9x10^{-5})$ were still well below $\alpha = 0.05$

Based upon these results, we can reasonably hypothesize that the orange juice delivery method is correlated with larger tooth growth than is the ascorbic acid method and that higher doses of either form of Vitamin C are correlated with larger tooth growth than are lower doses.

Appendix

Text 1

```
lapply(split(ToothGrowth,as.factor(ToothGrowth$dose)),summary)
```

```
## $`0.5`
##       len          supp         dose
##  Min.   : 4.200   OJ:10   Min.   :0.5
##  1st Qu.: 7.225   VC:10   1st Qu.:0.5
##  Median : 9.850           Median :0.5
##  Mean   :10.605           Mean   :0.5
##  3rd Qu.:12.250           3rd Qu.:0.5
##  Max.   :21.500           Max.   :0.5
##
## $`1`
##       len        supp         dose
##  Min.   :13.60   OJ:10   Min.   :1
##  1st Qu.:16.25   VC:10   1st Qu.:1
##  Median :19.25           Median :1
##  Mean   :19.73           Mean   :1
##  3rd Qu.:23.38           3rd Qu.:1
##  Max.   :27.30           Max.   :1
##
## $`2`
##       len        supp         dose
##  Min.   :18.50   OJ:10   Min.   :2
##  1st Qu.:23.52   VC:10   1st Qu.:2
##  Median :25.95           Median :2
##  Mean   :26.10           Mean   :2
##  3rd Qu.:27.82           3rd Qu.:2
##  Max.   :33.90           Max.   :2
```

Figure 1

```
suppressWarnings(library(ggplot2))
g <- ggplot(ToothGrowth,aes(len,fill=supp))
g + geom_histogram(binwidth = 3) + facet_grid(~dose)
```

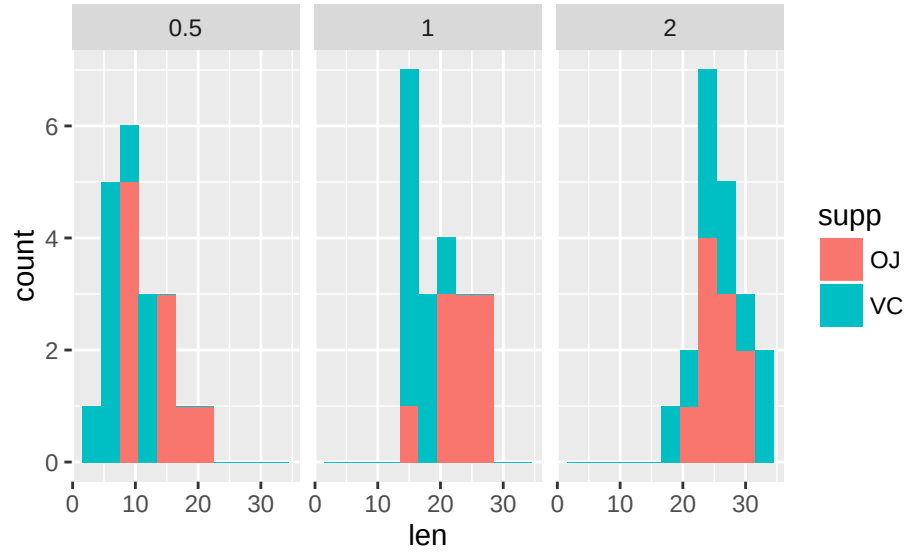

Figure 2

```
g <- ggplot(ToothGrowth,aes(supp,len))
g + geom_boxplot()
```

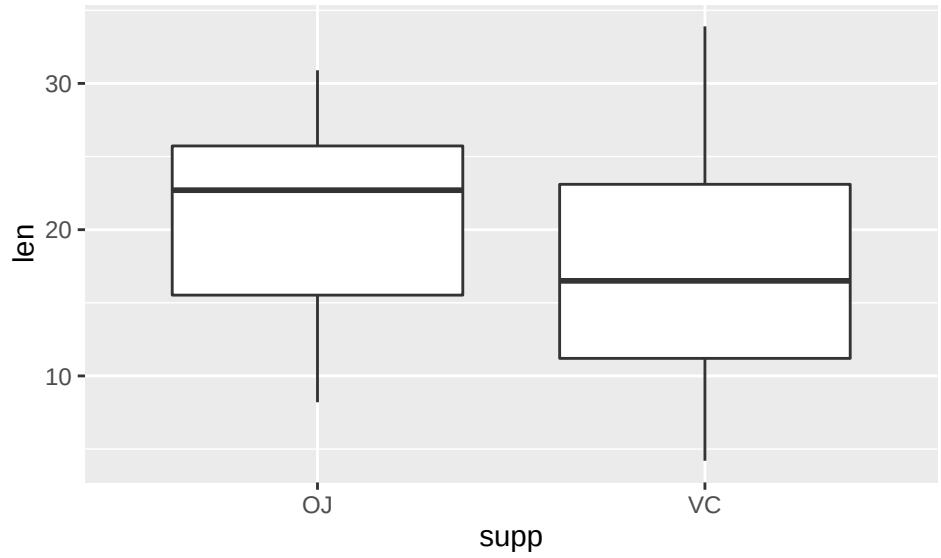

Figure 3

```
g <- ggplot(ToothGrowth,aes(dose,len))
g + geom_point(aes(color = factor(supp)),size=3) + geom_smooth(method = 'lm')
```