Statistical Inference Final Project - Part 2

Jerry Lakin

10/1/2020

Inferential Data Analysis

Exploratory Analysis

In this portion of the project we will analyze the ToothGrowth data set that comes with R and use statistical techniques to test the effectiveness of the two different supplements as well as the impact of dosage.

First we will load the data and look at a basic summary.

```
data(ToothGrowth)
summary(ToothGrowth)
```

```
##       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
```

As we can see there are three columns: supp and does which describe the supplement and dosage of the experiment, and len which gives us the resulting length.

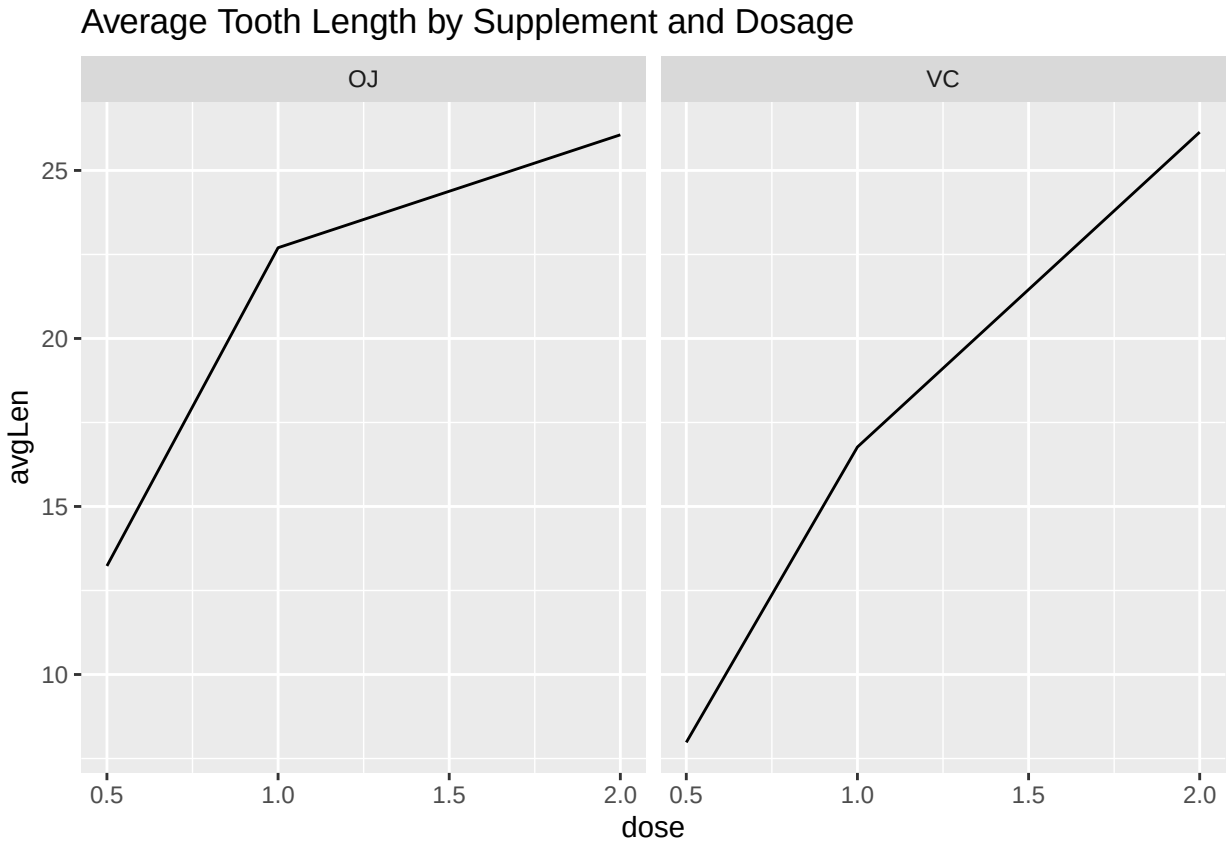
Hypothesis Testing

Next we'd like to visualize the data to see whether we can identify whether the supplement and dosage appear to have an impact on length. We will use the dplyr package to group the data by supplement and dosage and the ggplot2 package to visualize the results.

```
library(dplyr)
library(ggplot2)

suppDose <- ToothGrowth %>%
    group_by(supp, dose) %>%
    summarize(avgLen = mean(len))
```

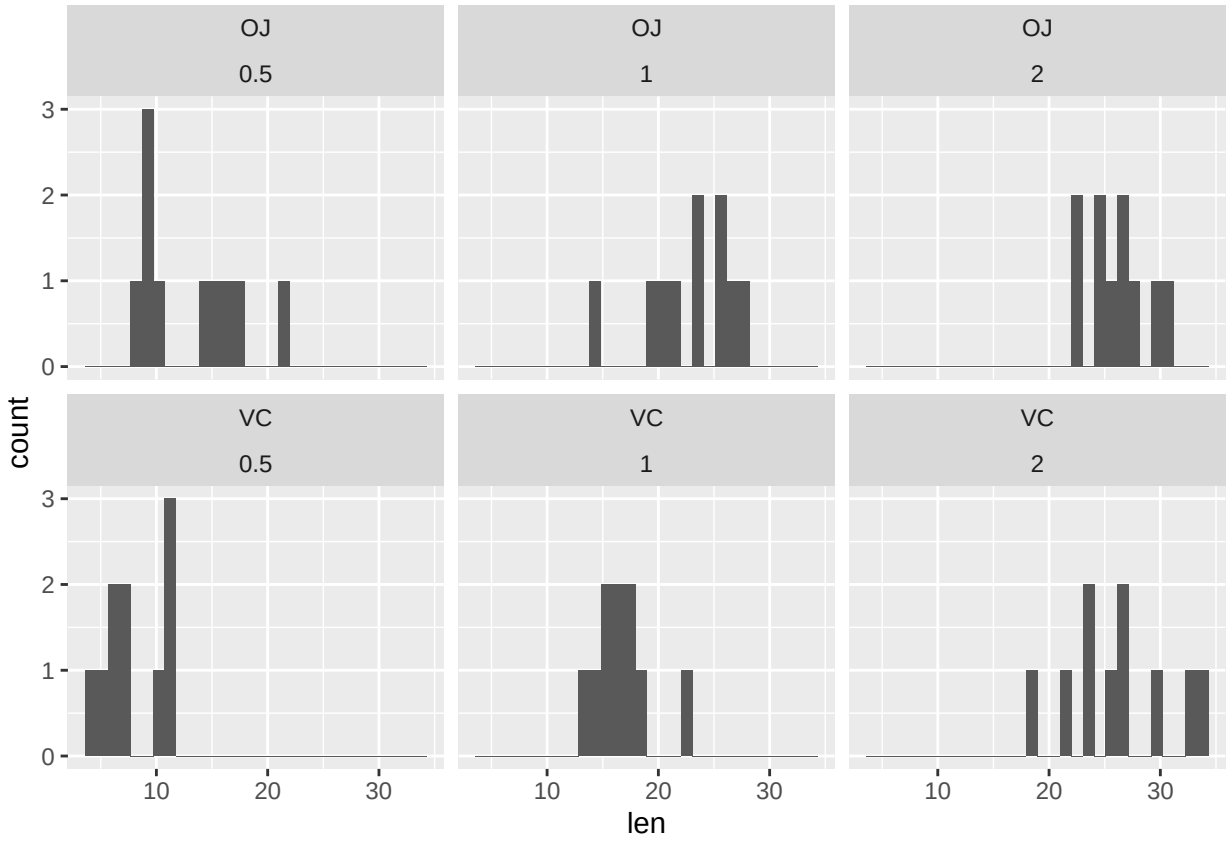
```
ggplot(suppDose, aes(x=dose, y=avgLen)) +
    geom_line() +
    facet_wrap(  ~ supp) +
    ggtitle("Average Tooth Length by Supplement and Dosage")
```



The average length across experiments clearly appears to increase with a larger dosage. Of the two supplements, OJ appears to have a larger impact, though the difference diminishes at higher dosages.

Next, we'd like to use hypothesis testing to determine whether the supplement and dosage impact the length variable. It would be nice to use the student's t-test to perform this analysis, but first we should confirm that our data are approximately normal. To do this we'll plot histograms of the lengths for each combination of supplement and dosage

```
ggplot(ToothGrowth, aes(len)) +
    geom_histogram(bins = 30) +
    facet_wrap(supp ~ dose)
```



Now that we've confirmed that our data are distributed in a roughly normal fashion, we can proceed to use the t-test. First, let's just see whether dosage and supplement impact length. We will perform hypothesis tests using the standard alpha value of 0.05.

```
p <- c(
    t.test(ToothGrowth[ToothGrowth$dose == 0.5,]$len, ToothGrowth[ToothGrowth$dose == 1,]$len)$p.value,
    t.test(ToothGrowth[ToothGrowth$dose == 0.5,]$len, ToothGrowth[ToothGrowth$dose == 2,]$len)$p.value,
    t.test(ToothGrowth[ToothGrowth$dose == 1,]$len, ToothGrowth[ToothGrowth$dose == 2,]$len)$p.value,
    t.test(ToothGrowth[ToothGrowth$supp == "OJ",]$len, ToothGrowth[ToothGrowth$supp == "VC",]$len)$p.va
)
```

Because we're doing multiple hypothesis tests, we will want to adjust our p-values. For this reason we'll hold off on looking at them for now.

In addition to the overall comparisons, we'd like to compare the two supplements at each dosage level. To do this, first we'll define new dataframes for each of the two supplements to simplify our code a bit.

```
OJSupp <- ToothGrowth[ToothGrowth$supp == "OJ",]
VCSupp <- ToothGrowth[ToothGrowth$supp == "VC",]
p <- c(
    p,
    t.test(OJSupp[OJSupp$dose == 0.5,]$len, VCSupp[VCSupp$dose == 0.5,]$len)$p.value,
    t.test(OJSupp[OJSupp$dose == 1,]$len, VCSupp[VCSupp$dose == 1,]$len)$p.value,
    t.test(OJSupp[OJSupp$dose == 2,]$len, VCSupp[VCSupp$dose == 2,]$len)$p.value
)
```

Now that we have our full list of p-values, we will adjust them using the Benjamini-Hochberg procedure and look at which tests yielded a p-value lower than our alpha of 0.05.

```
p.adjust(p, method = "BH") < 0.05
```

```
## [1]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
```

Looking back at our hypothesis tests, we see that each test comparing different dosages yielded a statistically significant difference in length. The test between the two supplements did not yield a statistically significant difference in length. As for the supplement comparisons at each dosage level, the difference was statistically significant at the lower two dosages but not the highest dosage.

Conclusions

In this analysis, we made several assumptions. We assumed that the populations of these experiments were identical outside of the supplements and dosages that they received and that there were not any significant differences in underlying factors that could have affected tooth growth. Secondly, we assumed that the samples were not paired. Finally, we assumed that the data were approximately normal and could be properly analyzed with t-tests.

Our conclusion is that the dosage has a significant impact on the tooth growth. In comparing supplements, we conclude that the "OJ" supplement causes larger growth than the "VC" supplement, but only at low dosages. For high dosages it appears that the difference in supplements disappears.