Exploratory Data Analysis of Tooth Growth

Jack Welch

June 6, 2017

Overview

Instructions taken from the Statistical Inference course offered by Johns Hopkins University on Coursera:

In the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package. This dataset contains observations related to the length of odontoblasts (cell response for tooth growth) in guinea pigs. We are to analyze the growth corresponding to a specific dose of Vitamin C delivered in 3 possible doses (0.5, 1.0 and 2.0 mg/day) and delivered via two possible different supplement types (OJ=Orange juice or VC=Ascorbic Acid).

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
4. State your conclusions and the assumptions needed for your conclusions.

Assumptions

This analysis and its conclusions are dependent on the following set of assumptions:

1. The sample populations are ramdomly selected and independent of each other.
2. The population was comprised of similar guinea pigs as was the methods for delivery of supplements, thus the variables are independent and identically distributed (iid).
3. Variances of tooth growth are different when using different supplements and delivery methods.
4. Tooth growth follows a normal distribution.
5. A confidence interval of 95% is satisfactory for our conclusions.

Exploratory Data Analysis of the ToothGrowth Dataset

We have not been given any instructions whatsoever related to the format of this data set or what it might represent. We have been asked to attempt to familiarize ourselves with this data with the use of exploratory data analysis techniques available with R programming. We will now conduct a few techniques so that we can become more familiar with this dataset. Comments are embedded within the R code below.

Initial Data Mining

```
# Load ToothGrowth dataset
tgdata <- ToothGrowth

# output a summary of the dataset
summary(tgdata)
```

```
##       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
```

```
# identify the structure of the existing dataset
str(tgdata)
```

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

```
# convert dose to a factor and look at structural change
tgdata$dose <- factor(tgdata$dose)
str(tgdata)
```

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

```
# visualize the dataset (since it is a small dataset)
tgdata
```

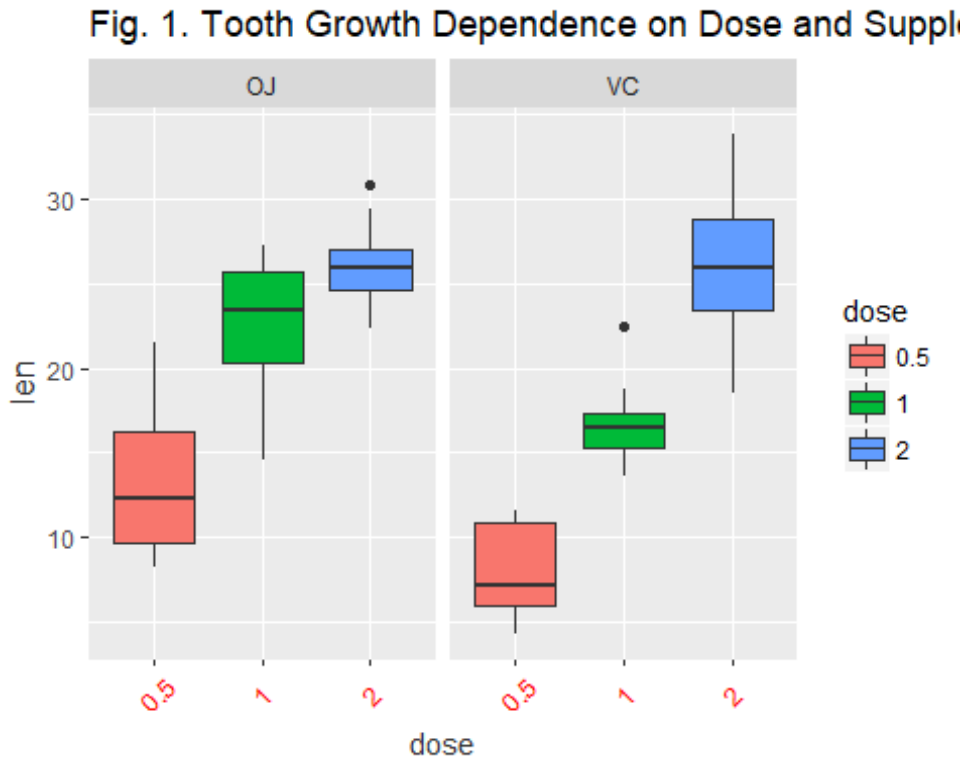
```
##     len supp dose
## 1   4.2   VC  0.5
## 2  11.5   VC  0.5
## 3   7.3   VC  0.5
## 4   5.8   VC  0.5
## 5   6.4   VC  0.5
## 6  10.0   VC  0.5
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
## 11 16.5   VC    1
## 12 16.5   VC    1
## 13 15.2   VC    1
## 14 17.3   VC    1
## 15 22.5   VC    1
```

```
## 16 17.3   VC    1
## 17 13.6   VC    1
## 18 14.5   VC    1
## 19 18.8   VC    1
## 20 15.5   VC    1
## 21 23.6   VC    2
## 22 18.5   VC    2
## 23 33.9   VC    2
## 24 25.5   VC    2
## 25 26.4   VC    2
## 26 32.5   VC    2
## 27 26.7   VC    2
## 28 21.5   VC    2
## 29 23.3   VC    2
## 30 29.5   VC    2
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
## 34  9.7   OJ  0.5
## 35 14.5   OJ  0.5
## 36 10.0   OJ  0.5
## 37  8.2   OJ  0.5
## 38  9.4   OJ  0.5
## 39 16.5   OJ  0.5
## 40  9.7   OJ  0.5
## 41 19.7   OJ    1
## 42 23.3   OJ    1
## 43 23.6   OJ    1
## 44 26.4   OJ    1
## 45 20.0   OJ    1
## 46 25.2   OJ    1
## 47 25.8   OJ    1
## 48 21.2   OJ    1
## 49 14.5   OJ    1
## 50 27.3   OJ    1
## 51 25.5   OJ    2
## 52 26.4   OJ    2
## 53 22.4   OJ    2
## 54 24.5   OJ    2
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

Box Plot Visualization

Let's now load the **ggplot2** library and visualize this data on a box plot.

```
library("ggplot2")
tgdata$dose=as.factor(tgdata$dose)
ggplot(tgdata, aes(x=dose,y=len)) +
    geom_boxplot(aes(fill = dose)) +
    ggtitle('Fig. 1. Tooth Growth Dependence on Dose and Supplement ') +
    facet_grid(.~supp) +
        theme(axis.title.y =
element_text(colour="gray20",size=12,angle=90,hjust=.5,vjust=1),
        axis.title.x = element_text(colour="gray20"),
        axis.text.x = element_text(colour="red",size=10,angle=45,hjust=.5,
vjust=.5))
```

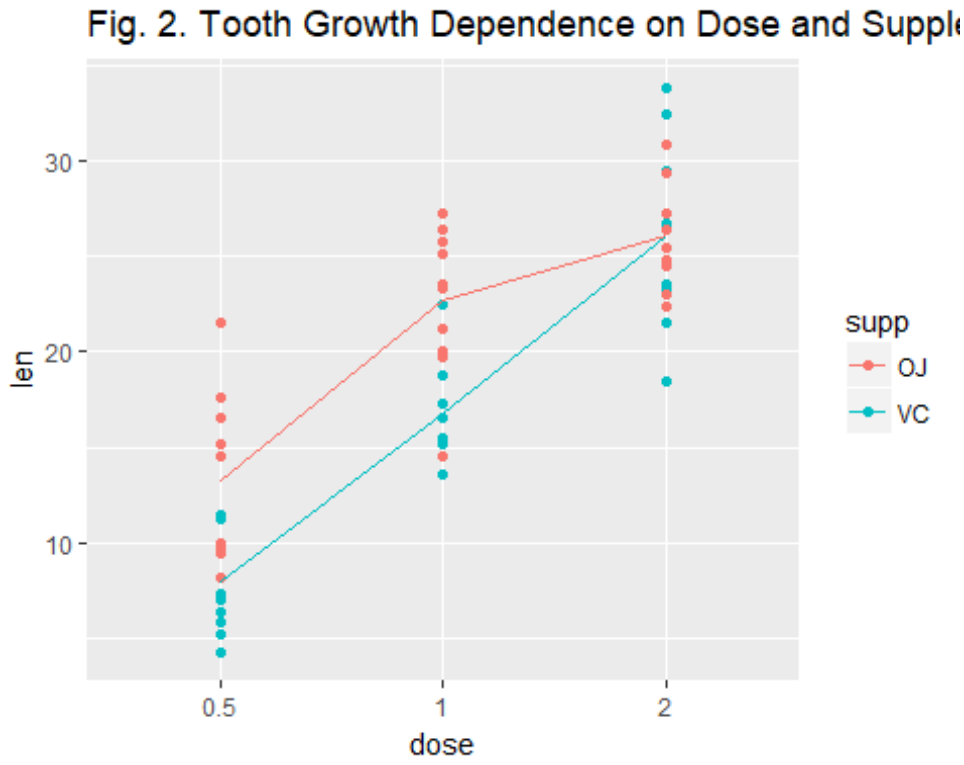


The box plots, as shown above in Fig. 1, allow us to visualize the distribution of the data along with five other important data points, including the min value, max value, mean, and the 1st and 3rd quartiles. Points on the lines outside the boxes are considered the outliers in the dataset. This visualization of the distribution of data clearly suggests that tooth growth in the guinea pigs clearly favored the delivery of the Vitamin C via the OJ with the smaller doses of 0.5 and 1 mg/day. The data further shows that while the means of the tooth growth related to the method of supplement delivery are equivalent with the higher dose of 2 mg/day, the tooth growth does show slightly higher normal variation with the ascorbic acid supplement in that higher dose of 2 mg/day.

Scatter Plot Visualization

Here is another visualization of the data using a scatterplot that I find very helpful.

```r
# Calculate the mean for every dose and supp
avg <- aggregate(len~., data=tgdata, mean)
g <- ggplot(aes(x=dose, y=len), data=tgdata) +
    geom_point(aes(color=supp)) +
    ggtitle('Fig. 2. Tooth Growth Dependence on Dose and Supplement ')
g <- g + geom_line(data=avg,aes(group=supp,colour=supp))
print(g)
```



This plot shows the same result as the boxplot in Fig. 1. In Fig. 2, once again, we can visualize that the tooth length associated with the OJ supplement are clearly higher with the lower doses and these means converge with the higher dose of 2 mg/day. At the higher dose of 2 mg/day, along with the convergence of the means, we can easily visualize that the variance of the data is greater with the VC supplement thus creating some events with greater tooth length along with events that are clearly shorter.

T Confidence Intervals and Hypothesis Testing

T Confidence Intervals is a statistical method for dealing with small datasets. In the Asymptotics lesson, we discussed confidence intervals using the Central Limit Theorem (CLT) and normal distributions. These methods required large sample sizes, and the formula for computing the confidence interval was mean +/- qnorm * std error(mean). Here, qnorm represents a specified quantile from a normal distribution. Also, in the study of Asymptotics, we also learned about the Z statistic Z =(X'-mu)/(sigma/sqrt(n)) which follows a standard normal distribution. This normalized Z statistic is especially nice because we know tha values of the mean and the variance which are 0 and 1, respectively.

The T statistic looks a lot like the Z. It's defined as t=(X'-mu)/(s/sqrt(n)). Like the Z statistic, the t is centered around 0. The only difference between the two is that the population standard deviation, sigma, in Z is replaced by the sample standard deviation in the t. So the distribution of the t statistic is independent on the population mean and variance. Instead, it depends on the sample size of n. The sample size of n now leads us into the definition of a concept of number of degrees of freedom, df, which is calculated simply as n-1.

In our lesson, we are encouraged, unless otherwise instructed, to consider that the default confidence interval should always be 95%. Our lecture took us through a series of exercises that concluded with the suggested use of the built-in R programming capabilities of t.test.

Subset Dosage as a Factor

```
dose1 <- subset(tgdata, dose %in% c(0.5, 1.0))
dose2 <- subset(tgdata, dose %in% c(0.5, 2.0))
dose3 <- subset(tgdata, dose %in% c(1.0, 2.0))
t.test(len ~ dose, paired = F, var.equal = F, data = dose1)

##
##  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 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 = dose2)

##
##  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 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 = dose3)

##
##  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 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
```

Supplement Dosage as a Factor

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

```
##
##  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
```

Conclusion

From the t confidence intervals calculated above, we can conclude that we should reject the null hypothesis, and confirm that there is a significant correlation between tooth length and dosage levels because the value of zero does not appear within any of the calculated confidence intervals. Alternatively, we would choose to accept the null hypothesis and confirm that there is no correlation between delivery method and tooth length as the value of zero does exist within the bounds of the calculated confidence interval.