Tooth Growth Analysis

Wesley Small (smallwesley)

May 7, 2016

OVERVIEW

This report shall conduct an analysis of the ToothGrowth data in available in the datasets package. The report shall progress as follows:

1. Data exploration of ToothGrowth dataset
2. Conduct an anaylsis; Utilizing confidence internal checks and hypothesis testing
3. Statement of Conclusions

Reference and Resources:

Please note the following resources for providing clarity in executing these report experiments:

- Confidence Intervals Module Lectures for this course:
- https://www.coursera.org/learn/statistical-inference/supplement/nM2Ak/confidence-intervals
- Hypothesis Testing Module/Lectures for this course:
- https://www.coursera.org/learn/statistical-inference/supplement/Qm5po/hypothesis-testing

1. Data Exploration

There is some background regarding the ToothGrowth dataset that can be obtain from command:

```
?ToothGrowth
```

The title of the study behind the dataset is **"The Effect of Vitamin C on Tooth Growth in Guinea Pigs"** The decription of which states, "The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC)."

The ToothGrowth data should have data-frame with 60 observations on 3 variables. The format is as follows:

#	column name	datatype	description
1.	len	numeric	Tooth length
2.	supp	factor	Supplement type (VC or OJ).
3.	dose	numeric	Dose in milligrams/day

```r
# LOAD LIBRARIES THAT WILL BE USED IN THIS ANALYSIS
library(datasets)
library(ggplot2)
# PREP AND LOAD DATASET
data("ToothGrowth")
```

We can confirm the dimensions/structure of the dataset as stated in the help summary as show below.

```r
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 ...
```

NO GUINEA PIG IDENTIER

It should be noted at this point, there is no subject/identifier field related to each observation. We know guinea pigs were given both treatments based on the summary, however, it is not recorded in this dataset. We cannot match observations per guinia pig to compare results. Thus while testing further we'll treat observations as independant or non-paired.

BY THE NUMBERS

Without iterated the entire dataset, exploring the unique distinct values, means, grouping, variance of the dataset. Particular emphasis on tallying results of lens to make assumptions about the study itself.

```r
# Example of formulas used to extra summaries from the data as shown in the tables below
nrow(ToothGrowth[ToothGrowth$supp == "OJ",])      # Length or Count of Observations
mean(ToothGrowth[ToothGrowth$dose == 0.5,]$len)  # Average / Mean
var(ToothGrowth[ToothGrowth$supp == "VC",]$len)  # Variance
length(unique(ToothGrowth$supp))                 # Unique
unique(ToothGrowth$dose)                         # Extracting Factors
```

UNIQUES:

column name	Distinct Factors?	Factors
supp	2	VC, OJ
dose	3	0.5, 1, 2
len	43	N/A -Lengths are variable

LENGTHS + AVERAGES (means)

Subset the the various factors; determine how many observations for each. Also included is the average per each factor grouping:

Factor	Length	Average (mean)	Variance
OJ	30	20.66	43.6334368
VC	30	16.96	68.3272299
O.5	20	10.605	
1.0	20	19.735	
2.0	20	26.1	

As shown, there are equal subsets of tests. We have not further information to determine if there dose/supp combinations were given to all guinea pigs evenly. Like the missing subject, we'll have to assume that all observations are independant. The variance between the two supplements is unequal, thus we shall definitely treat these observation as independant as well.

GROUP-AVERAGES

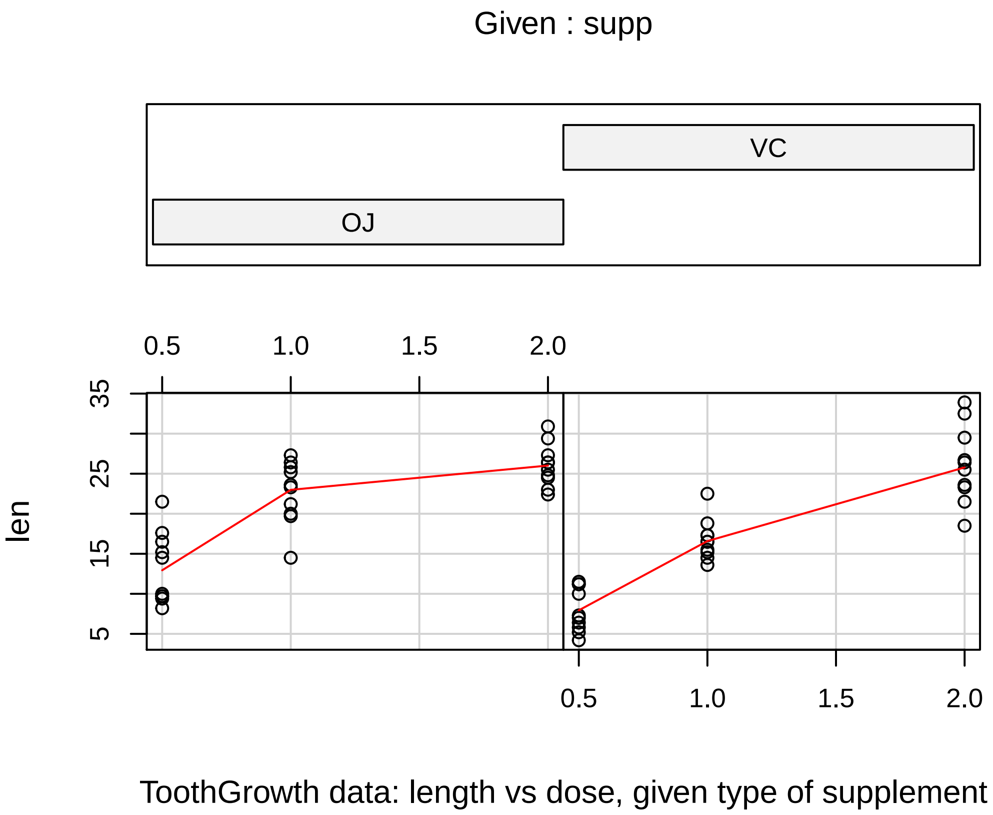
```
ToothGrowth$group <- paste0(as.character(ToothGrowth$supp),"_",as.character(ToothGrowth$dose))
ToothGrowth$group <-as.character(ToothGrowth$group)
```

GROUP	AVERAGE
VC_0.5	7.98
VC_1	16.77
VC_2	26.14
OJ_0.5	13.23
OJ_1	22.7
OJ_2	26.06

Visualization of the ToothGrowth Dataset

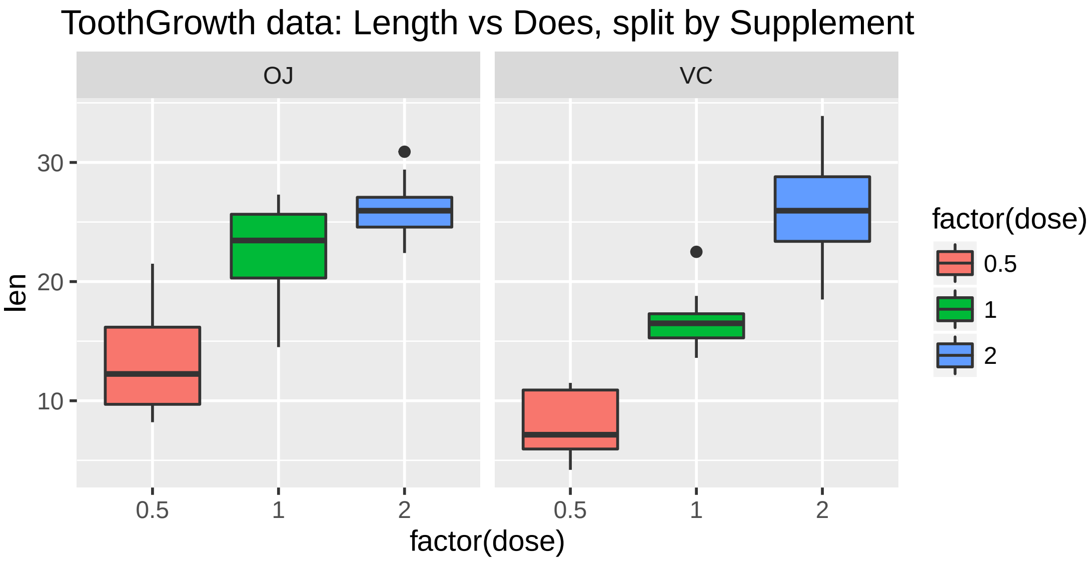
This first plot command comes directly from within ToothGrowth help summary available. It nicely shows a comparion of changes of length per dosage amount, faceted by supplement, of either OJ = Orange Juice, or VJ = ascorbic acid.

```
require(graphics)
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```

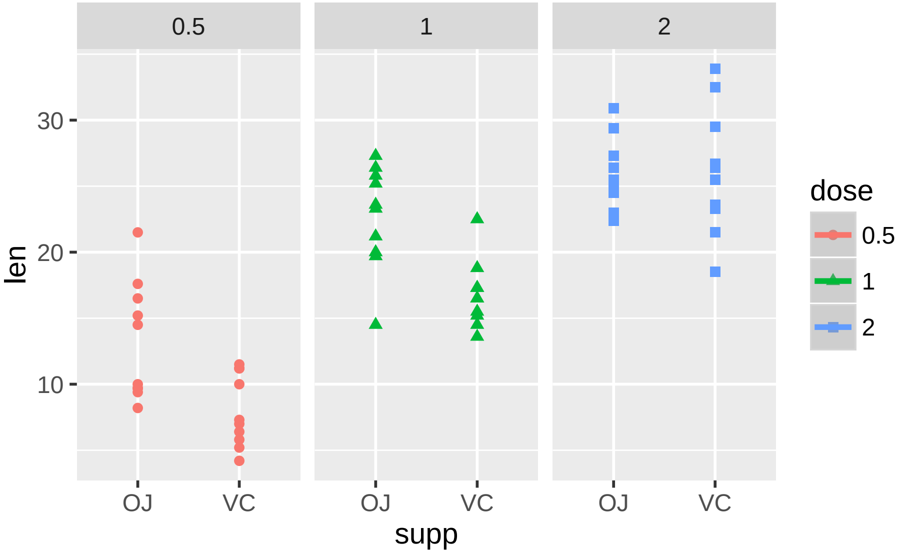


To further illustrate the patterns of the date, a boxplot is utilized to demonstrate ranges lengths per dosage, faceted by supplement type.

```r
ggplot(ToothGrowth, aes(x = factor(dose), y = len, fill = factor(dose) ) ) +
  geom_boxplot() + facet_grid(.~supp) +
  ggtitle("ToothGrowth data: Length vs Does, split by Supplement")
```



```r
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x=supp, y=len, color=dose, shape=dose)) +
geom_point() + facet_grid(.~dose)+ geom_smooth(method=lm, aes(fill=dose))
```



Via the visual anaylsis, we see a few trends: 1. ToothGrow is affected by the dosage, regardless of either supplement. The more dosage leads to more tooth growth. This is evident. The graphical nature of boxplots draws to look across at each dosage level to see the difference between either supplement. This is an area for anaylsis. 2. Note at the 2.0 dosage, it doesn't appear there is any significant difference. Certainly, at the lower dosages (0.5, 1.-), Orange Juice (OJ), has better effect on ToothGrowth.

2. Data Analysis

To compare tooth growth by supp and dose, we will perform some anaylsis using Confidence Intervals and Hypothesis Testing.

As stated in the study, guinea pigs were given combinations of supplement+dosage (i.e dose = 0.5 + supp = “VJ”. These groups will be used to study the effects between each other. As well saw in the boxplot above, we see a correlation between what is happening between each supplement at specific dosages of 0.5, 1 and 2. So, we will take subset groups of supplement/dose combinations (i.e. OJ_1, VC_1) and perform some statistical anaylsis.

Analysis Code Preparation

In our data exploration exercise, a **group** columns was added to the ToothGrowth dataset:

```
tgShuffled <- ToothGrowth[sample(nrow(ToothGrowth)),]
head(tgShuffled,n=3)
```

```
##     len supp dose group
## 27 26.7   VC    2  VC_2
## 30 29.5   VC    2  VC_2
## 24 25.5   VC    2  VC_2
```

Construct table for comparing statistic results:

```
op <- options(stringsAsFactors=F)
# EMPTY DATA.FRAMEOFR RESULTS
anaylsisDF <- data.frame(
  group1= character(0), group2= character(0),
  tEstimate = character(0),
  tConfInt= character(0), tStatistic = character(0),
  tPStatistic = character(0),stringsAsFactors=FALSE)

getAnaylsisRow <- function(g1Key, g2Key) {
   g1 <- ToothGrowth$len[ToothGrowth$group == g1Key]
   g2 <- ToothGrowth$len[ToothGrowth$group == g2Key]
   tanswer <- t.test(g1, g2, var.equals=FALSE, paired = FALSE)
   c(g1Key, g2Key, paste(round(tanswer$estimate,3),collapse=" vs "),
     paste(round(tanswer$conf.int,2),collapse=" <=> "),
     round(tanswer$statistic,3), round(tanswer$p.value,3))
}
```

Hypothesis Testing:

In all test cases below, we assume the following rules for our hypothesis testing:

- **H_0 (Null Hypothesis)**: *Group1_Mean = Group2_Mean*
- **H_a (Alternative Hypothesis)** is *Group1_Mean <> Group2_Mean*

We’ll execute logic to compare subsets each supplement (“OJ” and “VC”) with a dosage amount. We’ll subset by our groups as provided here:

```r
rbind(anaylsisDF, getAnaylsisRow("OJ_0.5","VC_0.5")) -> anaylsisDF
rbind(anaylsisDF, getAnaylsisRow("OJ_1","VC_1")) -> anaylsisDF
rbind(anaylsisDF, getAnaylsisRow("OJ_2","VC_2")) -> anaylsisDF
colnames(anaylsisDF) <-
  c("Group 1","Group 2", "Mean/T-Estimate", "Confidence Intervals","T-Statistic","P-Value")
```

```r
knitr::kable(anaylsisDF, format = "markdown")
```

Group 1	Group 2	Mean/T-Estimate	Confidence Intervals	T-Statistic	P-Value
OJ_0.5	VC_0.5	13.23 vs 7.98	1.72 <=> 8.78	3.17	0.006
OJ_1	VC_1	22.7 vs 16.77	2.8 <=> 9.06	4.033	0.001
OJ_2	VC_2	26.06 vs 26.14	-3.8 <=> 3.64	-0.046	0.964

Dosage @ 0.5 milligrams/day

We can conclude when dosage is 0.5 mg/day, we have a confidence interval (95%) that doesn't include zero, thus we should reject the null hypothesis. We have a set of mean estimates that show that OJ (Orange Juice) performs better than VC(Asorbic Acid).

Dosage @ 1.0 milligrams/day

We can conclude when dosage is 1.0 mg/day, we have a confidence interval (95%) that doesn't include zero, thus we should reject the null hypothesis. We have a set of mean estimates that show that OJ (Orange Juice) alos performs better than VC(Asorbic Acid). As you can see this is similar to 0.5 mg/day dosage results.

Dosage @ 2.0 milligrams/day

We can conclude when dosage is 0.5 mg/day, we have a confidence interval (95%) does include 0, accept the null hypothesis. In comparison of the between OJ_2 and VC_2, we see the, we clearly can see that there is little difference between the tooth growth rate to indicate which performs better.

NOTE: The larger the absolute value of the t-value, the smaller the p-value, provides greater the evidence against the null hypothesis. We see this in the case of 0.5 and 1.0 we should reject, and in the case of 2.0, we should acccept the null hypothesis.

Concluding Remarks

Our statistical analysis provides some concrete results, to what we observed after reviewing the inital boxplots in your exploratory stage. We see the averages computed in our anaylsis that there is an increase in tooth growth rate across all supplements. However, orange juice performs better at lower dosages to Guinea pigs.

We assumed given the data had no identifer per observation that we could not treat the observations as paired. Treated the observations of guinea tooth growth rates as independant.