Effect of Vitamin C on Tooth Growth in Guinea Pigs

Jim White
November 21, 2015

Overview:

This paper reviews the ToothGrowth dataset from the R datasets library. The name of the study from which the dataset was taken is "*The Effect of Vitamin C on Tooth Growth in Guinea Pigs*" 33(5): 491-504 by E.W. Crampton. The description of the dataset: "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)." From inside-R.org (http://www.inside-r.org/r-doc/datasets/ToothGrowth)

Basic exploratory analysis was completed and hypothesis testing was used to compare the tooth growth by the variables supp and dose.

The dataset has three variables: 1) *len*: length of teeth (odontoblasts (https://en.wikipedia.org/wiki/Odontoblast)), 2) *supp*: delivery methods [orange juice or ascorbic acid], and 3) *dose*: three dose levels [0.5, 1, and 2 mg]. (*note: references to VC and ascorbic acid will be interchangeable in this discussion.*)

Exploratory Analysis

After loading the dataset into a variable, exploratory analysis begins by running the functions str() and summary() to determine the number of observations, variables, and shape of the data.

Structure

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

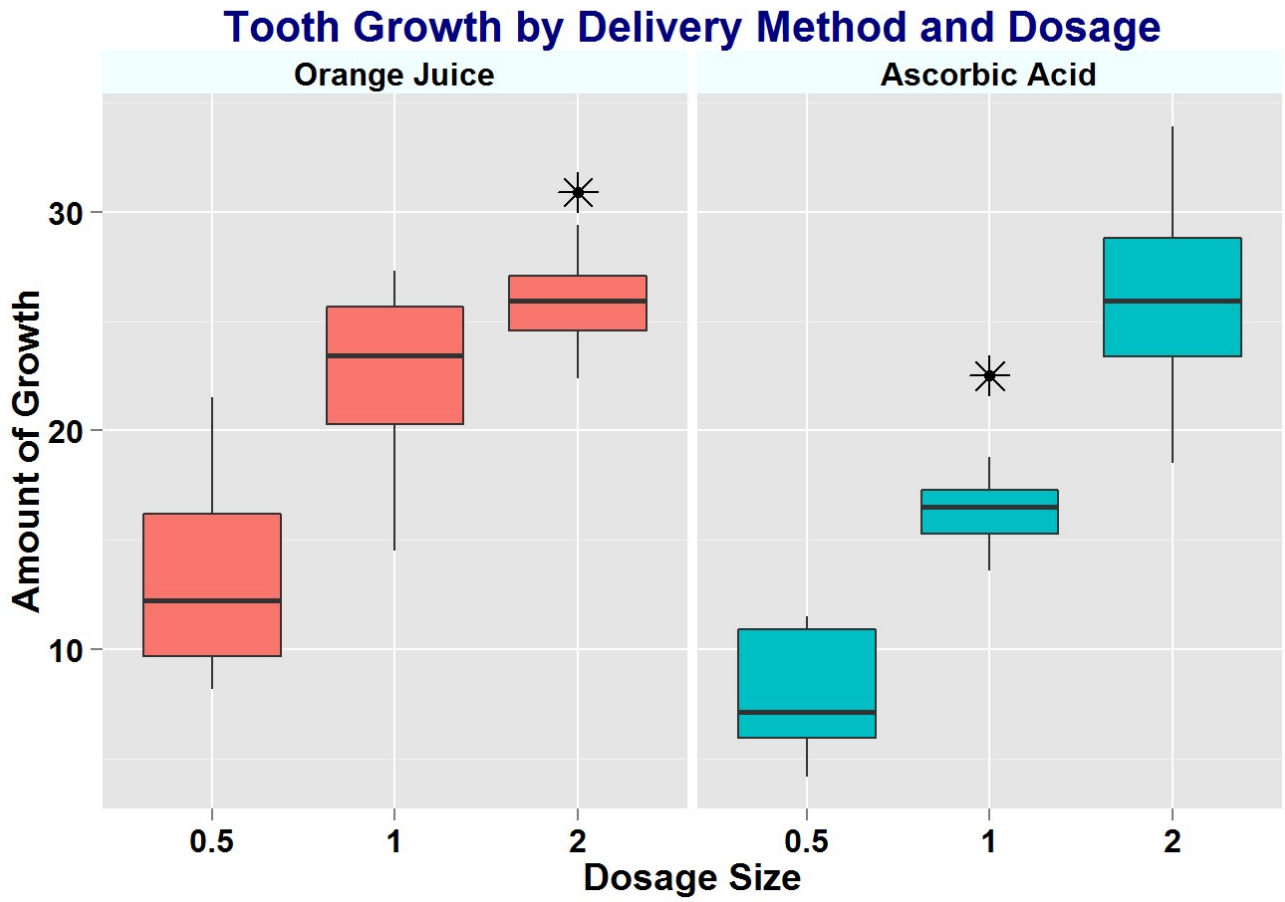
Summary

```
##       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
```

The dataset consists of 60 observations with three variables (as describe above). The len and dose variables are numeric and the delivery method variable (supp) is a factor variable. The range of the length (len) variable is 4.2 to 33.9, with an average of approximately 19. The two delivery methods each consists of 30 observations and each group of 30 observations consists of three groups of 10 each (for the doses of 0.5, 1.0, & 2.0).

Figure 1 (below) provides a boxplot view of the data with its averages and distributions across the two methods and three dose levels. A couple of outliers are noted by the star pattern dots (OJ - dose 2.0 and VC - dose 1.0)

Figure 1: Multiple Boxplots of Delivery Methods and Dosages



To conclude the exploratory analysis a table is included that summarizes the mean, standard deviation, and standard error for the groupings of the data by delivery method and dosage size.

```
## Source: local data frame [6 x 5]
## Groups: supp [?]
##
##     supp   dose  MEAN    SD    SE
##   (fctr) (fctr) (dbl) (dbl) (dbl)
## 1     OJ    0.5 13.23 4.460 1.410
## 2     OJ      1 22.70 3.911 1.237
## 3     OJ      2 26.06 2.655 0.840
## 4     VC    0.5  7.98 2.747 0.869
## 5     VC      1 16.77 2.515 0.795
## 6     VC      2 26.14 4.798 1.517
```

Based on this output, we might conclude that the $\overline{X}_{OJ}$ at the 2.0 dosage level may be equal to $\overline{X}_{VC}$. The other means at the 0.5 and 1.0 dosage levels to dnot appear to be equal. We verified via hypothesis testing.

Hypothesis Testing

The next step was to test the hypothesis that the means of the three pairs of groups (delivery method and dosage) are equal (or not).

The general form of the test is $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_a$

The t-test statistic is calculated as $T = \frac{\overline{X}_{OJ} - \overline{X}_{VC}}{\sqrt{\frac{s^2_{OJ}}{N_{OJ}} + \frac{s^2_{VC}}{N_{VC}}}}$

where:

1. $\overline{X}_{OJ}$ and $\overline{X}_{VC}$ are the means of the delivery methods
2. s^2_{OJ} and s^2_{VC} are the variances of the delivery methods
3. N_{OJ} and N_{VC} are the number of each of the delivery method samples

Assumptions

1. The student t-test is used due to the small sample sizes
2. For the calculation of the confidence interval α = 0.05 (at the 95% level)
3. These assumptions will apply to all dosage levels
4. "The assumption for the test is that both groups are sampled from normal distributions with equal variances."
   From Dept of Statistics - UC, Berkley (http://statistics.berkeley.edu/computing/r-t-tests)

Run the t-test for each level of the dosage (dose):

```
##       Dosage estimate1 estimate2  statistic      p.value parameter
## 1 Dosage 0.5     13.23      7.98  3.1697328 0.0053036613        18
## 2 Dosage 1.0     22.70     16.77  4.0327696 0.0007807262        18
## 3 Dosage 2.0     26.06     26.14 -0.0461361 0.9637097790        18
##    conf.low conf.high
## 1  1.770262  8.729738
## 2  2.840692  9.019308
## 3 -3.722999  3.562999
```

In the output, estimate1 is for OJ and estimate 2 is for VC (ascorbic acid). Based on the t-test and the associated p-values, we can reject H_0 for the 0.5 (p-value of 0.0053) & 1.0 (p-value of 0.0008) dosage levels. We cannot reject H_0 for the dosage level of 2.0 (p-value of 0.9637). Therefore, based on the means and the t-test the usage of orange juice for the lower dosage levels (0.5 & 1.0 mg) may provide greater tooth growth. The results for the 2.0 mg dosage level do not appear to be significantly different.

Appendix

Code Chunks

Load dataset and get structure

```
library(datasets) # load datasets package
str(ToothGrowth) # get structure of ToothGrowth dataset
```

Run the summary function

```
summary(ToothGrowth) # get summary of ToothGrowth dataset variables
```

Code chunk for Figure 1: Multiple Boxplots

```
library(ggplot2) #load plotting library
# function to change facet labels
my_labeller <- function(var, value){
    value <- as.character(value)
    if (var=="supp") {
        value[value=="OJ"] <- "Orange Juice"
        value[value=="VC"]   <- "Ascorbic Acid"
    }
    return(value)
}
# create plot
ToothGrowth$dose <- as.factor(ToothGrowth$dose) # convert dose to factor variable
g1 <- ggplot(ToothGrowth, aes(x = dose, y = len))
g1 <- g1 + geom_boxplot(outlier.colour = "black", outlier.shape = 8,
                        outlier.size = 5) + facet_wrap(~supp)
g1 <- g1 + geom_boxplot(aes(fill = factor(supp)))
g1 <- g1 + labs(title = "Tooth Growth by Delivery Method and Dosage")
g1 <- g1 + labs(x = "Dosage Size", y = "Amount of Growth")
g1 <- g1 + facet_grid(.~supp, labeller = my_labeller)
g1 <- g1 + theme(plot.title = element_text(size = 16, face = "bold", colour = "navy"),
                 axis.title.y = element_text(size = 14, face = "bold"),
                 axis.title.x = element_text(size = 14, face = "bold"),
                 axis.text.x = element_text(face = "bold", colour = "black", size = 12
),
                 axis.text.y = element_text(face = "bold", colour = "black", size = 12
),
                 strip.text = element_text(size = 12, face = "bold"),
                 strip.background = element_rect(fill = "azure1"),
                 strip.background = element_rect(linetype = 1),
                 legend.position = "none")
g1
```

Table of summaries of mean, standard deviation, and stand error by groupings

```
library(dplyr, warn.conflicts = FALSE) # load required library
grp <- group_by(ToothGrowth, supp, dose) # group data
# calucalte the mean, standard deviation, and standard error
sum1 <- summarize(grp, MEAN = mean(len), SD = round(sd(len), 3),
          SE = round(sd(len)/sqrt(length(len)), 3))
sum1
```

Subsetting data for t-tests

```
# prepared data for t-test
# convert to dose variable to numeric
ToothGrowth$dose <- as.numeric(as.character(ToothGrowth$dose))
# subset dose = 0.5
sub0.5 <- subset(ToothGrowth, dose == 0.5)
sub1.0 <- subset(ToothGrowth, dose == 1.0)
sub2.0 <- subset(ToothGrowth, dose == 2.0)
```

Running the t-tests

```
library(broom) # add package to tidy t-test output
# run t-test on each group and assign to variables
test0.5 <- tidy(t.test(len ~ supp, paired = FALSE, var.equal = TRUE, data = sub0.5))
test1.0 <- tidy(t.test(len ~ supp, paired = FALSE, var.equal = TRUE, data = sub1.0))
test2.0 <- tidy(t.test(len ~ supp, paired = FALSE, var.equal = TRUE, data = sub2.0))
# create tow labels for each t-test
row_labels <- c("Dosage 0.5", "Dosage 1.0", "Dosage 2.0")
row_labels <- as.data.frame(row_labels)
colnames(row_labels) <- c("Dosage")
# combine the outputs into a dataframe
results <- rbind(test0.5, test1.0, test2.0)
results <- cbind(row_labels, results)
results
```