Statistical Inference Course Project

Pradeep Babburi

10/31/2016

Part 2: Inferential Data Analysis

Introduction

This report outlines the basic statistical inferences on the analysis of ToothGrowth data. The data available in the R datasets package is "The Effect of Vitamin C on Tooth Growth in Guinea Pigs". The data contains 60 observations of 3 variables. The variables are

- *len* - The length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs.
- *supp* - supplement type used to deliver vitamin C. One of orange juice (coded as OJ) or ascorbic acid (coded as VC).
- *dose* - One of three dose levels of vitamin C [0.5, 1, 2] mg/day received by each animal.

For convenience the data is stored in a data frame "tg".

```
tg <- ToothGrowth               # ToothGrowth data
tg$dose <- factor(tg$dose)      # converting dose to a categorical variable
str(tg)                         # overview of data
```

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

Exploratory Analysis

In this report we will try to gain an understanding of the effects on length of odontoblasts with varying *supp* and *dose* which are read as factors for convenience in our analyis. The 60 responses of length can be segregated into 10 of each of the following groups.

```
Supplement OJ, dose 0.5 mg/day
Supplement OJ, dose 1 mg/day
Supplement OJ, dose 2 mg/day
Supplement VC, dose 0.5 mg/day
Supplement VC, dose 1 mg/day
Supplement VC, dose 2 mg/day
```

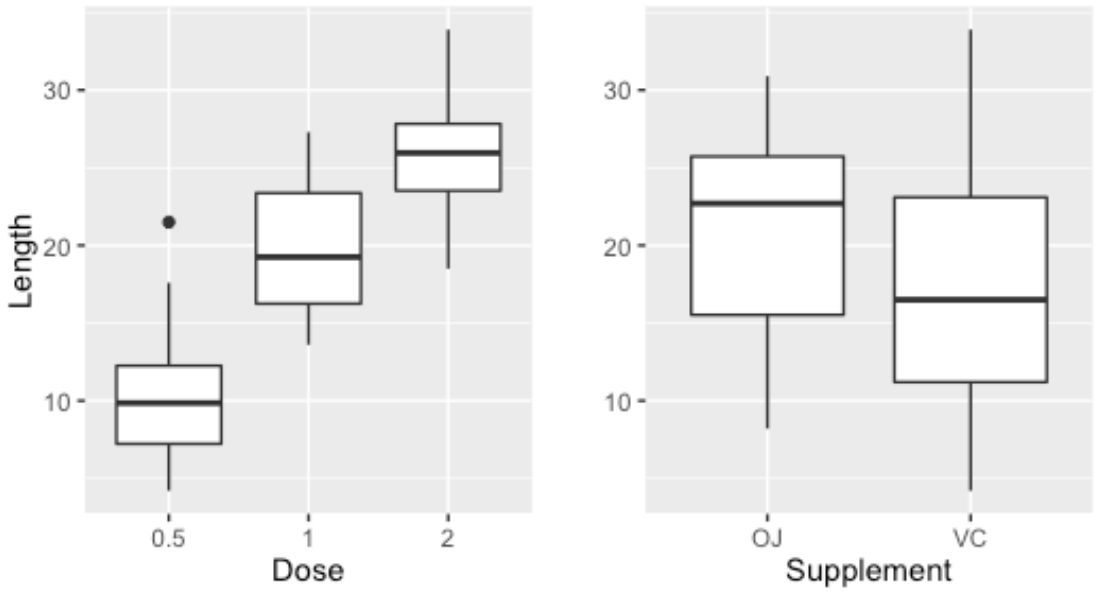
Six variables are created to store the lengths of all the above groups.

```
len.oj.d5 <- tg$len[tg$supp == "OJ"& tg$dose == 0.5]
len.oj.d1 <- tg$len[tg$supp == "OJ"& tg$dose == 1]
len.oj.d2 <- tg$len[tg$supp == "OJ"& tg$dose == 2]
```

```
len.vc.d5 <- tg$len[tg$supp == "VC"& tg$dose == 0.5]
len.vc.d1 <- tg$len[tg$supp == "VC"& tg$dose == 1]
len.vc.d2 <- tg$len[tg$supp == "VC"& tg$dose == 2]
cat("Mean length with dosage 0.5 mg/day =", mean(c(len.oj.d5, len.vc.d5)))

## Mean length with dosage 0.5 mg/day = 10.605

cat("Mean length with dosage 1 mg/day =", mean(c(len.oj.d1, len.vc.d1)))

## Mean length with dosage 1 mg/day = 19.735

cat("Mean length with dosage 2 mg/day =", mean(c(len.oj.d2, len.vc.d2)))

## Mean length with dosage 2 mg/day = 26.1

cat("Mean length using orange juice =", mean(c(len.oj.d5, len.oj.d1,
len.oj.d2)))

## Mean length using orange juice = 20.66333

cat("Mean length using ascorbic acid =", mean(c(len.vc.d5, len.vc.d1,
len.vc.d2)))

## Mean length using ascorbic acid = 16.96333
```

Let's create a quick boxplot to see how the length is affected by changes in dosage and delivery methods.

```
# boxplots of length by supp and dose
q1 <- qplot(dose, len, data = tg, geom = "boxplot", group = dose, xlab =
"Dose", ylab = "Length")
q2 <- qplot(supp, len, data = tg, geom = "boxplot", group = supp, xlab =
"Supplement", ylab = "")
grid.arrange(q1, q2, ncol = 2)
```



From the above results it can be said that both the mean and median lengths are greater with increased amounts of dosage and is also greater when the delivery method is Orange Juice (OJ) when compared to Ascorbic Acid (VC). Also the difference is significant between dosage amounts 0.5 and 1 mg/day.

Now, let's calculate some test statistics to see how likely is this observed difference in length with varying dosage and supplement type is true and that the observed statistic is not an extreme case.

Hypothesis Testing

Null Hypothesis - Let's consider a null hypothesis that there is no effect in tooth growth with change in dosage amounts and that the data we observed could be an error or just happened by chance.

Alternate Hypothesis - Let the alternate hypothesis be that the difference in mean lengths is indeed greater in subjects that received high amounts of vitamin C.

Assumptions - The assumptions for the hypothesis tests are that the calculated mean is a good estimate of the population mean and the difference in means is normally distributed.

```
# t-test between groups that received 0.5 and 1 mg/day dosage
t.test(c(len.oj.d1, len.vc.d1), c(len.oj.d5, len.vc.d5))

##
##  Welch Two Sample t-test
##
## data:  c(len.oj.d1, len.vc.d1) and c(len.oj.d5, len.vc.d5)
## 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:
##   6.276219 11.983781
## sample estimates:
## mean of x mean of y
##    19.735    10.605
```

The t-test results show that the difference in mean lengths is non zero and the t-value of 6.47 indicates that the difference between the two groups in comparision i.e., subjects that received 1 mg/day dosage and 0.5 mg/day is statistically significant in the positive direction. The results also show that the increase in average length falls between 6.27 and 11.98 with 95% confidence. In addition, considering a typical Type I error rate "α" value of 0.05, the low p-value is an indication that the null hypothesis could be false and that we can reject it under the stated assumptions.

Conclusion

To conclude the analysis from our preliminary investigation, the tooth growth data indicates that the length of odontoblasts in guinea pigs is higher when they are given higher amounts of vitamin C particularly when delivered using orange juice as the supplement. We have validated the observation by calculating the statistical significance

using t and p-values which indicates that the observed effect in our sample is inconsistent with the null hypothesis assuming that sample mean is representative of the population mean and the difference in means are in accord to the Central Limit Theorem.