Statistical Inference Part 2 : ToothGrowth Data Analysis

Bill Gourley

October 25, 2015

Overview

In this exercise, we will use the "ToothGrowth" data from the package "UsingR", perform a basic exploratory analysis of the data, and run a series of tests to determine whether supplements and dosage levels have an effect on tooth growth. The supplements are "Orange Juice" (OJ) and "Vitamin C" (VC), and there are 3 dosage levels, 0.5, 1.0 and 2.0 mg. The data contains 60 observations of the tooth growth length from a study performed on Guinea Pigs. The 60 subjects were split into 2 groups of 30 subjects each, one group receiving 'OJ' as a supplement, the other receiving 'VC'. Each group is independent so unpaired tests will be performed.

Analysis Requirements

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.

R Setup and Initialization

```
#packages
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.1
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.1
```

```
#Load data
data("ToothGrowth")
```

Exploratory Data Analysis

The structure of the raw dataset is shown below. The dose variable is set to be numeric but only has 3 unique values, therfore we will convert dose to be a factor with 3 levels (0.5, 1.0 and 2.0). The summary of the dataset after this change is also shown.

```
#data structure
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 ...
```

```
#get dose values
unique(ToothGrowth$dose)
```

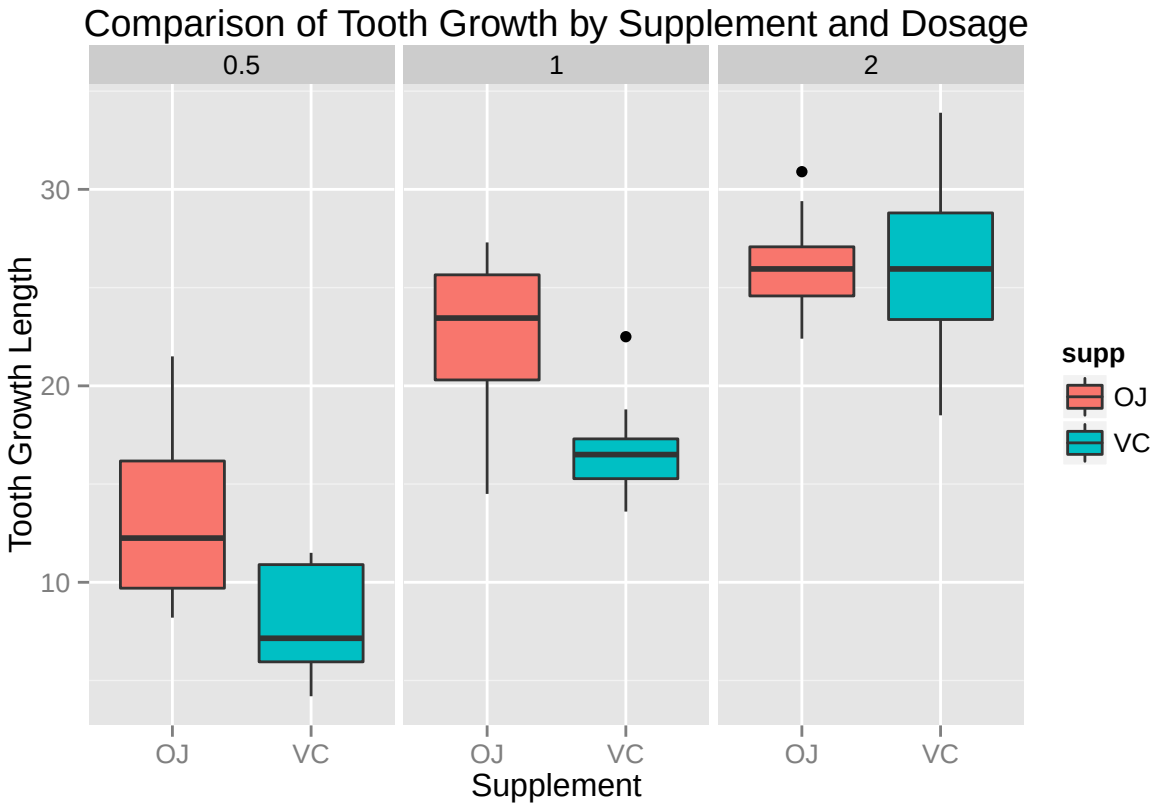
```
## [1] 0.5 1.0 2.0
```

```
#convert dose from num to factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

#data summary
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

The following figure shows a multiple boxplot which compares the Tooth Growth Length split by supplement



and dosage.

The following table summarizes the data.

```
#summary table of length
group_len <- group_by(ToothGrowth,dose,supp)
group_len.summary <- summarize(group_len,
          count = n(),
          mean = mean(len),
          median = median(len),
          std.dev = sd(len),
          variance = var(len))

group_len.summary
```

```
## Source: local data frame [6 x 7]
## Groups: dose
##
##   dose supp count  mean median  std.dev  variance
## 1  0.5   OJ    10 13.23  12.25 4.459709 19.889000
## 2  0.5   VC    10  7.98   7.15 2.746634  7.544000
## 3    1   OJ    10 22.70  23.45 3.910953 15.295556
## 4    1   VC    10 16.77  16.50 2.515309  6.326778
## 5    2   OJ    10 26.06  25.95 2.655058  7.049333
## 6    2   VC    10 26.14  25.95 4.797731 23.018222
```

From the table we can see that, for each of the dosage levels, 10 subjects received 'OJ' and 10 received 'VC' as a supplement. The means and medians for each dosage level are roughly equal, so we can surmise that the distributions are approximately normal. The variances, however, are markedly different so we will treat the variances as unequal when performing the tests. The sample sizes are small, so we will use t-tests rather than z-tests.

Comparison of Tooth Growth by Supplement and Dosage

For comparison purposes, we will carry out 3 sample t-tests, one for each dosage level. The data will be subset as follows:

```
#subset ToothGrowth by OJ and VC
OJ <- subset(ToothGrowth,supp == "OJ")
VC <- subset(ToothGrowth,supp == "VC")

#OJ subsets for dose levels 0.5, 1.0, 2.0
OJ_0.5 <- subset(OJ,dose == 0.5)
OJ_1.0 <- subset(OJ,dose == 1.0)
OJ_2.0 <- subset(OJ,dose == 2.0)

#VC subsets for dose levels 0.5, 1.0, 2.0
VC_0.5 <- subset(VC,dose == 0.5)
VC_1.0 <- subset(VC,dose == 1.0)
VC_2.0 <- subset(VC,dose == 2.0)
```

For each t-test, the null hypothesis H0 is 'the true difference in means is equal to 0'. The alternative hypothesis Ha is 'the true difference in means is not equal to 0'.

The three t-tests are :

1. OJ 0.5 dosage against VC 0.5 dosage

2. OJ 1.0 dosage against VC 1.0 dosage
3. OJ 2.0 dosage against VC 2.0 dosage

1. t-test supplement OJ 0.5mg dosage against supplement VC 0.5mg dosage

```
t.test(OJ_0.5$len,VC_0.5$len,paired = FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ_0.5$len and VC_0.5$len
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

The p-value for this test of 0.0064 indicates that we should reject the null hypothesis.

2. t-test supplement OJ 1.0mg dosage against supplement VC 1.0mg dosage

```
t.test(OJ_1.0$len,VC_1.0$len,paired = FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ_1.0$len and VC_1.0$len
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

The p-value for this test of 0.001 indicates that we should reject the null hypothesis.

3. t-test supplement OJ 2.0mg dosage against supplement VC 2.0mg dosage

```
t.test(OJ_2.0$len,VC_2.0$len,paired = FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  OJ_2.0$len and VC_2.0$len
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

The p-value for this test of 0.963 indicates that we should fail to reject the null hypothesis.

Assumptions

For the tests carried out above, we are assuming the following :

1. The groups are independent
2. The underlying distribution of each group is normal.
3. The sample set of guinea pigs selected is representative of the population.

Conclusions

Examining the results of the 3 t-tests carried out and the boxplot comparison carried out earlier, we can conclude the following :

1. Overall, administrating the supplement results in an increase in tooth growth.
2. At dosage levels of 0.5mg and 1mg, using Orange Juice as a supplement delivery system is more effective than by administrating Vitamin C alone.
3. However, there does not seem to be any marked difference in tooth growth between the two delivery systems when the dosage level is increased to 2mg.