Statistical Inference Project - Part 2

Bhaskar S

14th May 2016

Analysis of the ToothGrowth dataset in R

Synopsis

In this project, we analyze the **ToothGrowth** dataset in R to perform basic Exploratory Data Analysis and Hypothesis Tests to compare the tooth growth by supplements and dosage.

Exploratory Data Analysis

The **ToothGrowth** dataset in R contains 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).

For this analysis, we will be using the R packages `data.table` and `ggplot2`. The following code segment loads the desired libraries:

```
library(data.table)
library(ggplot2)
```

The following code segment loads the **ToothGrowth** dataset that is built into R. This automatically creates a data frame called `ToothGrowth`:

```
data('ToothGrowth')
```

We initialize a `data.table` variable called **TG** that will be referenced in this analysis. We use a `data.table` instead of the built-in data frame `ToothGrowth` for ease of use. Also, the `dose` data column is converted to a factor variable. Finally, we remove the data frame `ToothGrowth` once we initialize our `data.table` called **TG** for efficiency. The following code segment performs the desired initializations:

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
TG <- data.table(ToothGrowth)
rm(ToothGrowth)
```

As a first step in the exploratory data analysis, we want to compactly display the structure and contents of our `data.table` called **TG**. The following code displays the structure and contents of the `data.table` called **TG** using the R `str` function:

```
str(TG)
```

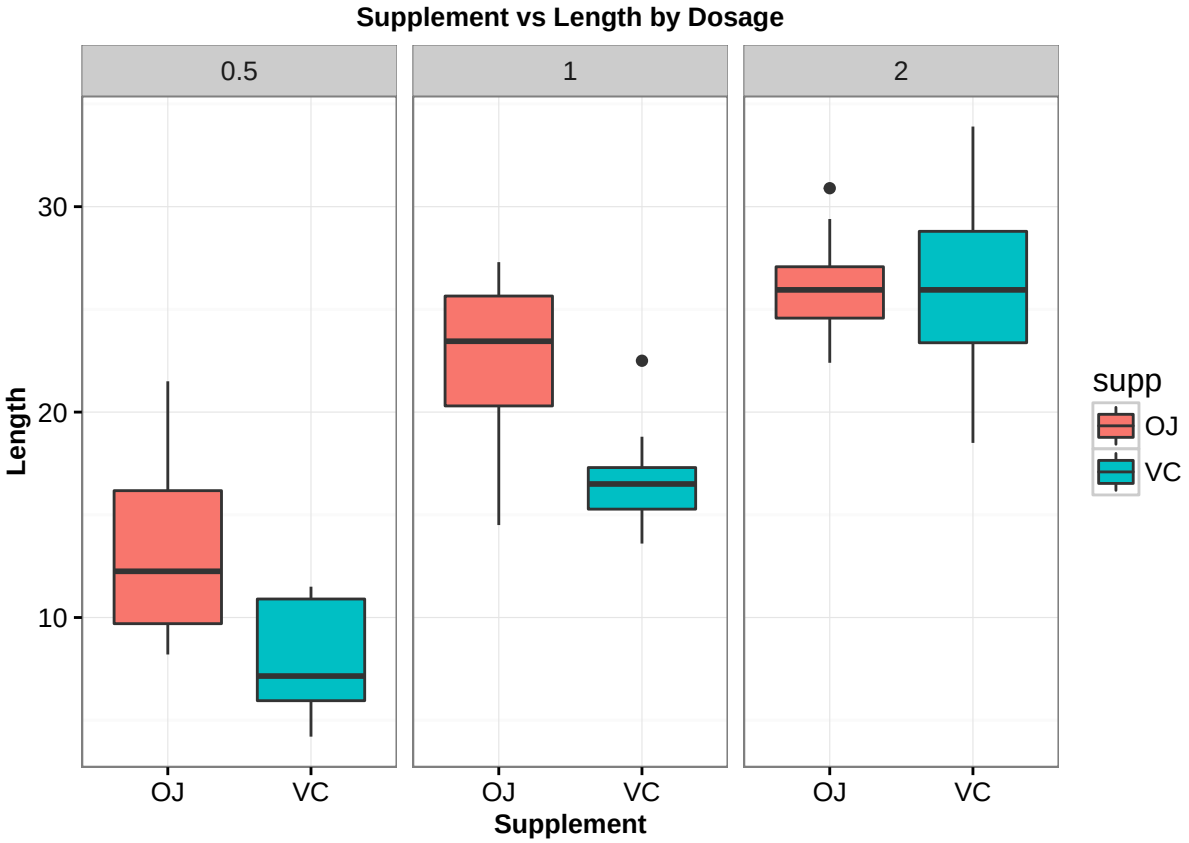
```
## Classes 'data.table' and 'data.frame':   60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

As we can see, the `data.table` called **TG** has **60** rows and **3** columns (`len`, `supp`, `dose`). The columns `supp` and `dose` are factor variables and represent the *supplement* and *dosage* respectively.

The next step in the exploratory data analysis is to display the data in the `data.table` called **TG** for visual analysis. We display the data in three sub-plots by `dose` with the `supp` along the x-axis and the `len` along the y-axis.

The following code plots the `data.table` called **TG** as a *boxplot*:

```
ggplot(TG, aes(x = supp, y = len, fill = supp)) +
    geom_boxplot(aes(fill = supp)) +
    labs(x = 'Supplement', y = 'Length') +
    ggtitle('Supplement vs Length by Dosage') +
    facet_grid(~dose) +
    theme_bw() +
    theme(plot.title = element_text(face = 'bold', size = 10),
          axis.title = element_text(face = 'bold', size = 10))
```



From the above plot, we can infer the following facts:

- For `dose` **0.5** and **1**, the *supplement* **OJ** seems to have a better growth result than the *supplement* **VC**
- For `dose` **2**, however, the *supplement* **VC** seems to have a better growth result than the *supplement* **OJ**

The next step in the exploratory data analysis is to display the `summary` statistics for the `len` data column in the `data.table` called **TG**.

For this, we initialize **6** subsets of the `data.table` called **TG** using the data columns `supp` and `dose`. The following code segment performs the desired initializations:

```
TG1.OJ <- TG[supp == 'OJ' & dose == 0.5]
TG1.VC <- TG[supp == 'VC' & dose == 0.5]
TG2.OJ <- TG[supp == 'OJ' & dose == 1]
TG2.VC <- TG[supp == 'VC' & dose == 1]
TG3.OJ <- TG[supp == 'OJ' & dose == 2]
TG3.VC <- TG[supp == 'VC' & dose == 2]
```

The following code displays the the `summary` statistics for the `len` data column from each of the **6** subsets we created above:

```
summary(TG1.OJ$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20    9.70   12.25   13.23   16.18   21.50
```

```
summary(TG1.VC$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20    5.95    7.15    7.98   10.90   11.50
```

```
summary(TG2.OJ$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.50   20.30   23.45   22.70   25.65   27.30
```

```
summary(TG2.VC$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   15.27   16.50   16.77   17.30   22.50
```

```
summary(TG3.OJ$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.40   24.58   25.95   26.06   27.08   30.90
```

```
summary(TG3.VC$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.50   23.38   25.95   26.14   28.80   33.90
```

From the above `summary` statistics, we can infer the following facts:

- For `dose` **0.5** and **1**, the *supplement* **OJ** seems to have a better growth result than the *supplement* **VC**
- For `dose` **2**, however, there does not seem to be any major advantage between *supplement* **OJ** or *supplement* **VC**

Hypothesis Testing

Hypothesis Testing on Supplement Types

Looking at the data, we have two supplement types (OJ and VC) and the sample size is equal to **30**. Also, we do not have any knowledge of the population variance. As a result, we will be conducting a **t** hypothesis test to find the **p-value** using the R `t.test` function.

We will test the *null* hypothesis that the mean tooth growth is equal between the supplements **OJ** and **VC** with a 95% **Confidence Interval**. Statistically, we are testing for **H0: mu1 = mu2**. The *alternate* hypothesis is **Ha: mu1 != mu2**. Since the *null* hypothesis is testing for equality, this is a **two-tail** test. If the **p-value** is < 0.05, we reject the *null* hypothesis.

For this, we initialize **2** subsets of the `data.table` called **TG** using the data column `supp`. The following code segment performs the desired initializations:

```
TG.OJ <- TG[supp == 'OJ']
TG.VC <- TG[supp == 'VC']
```

The following code performs a two-tailed **t** hypothesis test:

```
t.test(len ~ supp, data = TG, var.equal = FALSE, paired = FALSE, conf.level = 0.95)
```

```
##
##  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
```

From the above, we see the **p-value** is greater than **0.05** and hence we fail to reject the *null* hypothesis and conclude with 95% confidence level that there is no major difference between the mean growths from the two supplements **OJ** and **VC**.

Hypothesis Testing on Dosage Levels

Looking at the data, we have three dosage levels (0.5, 1, 2) and the sample size is equal to **20** for each level. Also, we do not have any knowledge of the population variance. As a result, we will be conducting a **t** hypothesis test to find the **p-value** using the R `t.test` function.

We will test the *null* hypothesis for each of the dosage levels that the mean tooth growth is equal between the supplements **OJ** and **VC** with a 95% **Confidence Interval**. Statistically, we are testing for **H0: mu1 = mu2**. The *alternate* hypothesis is **Ha: mu1 != mu2**. Since the *null* hypothesis is testing for equality, this is a **two-tail** test. If the **p-value** is < 0.05, we reject the *null* hypothesis.

For this, we initialize **3** subsets of the `data.table` called **TG** using the data column `dose`. The following code segment performs the desired initializations:

```r
TG1 <- TG[dose == as.factor(0.5)]
TG2 <- TG[dose == as.factor(1)]
TG3 <- TG[dose == as.factor(2)]
```

The following code performs a two-tailed **t** hypothesis test for dosage level **0.5**:

```r
t.test(len ~ supp, data = TG1, var.equal = FALSE, paired = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##            13.23             7.98
```

From the above, we see the **p-value** of **0.006359** is less than **0.05** and hence we reject the *null* hypothesis and conclude with 95% confidence level that there is difference between the mean growths from the two supplements **OJ** and **VC** for dosage **0.5**.

The following code performs a two-tailed **t** hypothesis test for dosage level **1**:

```r
t.test(len ~ supp, data = TG2, var.equal = FALSE, paired = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## 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 in group OJ mean in group VC
##            22.70            16.77
```

From the above, we see the **p-value** of **0.001038** is less than **0.05** and hence we reject the *null* hypothesis and conclude with 95% confidence level that there is difference between the mean growths from the two supplements **OJ** and **VC** for dosage **1**.

Finally, the following code performs a two-tailed **t** hypothesis test for dosage level **2**:

```r
t.test(len ~ supp, data = TG3, var.equal = FALSE, paired = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
```

```
## 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 in group OJ mean in group VC
##            26.06            26.14
```

From the above, we see the **p-value** of **0.9639** is greater than **0.05** and hence we fail to reject the *null* hypothesis and conclude with 95% confidence level that there is no difference between the mean growths from the two supplements **OJ** and **VC** for dosage **2**.

Summary

We conclude the following facts from our analysis:

- The supplements on their own do not seem to have an influence on the tooth growth
- The dosage levels seems to have an influence on the tooth growth