Analysis of the ToothGrowth Dataset

Ojo Oluwasegun

Sunday, January 25, 2015

Overview

This project report attempts to carry out some basic data analysis of the ToothGrowth dataset contained in the 'datasets' package in R [1]. The ToothGrowth dataset contains response of the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid) [2].

Summary and Exploratory Analysis

In this section, we attempt to summarize the dataset and carry out some exploratory analysis. We first load the dataset.

```
library(datasets)
data(ToothGrowth)
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
tail(ToothGrowth)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

Next we view the structure of the dataset and some summary statistics.

```
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 ...
```

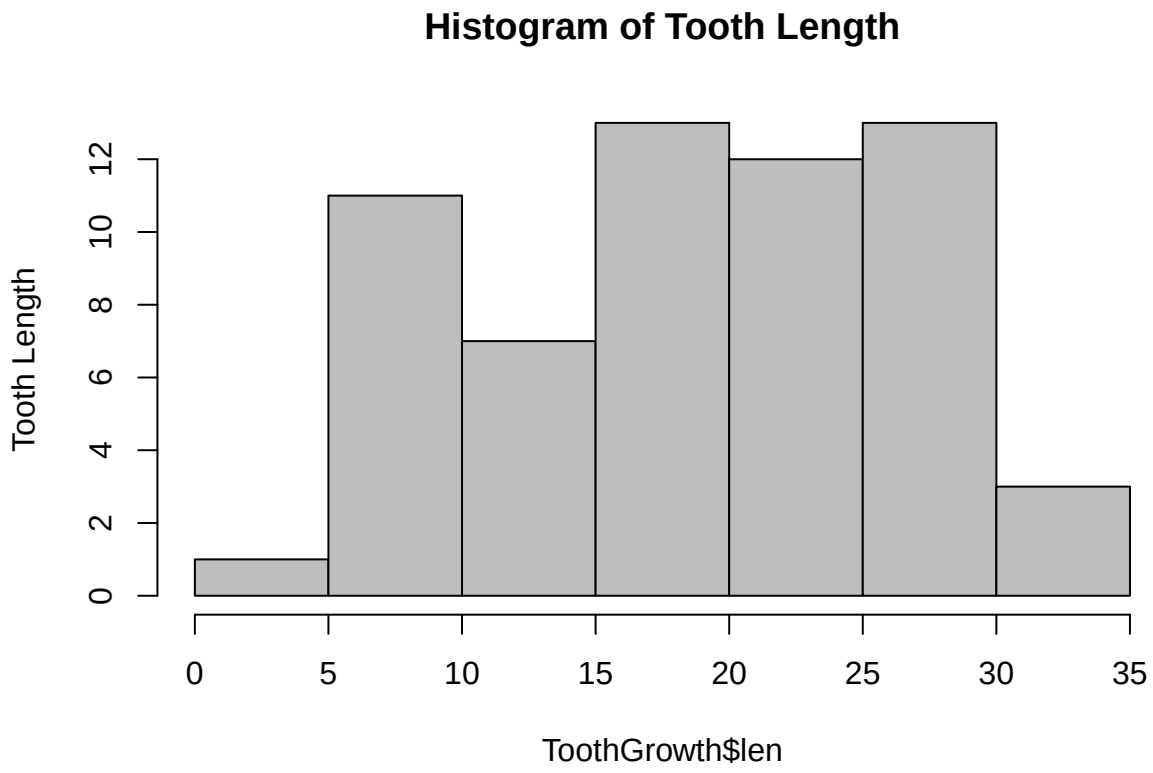
```
summary(ToothGrowth)
```

```
##       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
```

From the output above, the ToothGrowth dataset contains 3 variable viz 'len' which is the length of the teeth, 'supp' which is a factor variable indicating the supplement type and 'dose' indicating the dose in milligrams. Some basic summary statistics of the dataset is also shown.

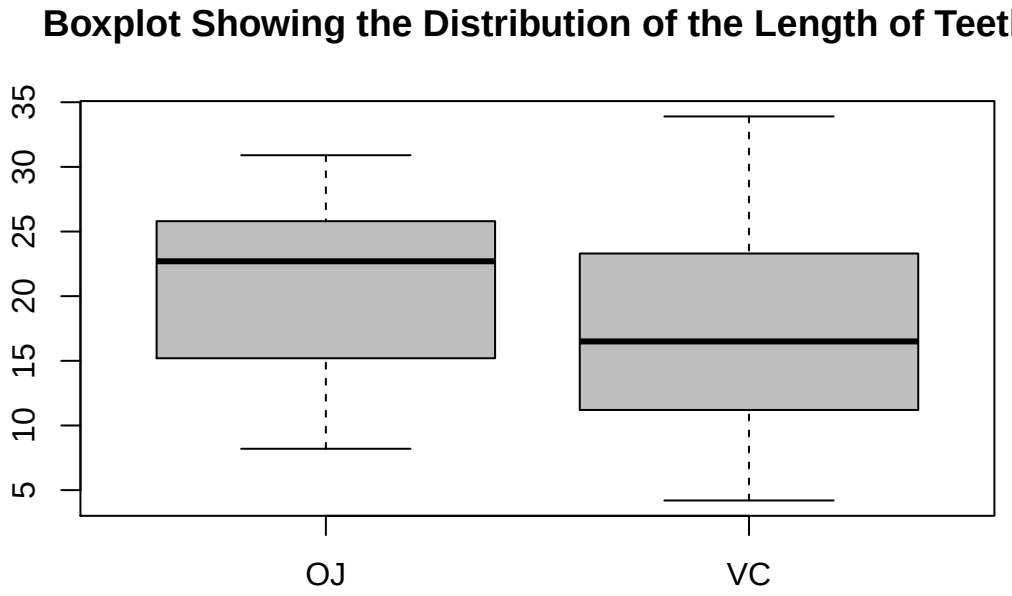
Next we try to explore the dataset using some plots. Shown below is a histogram showing the distribution of the length of the teeth.

```
hist(ToothGrowth$len, col = 'grey', main= 'Histogram of Tooth Length',
     ylab= 'Tooth Length')
```

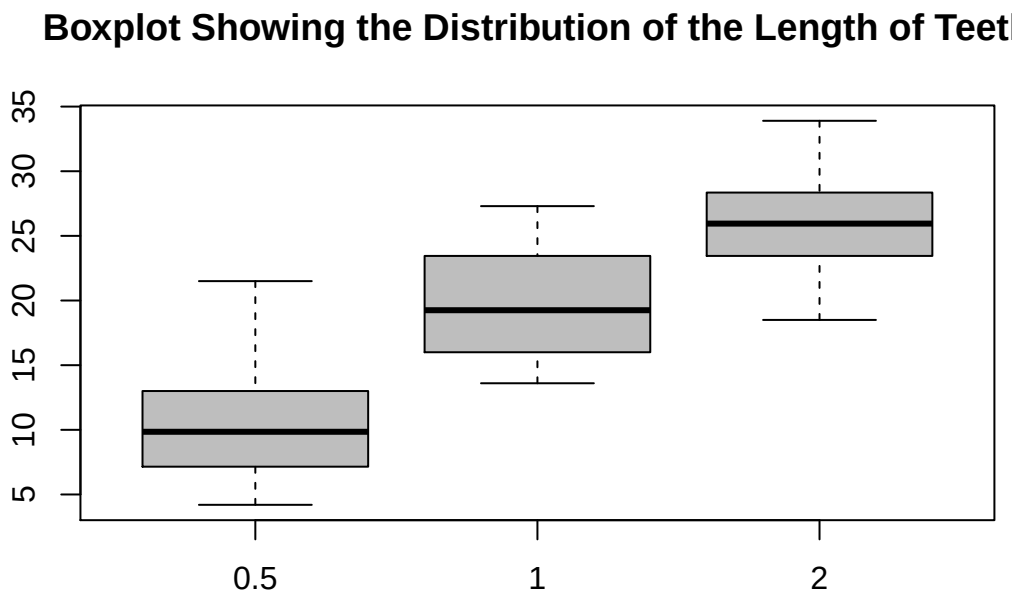


We can also plot a boxplot to show the distribution of the len variable with respect to the supp and dose variables.

```r
boxplot(ToothGrowth$len~ToothGrowth$supp, col = 'grey',
        main = 'Boxplot Showing the Distribution of the Length of Teeth')
```




```r
boxplot(ToothGrowth$len~ToothGrowth$dose, col = 'grey',
        main = 'Boxplot Showing the Distribution of the Length of Teeth')
```



The two figures above seem to show a lot of interesting leads we can explore further. First, there seem to be an increase in tooth growth with higher dosage of Vitamin C (from the second figure) and orange juice seem to be a better delivery method than ascorbic acid (from the first figure). We can explore these claims further using former tests of hypothesis. Specifically, we will test if there is a significant difference in ToothGrowth across the delivery method. We'll use paired sample T test for this because the same guinea pigs(subjects) were used across the two delivery methods (factors).

```
t.test(ToothGrowth$len~ToothGrowth$supp, paired = T, var.equal = T)
```

```
##
##  Paired t-test
##
## data:  ToothGrowth$len by ToothGrowth$supp
## t = 3.3026, df = 29, p-value = 0.00255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.408659 5.991341
## sample estimates:
## mean of the differences
##                     3.7
```

The above block of code performs a paired sample T test to test the null hypothesis that the true difference in mean of tooth lengths is zero (i.e there is no significant difference in tooth length across the levels of delivery methods) versus the alternate hypothesis that the true difference in means is not equal to zero (i.e, delivery methods have an effect on tooth growth). The reported p-value 0.00255 (which is the probability of having a t statistics as high as 3.3026 if the different types of delivery methods do not differ in tooth length) is less than 0.05, hence we reject the null hypothesis and conclude that there is indeed a significant difference in tooth length across the delivery methods. The reported confidence interval with lower bound 1.41 and upper bound 5.99 also agree with the conclusion of the test. This is because the 95% confidence interval does not

include the value 0. This means that if we were to repeatedly take samples, about 95% of the interval would contain the mean difference 3.7 (but not zero).

Assumptions.

1. Depedence: To use paired sample t test, the observations must be paired or dependent. This assumption is not violated as the same set of guinea pigs were used across the two delivery methods.

2. Equal Variance: The two delivery methods seem to have the same variability. This is confirmed by the width of the boxes in the first boxplot of the Exploratory Section. Since the two boxes seem to have approximately the same with, we can assume equal variance for the two delivery methods.

3. Normality Assumption: The histogram plot of the length variable does not seem to follow the bell shape of the Normal distribution. Also our sample size(30) for the test is not sufficiently large. This should be a source of concern but since we used t test, I think the conclusions of the test should be valid.

References

1. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
2. C. I. Bliss (1952). The Statistics of Bioassay. Academic Press.