ToothGrowth Data Analysis

Taedong Yun

Overview

In this article, we will analyze the ToothGrowth data in the R datasets package. We will perform basic exploratory data analyses by providing a summary of the data and elementary statistical inference.

Summary of Data

The ToothGrowth dataset contains the result of an experiment on Guinea Pigs about the correlation between Vitamin C dosage and the growth of tooth. Let us take a first look of the dataset.

```
library(datasets)
str(ToothGrowth)
```

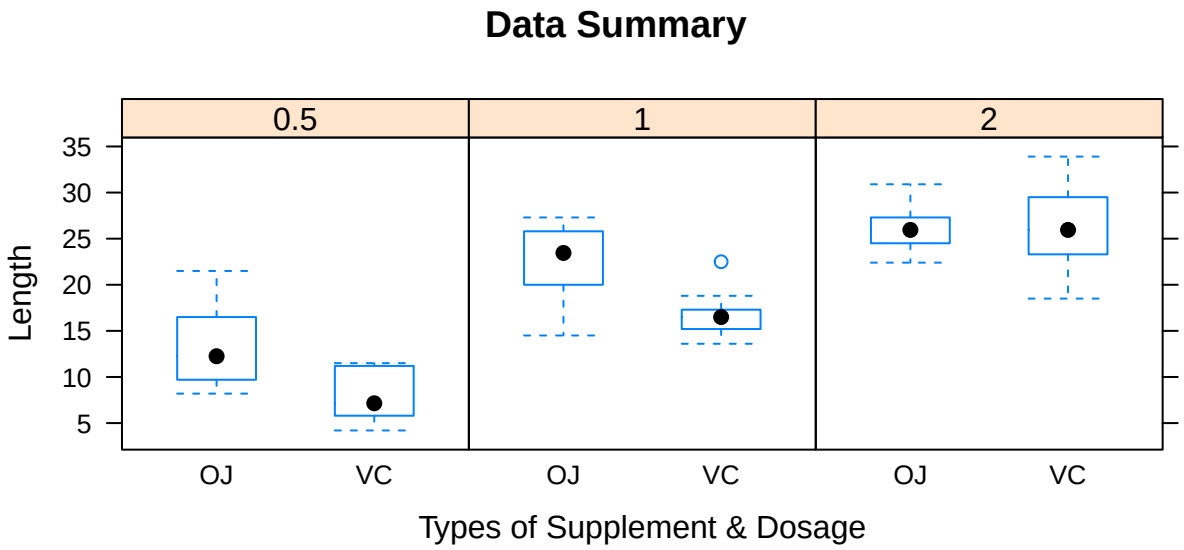
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

The dataset has three variables, `len`, `supp`, and `dose`. `len` is the response which is the length of a tooth, `supp` is a delivery type of Vitamin C, swhich is either "VC" (ascorbic acid) or "OJ" (orange juice), and `dose` is the amount of dosage in milligrams. Note that `dose` variable take a value in 0.5, 1, or 2 so we can make it a factor variable. This dataset contains 60 observations. More information about this dataset can be found at https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html. Here is a basic summary of each variables in the dataset.

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
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
```

Note that `supp` and `dose` factor variables are evenly distributed in their possible values. The following figure is a box plot of the dataset.



Inference

From the boxplot in the previous section, we study the following the problems.

1. Do two different types of supplement have different effects on the growth of a tooth?
2. Does more dosage means more growth of a tooth?

We will address these two questions with simple T tests.

Length vs Supplement Type

Given a null hypothesis that two different types of supplement have the *same* effect on the length of a tooth, we perform a two-sided T test against the variables `len` and `supp`. We assume unequal variance.

```
t.test(len ~ supp, data = ToothGrowth, paired = FALSE, var.equal = FALSE)
```

```
##
##  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 result of the test we can see that 95% confidence interval contains the value 0 (in other words the p-value is greater than 0.05). Hence, we **cannot** reject the null hypothesis.

Length vs Dosage

For our second quesion, we will compare the differences of tooth length for dosage 0.5 and 1mg, and for 1 and 2mg. First we compare the dosage of 0.5 and 1mg. Our alternative hypothesis here is that the tooth length for 0.5mg dosage is *less* than the tooth length for 1mg dosage.

```
suppressWarnings(library(dplyr))
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
TG1 <- filter(ToothGrowth, dose == 0.5 | dose == 1)
t.test(len ~ dose, data = TG1, paired = FALSE, var.equal = FALSE, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

Since the p-value is much less than 0.01, we can **reject** the null hypothesis and conclude that the tooth length for 0.5mg dosage is indeed less than the tooth length for 1mg dosage with 99% confidence.

Now let us compare 1mg and 2mg dosage with the same method.

```
TG2 <- filter(ToothGrowth, dose == 1 | dose == 2)
t.test(len ~ dose, data = TG2, paired = FALSE, var.equal = FALSE, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

Again, the p-value is less than 0.01 so we **reject** the null hypothesis and conclude that more dosage means more growth with 99% confidence.

Conclusion

In this report, we have examined the ToothGrowth dataset in R, studied basic summaries of the data, and performed a statistical inference on the correlations between the `len` variable and the other two variables.

Our first inference question was about the correlation between the tooth growth and the type of supplements, and even though the mean in "OJ" group was greater than the mean in "VC", we could **not** reject the null hypothesis that the mean of growth is equal for two different types of supplement with 95% confidence.

On the other hand, in our second statistical test we have concluded that more dosage of Vitamin C implies more tooth growth with more than 99% confidence.

Note that for the above T tests our assumption is that we took random samples from the population of Guinea pigs and the distribution of the tooth length is approximately iid normal, or at least roughly symmetric and mound shaped. We assumed unequal variances for all the tests we performed.

Appendix