Part 2 - Basic Inferential Data Analysis

Prafful Agrawal

July 7, 2020

Overview

In this project, we undertake a basic *inferential* data analysis to investigate the `ToothGrowth` data in R's `datasets` package. The `ToothGrowth` dataset consists of data on the effect of Vitamin C on the tooth growth in Guinea Pigs. It has details of **60** guinea pigs where each animal received one of the *three* dose levels of Vitamin C (**0.5**, **1** and **2** mg/day) by one of the *two* delivery methods, *orange juice* or *ascorbic acid* (coded as **OJ** and **VC** respectively).

Data Preprocessing

Load the `Toothgrowth` data and look at the first few rows.

```
##    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
```

Look at the structure of the dataset.

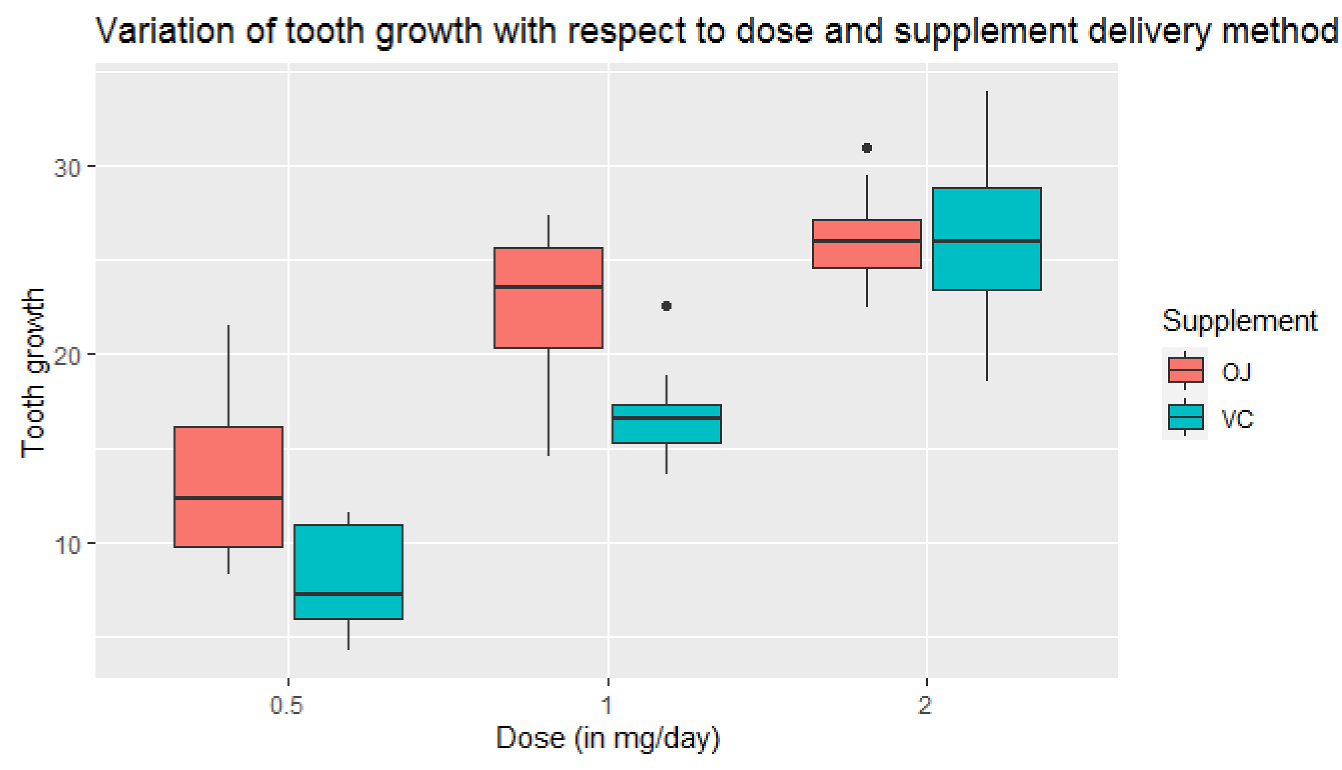
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Convert `dose` to a **factor** variable.

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

Basic Exploratory Data Analysis

Construct a plot showing the variation of tooth growth with respect to the *dose* and *supplement delivery method*.



From the above graph, we can observe that the tooth growth increases as the *dose* increases. Also, for low and medium *doses* (**0.5** and **1** mg/day), the *supplement delivery method* **OJ** shows a higher tooth growth than the *supplement delivery method* **VC**, whereas the tooth growth is nearly the same for high *dose* (**2** mg/day) irrespective of the *supplement delivery method*.

Let us compute the means of the tooth growth length with respect to the *dose* and *supplement delivery method*.

```
## # A tibble: 2 x 4
## # Groups:   supp [2]
##   supp  `dose 0.5` `dose 1` `dose 2`
##   <fct>      <dbl>    <dbl>    <dbl>
## 1 OJ         13.2      22.7     26.1
## 2 VC          7.98     16.8     26.1
```

This is in accordance with our previous observations.

Statistical Inference Analysis

Let us now test if our previous observations are statistically significant.

Since, the number of observations *n* are small, we will perform **One-sided T-test** to analyse the effect of *dose* and *supplement delivery method* on the tooth growth.

We will assume that the samples are **representative** of the population and the variance of the different samples are **different**. Also, the samples are **independent** from each other.

Variation of tooth growth with respect to dose

Our *Null* hypothesis states that the difference in the means of tooth growth with respect to *dose* is **Equal to Zero** while the *Alternative* hypothesis states that it is **Greater than Zero**, i.e.

$$H_0 : \Delta^1_{Mean} = 0$$

$$H_a : \Delta^1_{Mean} > 0$$

We will consider only the *high* and *low doses* (**2** and **0.5** mg/day respectively).

The 95% confidence interval and the p-value are:

```
## [1] "Confidence interval: 13.12 - Inf"
```

```
## [1] "p-value: 3.6e-10"
```

Thus, the *Null* hypothesis can be **rejected**, i.e. there is an increase in tooth growth with respect to increasing *dose*.

Variation of tooth growth with respect to supplement delivery method

Our *Null* hypothesis states that the difference in the means of tooth growth with respect to *supplement delivery method* is **Equal to Zero** while the *Alternative* hypothesis states that it is **Greater than Zero**, i.e.

$$H_0 : \Delta^2_{Mean} = 0$$

$$H_a : \Delta^2_{Mean} > 0$$

The 95% confidence interval and the p-value are:

```
## [1] "Confidence interval: 1.8 - Inf"
```

```
## [1] "p-value: 0.00127"
```

Thus, the *Null* hypothesis can be **rejected**, i.e. there is an increase in tooth growth with the use of *supplement delivery method* **OJ** over *supplement delivery method* **VC**.

Conclusion

1. Increasing the *dose* of the Vitamin C results in an increase of tooth growth.
2. Using **OJ** as *supplement delivery method* results in a higher tooth growth over using **VC** as *supplement delivery method*.

Possible shortcomings in the analysis

Since, there are two variables affecting the tooth growth simultaneously, the analysis will perform better for **paired** data.

Appendix

Load the packages.

```
library(ggplot2)
library(dplyr)
library(tidyr)
```

Load the data.

```
library(datasets)
data("ToothGrowth")

# Look at first few rows of data
head(ToothGrowth)
```

Structure of the dataset.

```
str(ToothGrowth)
```

Converting `dose` to a **factor** variable.

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

Code for Exploratory plot.

```
g <- ToothGrowth %>%
      ggplot(aes(x = dose, y = len, fill = supp)) +
        geom_boxplot() +
        labs(title = "Variation of tooth growth with respect to dose and supplement delivery method",
             x = "Dose (in mg/day)",
             y = "Tooth growth",
             fill = "Supplement")

print(g)
```

Code for Pivot tables of the means.

```
Tooth <- ToothGrowth %>%
          group_by(supp, dose) %>%
          summarize(mean = mean(len)) %>%
          pivot_wider(names_from = dose, values_from = mean,
                      names_prefix = "dose ")

print(Tooth)
```

Code for T-test 01.

```
g1 <- ToothGrowth$len[ToothGrowth$dose == '0.5']
g2 <- ToothGrowth$len[ToothGrowth$dose == '2']

t1 <- t.test(g2 - g1, paired = FALSE, var.equal = FALSE, alternative = "greater")
```

```
print(paste("Confidence interval:",
            paste(round(t1$conf.int, 2),
                  collapse = " - ")))
```

```
print(paste0("p-value: ", signif(t1$p.value, 3)))
```

Code for T-test 02.

```
g3 <- ToothGrowth$len[ToothGrowth$supp == 'VC']
g4 <- ToothGrowth$len[ToothGrowth$supp == 'OJ']

t2 <- t.test(g4 - g3, paired = FALSE, var.equal = FALSE, alternative = "greater")
```

```
print(paste("Confidence interval:",
            paste(round(t2$conf.int, 2),
                  collapse = " - ")))
```

```
print(paste0("p-value: ", signif(t2$p.value, 3)))
```