Basic Inferential Data Analysis of ToothGrowth Dataset

Jeevithiesh Duggani

07/05/2020

Contents

Abstract

This project analyzes the `ToothGrowth` data in the R datasets package.

The `aims` of the project are listed below :

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**.
4. State the `conclusions` and the `assumptions` needed for the **conclusions**.

Import the necessary packages and datasets

The following packages are to be imported.

- dataset
- ggplot2

```
    library(datasets)
    library(ggplot2)
```

Import the following datasets.

- ToothGrowth

```
    data("ToothGrowth")
```

Rename the values `OJ` as `Orange Juice` and `VC` as `Ascorbic Acid`.

```
    levels(ToothGrowth$supp) <- c("Orange Juice", "Ascorbic Acid")
```

Run Basic Exploratory Data Analyses

Check the `structure` of the dataset `ToothGrowth`.

```
    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 "Orange Juice",..: 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 ...
```

supp consists of **2** `unique` values. Identify number of unique values in `len` and `dose`.

```
    list(uniqLen  = length(unique(ToothGrowth$len)),
         uniqDose = length(unique(ToothGrowth$dose)))
```

```
## $uniqLen
## [1] 43
##
## $uniqDose
## [1] 3
```

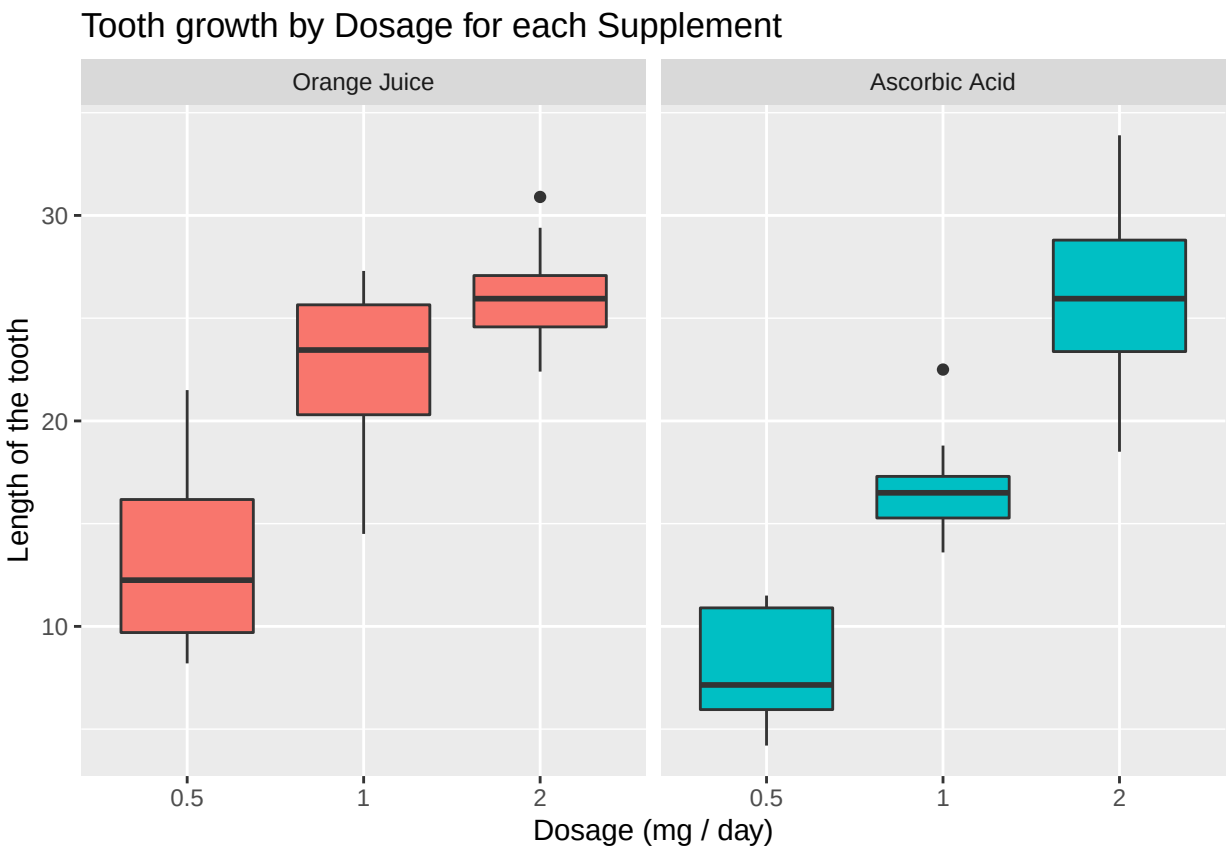
Convert dose into a factor as it consists of only **3** unique values.

```
    ToothGrowth$dose <- as.factor(ToothGrowth$dose)
    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 "Orange Juice",..: 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 ...
```

Plot a boxplot to visualize the `length of the teeth` based on the `supplement` and `dosage`.

```r
graph <- ggplot(ToothGrowth, aes(dose,len))
graph <- graph + facet_grid(.~supp)
graph <- graph + geom_boxplot(aes(fill = supp), show.legend = FALSE)
graph <- graph + labs(title = "Tooth growth by Dosage for each Supplement",
              y = "Length of the tooth", x = "Dosage (mg / day)")
graph
```



Observations

- It is observed that **both** `supplements` result in an **increase** in `growth` as the `dosage` is **increased**.
- Both supplements result in the **same** `length` when the `dosage` is at `2 mg / day`.
- `Orange Juice` is **more** `effective` at the **lower** `doses`.

Comparision of length of growth by supplement and dose

Use `Confidence Intervals` and `Hypothesis Testing` for the comparision.

Hypothesis 1

`Orange Juice` and `Ascorbic Acid` result in the same growth.

```r
    test1 <- t.test(len ~ supp, data = ToothGrowth)
    conf1 <- test1$conf.int
    pval1 <- test1$p.value
    test1
```

```
##
##  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 Orange Juice mean in group Ascorbic Acid
##                    20.66333                    16.96333
```

- The confidence interval is -0.1710156, 7.5710156 which contains the value `0`.
- The P value is 0.0606345 which is **greater** than `0.05`.

Therefore the null hypothesis cannot be rejected

Hypothesis 2

`Orange Juice` and `Ascorbic Acid` result in the same growth at a dosage of 0.5 mg / day.

```r
    test2 <- t.test(len ~ supp, data = ToothGrowth[ToothGrowth$dose == 0.5,])
    conf2 <- test2$conf.int
    pval2 <- test2$p.value
    test2
```

```
##
##  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 Orange Juice mean in group Ascorbic Acid
##                       13.23                       7.98
```

- The confidence interval is -0.1710156, 7.5710156 which does not contains the value `0`.
- The P value is 0.0606345 which is **much lesser** than `0.05`.

Therefore the null hypothesis can be rejected. The alternative hypothesis "The supplement Orange Juice results in more tooth growth at a dosage of 0.5 mg / day" is accepted instead.

Hypothesis 3

`Orange Juice` and `Ascorbic Acid` result in the same growth at a dosage of 1 mg / day.

```
    test3 <- t.test(len ~ supp, data = ToothGrowth[ToothGrowth$dose == 1,])
    conf3 <- test3$conf.int
    pval3 <- test3$p.value
    test3
```

```
##
##  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 Orange Juice mean in group Ascorbic Acid
##                       22.70                       16.77
```

- The confidence interval is 2.8021482, 9.0578518 which does not contains the value `0`.
- The P value is 0.0010384 which is **much smaller** than `0.05`.

Therefore the null hypothesis can be rejected. The alternative hypothesis "The supplement Orange Juice results in more tooth growth at a dosage of 1 mg / day" is accepted instead.

Hypothesis 4

`Orange Juice` and `Ascorbic Acid` result in the same growth at a dosage of 2 mg / day.

```
    test4 <- t.test(len ~ supp, data = ToothGrowth[ToothGrowth$dose == 2,])
    conf4 <- test4$conf.int
    pval4 <- test4$p.value
    test4
```

```
##
##  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 Orange Juice mean in group Ascorbic Acid
##                       26.06                       26.14
```

- The confidence interval is -3.7980705, 3.6380705 which contains the value `0`.
- The P value is 0.9638516 which is **much greater** than `0.05`.

Therefore the null hypothesis cannot be rejected

Assumptions

- `Normal Disibution` of the lengths of the tooth.
- `No other measured factors` affects the length of the tooth.
- For the `t tests` the variances of the two groups are different from one another.

Conclusions

- `Orange Juice` results in more growth at the **lower** doses of `0.5 and 1 mg / day`.
- `Orange Juice` and `Ascorbic Acid` results in the same growth at the **highest** doses of `2 mg / day`.
- For the overall picture, We cannot conclude that `Orange Juice` is more effective than `Ascorbic Acid`.

`**End**`