Data Analysis the ToothGrow dataset

2023-03-03

Overview

We load the 'ToothGrowth'data set from the datasets package in R, use the same to first perform an exploratory analysis and then use confidence intervals and hypothesis tests to compare tooth growth by supp and dose.

Loading DATA

First we load the ToothGrowth dataset and use str and summary functions to gather basic summary about the dataset.

```
data("ToothGrowth")
dt <- ToothGrowth
```

Basic expolatory data analysis

```
str(dt)
```

```
## '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(dt)
```

```
##       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
```

```
unique(dt$dose)
```

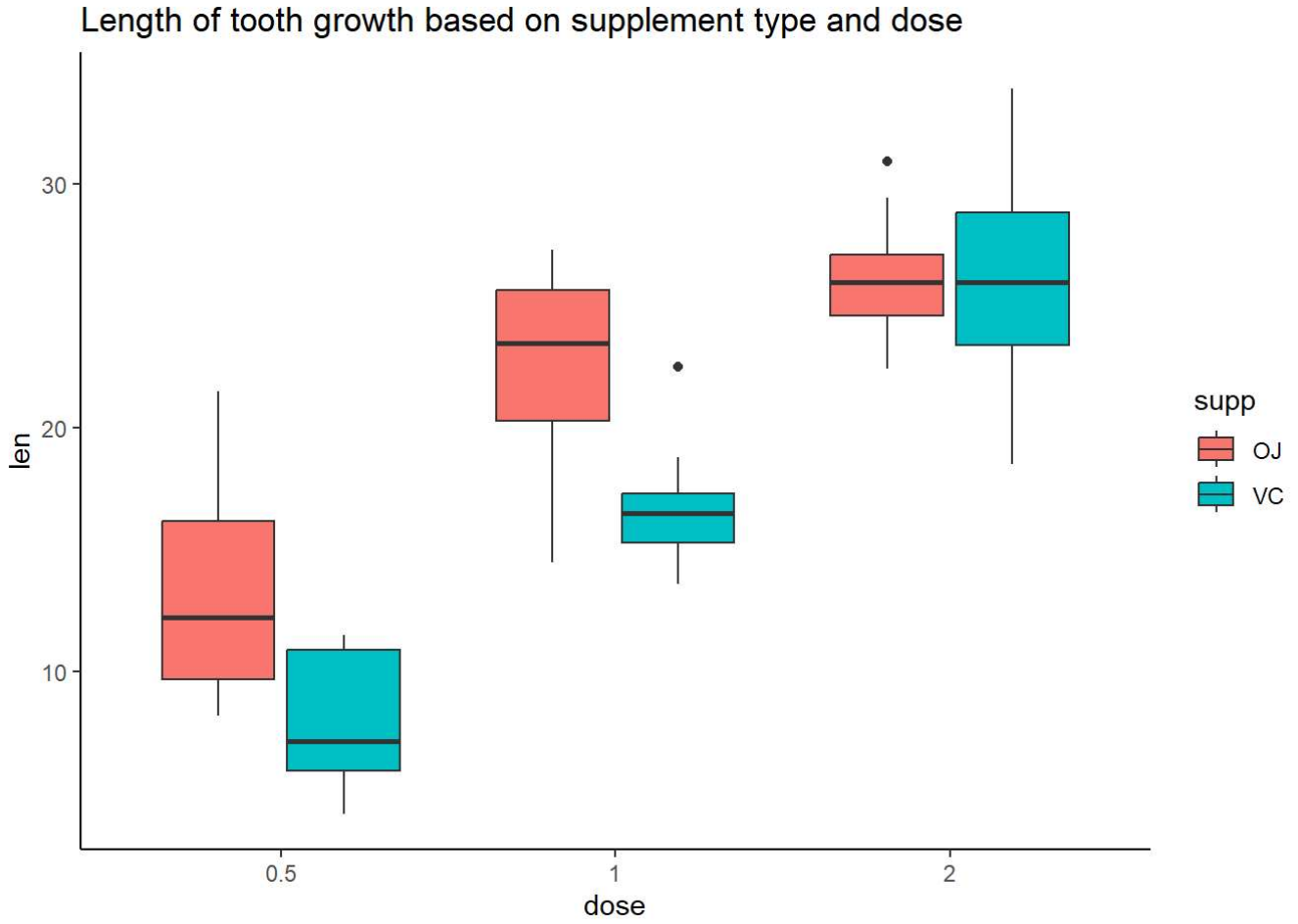
```
## [1] 0.5 1.0 2.0
```

```
dt$dose <- factor(dt$dose)
str(dt)
```

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

Ploting box plot

```
library(ggplot2)
ggplot(dt, aes(x = dose, y = len, fill = supp)) +
    geom_boxplot() +
    ggtitle("Length of tooth growth based on supplement type and dose") +
    theme_classic()
```



from the box plot, we can observe that the dosage for 0.5 and 1.0 has bigger differences in length as compared to the dosage for 2.0 mg/day. We can see a trend as the dose increases, the tooth length increases as well. Moreover, from the plot alone, we can tell supplement OJ is more effective for doses 0.5 and 1.0, where as for dose 2, there's not much difference between them.

Assumption

or the further analysis we assume that the ToothGrowth data follows normal distribution and also no other factors affect tooth growth other than dose and supp.

HypoThesis Testing

we would test three different hypothesis stating for a dose of (0.5 or 1 or 2) mg/day both supp deliver same growth.

Test for 0.5 mg/day

```
t.test(len ~ supp, data = subset( dt, dose == 0.5))
```

```
##
##  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 between group OJ and group VC is not equa
l to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

Test for 1 mg/day

```
t.test(len ~ supp, data = subset( dt, dose == 1))
```

```
##
##  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 between group OJ and group VC is not equa
l to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

Test for 2.0 mg/day

```
t.test(len ~ supp, data = subset( dt, dose == 2.0))
```

```
##
##  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 between group OJ and group VC is not equa
l to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

Forming a table to summarize the t.test results

```
dose <- c(0.5, 1.0, 2.0)
p_value <- c(0.0064, 0.0010, 0.9639)
conf.int <- c("1.72, 8.78", "2.80, 9.06", "-3.80, 3.64")
decision <- c("Reject null", "Reject null", "Do not reject null")
data.frame(dose, conf.int, p_value, decision)
```

```
##   dose    conf.int p_value           decision
## 1  0.5  1.72, 8.78  0.0064        Reject null
## 2  1.0  2.80, 9.06  0.0010        Reject null
## 3  2.0 -3.80, 3.64  0.9639 Do not reject null
```

Conclusion

The central assumption for the results is that the sample is representative of the population, and the variables are IID random variables.

For the t.test, two assumptions are made,

The data isn’t paired, meaning they’re independent The variance are different. With that, in reviewing the t.test, supplement type OC are more effective than VC for doses less than 1.0. But for dose at 2.0 mg/day, there is no difference between the supplement types.