ToothGrowth Data - confidence intervals & hypothesis testing.

Vijay (VIJAYASARADHI GANNAVARAM)

==========

```
library(ggplot2)
library(graphics)
```

Overview

- In this document, we investigate the ToothGrowth dataset
- For this we initally perform an exploratory data analysis.
- Then, we scope out a problem definition which will make the analysis interesting.

Exploratory Data Analysis

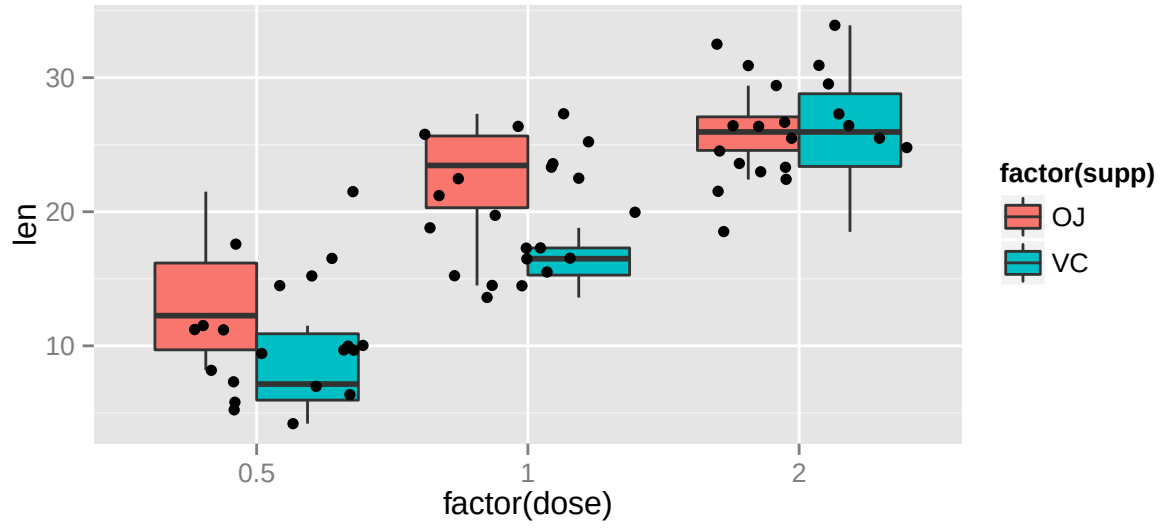
We will load the ToothGrowth Data

```
data(ToothGrowth)
columns= names(ToothGrowth)
numrows= nrow(ToothGrowth)
```

The loaded ToothGrowth data has: - len, supp, dose columns - 60 number of rows

We observe that there are equal number of observations for each supplement and each dose level. Let us explore how the length changes with dose and supplement and understand the spread using a box plot.

```
p= ggplot(ToothGrowth, aes(factor(dose),len))
p + geom_boxplot(aes(fill = factor(supp))) + geom_jitter()
```



From the above plots, we make the following observations:

- This shows that length increases with dose levels in general for both the supplements
- For low and medium dosage levels (0.5 & 1), on an average, OJ results in higher length than VC
- We observe that at high dosage levels(2), the medians of len for both supplements are almost equal
- Above observations can be visualized using plots in the Appendix.

Problem definition

We would like to estimate the mean for two groups when **dosage is high**. We define null hypothesis as **True difference in means of 'len' variable is equal to zero for high dose, for both the supplements**

Scope out the input data for analysis: High dose observations for subjects treated with both supplements

- We will first subset the data to get only high-dose specific rows
- Next, we split the data first by supplement and look at the distribution of len for each supplement separately

```
ToothGrowth_high_dose= ToothGrowth[ToothGrowth$dose==2.0, ]
VC_TG= ToothGrowth_high_dose[ToothGrowth_high_dose$supp=='VC' , ]
OJ_TG= ToothGrowth_high_dose[ToothGrowth_high_dose$supp=='OJ', ]
```

Paired T-interval

First, we will assume that these tests are paired, i.e., they are done on the same group at different times, and estimate the difference in the mean of len, modeling them using t-distribution and getting the confidence interval using t-test.

```
difference= OJ_TG$len - VC_TG$len
paired_tobj= t.test(difference)
paired_conf_interval= paired_tobj$conf
paired_tstat= paired_tobj$statistic
paired_pval= paired_tobj$p.value
```

When paired, we observe that:

- The difference in mean has confidence interval -4.3289765, 4.1689765
- The interval contains zero
- the t-statistic is -0.042592
- The pvalue is 0.9669567

Independent Group T-interval

Now, we will assume that these groups are independent, where one group is treated with supplement VC and the other is treated with OJ.

```
tobj= t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data= ToothGrowth_high_dose)
independent_conf_interval= tobj$conf
independent_tstat= tobj$statistic
independent_pval= tobj$p.value
```

When the groups are independent, we observe that:

- The difference in mean has confidence interval -3.7980705, 3.6380705
- The interval contains zero
- the t-statistic is -0.0461361
- The pvalue is 0.9638516

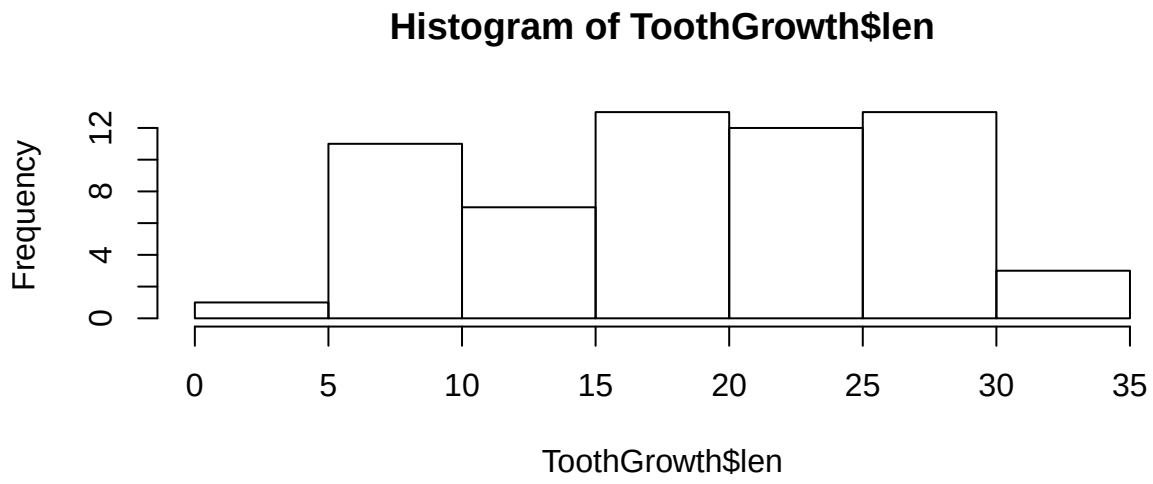
Conclusions

- The length increases with increase in dosage for both the supplements.
- When treated with supplement OJ, the increase in length is high on average for dose levels of 0.5 and 1
- However, it looks like the average increase in length is similar when the dose levels for both OJ and VC are same.
- We verified this using both paired and independent group t-tests.
- The confidence intervals for the difference contains 0 and the t-statistic and p-values indicate that the null hypothesis can be accepted.

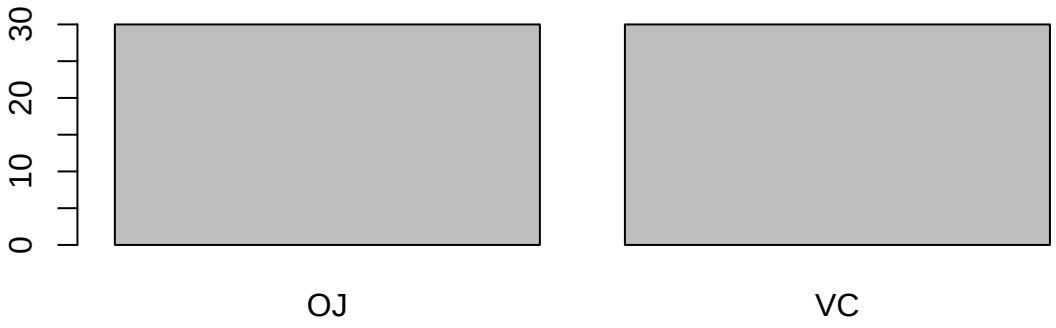
Appendix

A quick look at the frequencies of supp and dose variables along with the histogram of len variable.

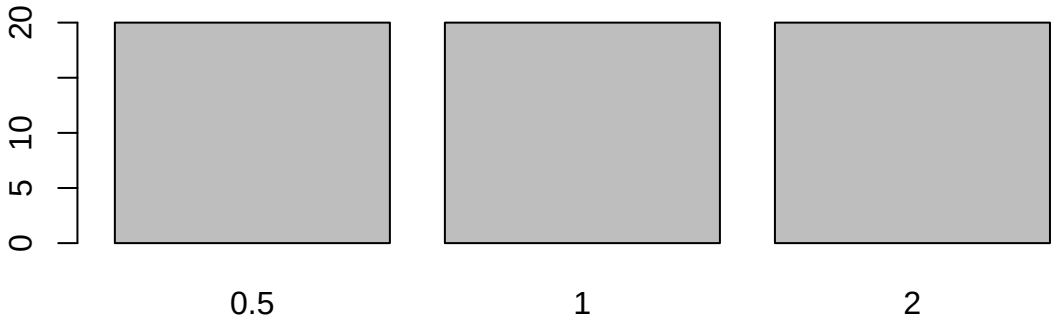
```
hist(ToothGrowth$len)
```




```
barplot(table(ToothGrowth$supp)) #equal number of subjects for both supplements
```

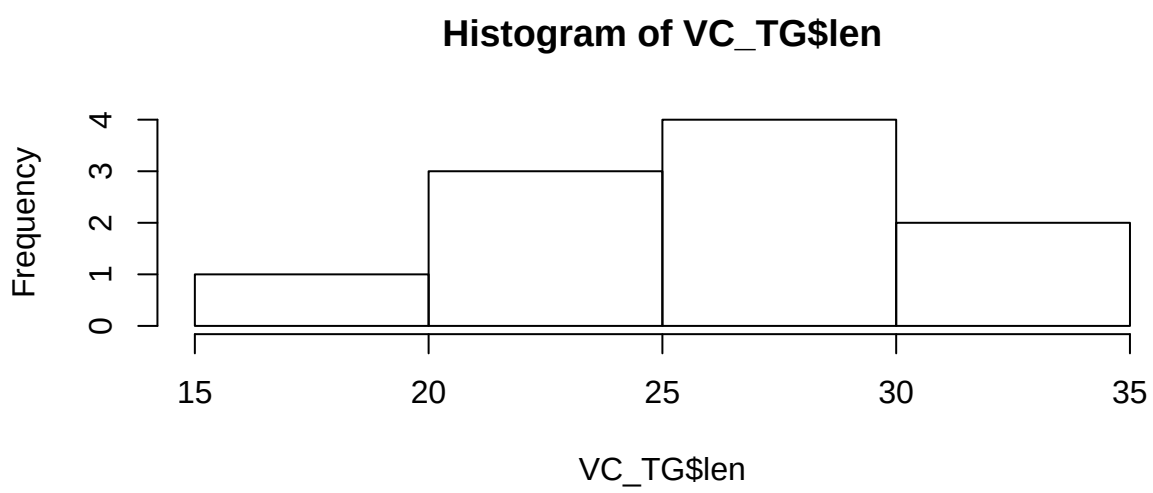



```
barplot(table(ToothGrowth$dose))
```

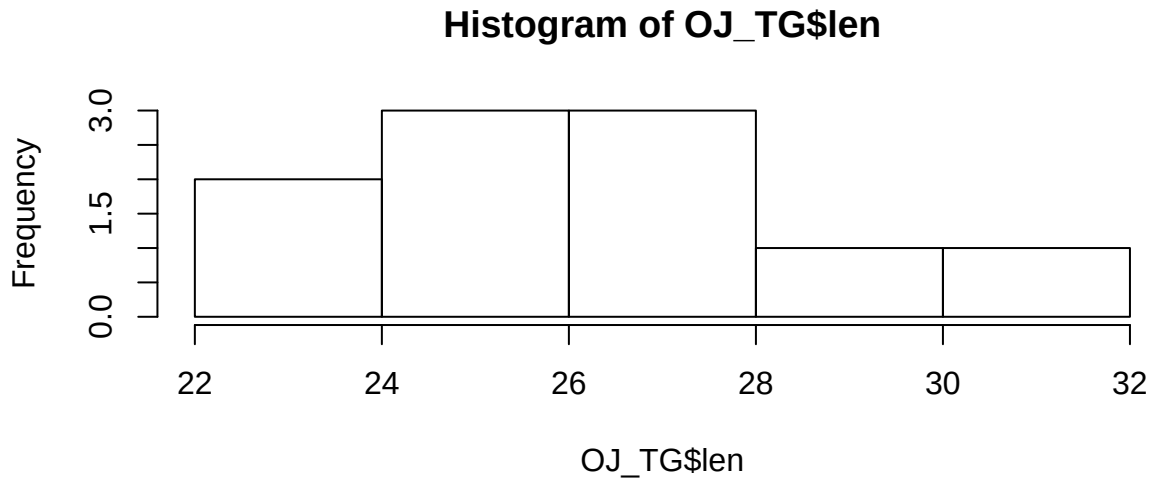



```
ToothGrowth_high_dose= ToothGrowth[ToothGrowth$dose==2.0, ]

hist(VC_TG$len)
```




```
hist(OJ_TG$len)
```




```
mean(VC_TG$len)
```

```
## [1] 26.14
```

```
mean(OJ_TG$len)
```

```
## [1] 26.06
```