Part Two: Basic inferential data analysis

Samir N. Hag Ibrahim

7/12/2020

PART TWO:BASIC INFERENTIAL DADA ANALYSIS

1. Load the data and perform basic EDA.

first, laod the libraries and the dataset

EDA is useful to get a glimpse of the provided data. EDA was performed as follow:

Data Validation and quality

A. data structure

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

```
## [1] "len"  "supp" "dose"
```

The dataset is made up of 60 observation and 3 variables namely: “len”, “supp” and “dose”. in the dataset,we have 2 numeric variables and 1 categorical variable with two levels (“OJ” and “VC”) and no missing values were recorded.

B. Missing Values

Accordingly, there is no missing values in the dataset.

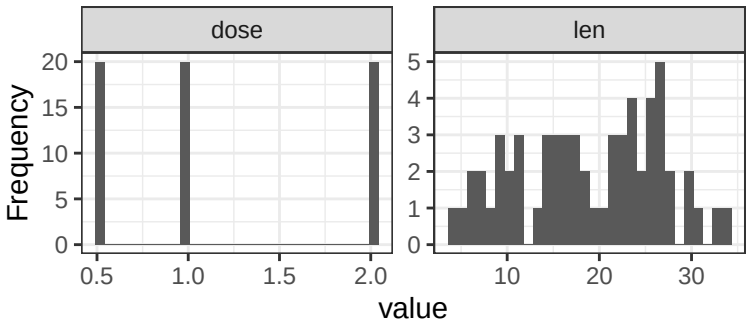
C. Data summary

for more details about the dataset

The dataset contains 2 numeric variables and 1 factor variable (with 2 levels) each with 60 observation and no missing values recorded. chacking the frequency of the catergorical variables “supplement”:

```
## supp
## OJ VC
## 30 30
```

we have 30 obervations for each level. Now plot the frequencies.

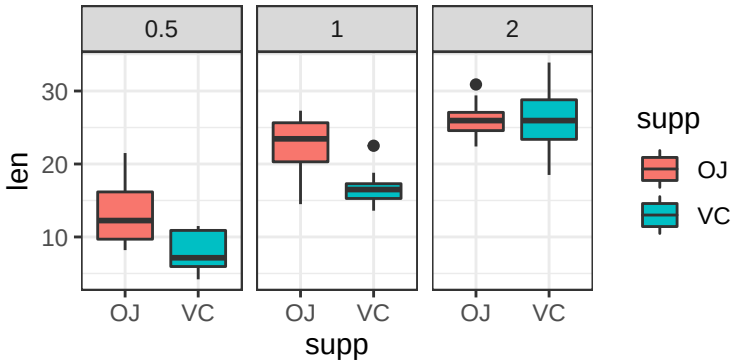


it seems that the dose should be considered as a *Factor* variable and not numeric variable. we need to convert dose variable into *Factor*.

Accordingly, we have 2 supp levels (30 obs each) and 3 dose levels (20 obs each). sounds like we have 10 obs per dose for each supp. to check this:

```
##
##       OJ VC
##   0.5 10 10
##   1   10 10
##   2   10 10
```

D. effect on tooth length



So, the over all conclusion from EDA is that: 1. There are 3 variables (2 numeric and 1 categorical) 2. *dose* variable was set as a numeric variable where it is actually a *Factor*. 3. There are 10 obs per dose for each supplement. 4. There is no missing values in the data set. 5. Possibly there an effect of dose on the tooth growth, since increasing the dose leads to an increase in the teeth legnth, but it is not yet clear if there is an effect of supplement type on that.

2. Provide a basic summary of the data

Here bellow is the statistical summary of the data.

3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose

A- t-test

I- Effect of Supplement on Toothgrowth

```
##
##  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
```

According to the t-test, the p-value forlen by supp was found to be 0.0606345 "i.e. *p>0.05*", therefore, we can not reject the null hypothesis (i.e. there is no differece between the means). But since _p -value < 0.1, so, we can not be sure if the we truely can reject the null hypothesis, therefore, the power was calculated in the next section.

II- The effect of DOSE on Toothgrowth

Since we have 3 levels for *dose*, we will subset the main dataset into 3 subsets

```
##         Dose 0.5 & 1 Dose 0.5 & 2   Dose 1 & 2
## p-value 0.0002951967 9.361615e-07 9.361615e-07
```

from the t.test, there was a significant difference between all doses examined and the teeth length (i.e. the dose has a significnat effect on the teeth length).

B- power:

Power was used here to evaluate the power (or probability) of rejecting the null hypothesis "that there is a significant difference between the means of *supp*" under the current test conditions.

Tets conditions: number of samples = 30 for each treatment, the difference between the means is 3.7 sd of the data = 7.64

Accordingly, under the current stats, there is 58.18% probability to reject the null hypothesis (relatively low probability for reject the null hypothesis). in order to increase this probability, we have either one of two options: 1- increase the number of samples 2- increase delta (i.e. the means difference)

I- Number of samples

what is the number of samples required that can increase the probability (up to 80%) for rejecting the null hypothesis???

Accordingly, we need 54 sample (compared with the current no. of observation '30') to increase the probability of rejecting the null hypothesis up to 80%.

II- Increase delta

Also we can increase the power of rejecting the null hypothesis if the means difference increased.

from the calculations above, we can observe that the mininum value of *delta* at which there is 80 % probability for rejecting the null hypothesis is 4.97. This value is greater that the actual means difference 3.7. Hence, the only possible way to increase the probability of rejecting the null hypothesis is by increaseing the number of observations up to 54 sample.

APPENDIX

Codes

```
library(ggplot2)
library(DataExplorer)
library(mlbench)
#library(skimr)
library(dplyr)
data("ToothGrowth")
```

```
str(ToothGrowth); names(ToothGrowth)
```

```
sum(is.na(ToothGrowth))
```

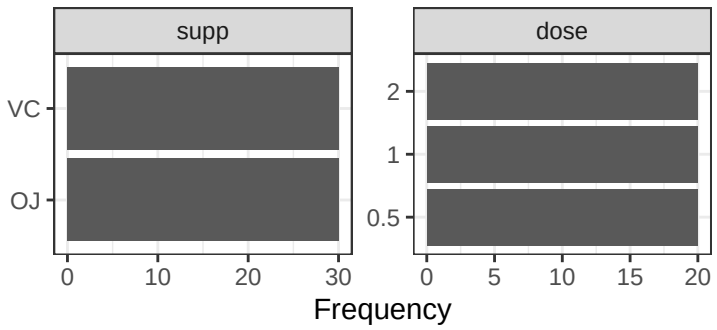
```
summary(ToothGrowth)
```

```
with(ToothGrowth, table(supp))
```

```
plot_histogram(ToothGrowth, ggtheme = theme_bw())
```

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

```
plot_bar(ToothGrowth, ggtheme = theme_bw())
```




```
table(ToothGrowth$dose, ToothGrowth$supp)
```

```
ggplot(ToothGrowth, aes(x = supp, y = len))+
        geom_boxplot(aes(fill = supp))+
        facet_wrap(. ~dose)+
        theme_bw()
```

```
library(pastecs)
summary(ToothGrowth)
stat_tooth <- stat.desc(ToothGrowth)
```

```
t.test_supp <- t.test(data=ToothGrowth, len~supp, paired = FALSE)
t.test_supp
```

```
d1 <- ToothGrowth[ToothGrowth$dose == c(0.5, 1), ]
d2 <- ToothGrowth[ToothGrowth$dose == c(0.5, 2), ]
d3 <- ToothGrowth[ToothGrowth$dose == c(1, 2), ]
```

```
t.test_d1<- t.test(data = d1, len~dose, paired = FALSE)
t.test_d2<- t.test(data = d2, len~dose, paired = FALSE)
t.test_d3<- t.test(data = d2, len~dose, paired = FALSE)
t.test_dose <- as.data.frame(row.names = "p-value", cbind("Dose 0.5 & 1" = t.test_d1$p.value,"Dose 0.5 &
t.test_dose
```

```
stat_OJ <- stat.desc(ToothGrowth[ToothGrowth$supp == "OJ", ])
stat_VC <- stat.desc(ToothGrowth[ToothGrowth$supp == "VC", ])
```

```
power_supp <- power.t.test(n=30, delta = stat_OJ$len[9]-stat_VC$len[9], sd= stat_tooth$len[13], type = '
power_supp
```

```
power_n <- power.t.test(power = 0.8, delta = stat_OJ$len[9]-stat_VC$len[9], sd = stat_tooth$len[13], typ
power_n$n
```

```
power_delta <- power.t.test(power = 0.8, n=30, sd = stat_tooth$len[13], type = "two.sample", alternativ
power_delta
```

DONE *********************