Project-02 - EDA for ToothGrowth data

jose Antonio (joseantonio@me.com)

github: Project

Title: Analize the ToothGrowth Data

Overview

In this Project we're going to analyze the ToothGrowth data in the R datasets package.

The objectives are:

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 conclusions and the assumptions needed for your conclusions.

Libraries

```
getwd()
```

```
## [1] "/Volumes/Dados/Coursera/0-A-STATISTIC-INFERENCE"
```

```
library(knitr)
library(markdown)
library(rmarkdown)
library(plyr)
library(stats)
library(ggplot2)
```

1 - Load the ToothGrowth data and perform some basic EDA

1.1 - Loading the data

```
library(datasets)
data(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 "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 ...
```

```
head(ToothGrowth)
```

```
##    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
```

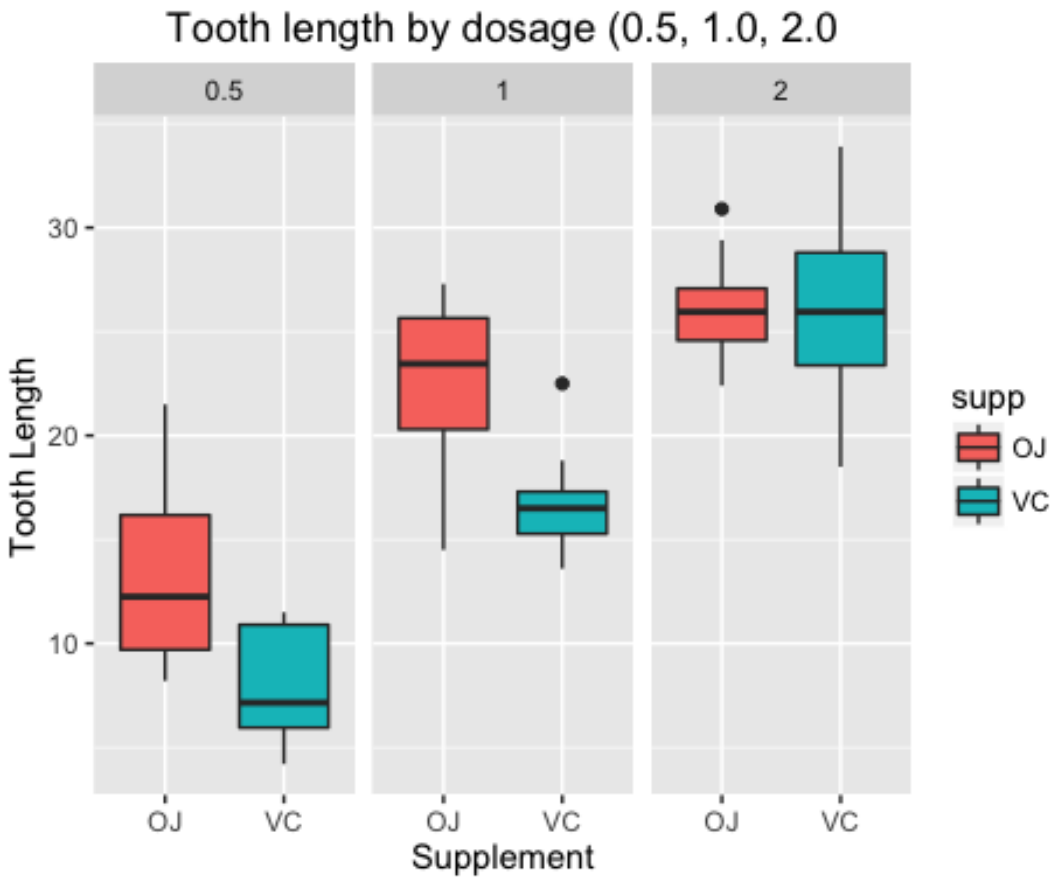
```
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

1.2 - Perform some basic EDA

The dataset has 60 observations and 3 variables as describe:

- len (numeric) = Tooth lenght
- supp (factor) = Vitamin supplement (VC = Vitamin C, OJ - Oranje Juice)
- dose (numeric) = Dose in milligrams (0.5, 1.0, 2.0) - (This variable could be factor if necessary).
- Graphic per Dose and Kind of Vitamin

```
g <- ggplot(ToothGrowth, aes(x=supp,y=len, fill=supp))
g <- g + geom_boxplot()+facet_wrap(~dose)
g <- g + xlab('Supplement')+ylab('Tooth Length')
g <- g + ggtitle("Tooth length by dosage (0.5, 1.0, 2.0")
g
```



As demonstrated in the Graphic, "apparently" Orange Juice (OJ) supplement seems to provide a better tooth growth with smaller doses.

```
aggregate(ToothGrowth$len, by=list('Dose'=ToothGrowth$dose,'Supp'=ToothGrowth$supp),
FUN = mean)

##   Dose Supp     x
## 1  0.5   OJ 13.23
## 2  1.0   OJ 22.70
```

```
## 3  2.0   OJ 26.06
## 4  0.5   VC  7.98
## 5  1.0   VC 16.77
## 6  2.0   VC 26.14
```

Using aggregate function to mean group by dose, we can see that for Dose 0.5, OJ has 13.23 against 7.98 from VC. With Dose 1.0, OJ has 22.70 against 16.77. At a dosage of 2 mg, the impact becomes near the same value.

2 - Provide a basic summary of the data.

```
summary(ToothGrowth)
```

```
##       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
```

```
table(ToothGrowth$dose, ToothGrowth$supp)  ## table view of data
```

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

- Transforming the variable dose to a factor variable and making a summary

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

```
## 0.5   1   2
##  20  20  20
```

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

To confirm in our graphic, the higher the dosage, the longer the teeth.

```
t.test(len~dose, paired=F, var.equal=F, data=subset(ToothGrowth, dose %in% c(0.5,2)))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

Looking for basic EDA and summaries, there seems to be a correlation between the dose and tooth growth. OJ is more effective for lower doses, but near equal to VC on 2 milligram.

To verify this, we need indentify the confidence interval for supp. Lets test if OJ is better than VC for tooth growth, looking for the Ho (Hypothesis Null) as that OJ and VC realize the same.

```
t.test(len~supp, paired=F, var.equal=F, data=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
```

Based on 95% confidence interval, the Ho cannot be rejected. So, OJ is not better than VC.

3.1 Testing each dosage (0.5, 1.0 and 2.0)

```
t.test(len ~ supp, paired=F, var.equal=F, data=subset(ToothGrowth, 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 is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

At dosage=.5, OJ is better than VC (Ho is rejected base on 95% confidence interval).

```
t.test(len ~ supp, paired=F, var.equal=F, data=subset(ToothGrowth, dose == 1.0))
```

```
##
##  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 OJ mean in group VC
##            22.70            16.77
```

At dosage=1.0, OJ is better than VC (Ho is rejected based on 95% confidence inteval)

```
t.test(len ~ supp, paired=F, var.equal=F, data=subset(ToothGrowth, 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 is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

At dosage= 2.0, OJ is as good as VC (Ho cannot be rejected based on 95% confid ence interval)

4 - State conclusions and the assumptions for conclusions

The tests using confidence intervals assumed that the sample are independent (randomily chosen), the population variance is the same and that the distribution is normal.

OC and VC have different effects on tooth growth, and dosage is a key factor. OC dosage have a larger impact than VC on tooth growth with a smaller dosage, but at 2mg, the impact is the same. So, there is correlation between the supplement type used and teeth growth in guinea pigs.