Analyzing ToothGrowth Data

Harish Kumar Rongala

October 23, 2016

Overview

Can we draw conclusions using statistical techniques ? Let's explore more about it by analyzing ToothGrowth data set, available in R. We will follow these steps to make our conclusions

1. Load data and look at its basic features
2. Perform Exploratory Data Analysis to find any interesting features
3. Perform T-tests
4. Draw conclusions
5. List our assumptions

Data set Description The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC).

1. What's in ToothGrowth Data

Using R's **str** function, we can look at the structure of ToothGrowth data

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

So, we are dealing with rather smaller data set. We have 60 observations with 3 variables to deal with. Among them, 2 are numeric/continuous and 1 is categorical/discrete. Let's dive in and look at more quantitative details, using R's **summary** function.

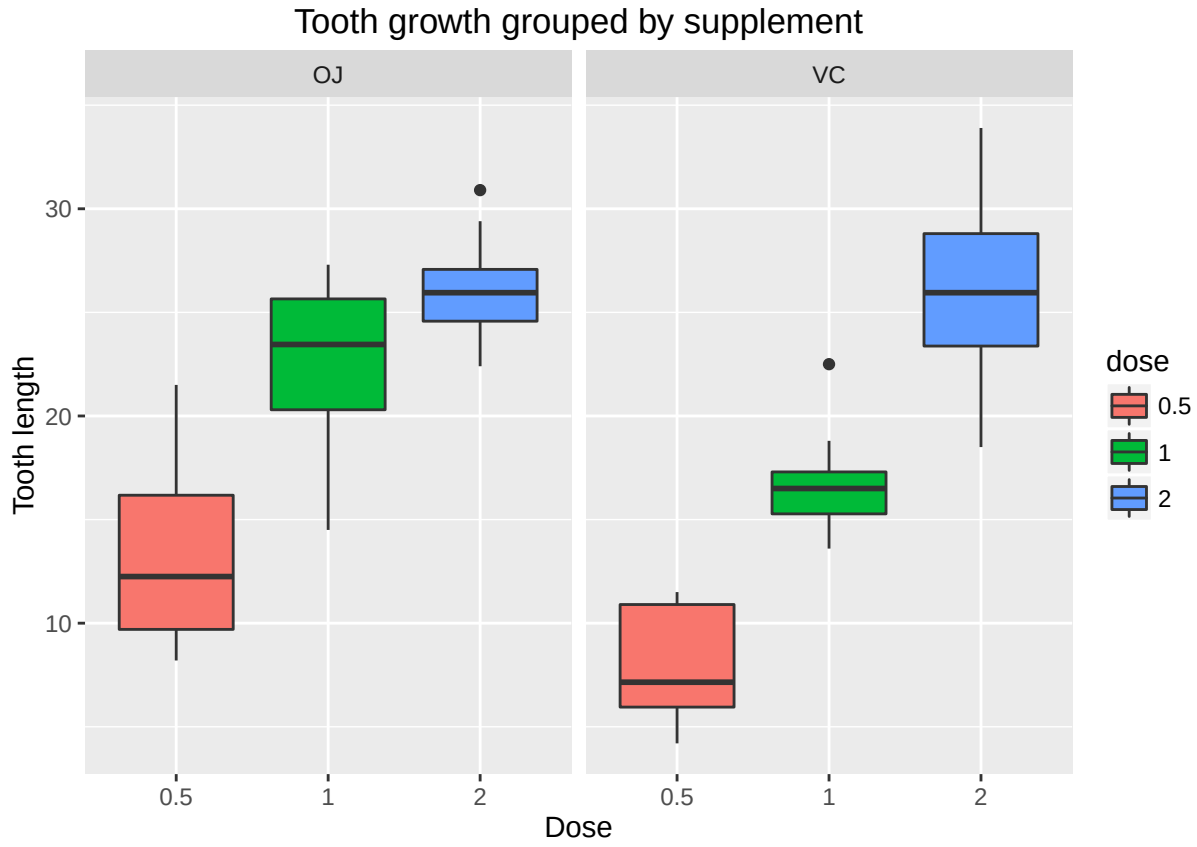
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

We have 30 observations for each type of supplement. Interesting values like mean, median, quantiles of the variables are listed. That's pretty detailed, however this may not make sense when observed individually. So, we move further and combine these variables to unearth any interesting findings.

2. Exploratory data analysis

If we understood the experiment, **len** (Tooth length) is something we are interested in. We are trying to discover the affects of the supplements, **Orange Juice (OJ)** and **Vitamin C (VC)** given in different doses, on tooth growth.

Let's quickly plot the affect of supplements in different doses on tooth length. we group the tooth growth by supplement to closely observe the affect of each supplement with increase in the dose.



Observation: From the above plot, we notice that both Orange Juice (OJ) and Vitamin C (VC) increase the tooth growth, as the dose increases. Although for dose 0.5 and 1, Orange Juice (OJ) seems to contribute more to tooth length than Vitamin C (VC). We cannot conclude as the results are not consistent.

3. Confidence Intervals

Looking at their confidence intervals can help us conclude which factor has actual affect on the tooth growth. We have a total of observations 60 observations, which is relatively very less to make proper estimate. However, Gosset's (Student's) t-test can help us deal with this type of less sample size.

T-test for length and supplement

```
#T-test for length and supplement
t.test(data=ToothGrowth, len~supp, var.equal=FALSE, paired=FALSE)$conf
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

Result: Our 95% Confidence interval **-0.17 to 7.57** contains **0**, which means supplementary has no significant affect on the tooth growth.

T-test for length and dose

As the grouping factor must have exactly two levels, we cannot test for all 3 dose values (0.5, 1, 2) at once. We will consider two at a time

```
#T-test for length and dose (0.5,1)
t.test(data=subset(ToothGrowth,dose!=2), len~dose, var.equal=FALSE, paired=FALSE)$conf
```

```
## [1] -11.983781  -6.276219
## attr(,"conf.level")
## [1] 0.95
```

Result: Our 95% Confidence interval is **-11.98 to -6.27**, which means dose **1** has more affect than **0.5** on the tooth growth.

```
#T-test for length and dose (1,2)
t.test(data=subset(ToothGrowth,dose!=0.5), len~dose, var.equal=FALSE, paired=FALSE)$conf
```

```
## [1] -8.996481 -3.733519
## attr(,"conf.level")
## [1] 0.95
```

Result: Our 95% Confidence interval is **-8.99 to -3.73**, which means dose **2** has more affect than **1** on the tooth growth.

```
#T-test for length and dose (0.5,2)
t.test(data=subset(ToothGrowth,dose!=1), len~dose, var.equal=FALSE, paired=FALSE)$conf
```

```
## [1] -18.15617 -12.83383
## attr(,"conf.level")
## [1] 0.95
```

Result: Our 95% Confidence interval is **-18.15 to -12.83**, which means dose **2** has more affect than **0.5** on the tooth growth.

4. Conclusion

From our Exploratory data analysis and T-tests, we can draw following conclusions about the Tooth Growth data

- There is no significant difference between Orange Juice (OJ) and Vitamin C (VC) in the tooth growth scenario.
- Increase in the dosage of supplement increases the tooth growth.

5. Assumptions

The above conclusions are made on the following assumptions

- Data is independent, Guinea pigs are randomly selected
- Data is not paired, No guinea pig took both supplements

6. Appendix

Code for plot in the Exploratory data analysis section can be found here.

```
library(ggplot2)
tdata<-ToothGrowth
# Convert variable dose into factor to better group in the plots
tdata$dose<-as.factor(tdata$dose)
#Tooth growth grouped by supplement
plot1<-ggplot(tdata,aes(dose,len))+geom_boxplot(aes(fill=dose))+facet_grid(.~supp)+labs(x="Dose",y="Too
print(plot1)
```

To closely compare the tooth growth by supplement, we could consider the following plot grouped by dose

```
library(ggplot2)
tdata<-ToothGrowth
# Convert variable dose into factor to better group in the plots
tdata$dose<-as.factor(tdata$dose)
#Tooth growth grouped by dose
plot2<-ggplot(tdata,aes(supp,len))+geom_boxplot(aes(fill=supp))+facet_grid(.~dose)+labs(x="Supplement",
print(plot2)
```