Basic inferential data analysis

Zanin Pavel

February 28, 2016

Link to project on GitHUB

Overview

In this project we are going to analyze the ToothGrowth data from the R datasets package.
This dataset describes how changes the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

1. Getting data and exploratory analysis

For start working load the dataset:

```
library(datasets) #loading neccesary library
data(ToothGrowth) #loading specified dataset
```

Let's see what are these dataset:

```
str(ToothGrowth) #compactly displaying the internal 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 ...
```

```
head(ToothGrowth) #showing the first 6 rows of dataset
```

```
##    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
```

```
summary(ToothGrowth) #showing dataset's summary
```

```
##       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
```

```
ToothGrowth$dose # showing the list of doses
```

```
##  [1] 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 1.0 1.0 1.0 1.0 1.0 1.0 1.0
## [18] 1.0 1.0 1.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 0.5 0.5 0.5 0.5
## [35] 0.5 0.5 0.5 0.5 0.5 0.5 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 2.0
## [52] 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0
```

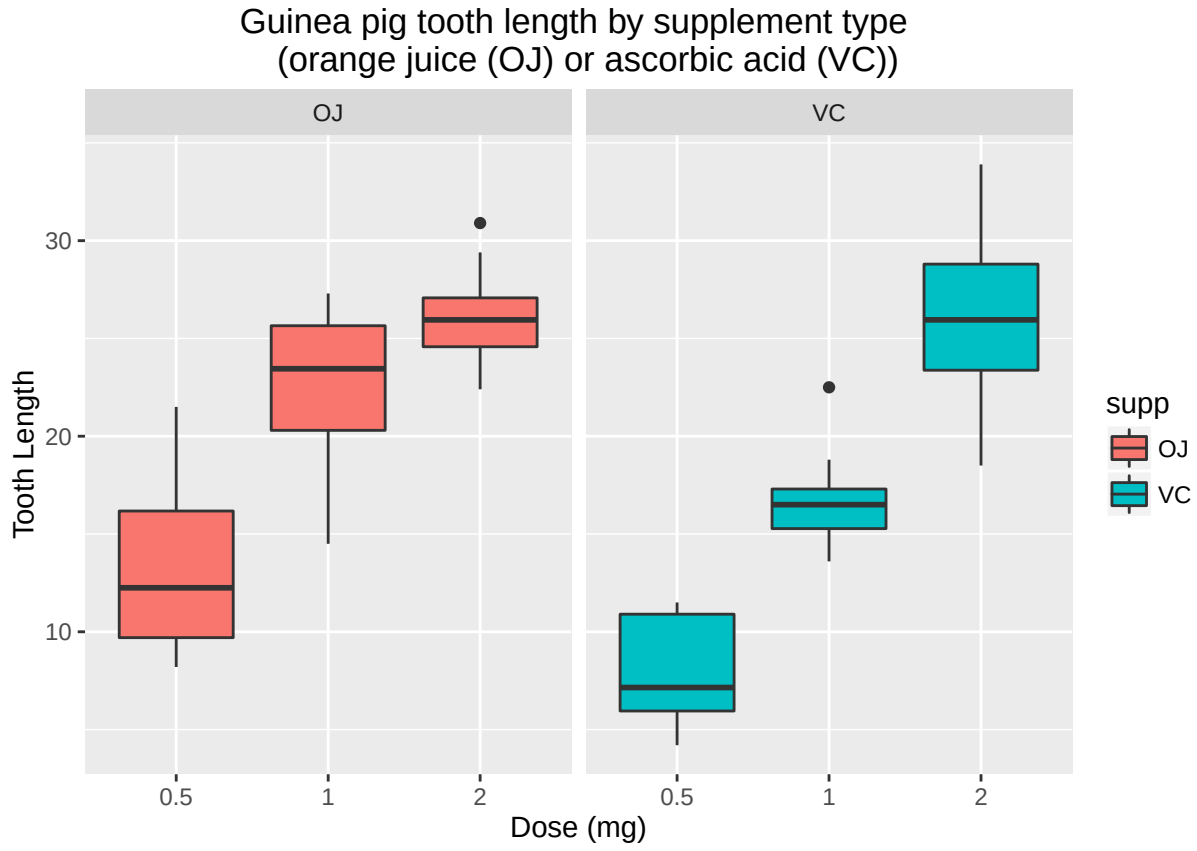
So, as we can see, we've got 60 observations, for 2 supplement types (VC or OJ) and 3 dose levels of Vitamin C (0.5, 1, and 2). Dataset's description not lied to us =)

For the next step let's make exploratory plot's for this data:

```
library(ggplot2) #loading neccesary library
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```
ggplot(ToothGrowth, aes(x=factor(dose), y=len)) +
  facet_grid(.~supp) +
  geom_boxplot(aes(fill = supp)) +
  labs(title="Guinea pig tooth length by supplement type
  (orange juice (OJ) or ascorbic acid (VC))",
    x="Dose (mg)",
    y="Tooth Length")
```



2. Hypothesis testing

Hypotesis 1: There is no difference between supplement types (orange juice or ascorbic acid), regardless from doses

For testing this let's try t.test:

```
t.test(len ~ supp, 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
```

So, if confidence interval include zero and p-value is bigger than usual α level (.05) then we our hypotesis is true and we cannot reject it.

Hypotesis 2: Effect from 0.5 mg dose for both supplement types is equal

For testing this try t.test again:

```
t.test(len ~ supp, 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
```

As p-value is lower than usual α level (.05) then our hypotesis isn't true and we reject it. Orange juice has much effectiveness for this dose than ascorbic acid.

Hypotesis 3: Effect from 1 mg dose for both supplement types is equal

For testing this try t.test again:

```r
t.test(len ~ supp, data = subset(ToothGrowth, 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 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
```

As p-value is lower than usual α level (.05) then our hypotesis isn't true and we reject it. Orange juice has much effectiveness for this dose than ascorbic acid.

Hypotesis 4: Effect from 2 mg dose for both supplement types is equal

For testing this try t.test again:

```r
t.test(len ~ supp, data = subset(ToothGrowth, dose == 2))
```

```
##
##  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
```

So, if confidence interval include zero and p-value is bigger than usual α level (.05) then we our hypotesis is true and we cannot reject it.

3. Conclusions

Dataset ToothGrowth allows to us make next conclusions:
* Vitamin C consumption results to increasing pig's tooth growth.
* In small doses (0.5 and 1 mg) orange juice much effective than ascorbic acid.
* In big dose (2 mg) both supply types have same effectiveness.