Basic Inferential Analysis of the Data on the Effect of Vitamin C on Tooth Growth in Guinea Pigs

Mykola Herasymovych

13.08.2015

Overview

The aim of the project is to perform basic inferential analysis on the ToothGrowth data from the R datasets package. The analysis includes exploring some simple descriptive statistics and building simple linear model with confidence intervals for its coeficients in order to draw conclusions about influence of vitamin C in various forms and doses on Guinea pigs' tooth growth.

1. Load the ToothGrowth data and perform some basic exploratory data analyses

Firstly, we load the dataset and look at data format using basic R commands like str() and head().

```
# load and look at the data
library("datasets")
data <- ToothGrowth
str(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 ...
```

```
head(data)
```

```
##    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
```

As R has read "dose" variable as a number, we have to convert it to the factor format.

```
# convert dose variable to the factor format
data$dose <- as.factor(data$dose)
str(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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

Our data has 60 observations and 3 variables: len, supp and dose, last two of which are factor variables. "Supp" variable is supplement type and it has two levels: VC (ascorbic acid) and OJ (orange juice). "Dose" variable is dose level of Vitamin C and it has three levels: 0.5, 1, and 2 mg.

2. Provide a basic summary of the data.

To explore the data more deeply we get the table with some descriptive statistics, calculate the standard deviation for numeric variable and visualise the data using simple plot.

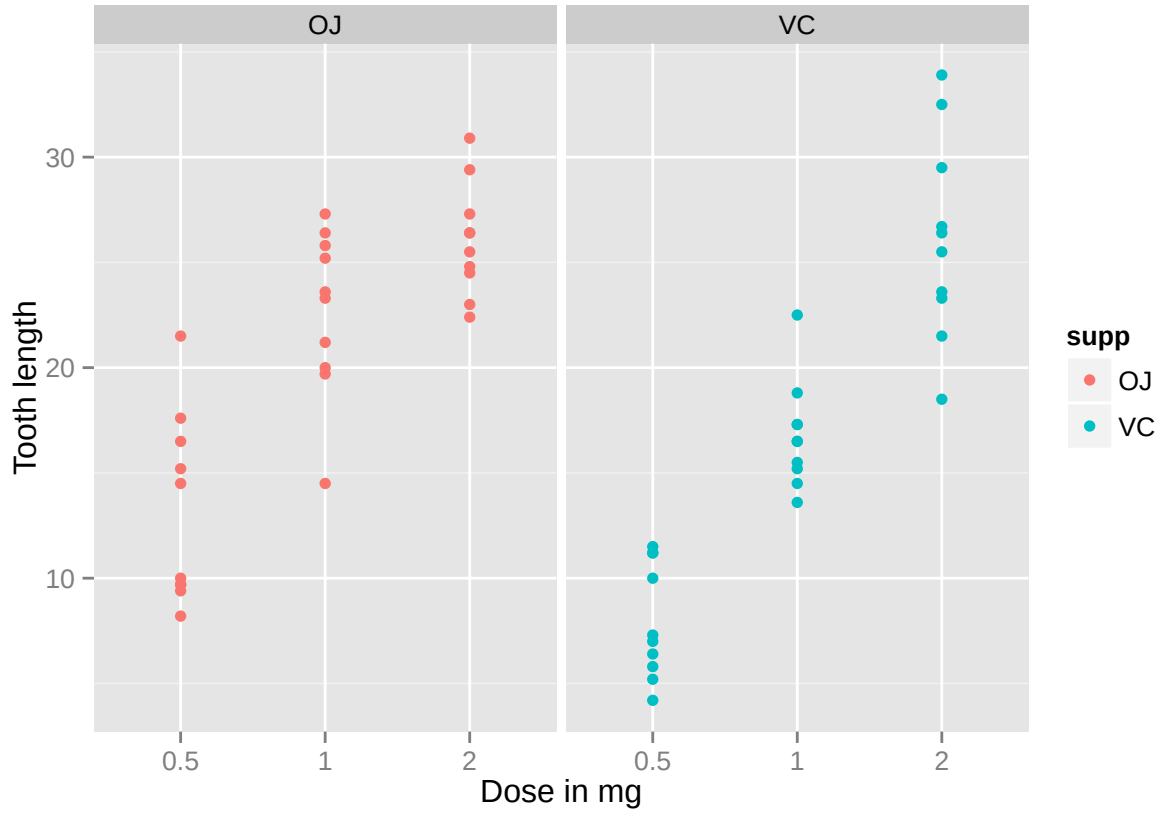
```
# get description statistics
summary(data)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

```
# calculate standard deviation of tooth length
sd(data$len)
```

```
## [1] 7.649315
```

```
# plot the data
library("ggplot2")
ggplot(data = data, aes(x = as.factor(dose), y = len, col = supp)) +
        geom_point(stat = "identity") +
        facet_grid(. ~ supp) +
        xlab("Dose in mg") +
        ylab("Tooth length") +
        guides(fill = guide_legend(title = "Supplement type"))
```



Thus, average tooth length is 18.81 and meadian length is 19.25. Standard deviation is 7.65. From the plot we can see that there is positive correlation between tooth length and dose of vitamin C. It also appears that the variance of tooth length differs with the supplement type: VC (ascorbic acid) is more variable. Let's calculate separate ratios for separate supplementary type to prove the hypothesis.

```
# calculate mean and standard deviation for OJ factor
mean(data$len[which(data$supp == "OJ")])
```

```
## [1] 20.66333
```

```
sd(data$len[which(data$supp == "OJ")])
```

```
## [1] 6.605561
```

```
# calculate mean and standard deviation for VC factor
mean(data$len[which(data$supp == "VC")])
```

```
## [1] 16.96333
```

```
sd(data$len[which(data$supp == "VC")])
```

```
## [1] 8.266029
```

Consequently, tooth length of VC factor is actually more variable and has smaller mean than of OJ factor.

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

In this part of investigation we build a linear model that explains how dose and supplement type influence tooth length. We also calculate 95% confidence intervals to check the hypothesis of coeficient significance.

```
# make a linear model of tooth length by dose and supplement type factors
model <- lm(len ~ dose + supp, data = data)

# get a summary of the model and calculate 95% confidence intervals for coeficients
summary(model)
```

```
##
## Call:
## lm(formula = len ~ dose + supp, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.085 -2.751 -0.800  2.446  9.650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4550     0.9883  12.603  < 2e-16 ***
## dose1         9.1300     1.2104   7.543 4.38e-10 ***
## dose2        15.4950     1.2104  12.802  < 2e-16 ***
## suppVC       -3.7000     0.9883  -3.744 0.000429 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 56 degrees of freedom
## Multiple R-squared:  0.7623, Adjusted R-squared:  0.7496
## F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

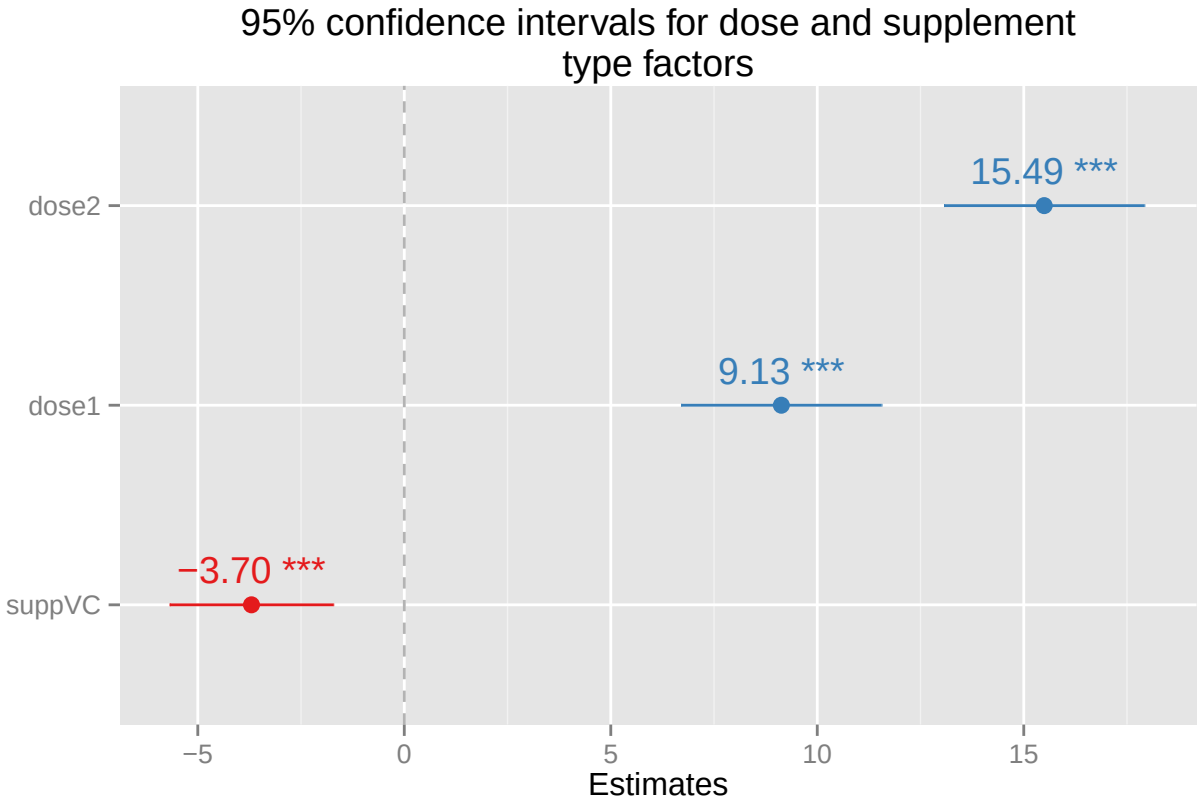
```
confint(model)
```

```
##                 2.5 %    97.5 %
## (Intercept) 10.475238 14.434762
## dose1        6.705297 11.554703
## dose2       13.070297 17.919703
## suppVC      -5.679762 -1.720238
```

```
# visualize 95% confidence intervals for coeficients
library("sjPlot")
```

```
## Visit http://strengejacke.de/sjPlot for illustrative examples of sjPlot-functions.
```

```
sjp.lm(model, title = "95% confidence intervals for dose and supplement type factors")
```



Let's now interpret summary tables and confidence intervals' plot of the model. Note that R automatically converted dose and supp factor variables into several dummy variables, and now dose1 means 1 mg dose, dose2 means 2 mg dose, suppVC means usage of ascorbic acid supplement. Thus, intercept shows us tooth length received by 0.5 mg orange juice supplement usage.

First of all, all the factors of the model are significant at 0 level and definitely influence tooth length variable. If you use 0.5 mg dose of orange juice, tooth length is 12.455, which is showed by the intercept value. If you use 1 mg dose of it, tooth length grows by 9.13 (dose1 coef). If you use 2 mg dose of it, length incereases by 15.495 (dose2 coef). If you use ascorbic acid instead of orange juice, tooth length drops by -3.7.

If we look at confidence intervals, it seems clear that coeficients are significant at the 95% confidence level, as none of them includes 0 point. Moreover, 2 mg dose makes definitely stronger influence than 1 mg, and 1 mg dose makes stronger influence than 0.5 mg at the 95% confidence level, as the intervals don't cross. Ascorbic acid has clearly less influence than orange juice.

4. State your conclusions and the assumptions needed for your conclusions.

Overall, *assuming that all the values from the ToothGrowth dataset were independently and randomly sampled from a population whose values are distributed according to a Gaussian distribution*, we can draw some conclusions (valid at the 95% confidence level):

- Vitamin C positively influences tooth growth of Guinea pigs;
- The bigger the dose, the stronger is the effect of vitamin C. The 2 mg dose has the most influence;
- Orange juice has stronger and more stable effect on tooth growth than ascorbic acid.