Statistical Inference course project - The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Evgeniy Zabrodskiy

November, 2015

Overview

The goal of this analysis is to explore the ToothGrowth data, make assumptions and apply statistical inference methods to confirm or reject those assumptions.
The dataset ToothGrowth contains the length of odontoblasts (teeth) in each of 60 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).

Exploratory analysis

Basic summary of the data

ToothGrowth dataframe 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 ...
```

We can see that there are three variables:
- len contains the length of the odontoblasts in microns;
- supp is a factor variable showing the type of the supplement containing Vitamin C ("OJ" stands for "Orange Juice""","VC" stands for "Ascrobic Acid");
- dose stores the dose of Vitamin C in miligrams. The variable type will be converted to factor for convenience.

Dataframe summary:

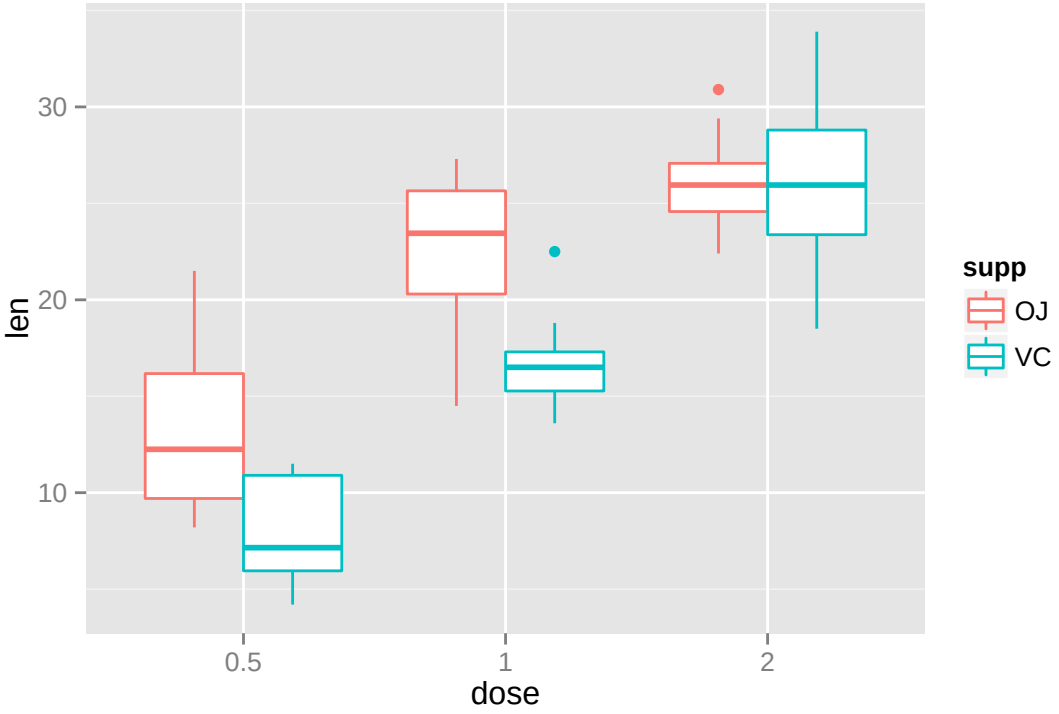
```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

Two first lines of the data:

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
```

Exploratory plots

The boxplot below shows the distribution of length for each *supp* per *dose*:



Based on this boxplot we can make different assumptions. Some of them will be described and tested below.

Assumptions

Assumption 1: Mean length of odontoblasts is different in two groups with the following properties:
Group A: pigs which received Vitamin C with **orange juice**, dose **0.5 mg**;
Group B: pigs which recieved Vitamin C as **ascrobic acid**, dose **0.5 mg**.
(same dose, different supplements)

Assumption 2: Mean length of odontoblasts is different in two groups with the following properites:
Group A: pigs which received Vitamin C with **orange juice**, dose **0.5 mg**;
Group B: pigs which received Vitamin C with **orange juice**, dose **1 mg**.
(same supplement, different doses)

There are more assumptions that can be made but they are all similar and are tested the same way.

Testing assumptions and making conclusions

Talking about lengths of odontoblasts we may assume that the lengths population is normally distributed even though we cannot clearly see that from the sample data. Based on this assumption and taking into account that the sample sizes of the groups we are going to compare is quite small (10 elements each), we should use Student's t-test to confirm or reject our assumptions. Also it is assumed that the variances of the populations are different, thus Welch Two Sample t-test modification of Student's t-test will be used.

Hypothesis testing for assumption 1:

$H_0 : \mu_{oj,0.5} = \mu_{vc,0.5}$
$H_1 : \mu_{oj,0.5} \neq \mu_{vc,0.5}$

```
##
##  Welch Two Sample t-test
##
## data:  len.VC.05 and len.OJ.05
## 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:
##  -8.780943 -1.719057
## sample estimates:
## mean of x mean of y
##      7.98     13.23
```

From the T-test output we can see that p-value = 0.0063586 which represents the probability of H_0 being true and since it is less than $\alpha = 0.05$ for a two-sided test we may reject H_0. Also the confidence interval which does not contain zero can be used to make the same conclusion. It means that with 95% probability the true difference of the means of the populations will be within the borders of the interval.

Conclusion 1: Assumption 1 is confirmed: mean length of odontoblasts is different for the groups of pigs which recieved 0.5 mg dose of Vitamin C in forms of orange juice and ascorbic acid.

Hypothesis testing for assumption 2:

$H_0 : \mu_{oj,0.5} = \mu_{oj,1}$
$H_1 : \mu_{oj,0.5} \neq \mu_{oj,1}$

```
##
##  Welch Two Sample t-test
##
## data:  len.OJ.05 and len.OJ.1
## t = -5.0486, df = 17.698, p-value = 8.785e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.415634  -5.524366
## sample estimates:
## mean of x mean of y
##     13.23     22.70
```

From the T-test output we can see that p-value = 8.7849191×10^{-5} which represents the probability of H_0 being true and since it is less than $\alpha = 0.05$ for a two-sided test we may reject H_0.

Conclusion 2: Assumption 2 is confirmed: mean length of odontoblasts difference is significant for subgroups of pigs which recieved 0.5 mg and 1 mg dose of Vitamin C with orange juice.

Appendix - source code and additional plots

```
data("ToothGrowth")
str(ToothGrowth)
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
summary(ToothGrowth)
head(ToothGrowth, 2)
```

Source code for the boxplot which is used in the analysis part of the document:

```
qplot(data = ToothGrowth,
      geom = "boxplot",
      x = dose,
      y = len,
      color = supp)
```

The following boxplot is not included in the analysis part due to document length restrictions but it can be useful to view some properties of the data distribution:

```
qplot(data = ToothGrowth,
      geom = "histogram",
      binwidth = 2,
      x = len,
      fill = dose,
      facets = dose~supp)
```