Part 2: Basic Inferential Data Analysis

Ivo Georgiev

July 28, 2016

Overview

This document presents basic inferential analysis performed on the *ToothGrowth* dataset, as part of the final project for the *Statistical Inference* course of the Coursera *Data Science Specialization*. The dataset explores the effect of vitamin C on tooth growth in guinea pigs. This is an R dataset included in the *datasets* package of the basic R distribution.

Original source: C.I.Bliss (1952) *The Statistics of Bioassay.* Academic Press.

Reference: Crampton, E.W. (1947) *The growth of the odontoblast of the incisor teeth as a criterion of vitamin C intake of the guinea pig.* The Journal of Nutrition **33(5)**: 491-504. (link)

Basic exploratory data analysis

Let's look at the data.

```r
require(datasets)
data("ToothGrowth")
dim(ToothGrowth)
```

```
## [1] 60  3
```

We have 60 observations and three data points for each.

```r
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  
```

We have two numerical columns and one factor. There are 30 observations for each value of the factor.

```r
?ToothGrowth
```

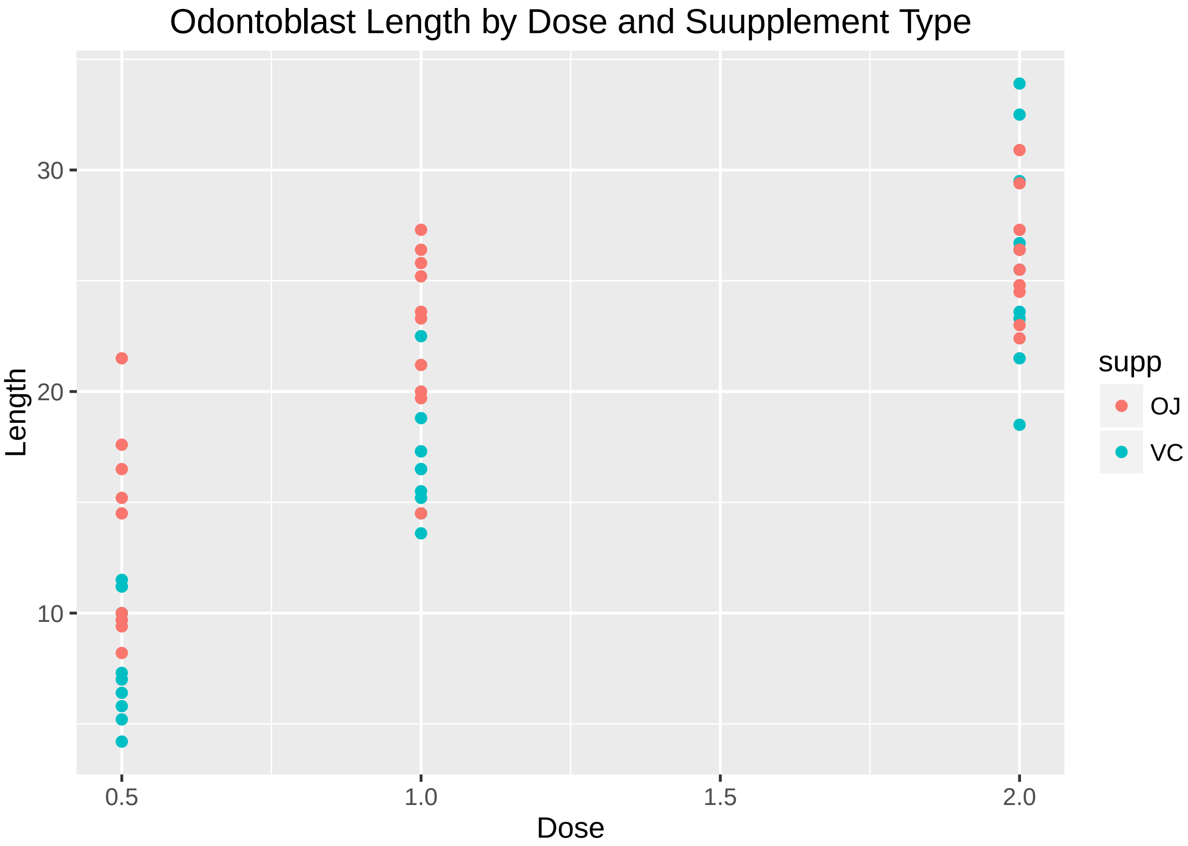
From the dataset documentation we can see that the dependent variable is the odontoblast length (**len**) and the independent variables describe the treatment with vitamin C by dose (**dose**) and type of supplement (**supp**), orange juice (**"OJ"**) or ascorbic acid (**"VC"**).

How is the data distributed? Let's look at a quick plot.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

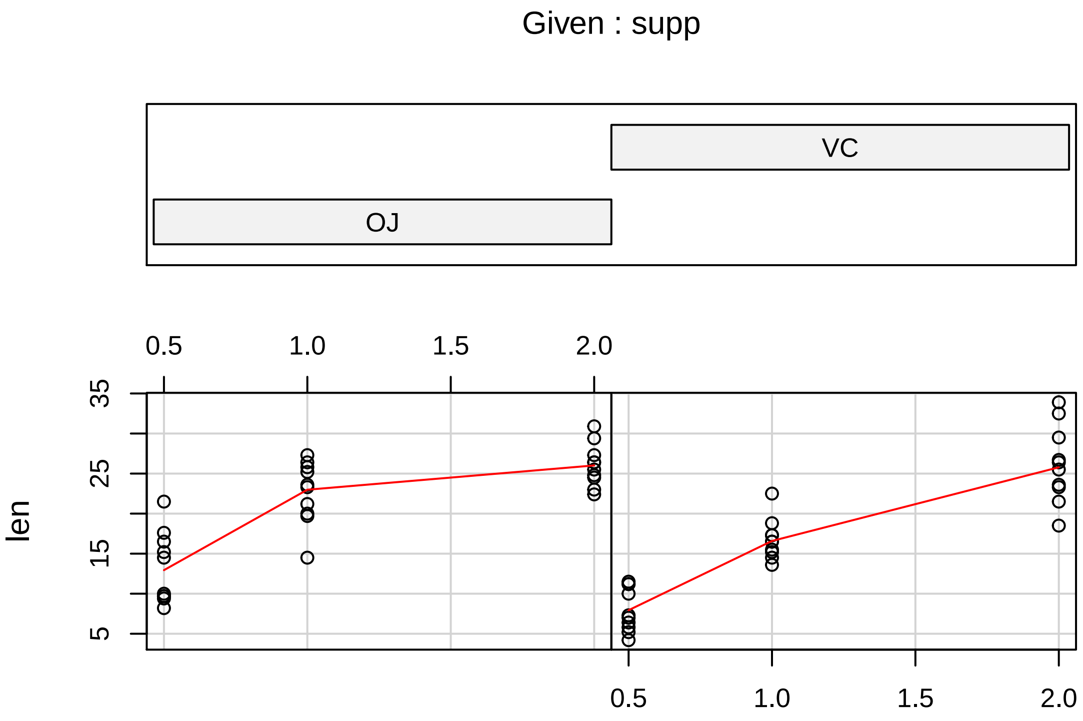
```
qplot(dose, len, col=supp, data=ToothGrowth, main = "Odontoblast Length by Dose and Suupplement Type", x
```



We can see that the data is clustered in three groups over the corresponding doses for the two vitamin C supplements. Within a supplement, the response increases monotonically with the dosage. However, there is no clear within-cluster response pattern for the two supplements, so they look like the natural basis of comparison with statistical inference.

The dataset documentation also suggests a conditioning plot that compares the data side by side for the two supplements (conditioning variable).

```
require(graphics)
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```



ToothGrowth data: length vs dose, given type of supplement

Indeed, we can see clearly the variance in length response at the different dosages for the two supplement types.

Statistical Analysis

It is obvious from the plots in the previous section that there are two goals for our inferential analysis:

1. Verify that upward trend of the response vs. dosage within each supplement.
2. Compare the response within the same dose for the different supplements.

For the comparison we can use the **t** test (**t.test**), the primary statistical test for comparing two data vectors that was taught in the course. This test, based on the heavy-tailed *Student T distribution*, is particularly suited for our small dataset.

Assumptions

There are several assumptions that need to be stated, for they will determine the correct arguments to **t.test**:

1. The dataset slices corresponding to the two different supplements (supplement slices) are **independent**. This is clear from the documentation which states that 60 different animals were used in the trial. This is a basic requirement for the use of the **t** test.
2. The supplement and dosage slices are **not paired**. The documentation states that each of the 60 animals was give only *one* of the three doses of *one* of the two supplements, so this is clearly the case.
3. The slices are **not normally distributed**. (to verify)
4. The slices have **unequal variances**. (to verify)

Data preparation

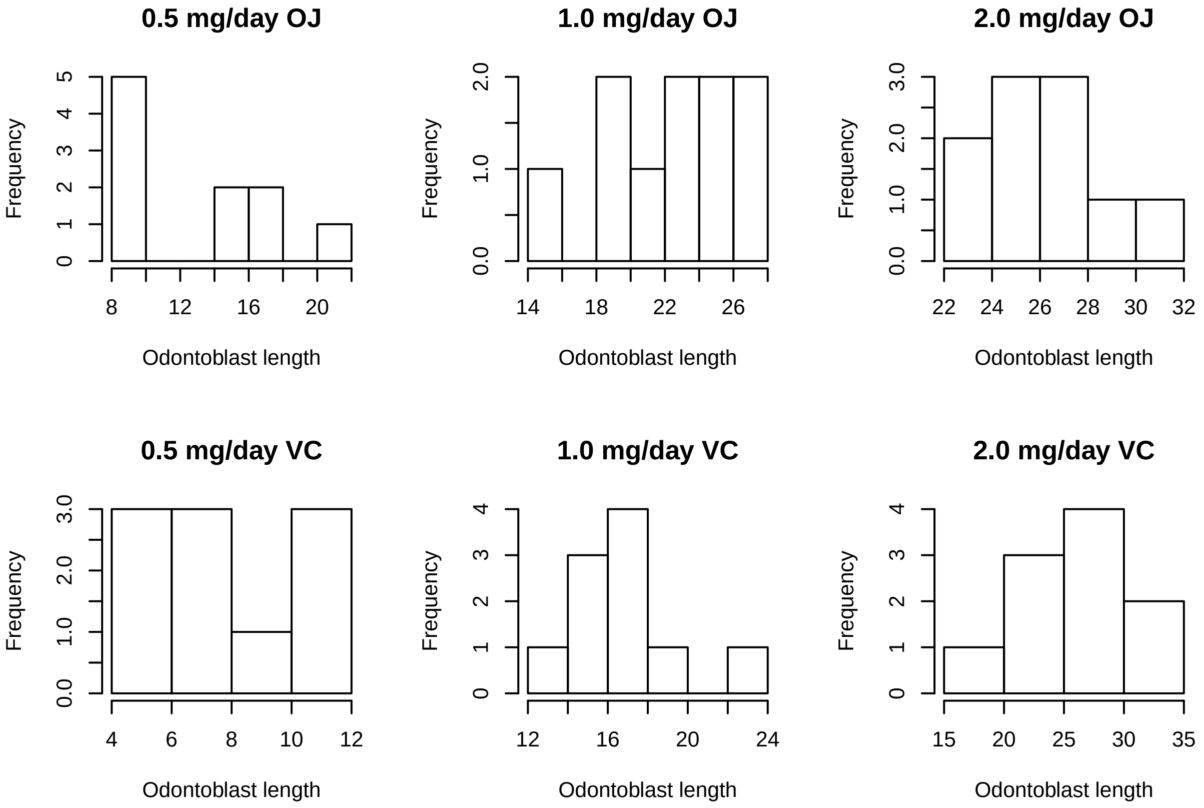
We need to slice the dataset to obtain the different data vectors for our analysis.

```
attach(ToothGrowth)
dose1oj <- ToothGrowth[dose ==0.5 & supp == "OJ",]
dose2oj <- ToothGrowth[dose ==1.0 & supp == "OJ",]
dose3oj <- ToothGrowth[dose ==2.0 & supp == "OJ",]

dose1vc <- ToothGrowth[dose ==0.5 & supp == "VC",]
dose2vc <- ToothGrowth[dose ==1.0 & supp == "VC",]
dose3vc <- ToothGrowth[dose ==2.0 & supp == "VC",]
detach(ToothGrowth)
```

Assumption check

Let's look at the distributions of lengths in the six slices. Since each has only 10 data points, establishing normality is highly unlikely.



Evidently, the lengths are not normally distirbuted even within the dose and supplement dataset slices. Now, let's calucate the variances.

```
var(dose1oj$len)
```

```
## [1] 19.889
```

```r
var(dose2oj$len)
```

```
## [1] 15.29556
```

```r
var(dose3oj$len)
```

```
## [1] 7.049333
```

```r
var(dose1vc$len)
```

```
## [1] 7.544
```

```r
var(dose2vc$len)
```

```
## [1] 6.326778
```

```r
var(dose3vc$len)
```

```
## [1] 23.01822
```

The variances differ widely.

Length increase with dosage

The **t** tests between dosages within each supplement group clearly show that the tooth growth increases monotonically with supplement dose. Ascorbic acid shows a slightly more pronounced increase in tooth length with the growing dosages. (Appendix A)

In hypothesis testing terms, we are assuming **no growth** as the **null hypothesis** and comfortably reject it every time.

Comparison of supplements

Here the **null hypothesis** is that the there is **no difference between the effects** (equal means) of the two different supplements when controlled for dose.

The **t** test is inconclusive at the 0.5 mg/day and 1.0 mg/day doses, showing a slight increase in tooth length from orange juice to ascorbic acid. The null hypothesis can be rejected only for the 1.0 mg/day dose, where ascorbic acid shows more growth. The 95% confidence interval is, nevertheless, too wide.

At the 2.0 mg/day dose, the effects of the two different supplements is indistinguishable. Notice that the 95% confidence interval contains 0, making the rejection of the null hypothesis impossible.

Conclusion

We have seen and verified that:

1. Both vitamin C supplements make the teeth of guinea pigs grow, and the length increases monotonically with dosage.
2. There is no clear distinction between the effects of the two supplements on the tooth length, though ascorbic acid shows slightly stronger effects.

Appendices

Appendix A: Growth w/ dosage within supplement

Growth difference with orange juice betweem the 0.5 and 1.0 doses.

```
t.test(dose1oj$len, dose2oj$len, paired = FALSE, var.equal = FALSE)$conf.int
```

```
## [1] -13.415634  -5.524366
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(dose1oj$len, dose2oj$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 8.784919e-05
```

Growth difference with orange juice betweem the 1.0 and 2.0 doses.

```
t.test(dose2oj$len, dose3oj$len, paired = FALSE, var.equal = FALSE)$conf.int
```

```
## [1] -6.5314425 -0.1885575
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(dose2oj$len, dose3oj$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 0.03919514
```

Growth difference with ascorbic acid betweem the 0.5 and 1.0 doses.

```
t.test(dose1vc$len, dose2vc$len, paired = FALSE, var.equal = FALSE)$conf.int
```

```
## [1] -11.265712  -6.314288
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(dose1vc$len, dose2vc$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 6.811018e-07
```

Growth difference with ascorbic acid betweem the 1.0 and 2.0 doses.

```
t.test(dose2vc$len, dose3vc$len, paired = FALSE, var.equal = FALSE)$conf.int
```

```
## [1] -13.054267  -5.685733
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test(dose2vc$len, dose3vc$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 9.155603e-05
```

Appendix B: Growth comparsion between supplement per dose

Comparison of growth of orange juice vs. ascorbic acid at the 0.5 dose.

```r
t.test(dose1oj$len, dose1vc$len, paired = FALSE, var.equal = FALSE)$conf.int
```

```
## [1] 1.719057 8.780943
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test(dose1oj$len, dose1vc$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 0.006358607
```

Comparison of growth of orange juice vs. ascorbic acid at the 1.0 dose.

```r
t.test(dose2oj$len, dose2vc$len, paired = FALSE, var.equal = FALSE)$conf.int
```

```
## [1] 2.802148 9.057852
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test(dose2oj$len, dose2vc$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 0.001038376
```

Comparison of growth of orange juice vs. ascorbic acid at the 2.0 dose.

```r
t.test(dose3oj$len, dose3vc$len, paired = FALSE, var.equal = FALSE)$conf.int
```

```
## [1] -3.79807  3.63807
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test(dose3oj$len, dose3vc$len, paired = FALSE, var.equal = FALSE)$p.value
```

```
## [1] 0.9638516
```