Statistical Inference Coursera – The Data Science Specialization

Course Project Report by Konstantinos Papastamos
Part 2

All the code written for the project can be found on the following Page*

*I didn't use github for obvious reasons (it's public)

Exercise 2

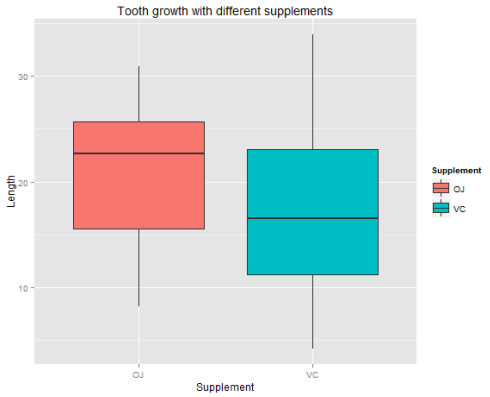
1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
4. State your conclusions and the assumptions needed for your conclusions.

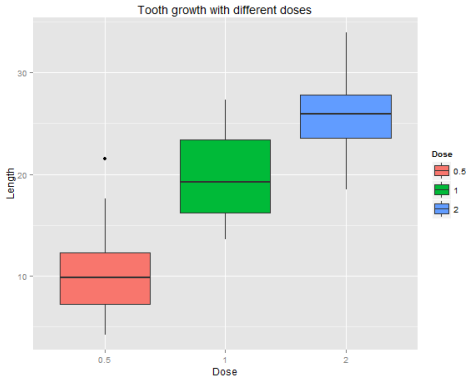
So in the given dataset we have data about the response in the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC). For more information on the dataset visit this site.
I am going to use t testing to analyze the data. In order to do that I will breakdown our initial data by supplement and by dose so as to make a more in-depth analysis.

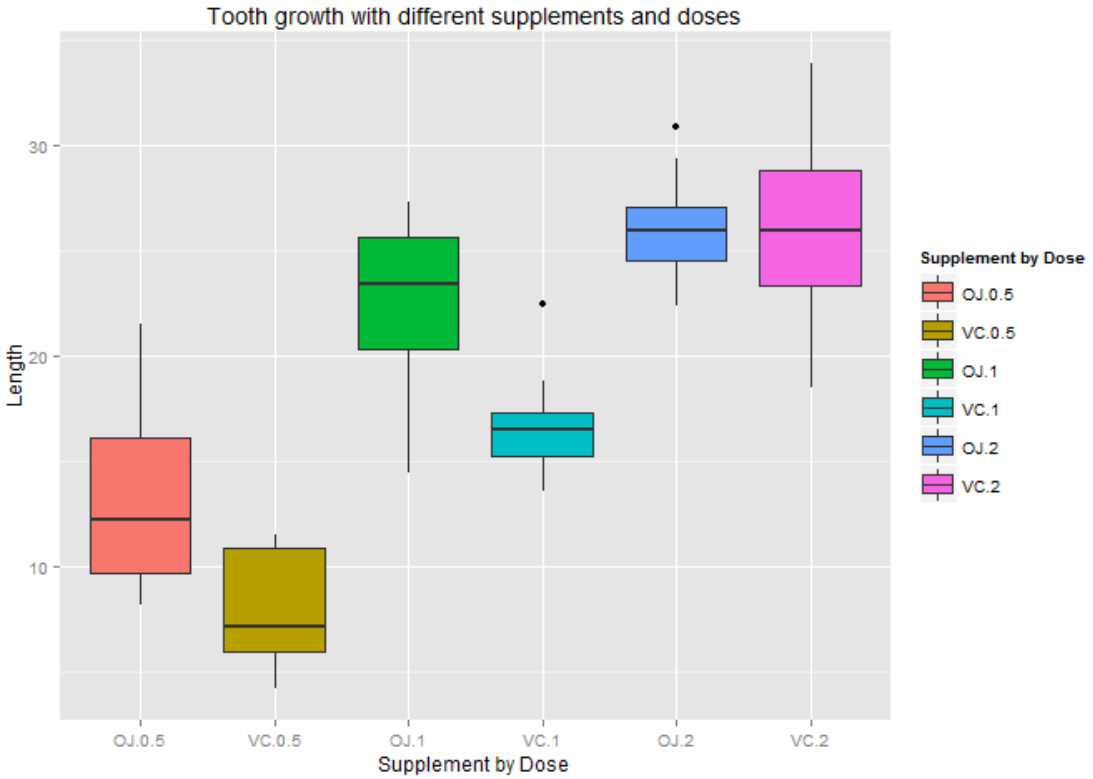
First lets see what this data looks like:

```
> 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
```

So we can see that we have 60 observations (60 guinea pigs). In 30 received the OJ supplement and the other 30 the VC. Each dosage level was received by 1/3 of the population. Here we see that even though the dosage is a numeric variable, we can turn it to a factor since there are only 3 different levels. Therefore I converted the dose variable to a factor.







So let's breakdown the data and start testing.

```
# Supplement type OJ and Dose level 0.5
doj0.5=subset(data,dose==.5 & supp=="OJ")

# Supplement type OJ and Dose level 1
doj1=subset(data,dose==1 & supp=="OJ")

# Supplement type OJ and Dose level 2
doj2=subset(data,dose==2 & supp=="OJ")

# Supplement type VC and Dose level 0.5
dvc0.5=subset(data,dose==.5 & supp=="VC")

# Supplement type VC and Dose level 1
dvc1=subset(data,dose==1 & supp=="VC")

# Supplement type VC and Dose level 2
dvc2=subset(data,dose==2 & supp=="VC")
```

Testing

Assumptions:

1. The data used for testing is not paired
2. The variances are not equal
3. All tests were performed at confidence level 0.95

Test 1: Is there any statistically significant difference in growth between the two supplements with dose level 0.5?

```
> t.test(doj0.5$len,dvc0.5$len, conf.level=0.95)

        Welch Two Sample t-test

data:  doj0.5$len and dvc0.5$len
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 of x mean of y
    13.23      7.98
```

The p-value is lower than 0.05 plus the confidence interval doesn't contain zero, hence we can reject the null hypothesis (that there is no difference in means) and assume that in dose level 0.5 the OJ supplement is more effective than the VC.

Test 2: Is there any statistically significant difference in growth between the two supplements with dose level 1?

```
> t.test(doj1$len,dvc1$len, conf.level=0.95)

        Welch Two Sample t-test

data:  doj1$len and dvc1$len
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 of x mean of y
    22.70     16.77
```

The p-value is lower than 0.05 plus the confidence interval doesn't contain zero, hence we can reject the null hypothesis (that there is no difference in means) and assume that in dose level 1 the OJ supplement is more effective than the VC.

Test 3: Is there any statistically significant difference in growth between the two supplements with dose level 2?

```
> t.test(doj2$len,dvc2$len, conf.level=0.95)

        Welch Two Sample t-test

data:  doj2$len and dvc2$len
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 of x mean of y
    26.06     26.14
```

The p-value is greater than 0.05 plus the confidence interval contains zero, hence we fail to reject the null hypothesis (that there is no difference in means) and assume that in dose level 2 the OJ supplement is as effective as the VC.

Test 4: Is there any statistically significant difference in growth between dose levels 0.5 and 1 for the supplement OJ?

```
> t.test(doj0.5$len,doj1$len, conf.level=0.95)

        Welch Two Sample t-test

data:  doj0.5$len and doj1$len
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
```

The p-value is quite lower than 0.05 plus the confidence interval doesn't contain zero, hence we reject the null hypothesis (that there is no difference in means) and assume that dose level 1 is more effective than dose level 0.5 for the supplement OJ.

Test 5: Is there any statistically significant difference in growth between dose levels 1 and 2 for the supplement OJ?

```
> t.test(doj1$len,doj2$len, conf.level=0.95)

        Welch Two Sample t-test

data:  doj1$len and doj2$len
t = -2.2478, df = 15.842, p-value = 0.0392
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.5314425 -0.1885575
sample estimates:
mean of x mean of y
    22.70     26.06
```

The p-value is lower than 0.05 plus the confidence interval doesn't contain zero, hence we reject the null hypothesis (that there is no difference in means) and assume that dose level 2 is more effective than dose level 1 for the supplement OJ.

Test 6: Is there any statistically significant difference in growth between dose levels 0.5 and 1 for the supplement VC?

```
> t.test(dvc0.5$len,dvc1$len, conf.level=0.95)

        Welch Two Sample t-test

data:  dvc0.5$len and dvc1$len
t = -7.4634, df = 17.862, p-value = 6.811e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.265712  -6.314288
sample estimates:
mean of x mean of y
     7.98     16.77
```

The p-value is quite lower than 0.05 plus the confidence interval doesn't contain zero, hence we reject the null hypothesis (that there is no difference in means) and assume that dose level 1 is more effective than dose level 0.5 for the supplement VC.

Test 7: Is there any statistically significant difference in growth between dose levels 1 and 2 for the supplement VC?

```
> t.test(dvc1$len,dvc2$len, conf.level=0.95)

        Welch Two Sample t-test

data:  dvc1$len and dvc2$len
t = -5.4698, df = 13.6, p-value = 9.156e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.054267  -5.685733
sample estimates:
mean of x mean of y
    16.77     26.14
```

The p-value is quite lower than 0.05 plus the confidence interval doesn't contain zero, hence we reject the null hypothesis (that there is no difference in means) and assume that dose level 2 is more effective than dose level 1 for the supplement VC.

Conclusions:

1. At dose levels 0.5 and 1 the supplement OJ is more effective than VC.
2. At dose level 2 the two supplements are equally effective.
3. Dose level greatly affects growth for both supplements.