Statistical Inference Course Project: Part 2

by Maxim Podkolzine

Problem: Analyze the `ToothGrowth` data in the R datasets package.

Q1-Q2: First of all let's load the data and provide a basic summary.

```
data(ToothGrowth)
head(ToothGrowth)
```

```
##    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
```

The response is 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).

```
str(ToothGrowth)
```

```
## '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 ...
```

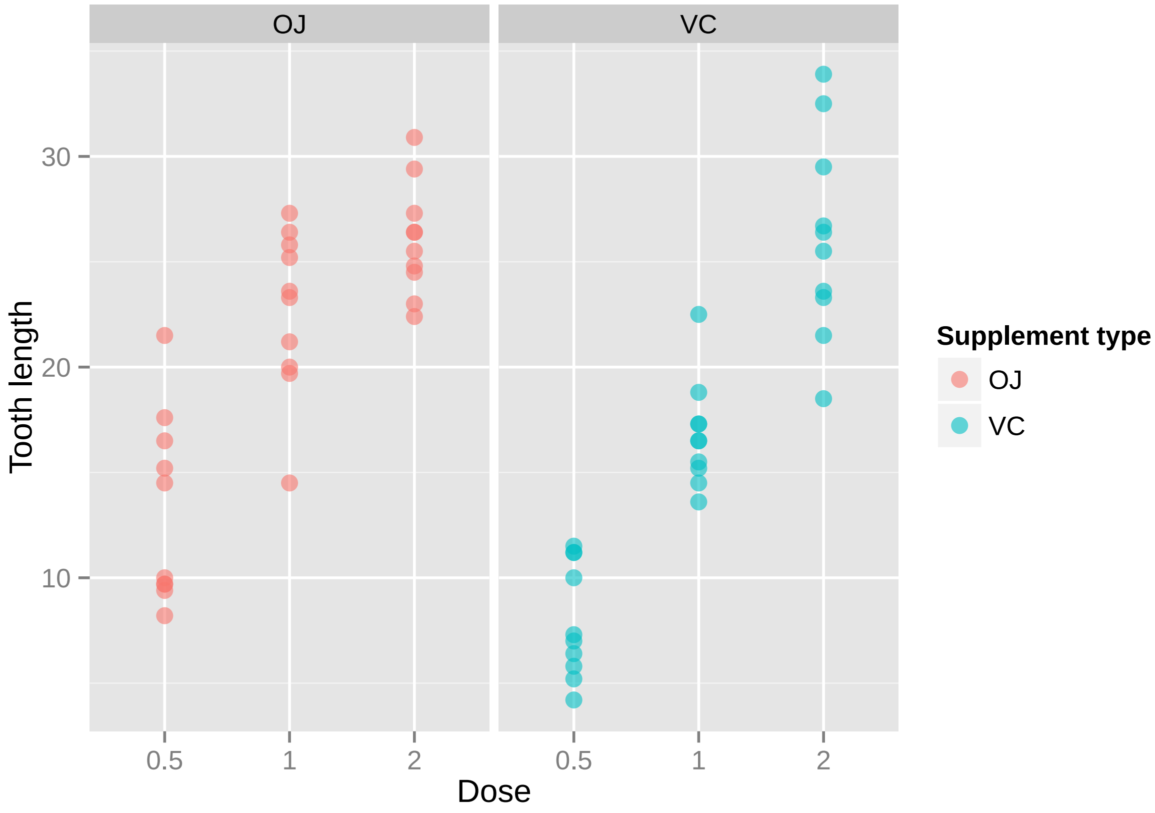
A data frame contains 60 observations of 3 variables: `len` is a tooth length (numeric), `supp` is a supplement type (a factor, VC or OJ), `dose` is a dose in milligrams (numeric, 0.5, 1 or 2).

```
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  
```

A visual presentation might be helpful.

```
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, color=supp)) +
  geom_point(size=3, alpha=0.6) +
  facet_grid(. ~ supp) +
  xlab("Dose") +
  ylab("Tooth length") +
  guides(color=guide_legend(title="Supplement type"))
```



Q3: Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

As can be seen in the plot, there is a clear positive correlation between the tooth length and the dose levels of Vitamin C, for both delivery methods. Let's test that formally, but start with a simple regression.

```
fit <- lm(len ~ dose + supp, data=ToothGrowth)
summary(fit)
```

```
##
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## suppVC       -3.7000     1.0936  -3.383   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

`R-squared` is equal to 70%, which means most of the variance in tooth length is explained by the dose and supplement.

Now let's t-test that the tooth length depends on the dosage.

```
dose.0.5 = ToothGrowth$len[ToothGrowth$dose == 0.5]
dose.1 = ToothGrowth$len[ToothGrowth$dose == 1]
dose.2 = ToothGrowth$len[ToothGrowth$dose == 2]
```

```
t.test(dose.1, dose.0.5, paired=FALSE, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dose.1 and dose.0.5
## t = 6.4766, df = 38, p-value = 1.266e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.276252 11.983748
## sample estimates:
## mean of x mean of y
##    19.735    10.605
```

```
t.test(dose.2, dose.1, paired=FALSE, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dose.2 and dose.1
## t = 4.9005, df = 38, p-value = 1.811e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.735613 8.994387
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

95% confidence intervals for both tests do not contain 0. So we reject the null hypothesis (which is the difference in length is 0), concluding that the dosage effect is significant.

Finally let's t-test that the tooth length depends on the delivery method.

```
oj.group = ToothGrowth$len[ToothGrowth$supp=="OJ"]
vc.group = ToothGrowth$len[ToothGrowth$supp=="VC"]
```

```
t.test(oj.group, vc.group, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj.group and vc.group
## 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 of x mean of y
##  20.66333  16.96333
```

95% confidence interval does contain 0, so we cannot reject the null hypothesis (which is again the difference in length is 0). But we can with 90% interval, since the p-value is 6%.

Q4: Conclusions.

- Dosage has positive impact on tooth length: the length increases with higher dosage of Vitamin C.
- Orange juice is more effective than ascorbic acid, but the influence is not that significant, mostly because the difference with 2mg dosage is very small.