Exploratory analyses of tooth growth data

Yevgeny V. Yorkhov

04/26/2015

Overview

Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.

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.

Load data and perform some basic exploratory data analysis

```
library(ggplot2)
library(datasets)
data(ToothGrowth)

dim(ToothGrowth) # dimension of the dataframe
```

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

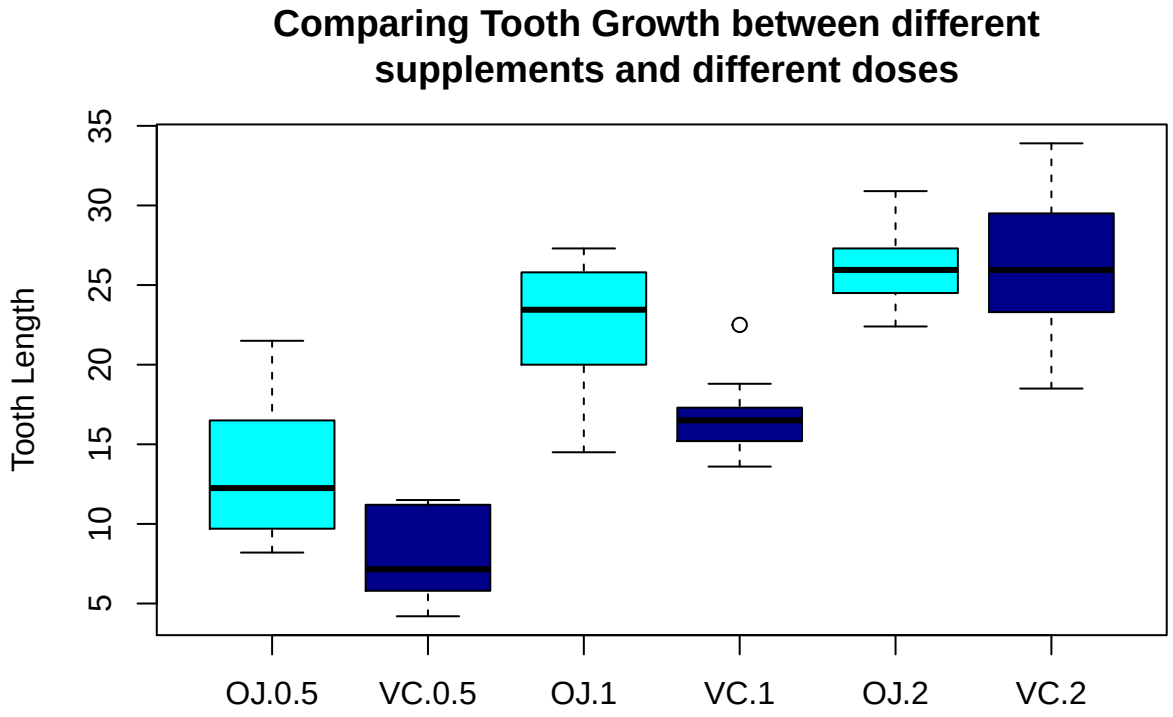
```
head(ToothGrowth) # first some rows from the dataframe
```

```
##    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
```

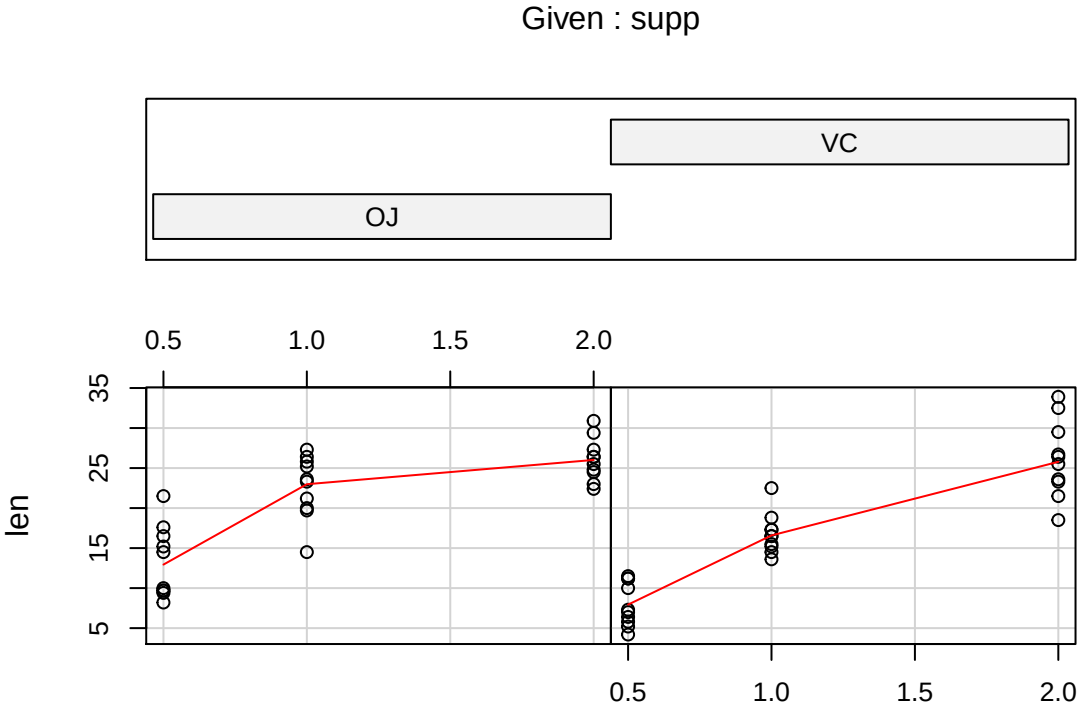
Descrition Dataframe consist of 60 observations on three variables: the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

1. len - numeric, Tooth length
2. supp - factor, Supplement type (VC or OJ)
3. dose - numeric, Dose in milligrams

```
# Boxplot
boxplot(len ~ supp * dose, data=ToothGrowth, ylab="Tooth Length",
        main="Comparing Tooth Growth between different \n supplements and different doses",
        col=c("cyan", "darkblue"))
```




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



Conclusion

As we can see there are some basic assumptions that we can draw after plotting.

1. Averages of tooth length seems to increase with the supplement doses. In other words, there seems to be a relationship between applying a supplement doses and the tooth growth
2. The tooth length averages for doses 0.5 and 0.1 differ with supplements. Both averages of supplement OJ are bigger than averages of supplemet VC.
3. The tooth lenght averages for dose 2.0 seem to be equal for the supplements, but it looks like variances are completely different.
4. OJ supplement is more effective than VC.

Provide a basic summary of the data

```
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
```

```
table(ToothGrowth$supp,ToothGrowth$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

Confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose

As the sample size is small I am going to use T-test to make a 95% confidence interval for the difference in tooth growth depending on supplement type. The assumption here is that this is a paired test so there is no other cause of tooth length differences other than supplement.

Null-hypothesis: OJ supplement has more impact than VC.

Confidence interval manual calculation. From the ***Lecture 8*** we know that the confidence interval for different groups must be calculated as follows:

- a $(1-\alpha)\times 100\%$ confidence interval for $\mu_y - \mu_x$ is

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$$

- The pooled variance estimator is

$$S_p^2 = \{(n_x-1)S_x^2 + (n_y-1)S_y^2\}/(n_x+n_y-2)$$

- Assuming a constant variance across the two groups

Let's calculate this manually.

```
lenOJ<-ToothGrowth[ToothGrowth$supp=="OJ",]$len
lenVC<-ToothGrowth[ToothGrowth$supp=="VC",]$len

nOJ <- length(lenOJ)
nVC <- length(lenVC)

# the pooled variance estimator for independent groups is:
sp <- sqrt(((nOJ - 1)*sd(lenOJ)^2 + (nVC-1)*sd(lenVC)^2) / (nOJ + nVC - 2))

# find out mean difference
mean_diff <- mean(lenOJ) - mean(lenVC)
semd <- sp * sqrt(1 / nOJ + 1/nVC)
mean_diff + c(-1, 1) * qt(.975, nVC + nOJ - 2) * semd
```

```
## [1] -0.1670064  7.5670064
```

Assuming that these two groups don't have constant variances, calculate the confidence interval and test the null-hypothesis again.

```
# As these two groups are independent and don't have the same variances
# we use both paried and var.equal to be FALSE
t.test(lenOJ,lenVC, paired=FALSE,var.equal=FALSE)$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

Conclusion

As these both intervals above cover 0, the assumption that the supplement OJ has more impact than VC on the length of tooth failed, i.e. ***Null-hypothesis is not true***.

Confidence intervals for particular doses

Let's calculate confidence interval for particular doses: 0.5, 1.0, 2.0 . We assume that groups are independent and don't have constant variances.

Dose 0.5 Null-hypothesis is that summplement OJ has more impact than VC on tooth length if using dose 0.5

```
lenOJ<-ToothGrowth[ToothGrowth$supp=="OJ" & ToothGrowth$dose == .5,]$len
lenVC<-ToothGrowth[ToothGrowth$supp=="VC" & ToothGrowth$dose == .5,]$len

t.test(lenOJ,lenVC, paired=FALSE,var.equal=FALSE)$conf.int
```

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

Conclusion

Confidence interval doesn't cover zero, hence Null-hypothesis is true if we use dose 0.5 .

Dose 1.0 Null-hypothesis is that summplement OJ has more impact than VC on tooth length if using dose 1.0

```
lenOJ<-ToothGrowth[ToothGrowth$supp=="OJ" & ToothGrowth$dose == 1.0,]$len
lenVC<-ToothGrowth[ToothGrowth$supp=="VC" & ToothGrowth$dose == 1.0,]$len

t.test(lenOJ,lenVC, paired=FALSE,var.equal=FALSE)$conf.int
```

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

Conclusion

Null-hypothesis is true if we use dose 1.0 .

Dose 2.0 Null-hypothesis is that summplement OJ has more impact than VC on tooth length if using dose 2.0

```
lenOJ<-ToothGrowth[ToothGrowth$supp=="OJ" & ToothGrowth$dose == 2.0,]$len
lenVC<-ToothGrowth[ToothGrowth$supp=="VC" & ToothGrowth$dose == 2.0,]$len

t.test(lenOJ,lenVC, paired=FALSE,var.equal=FALSE)$conf.int
```

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

Conclusion

Confidence interval does cover zero, hence Null-hypothesis is false if we use dose 2.0 .

Conclusion

1. We have strong statistical reasons to beleive that the supplement OJ is not more effective than VC on the whole.
2. In addition we have shown in the report that OJ supplement is more effective if we use doses 0.5 and 1.0 .