Johns Hopkins - Statistical inference

Benjamin Berhault

September 12, 2015

Basic inferential data analysis - Peer Assessment 2/2

Synopsis

We're going to analyze the ToothGrowth data in the R datasets package. In this dataset, the length of teeth is examined for 10 pigs at three Dose levels of two different supplements. The supplements are orange juice (OJ) or ascorbic acid (VC) which contain high level of Vitamin C.

URL

- Coursera - Johns Hopkins : Statistical Inference (https://class.coursera.org/statinference-032/)
- Code source can be found here : github.com/benjamin-berhault/statistical-inference (https://github.com/benjamin-berhault/statistical-inference)

Mission

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 Supplement 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 necessary library
library(datasets)
library(ggplot2)
library(gridExtra)
library(GGally)
library(data.table)
```

1. Load the ToothGrowth data and perform some basic exploratory data analyses

Data loading

```
# The Effect of Vitamin C on Tooth Growth in Guinea Pigs
data(ToothGrowth)
toothGrowth <- data.table(ToothGrowth)
setnames(toothGrowth, c('len','supp','dose'),c('Length','Supplement','Dose'))
toothGrowth$Dose <- as.factor(toothGrowth$Dose) # convert to factor
```

Basic Summary of the data

```
str(toothGrowth)
```

```
## Classes 'data.table' and 'data.frame':   60 obs. of  3 variables:
##  $ Length    : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ Supplement: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Dose      : Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(toothGrowth)
```

```
##      Length      Supplement  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
```

```
head(toothGrowth)
```

```
##    Length Supplement 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
```

Contingency table

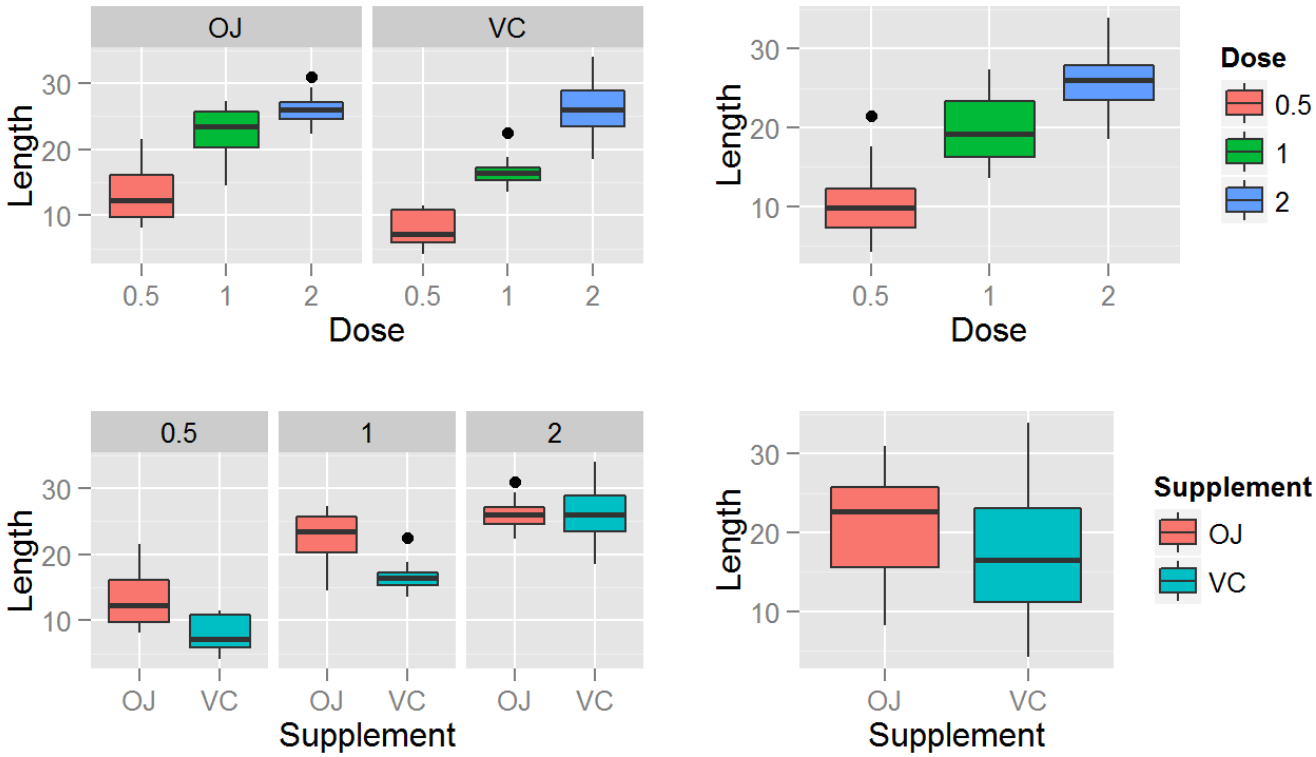
```
table(toothGrowth$Supplement, toothGrowth$Dose)
```

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

2. Provide a basic summary of the data.

Dose comparison

Supplements comparison



3. Use confidence intervals and/or hypothesis tests to compare tooth growth by Supplement and Dose. (Only use the techniques from class, even if there's other approaches worth considering)

```
dataVC <- toothGrowth[toothGrowth$Supplement=='VC',]
dataOJ <- toothGrowth[toothGrowth$Supplement=='OJ',]
lengthComparison <- (dataOJ$Length - dataVC$Length)
comp_data <- data.frame(lengthComparison, dataOJ$Dose)
names(comp_data) <- c("lengthComparison", "Dose")
```

Now, we have 2 subsets by kind of supplement tested :

- orange juice (dataVC)
- ascorbic acid (dataVC)

Dose : 0.5 mg

```
n <- subset(comp_data, Dose == 0.5)$lengthComparison
t.test(n)$conf
```

```
## [1] 1.263458 9.236542
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval $[1.26; 9.2]$ does not include zero, indicating that we should reject the null hypothesis that there is no difference between OJ and VC using a 0.5 mg dose, in favour of the alternative that there is a difference.

Dose : 1 mg

```
n <- subset(comp_data, Dose == 1)$lengthComparison
t.test(n)$conf
```

```
## [1] 1.951911 9.908089
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval $[1.95; 0.9]$ does not include zero, indicating that we should reject the null hypothesis that there is no difference between OJ and VC using a 1.0 mg dose, in favour of the alternative that there is a difference.

Dose : 2 mg

```
n <- subset(comp_data, Dose == 2)$lengthComparison
t.test(n)$conf
```

```
## [1] -4.328976  4.168976
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval is both positive and negative meaning that the two methods have similar effectiveness at dose == 2.0.

Analysis based on Analysis of Variance (ANOVA)

```
anova.out <- aov(Length ~ Supplement * Dose, data=toothGrowth)
summary(anova.out)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Supplement      1  205.4   205.4  15.572 0.000231 ***
## Dose            2 2426.4  1213.2  92.000  < 2e-16 ***
## Supplement:Dose  2  108.3    54.2   4.107 0.021860 *
## Residuals      54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The results show a strong relationship between the length of teeth and dosage ($p - value < 2^{-16} < 0.01$)
- Also a very clear effect on length by supplement type ($p - value < 0.000231 < 0.01$).
- Finally, a minor interaction between the combination of supplement type and dosage compared to the length ($p - value < 0.021860 < 0.05$).

Table means (Supplement x Dose)

```
print(model.tables(anova.out,"means"),digits=3)
```

```
## Tables of means
## Grand mean
##
## 18.81333
##
##  Supplement
## Supplement
##    OJ    VC
## 20.66 16.96
##
##  Dose
## Dose
##   0.5     1     2
## 10.60 19.73 26.10
##
##  Supplement:Dose
##           Dose
## Supplement 0.5   1     2
##         OJ 13.23 22.70 26.06
##         VC  7.98 16.77 26.14
```

4. State your conclusions and the assumptions needed for your conclusions.

We see clearly that each drugs have an effect concerning the pig teeth length. A greater quantity increase the size significantly. There is some kind of difference between the 2 drugs. 0.5 mg and 1 mg, orange juice supplements seems having a greater impact but the effect become similar for a quantity of 2 mg.

These assumptions are based on the facts :

- the guinea pigs are representative for the population of guinea pigs,
- dosage and supplement were randomly assigned