Final Project- Statistical Inference (Part 2)

06/27/2019

Part 2: Basic inferential data analysis.

Overview:

We're going to analyze the ToothGrowth data in the R datasets package. As the dataset shows, the length of teeth is examined for 60 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.

The following discussions will be addressed in this report:

1- Load the ToothGrowth data and perform some basic exploratory data analyses 2- Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. 3- State your conclusions and the assumptions needed for your conclusions.

```
# load needed libraries

library(datasets)
library(ggplot2)
library(data.table)
library(data.table)
```

Question 1: Load the ToothGrowth data and perform some basic exploratory data analyses

```
data(ToothGrowth)
ToothGrowth<- data.table(ToothGrowth)
#change the name of variables to make those more descriptive:
setnames(ToothGrowth, c('len','supp','dose'),c('Length','Supplement','Dose'))
#convert the type of dose variable to factor!
#ToothGrowth$Dose <- as.factor(ToothGrowth$Dose)
```

Some basic Summary of data table

```
# Structute of variables in data table
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      : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
# See the first 10 rows of data table
head(ToothGrowth,10)
```

```
##     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
##  7:   11.2         VC  0.5
##  8:   11.2         VC  0.5
##  9:    5.2         VC  0.5
## 10:    7.0         VC  0.5
```

```r
# See the summary of data table
summary(ToothGrowth)
```

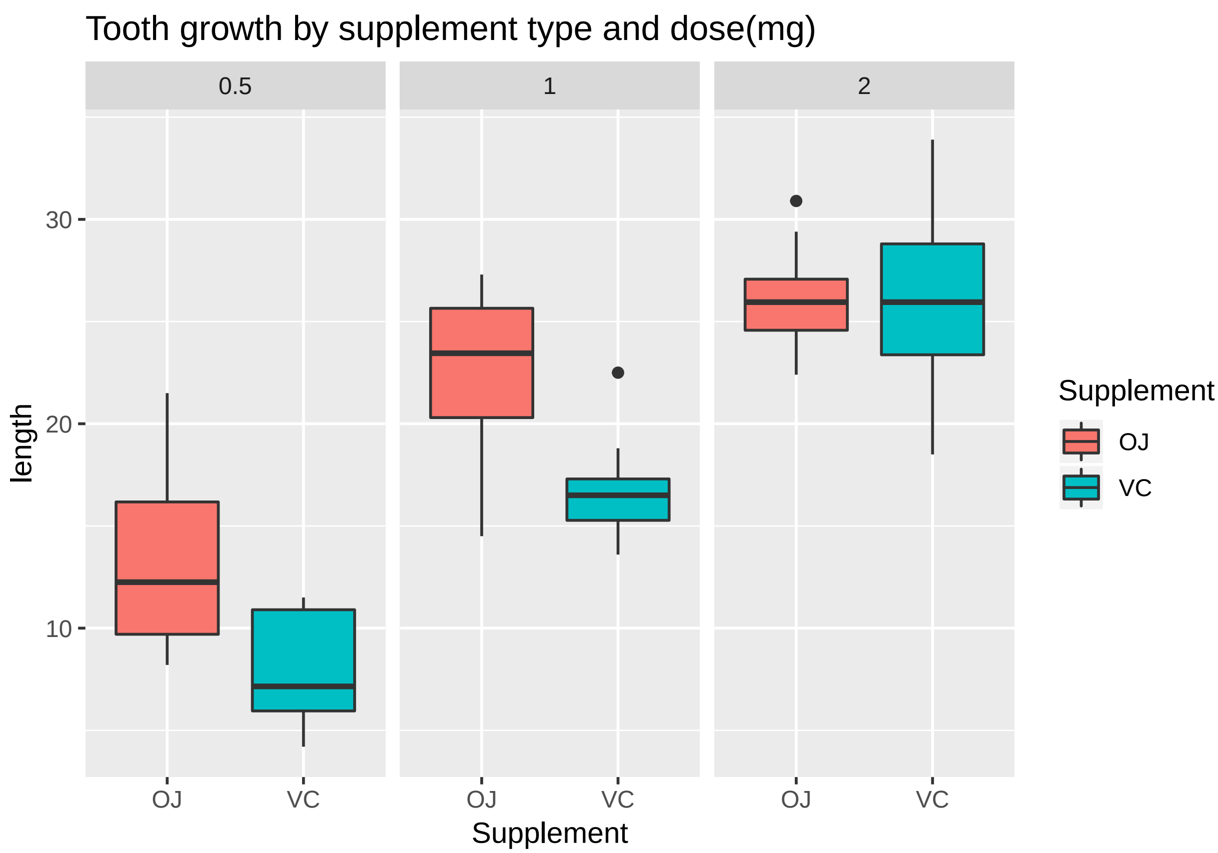
```
##      Length      Supplement      Dose      
##  Min.   : 4.20   OJ:30      Min.   :0.500  
##  1st Qu.:13.07   VC:30      1st Qu.:0.500  
##  Median :19.25              Median :1.000  
##  Mean   :18.81              Mean   :1.167  
##  3rd Qu.:25.27              3rd Qu.:2.000  
##  Max.   :33.90              Max.   :2.000  
```

```r
# To gain an overall summary of the original data, we create a contingency table from the original data
table(ToothGrowth$Supplement, ToothGrowth$Dose)
```

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

Perform basic exploratory data analysis

```r
ggplot(aes(x = Supplement, y = Length), data = ToothGrowth) +
    geom_boxplot(aes(fill = Supplement)) + facet_wrap(~ Dose)+ labs(title="Tooth growth by supplement ty
```



Summary of the boxplot:

Based on this basic exploratory data analysis using the boxplot, the dosage appears to significantly affect tooth length. As, the plot shoes the more dosage is co-related to more tooth growth. While the other variable (supplement) is less clear to see its impact to tooth growth. The more Vitamin C given as a supplement, the longer the teeth grew. However, at low dose levels (0.5 and 1.00), orange juice appears to correlate with longer teeth, but at 2.0mg there is no difference between OJ and VC!

Question 2: Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

Based on our observation from exploratory data analysis, we think it is worthful to test the tooth growth as function of both supplement and dosage. Before constructing our hypothesis tests the following assumptions are taken:

1- The variables are independent and identically distributed (iid).

2- Variances of the distributions of tooth growth according to each supplement and dosage are different.

3- Tooth growth observations are normal distributed.

Based on the above assumptions and the fact that number of observations are relatively small (n=60), We will use the t-test for testing hypotheses. The level of significancy is set as alpha=0.05.

Testing influence on tooth growth according to supplement and dose:

1- H0: mu(VC, 0.5) = mu(OJ, 0.5)

2- H0: mu(VC, 1.0) = mu(OJ, 1.0)

3- H0: mu(VC, 2.0) = mu(OJ, 2.0)

For each of these hypothesis the H0 will be tested against the respective alternate hypothesis (~=). So that the two.sided t-test will be run for each case.

1: H0: mu(VC, 0.5) = mu(OJ, 0.5)

```r
VC1<-ToothGrowth[which(ToothGrowth$Dose==0.5 & ToothGrowth$Supplement=='VC'),]$Length
OJ1<-ToothGrowth[which(ToothGrowth$Dose==0.5 & ToothGrowth$Supplement=='OJ'),]$Length
t.test(VC1, OJ1, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC1 and OJ1
## 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:
##  -8.780943 -1.719057
## sample estimates:
## mean of x mean of y
##      7.98     13.23
```

Result: The p-value=0.006 smaller than alpha=0.05. And as we can see the 95% confidence interval doesn't contain 0 so **REJECT** the null hypothesis.

2: H0: mu(VC, 1.0) = mu(OJ, 1.0)

```r
VC2<-ToothGrowth[which(ToothGrowth$Dose==1.0 & ToothGrowth$Supplement=='VC'),]$Length
OJ2<-ToothGrowth[which(ToothGrowth$Dose==1.0 & ToothGrowth$Supplement=='OJ'),]$Length
t.test(VC2, OJ2, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC2 and OJ2
## 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:
##  -9.057852 -2.802148
## sample estimates:
## mean of x mean of y
##     16.77     22.70
```

Result: The p-value=0.001 smaller than alpha=0.05. And as we can see the 95% confidence interval doesn't contain 0 so **REJECT** the null hypothesis.

3: H0: mu(VC, 2.0) = mu(OJ, 2.0)

```r
VC3<-ToothGrowth[which(ToothGrowth$Dose==2.0 & ToothGrowth$Supplement=='VC'),]$Length
OJ3<-ToothGrowth[which(ToothGrowth$Dose==2.0 & ToothGrowth$Supplement=='OJ'),]$Length
t.test(VC3, OJ3, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC3 and OJ3
## 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:
##  -3.63807  3.79807
## sample estimates:
## mean of x mean of y
##     26.14     26.06
```

Result: The p-value=0.96 greater than alpha=0.05. And as we can see the 95% confidence interval contains 0 so we **FAIL to REJECT** the null hypothesis. Therefore, using Dose=2.0, the effect of OJ compare to VC on tooth growth is not distingushable! as oppose to using Dose=0.5 and Dose=1.0, as the above two hypotheses show, OJ is correlated to more tooth growth compare with VC.

Testing the impact of different doses

We set the null hypothesis (H0) as to be: the dose has no influence and the tooth growth is always the sam, and the alternative hypothesis (Ha) as higher dose results in an greater tooth growth.

1- compare the 0.5mg and 1.0mg dose

```r
#H0: 0.5 = 1.0
#Ha: 0.5 < 1.0
d05 = ToothGrowth$Length[ToothGrowth$Dose == 0.5]
d1 = ToothGrowth$Length[ToothGrowth$Dose == 1.0]
t.test(d05, d1, alternative = "less", paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  d05 and d1
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Result: According to the realy small p-value = 6.342e-08, we have to reject the null hypothesis and conclude that the 0.5mg dose results in less tooth growth than the 1.0mg dose.

2- compare the 1.0mg and 2.0mg dose

```r
#H0: 1.0 = 2.0
#Ha: 1.0 < 2.0
d1 = ToothGrowth$Length[ToothGrowth$Dose == 1.0]
```

```r
d2 = ToothGrowth$Length[ToothGrowth$Dose == 2.0]
t.test(d1, d2, alternative = "less", paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  d1 and d2
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

Result: According to the realy small p-value = 9.532e-06, we have to reject the null hypothesis and conclude that the 1.0mg dose results in less tooth growth than the 2.0mg dose.

Conclusion:

With alpha=0.05 as the level of significance, we can conclude the followings:

1) In general, the OJ supplement results in a greater tooth growth. However, There is one exception: With the 2.0mg dose the supplement type dosn't make any difference.
2) A higher dose results in a greater tooth growth.