Statistical Inference - Data Analysis (Tooth Growth)

Jeff Sternin

January 4, 2018

1.Report

1.0 Synopsis

This report provides data analysis of the data set Guinea Pig Tooth Grows experiment.
It has exploratory data analysis through basic functions (dim, str, summary) and basic graphs.
Based on exploratory data analysis we do deeper analysis such as influence of factors "supp" and "dose" on length "len" of Guinea Pig Tooth Growth.
RStudio console ?ToothGrowth provides full description of data fields.
Appendix contains necessary code, results and plots.

1.1 Basic exploratory data analysis (EDA)

We start with basic summary of the data and based on it do some exploratory graphs.
Appendix 2.1 provides code and results.

1.1.1 Basic summary of the data

1. dim

- data set has 60 rows and 3 columns

2. summary

- columns are : len , supp (supplement), dose
- only 2 supplements VC and OJ - each 30 rows
- dose - has 3 options - 0.5,1.0,2.0 - each 10 rows

3. str

- confirms basic information from summary
  Appendix 2.1.1 contains code and results

1.1.2 Basic exploratory analysis - boxplot by dose

By fixing supplement and changing dose we can see how dose influence length. From the graph we can deduce that increasing of dose increase the length for both supplements - VC and OJ.
Appendix 2.1.2 contains code and plot.

1.1.3 Basic exploratory analysis - boxplot by supplement

By fixing dose we compare different supplements with the same dose. We can see that OJ (orange juice) is more effective than VC (vitamin C) with doses 0.5 and 1.0. For dose 2.0 both supplements approximately the same.
Appendix 2.1.3 contains code and plot.

1.2 Confidence intervals and hypothesis testing

Based on EDA we have 3 values of dose (0.5,1,2), and 2 values of supplement (OJ and VC). All together it is 3x2 matrix of factors.
We walk through factor matrix first by column comparing supplement (column) and then by row (same dos different supplement).
All processing done twice - first by code and then using t.test (as in lecture 8).
We make sure that they identical. Also calculated twice t-statistic and we test H0 hypothesis that means of 2 samples equal vs. they not.

1.2.1 Factor matrix (dose by supp)

We build 3x2 matrix and each cell contains sample (vector) of "len" filtered from original data set. This matrix makes easier to compare different slices of factors. Code is Appendix 2.1.1

1.2.2 Main processing function

Function process1 does main work. We pass both samples x,y also formatting string and name of constant factor (dose or supplement) and two names of factors we compare. Output of the function contains 3 lines
-Line 1: We calculate means of both samples and print them and their difference with the names of factors.
-Line 2: We calculate both standard deviations, combined standard deviation, t-statistic and threshold to test hypothesis H0 (means are equal) vs H1 (they not equal).
We calculate confidence interval as well. Line 2 starts with: My confidence interval; print t-stat, df - degrees of freedom , threshold and result of H0 test: Accepted or Rejected
-Line 3: Does the same calculations using t.test.
We print confidence interval (it is the same as in Line 2) t-stat and p-value.
Code for this function in Appendix 2.2.2

1.2.3 Compare samples for different dose with the same supplement

By scanning factor matrix by column (supplement) and by row within same column (by dose).
Calling main processing function we get 4 comparisons:
- supp VC; dose 1.0 vs 0.5 - H0 : mean equality rejected
- supp VC; dose 2.0 vs 1.0 - H0 : mean equality rejected
- supp OJ; dose 1.0 vs 0.5 - H0 : mean equality rejected
- supp OJ; dose 2.0 vs 1.0 - H0 : mean equality rejected

All hypothesis of H0 (means are equal) are rejected. It is clearly visible that larger dose give larger mean. Code and results are in Appendix 2.2.3

1.2.4 Compare samples for different supplements with the same dose

By scanning factor matrix by row (dose) and comparing different columns (supplement).
Calling main processing function we get 3 comparisons:
- dose 0.5; supp OJ vs VC - H0 : mean equality rejected
- dose 1.0; supp OJ vs VC - H0 : mean equality rejected
- dose 2.0; supp OJ vs VC - H0 : mean equality accepted

Dose 0.5 and 1.0 hypothesis of H0 (means are equal) are rejected. OJ mean is larger than VC mean.
Dose 2.0 H0 is accepted - means for OJ and VC are equal. Code and results are in Appendix 2.2.4

1.3 Conclusions.

Based on ToothGrowth exploratory data analysis, analysis of factors that influence length (len) of the tooth we come to the following conclusion.
Higher dose of supplement causes larger mean of the sample (different dose, same supplement).
All hypothesis of mean equality (error I alpha is 0.05 two sided rejected).
For doses of supplement 0.5 and 1.0 supplement "OJ" causes larger mean of length than VC.
For dose 2.0 difference between means of samples for supplement OJ and VC is insignificant (H0 hypothesis of equality is accepted)
Conclusions based on lectures material and the following link:
http://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm

2.Appendix

Appendix 2.1 Basic exploratory data analysis

Appendix 2.1.1 Basic summary of the data - dim,summary,str

```
library(datasets)
dim(ToothGrowth)
```

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

```
summary(ToothGrowth)
```

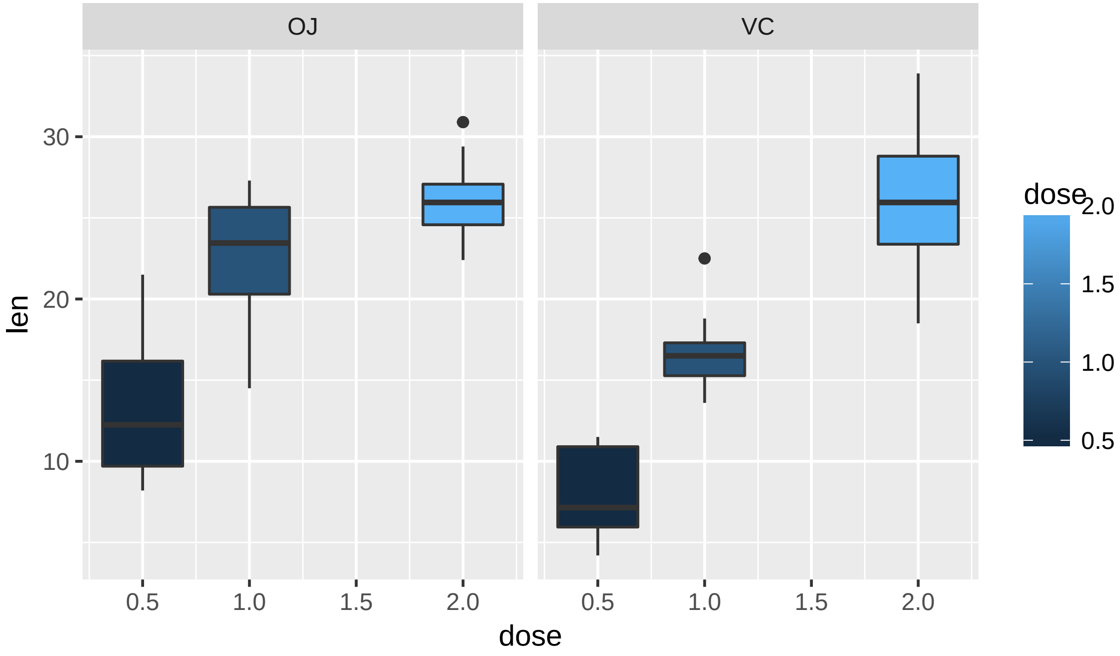
```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.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

```
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 ...
```

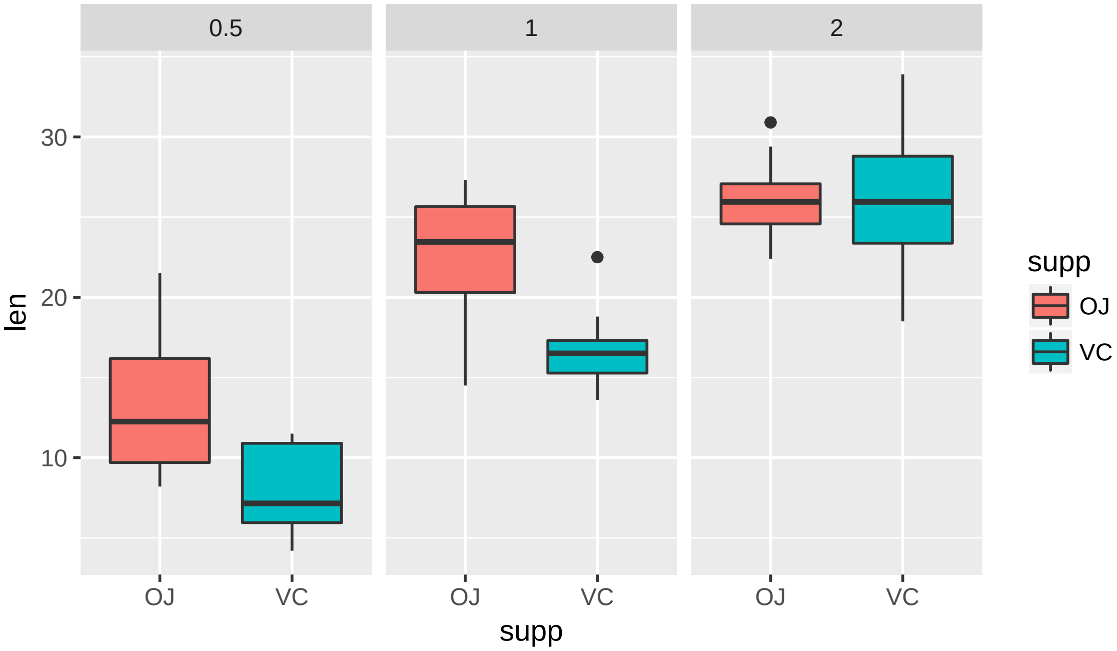
Appendix 2.1.2. Exploratory Data Analysis - plots by dose (same supplement)

```
library(ggplot2)
ggplot(ToothGrowth, aes(x=dose, y=len, group=dose)) + geom_boxplot(aes(fill=dose)) +  facet_grid(. ~ supp)
```



Appendix 2.1.3. Exploratory Data Analysis - plots by supplement (same dose)

```
ggplot(ToothGrowth, aes(x=supp, y=len, group=supp)) + geom_boxplot(aes(fill=supp)) +  facet_grid(. ~ dose)
```



Appendix 2.2 Confidence intervals and hypothesis testing

Appendix: 2.2.1 Building matrix of factors(dose by supp) with corresponding samples of tooth length

```
res <- matrix(data=data.frame(c(1:10)),nrow=3,ncol=2)
colnames(res) <- c("VC","OJ")
rownames(res) <- c(0.5,1.0,2.0)
for(i in 1:nrow(res))
  for(j in 1:ncol(res))
    res[[i,j]] <- as.numeric(filter(ToothGrowth,supp == colnames(res)[j] & dose == rownames(res)[i])$len)
```

Appendix 2.2.2 Function to calculate confidence intervals and hypothesis testing

```
process1<-function(x,y,sFormat,sConst,s1,s2) {
    m1<-mean(x)
    m2<- mean(y)
    n1 = length(x)
    n2=length(y)
    sd1 <- sqrt(((n1-1) *sd(x)^2 + (n2-1)*sd(y)^2)/(n1+n2-2)) * (sqrt(1/n1+1/n2))
    conf <- m1-m2 + c(-1,1)*qt(0.975,(n1+n2-2))*sd1
    tstat <- (m1-m2)/ sd1
    qt1 <- qt(0.975,(n1+n2-2))
    if (tstat>qt1) { sH0 <- "Rejected" } else {sH0 <- "Accepted" }
    s <- sprintf("%4.2f", t.test(x,y,conf.level = 0.95,var.equal=TRUE)$conf.int)
    sp <- sprintf("%7.6f", t.test(x,y,conf.level = 0.95,var.equal=TRUE)$p.value)
    st <- sprintf("%4.2f", t.test(x,y,conf.level = 0.95,var.equal=TRUE)$statistic)
    print("---------------------------------------------------------------------------------------------- ")
    print(sprintf(sForm,sConst,s1,m1,s2,m2,m1-m2))
    print(sprintf("My conf interval: (%4.2f: %4.2f) t-stat:%4.2f df:%2.0f qt(0.975,df):%4.2f H0: %s",
                  conf[1],conf[2],tstat,(n1+n2-2),qt1,sH0))
    print(sprintf("t.test: mean equal alpha=.05%%, two sided, conf.int (%s:%s), t-stat:%s,p-val:%s",
                 s[1],s[2],st[1],sp[1]))
  return
}
```

Appendix 2.2.3 Comparison for different dose with the same supplemet.

```
## compare same supp with different dose

for(k in 1:ncol(res))
  for(i in 2:nrow(res)) {
    x<- res[[i,k]]
    y<- res[[i-1,k]]
    sForm<-"Compare suppplement %s, dose :%s (mean:%5.2f) vs dose:%s (mean:%4.2f) (mean diff:%5.2f)"
    process1(x,y,sForm,colnames(res)[k],rownames(res)[i],rownames(res)[i-1])
}
```

```
## [1] "---------------------------------------------------------------------------------------------- "
## [1] "Compare suppplement VC, dose :1 (mean:16.77) vs dose:0.5 (mean:7.98) (mean diff: 8.79)"
## [1] "My conf interval: (6.32: 11.26) t-stat:7.46 df:18 qt(0.975,df):2.10 H0: Rejected"
## [1] "t.test: mean equal alpha=.05%, two sided, conf.int (6.32:11.26), t-stat:7.46,p-val:0.000001"
## [1] "---------------------------------------------------------------------------------------------- "
## [1] "Compare suppplement VC, dose :2 (mean:26.14) vs dose:1 (mean:16.77) (mean diff: 9.37)"
## [1] "My conf interval: (5.77: 12.97) t-stat:5.47 df:18 qt(0.975,df):2.10 H0: Rejected"
## [1] "t.test: mean equal alpha=.05%, two sided, conf.int (5.77:12.97), t-stat:5.47,p-val:0.000034"
## [1] "---------------------------------------------------------------------------------------------- "
## [1] "Compare suppplement OJ, dose :1 (mean:22.70) vs dose:0.5 (mean:13.23) (mean diff: 9.47)"
## [1] "My conf interval: (5.53: 13.41) t-stat:5.05 df:18 qt(0.975,df):2.10 H0: Rejected"
## [1] "t.test: mean equal alpha=.05%, two sided, conf.int (5.53:13.41), t-stat:5.05,p-val:0.000084"
## [1] "---------------------------------------------------------------------------------------------- "
## [1] "Compare suppplement OJ, dose :2 (mean:26.06) vs dose:1 (mean:22.70) (mean diff: 3.36)"
## [1] "My conf interval: (0.22: 6.50) t-stat:2.25 df:18 qt(0.975,df):2.10 H0: Rejected"
## [1] "t.test: mean equal alpha=.05%, two sided, conf.int (0.22:6.50), t-stat:2.25,p-val:0.037363"
```

Appendix 2.2.4 Comparison: different supplements with the same dose.

```
for(i in 1:nrow(res)) {
   x<- res[[i,"OJ"]]
   y<- res[[i,"VC"]]
   sForm<-"Compare dose :%s supplement: %s (mean:%5.2f) vs %s (mean:%5.2f) (mean diff:%4.2f) "
   process1(x,y,sForm,rownames(res)[i],"OJ","VC")

}
```

```
## [1] "---------------------------------------------------------------------------------------------- "
## [1] "Compare dose :0.5 supplement: OJ (mean:13.23) vs VC (mean: 7.98) (mean diff:5.25) "
## [1] "My conf interval: (1.77: 8.73) t-stat:3.17 df:18 qt(0.975,df):2.10 H0: Rejected"
## [1] "t.test: mean equal alpha=.05%, two sided, conf.int (1.77:8.73), t-stat:3.17,p-val:0.005304"
## [1] "---------------------------------------------------------------------------------------------- "
## [1] "Compare dose :1 supplement: OJ (mean:22.70) vs VC (mean:16.77) (mean diff:5.93) "
## [1] "My conf interval: (2.84: 9.02) t-stat:4.03 df:18 qt(0.975,df):2.10 H0: Rejected"
## [1] "t.test: mean equal alpha=.05%, two sided, conf.int (2.84:9.02), t-stat:4.03,p-val:0.000781"
## [1] "---------------------------------------------------------------------------------------------- "
## [1] "Compare dose :2 supplement: OJ (mean:26.06) vs VC (mean:26.14) (mean diff:-0.08) "
## [1] "My conf interval: (-3.72: 3.56) t-stat:-0.05 df:18 qt(0.975,df):2.10 H0: Accepted"
## [1] "t.test: mean equal alpha=.05%, two sided, conf.int (-3.72:3.56), t-stat:-0.05,p-val:0.963710"
```