Statistical Inference Course : Project 2

Han Yunxi

18 January 2015

Analysis of ToothGrowth data in R dataset

This project requires us to load the **ToothGrowth** data from the default dataset in R and perform to some basic exploratory data analyses and provide a basic summary of the data.

Basic Data Exploration & Summary

For basic data exploration, We first load the **R** library **ToothGrowth** dataset and use the command `library(help = "datasets")` to understand the dataset. From the description, we understand the dataset title is: **The Effect of Vitamin C on Tooth Growth in Guinea Pigs**.

```
library(datasets)
str(ToothGrowth)      #structure of the dataset
```

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

```
summary(ToothGrowth)  #table summary
```

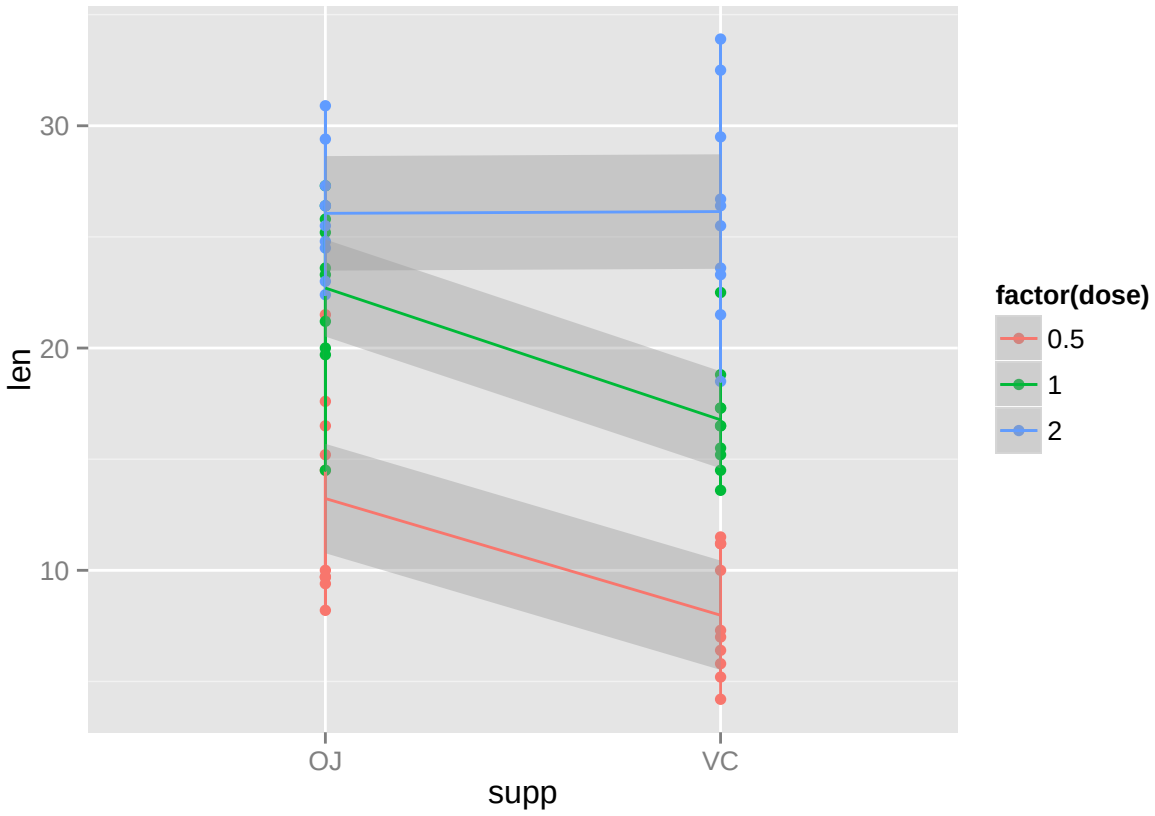
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

The dataset is a [60 rows by 3 columns] dataframe and contains information on (a) Length of tooth - `len`, (b) Supplement Type - `supp` and (c) Dosage given - `dose`.

In addition, we find that the maximum tooth length is *33.9* and the minimum is *4.2*, with the average tooth growth length at *18.81*. There are only 2 types of supplement (Orange Juice or Vitamin C). For the dosage level, we find that there are only 3 levels of dosage (0.5,1 and 2).

We also plot the dataset below to understand the effects of supplement and dosage on tooth growth. From the plot, we can see that Vitamin C results in larger range of tooth growth **(i.e. larger variance)** and Orange Juice gives limited but more predictable growth **(i.e. more points centred around a particular region)**

```
library(ggplot2)
qplot(supp, len, data=ToothGrowth, colour=factor(dose)) +
geom_line(data=ToothGrowth) +
geom_smooth(aes(group=dose), method="lm")
```



Statistical Inference

For the next section, we use Student's t-test for paired samples to compare tooth growth by supplement and dose. The aim is to see if for the same dosage of 2 different supplements, there was an improvement, deterioration, or if the means of tooth length have remained substantially the same (hypothesis H_0).

Assumption For small samples of only 10, we assume that Gosset's t distribution. We also assume the distribution of the data is roughly symmetric and mound shaped.

Paired observations are often analyzed using the t interval by taking differences. We assume that the ToothGrowth dataset is paired, i.e. OJ and VC is given to same 10 test subjects or between pairs of guinea pigs matched into meaningful groups.

Degree of freedom used is *(n-1)* or 9.

R function `myCI` takes in the *dosage* parameter and calculates the 95% confidence interval of the tooth growth length. The function is as seen below.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
myCI <- function(dosage){
  vitaminGroup <- filter(ToothGrowth, supp == "VC", dose == dosage)
  orangeGroup <- filter(ToothGrowth, supp == "OJ", dose == dosage)

  #Method 1: Using Mathematical Formula
  difference <- orangeGroup$len - vitaminGroup$len
  mn <- mean(difference)
  s <- sd(difference)
  n <- 10

  CI <- round(mn + c(-1, 1) * qt(.975, n-1) * s / sqrt(n) , 2)
  cat("We are 95% confident that for dosages of", dosage ,"mg
  Orange Juice supplements affects tooth growth than Vitamin C by (", CI , ")mm.")

  #Method 2: Using R
  t <- t.test(orangeGroup$len, vitaminGroup$len, paired=TRUE, var=FALSE,
              conf.level = 0.95)
  t
}
```

```
myCI(0.5)
```

Results - Confidence Intervals for 0.5, 1.0 and 2.0 mg dosage

```
## We are 95% confident that for dosages of 0.5 mg
##     Orange Juice supplements affects tooth growth than Vitamin C by ( 1.26 9.24 )mm.

##
##  Paired t-test
##
## data:  orangeGroup$len and vitaminGroup$len
## t = 2.9791, df = 9, p-value = 0.01547
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.263458 9.236542
## sample estimates:
## mean of the differences
##                    5.25
```

```
myCI(1.0)
```

```
## We are 95% confident that for dosages of 1 mg
##     Orange Juice supplements affects tooth growth than Vitamin C by ( 1.95 9.91 )mm.

##
```

```
##  Paired t-test
##
## data:  orangeGroup$len and vitaminGroup$len
## t = 3.3721, df = 9, p-value = 0.008229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.951911 9.908089
## sample estimates:
## mean of the differences
##                    5.93
```

```
myCI(2.0)
```

```
## We are 95% confident that for dosages of 2 mg
##     Orange Juice supplements affects tooth growth than Vitamin C by ( -4.33 4.17 )mm.
```

```
##
##  Paired t-test
##
## data:  orangeGroup$len and vitaminGroup$len
## t = -0.0426, df = 9, p-value = 0.967
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of the differences
##                   -0.08
```

Conclusion There are 2 ways to interpret the results, by p-values or by t-value.

First, if the p-value is greater than 0.05, then we can accept the hypothesis H_0 there is no difference in tooth length be it taking Vitamin C or Orange Juice.

Secondly, we can calculate the t-tabulated value: `{r} qt(0.975, 9)`. If (t-computed $<$ t-tabulated), so we accept the null hypothesis H_0

In conclusion, Orange Juice has significant improvement to tooth growth over Vitamin C at doses of 0.5 and 1.0 mg (reject H_0), whereas there is no significant difference at 2.0mg(accept H_0). This is in line with the results in the graph on page 1.

Reference

http://docs.ggplot2.org/current/geom_smooth.html http://www.r-bloggers.com/paired-students-t-test/