Statistical Inference - Basic Inferential Data Analysis Project - Part 2

Pierre Baudin

February 8, 2017

Basic Inferential Data Analysis Instructions

In this project, we're going to analyze the ToothGrowth data in the R datasets package.

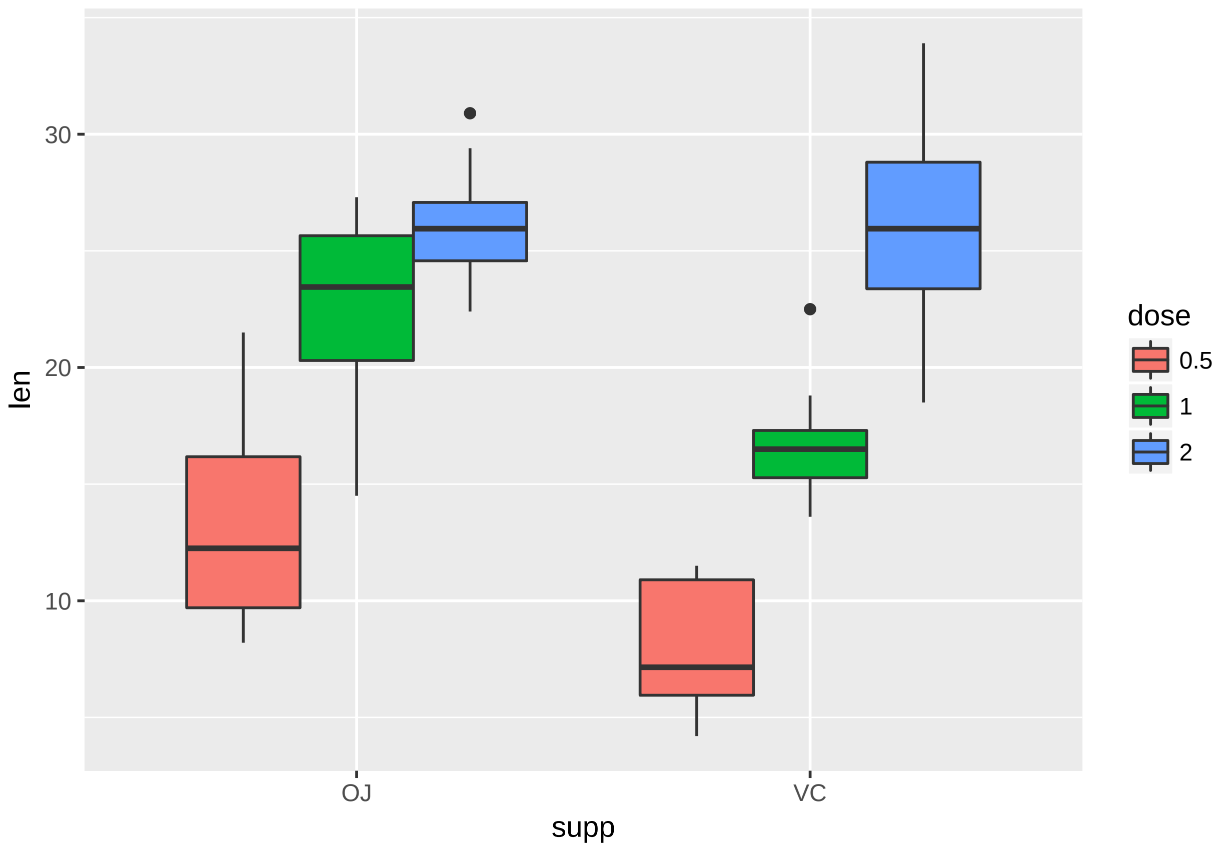
Load the ToothGrowth data and perform some basic exploratory data analyses

Note: The dataset contains 60 observations and three variables where the response is the tooth length 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 (OJ) or ascorbic acid (VC)).

```r
# load packages
library(dplyr)
library(datasets)
library(ggplot2)

# Import dataset
toothdf <- ToothGrowth

# For easier handling, convert the dose into factor
toothdf$dose <- as.factor(toothdf$dose)

# Add ID to each observation (10 observation for each dose)
toothdf[,"ID"] <- rep(1:10,6)

# Basic exploratory data analysis (class and observations)
glimpse(toothdf)
```

```
## Observations: 60
## Variables: 4
## $ len  <dbl> 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5.2, 7.0, 16....
## $ supp <fctr> VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, VC, ...
## $ dose <fctr> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1, 1, ...
## $ ID   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9,...
```

```r
# Basic graph to understand the data
ggplot(toothdf, aes(x = supp, y = len, fill = dose)) +
      geom_boxplot()
```



Basic summary of the data.

```
summary(toothdf)
```

```
##       len        supp     dose          ID      
##  Min.   : 4.20   OJ:30   0.5:20   Min.   : 1.0  
##  1st Qu.:13.07   VC:30   1  :20   1st Qu.: 3.0  
##  Median :19.25           2  :20   Median : 5.5  
##  Mean   :18.81                    Mean   : 5.5  
##  3rd Qu.:25.27                    3rd Qu.: 8.0  
##  Max.   :33.90                    Max.   :10.0
```

Confidence intervals and t-test

Comparison of tooth growth by supp

```
# two-sided t.test for len vs supp
# here we assume that the variables are paired
t.test(toothdf$len ~ toothdf$supp, alternative = "two.sided",
       paired = TRUE, var.equal = FALSE, conf.level = 0.95)
```

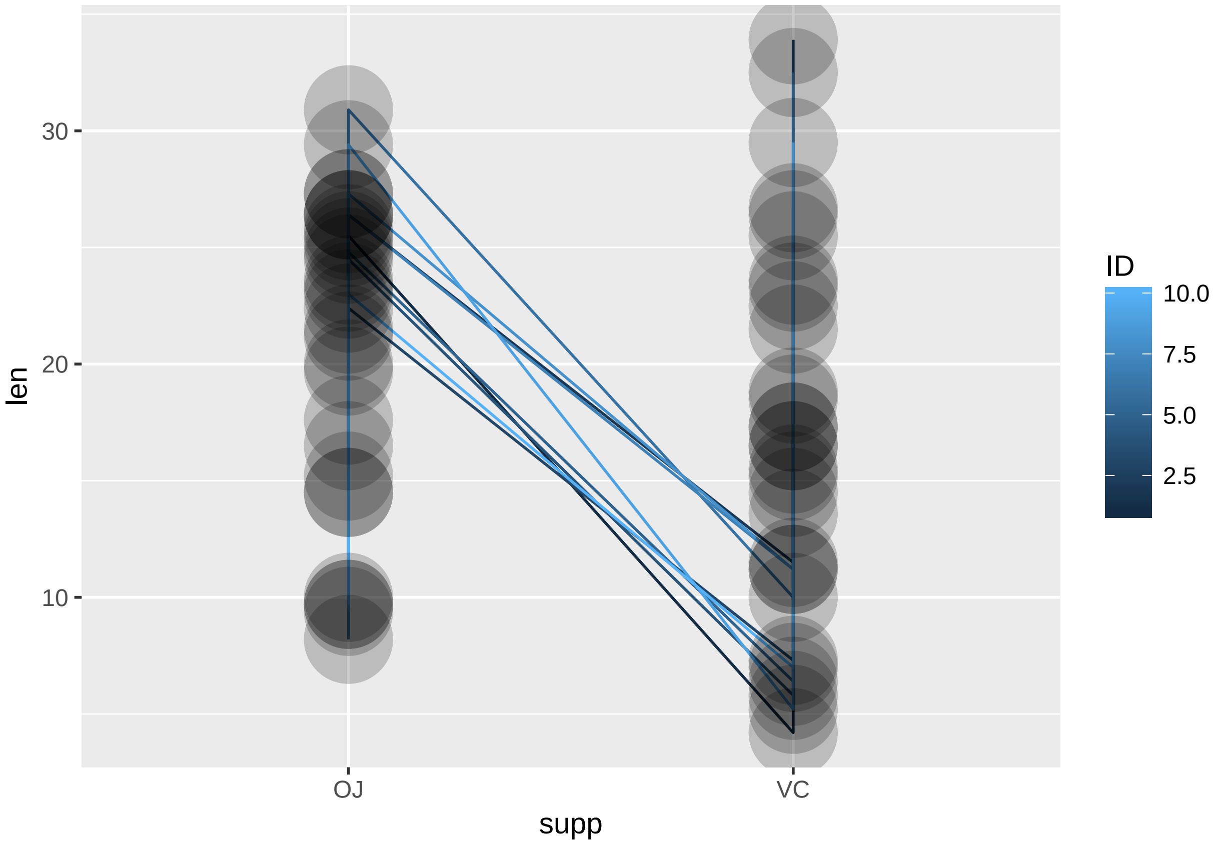
```
## 
##  Paired t-test
```

```
##
## data:  toothdf$len by toothdf$supp
## t = 3.3026, df = 29, p-value = 0.00255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.408659 5.991341
## sample estimates:
## mean of the differences
##                     3.7
```

The entire confidence interval is above zero which means that OJ has more effect on Tooth Growth for the rats as compared to VC.

Below is a visualization of the data used in this t-test:

```
# Plot of the observation
ggplot(toothdf, aes(x = supp, y = len, group=factor(ID))) +
      geom_line(aes(color=ID)) +
      geom_point(size = 15, pch = 19, fill = "steelblue", alpha = 0.2)
```



Comparison of tooth growth by dose

```
# two-sided t.test for dose 1 and 0.5
# we assume that the variables are paired
t.test(x = toothdf[toothdf$dose == "1", ]$len,
       y = toothdf[toothdf$dose == "0.5", ]$len, paired = TRUE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  toothdf[toothdf$dose == "1", ]$len and toothdf[toothdf$dose == "0.5", ]$len
## t = 6.9669, df = 19, p-value = 1.225e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.387121 11.872879
## sample estimates:
## mean of the differences
##                    9.13
```

```
# two-sided t.test for dose 2 and 1
# we assume that the variables are paired
t.test(x = toothdf[toothdf$dose == "2", ]$len,
       y = toothdf[toothdf$dose == "1", ]$len, paired = TRUE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  toothdf[toothdf$dose == "2", ]$len and toothdf[toothdf$dose == "1", ]$len
## t = 4.6046, df = 19, p-value = 0.0001934
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.471814 9.258186
## sample estimates:
## mean of the differences
##                   6.365
```

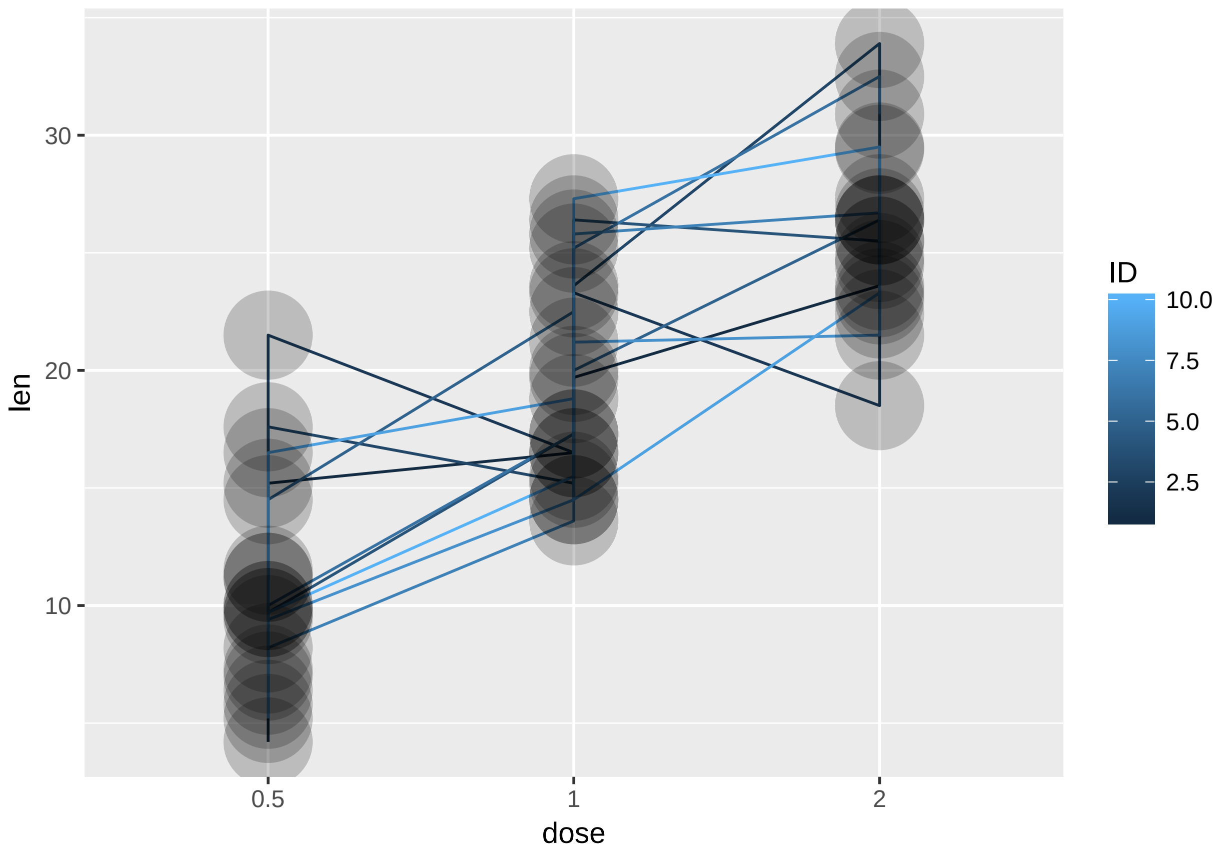
```
# two-sided t.test for dose 2 and 0.5
# we assume that the variables are paired
t.test(x = toothdf[toothdf$dose == "2", ]$len,
       y = toothdf[toothdf$dose == "0.5", ]$len, paired = TRUE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  toothdf[toothdf$dose == "2", ]$len and toothdf[toothdf$dose == "0.5", ]$len
## t = 11.291, df = 19, p-value = 7.19e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.6228 18.3672
## sample estimates:
## mean of the differences
##                  15.495
```

The three t-tests have been performed for each set of dosage (0.5, 1 and 2). We can see that all the resulting confidence intervals are above zero. This mean that an increase in dosage result in higher tooth growth in the rat.

Below is a visualization of the data used in this t-tests:

```
# Plot of the observation
ggplot(toothdf, aes(x = dose, y = len, group=factor(ID))) +
      geom_line(aes(color=ID)) +
      geom_point(size = 15, pch = 19, fill = "steelblue", alpha = 0.2)
```



Conclusions and assumptions

In conclusion, the analysis shows that the orange juice is more effective for the teeth growth in the rat. In addition, we presented an analysis showing that the higher the dosage, no matter the supplement, the higher the growth rate. We assume that the values are paired for the t-test.