Statistical Inference Project - Part2

A. Swarup

May 20, 2018

Overview

In this second part of the Coursera Statistical Inference Course project, we'd analyze the ToothGrowth data in the R datasets package. We'd Provide a basic summary of the data. Next, using confidence intervals and/or hypothesis tests we'd compare tooth growth by supplement (delivery method) and dose. We'd only be using the techniques from class, even if there are other approaches worth considering. We'd then present the conclusions and assumptions used therein.

Load and visualize the data

Prepare Environment and Load the ToothGrowth data

We use following code to load the ToothGrowth dataset in our work environment.

```
rm(list=ls())
library(knitr)
library(datasets)
library(ggplot2)
data("ToothGrowth")
```

We know from R Help that the ToothGrowth is a data set of 60 observations, which represent the response in the length of odontoblasts (teeth), 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 or ascorbic acid). An Example taken from Help shows a plot of the data as Figure 1 in the Appendix.

Basic Summary of Data

Preliminary Look section in the Appendix lists commands to take a preliminary look at the data. ToothGrowth data structure and summary overview show that:

1. The data set has 60 observations of 3 variables, representing **length** (len), **supplement** (supp) and **dosage** (dose). Variables *len* and *dose* are numeric, while *supp* is a factor variable.

2. Summary statistics shows that:
   - Variable *len* has a max value - 33.9, min value - 4.2 and mean - 18.8133.
   - Variable *supp* has only two unique values (delivery types): **OJ** (Orange Juice) and **VC** (Vitamin C), with 30 observations each.
   - Variable *dose* has a max value - 2, min value - 0.5 and mean - 1.1667.
     Furthermore, since *dose* has only three unique values, **0.5**, **1** & **2**, it can also be converted to a factor variable.

Exploratory Data Analysis

We can visualize the dataset using ggplot. Since we are to analyze tooth growth by supplement and dosage, we plot in Figure 2 of the Appendix the Tooth Length as a function of Delivery Method and Dosage.

From the plot of Figure 2 we observe the following: + Tooth length increases with dosage for all dose levels (0.5 -> 1 -> 2). + Furthermore, we notice that the supplement type OJ increases tooth length more than VC when dose amounts are 0.5 and 1.0 mg/day. + However, with a dose of 2.0 mg/day, the tooth growths for the two supplement types seem almost equal.

We would investigate whether the observations drawn here are statistically significant using hypothesis test.

Analysis of ToothGrowth data:

Analysis using confidence intervals and/or hypothesis tests to compare tooth growth by supplement and dose

1. t-test of supplement types

```
t.test(len ~ supp, paired = F, var.equal = F, data =ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

The p-value is 0.06. This indicates that we cannot reject the null hypothesis. In other words different supplement types (or delivery methods) have no effect on tooth growth. However, we need to analyze this further.

2. t-test for effect of dosage amounts

Evaluate impact of control variable *dose* on target variable *len*. Hypotheses are: Higher the dose, higher is the impact (i.e., higher *len* or more tooth growth)

- Subset data

```
dose_set1 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
dose_set2 <- subset(ToothGrowth, dose %in% c(0.5, 2.0))
dose_set3 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))
```

- Perform t-tests - for the sake of brevity, we hide full outputs here

```
td1 <- t.test(len ~ dose, paired=F, var.equal=F, data=dose_set1)
td2 <- t.test(len ~ dose, paired=F, var.equal=F, data=dose_set2)
td3 <- t.test(len ~ dose, paired=F, var.equal=F, data=dose_set3)
```

For the above three cases, respective hypotheses being:
a. Dosage = 1 has higher impact than dosage = 0.5
b. Dosage = 2 has higher impact than dosage = 0.5
c. Dosage = 2 has higher impact than dosage = 1

we find that the p-values are:

```
td1$p.value; td2$p.value;  td3$p.value
```

```
## [1] 1.268301e-07
```

```
## [1] 4.397525e-14
```

```
## [1] 1.90643e-05
```

1.27E-07, 4.40E-14, and 1.91E-05, respectively. These are all less than α=0.05. So in all three cases we can reject the null hypothesis, i.e., fail to reject the hypotheses. In other words, increasing the dose level leads to increased tooth growth.

3. t-tests for effect of supplement-type for the three dosage levels

- Subset data

```
dose1 <- subset(ToothGrowth, dose == 0.5)
dose2 <- subset(ToothGrowth, dose == 1)
dose3 <- subset(ToothGrowth, dose == 2)
```

- Perform t-tests - again, we hide output details here

```
td1s <- t.test(len ~ supp, data=dose1) # Small Dosage = 0.5
td2s <- t.test(len ~ supp, data=dose2) # Medium Dosage = 1
td3s <- t.test(len ~ supp, data=dose3) # Large Dosage = 2
```

The p-values from above three tests are:

```
td1s$p.value; td2s$p.value;  td3s$p.value
```

```
## [1] 0.006358607
```

```
## [1] 0.001038376
```

```
## [1] 0.9638516
```

We see above that the p-values for three dosage levels of 0.5, 1.0, and 2.0 are .006, .001 and .964, respectively. For the first two cases, they are less than α=0.05, whereas for the third case greater than α=0.05. This confirms our exploratory observation that for dosage amounts 0.5 and 1, tooth growth is higher for the delivery method VC than for the delivery method OJ, whereas for the dosage amount of 2, supplement type does not make a difference in tooth growth, i.e., either OJ (Orange Juice) or VC (Vitamin C) could be used.

Conclusion and Assumptions

- For both the supplement types 'Orange Juice' and 'Vitamin C', the tooth length of guinea pigs increases with dosage: 0.5 -> 1 -> 2
- For dosage levels 0.5 and 1, Orange Juice causes greater increase in the tooth length than Vitamin C. For dosage level 2, there are so such difference between Orange Juice and Vitamin C.

Assumptions for this analysis are as follows:

- Guniea pigs were randomly assigned to a combination of dosage and treatment type, allowing us to treat samples as independent.
- Data follows t-distribution, as the observations are limited.
- The data of 60 samples for 10 guinea pigs is assumed to be representative of all guinea pigs, so as to generalize the conclusions to the population.

APPENDIX

Basic Plot of ToothGrowth Data

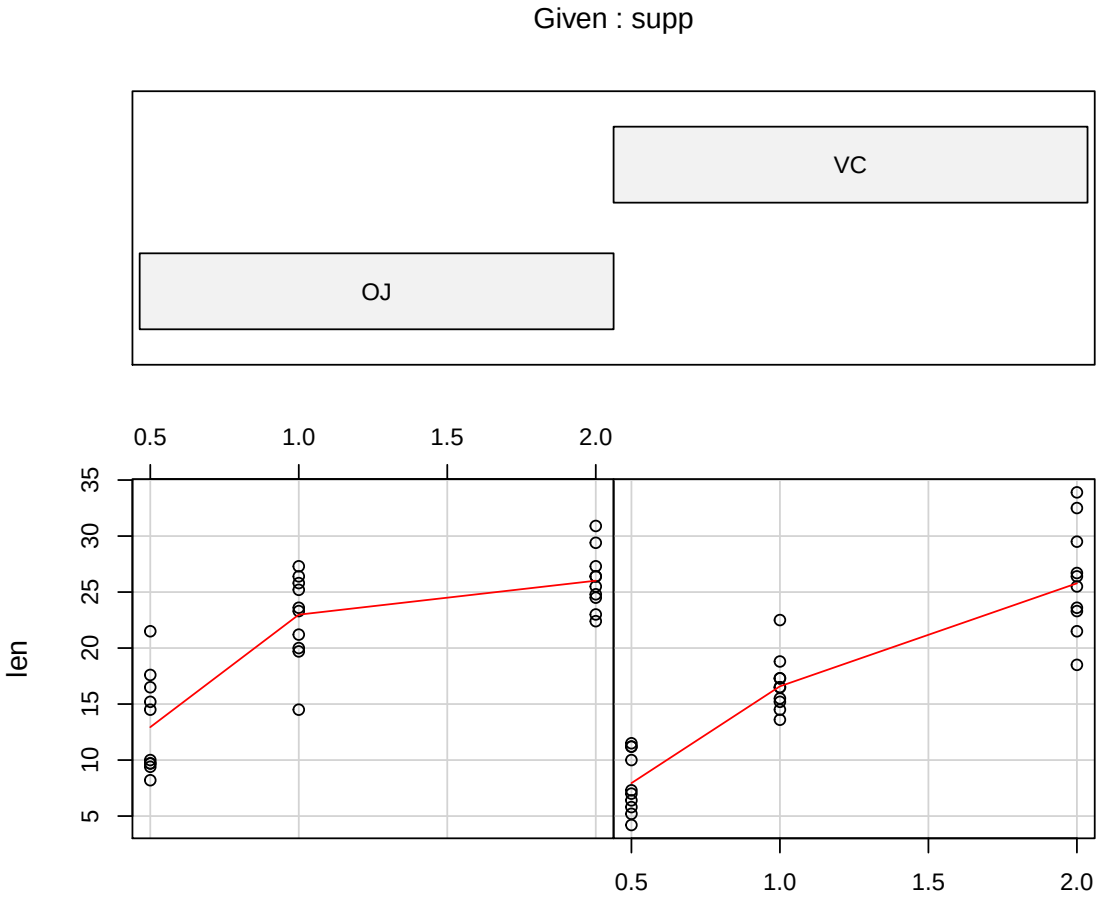

Figure 1. ToothGrowth data: length vs dose, given type of supplement

Preliminary Look at the ToothGrowth Data

```
# A look at first few rows
head(ToothGrowth)
```

```
##    len supp 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
```

```r
# A look at the last few rows
tail(ToothGrowth)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

```r
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
```

```r
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 ...
```

```r
table(ToothGrowth$supp)
```

```
##
## OJ VC
## 30 30
```

```r
table(ToothGrowth$dose)
```

```
##
## 0.5   1   2
##  20  20  20
```

Boxplot of the ToothGrowth Data

```
# Data visualization using ggplot
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x=dose,y=len)) + geom_boxplot(aes(fill = dose)) +
xlab("Dosage") + ylab("Tooth Length") + facet_grid(~ supp) +
ggtitle("Figure 2. Tooth Length as a Function of Delivery Method and Dosage") +
theme(plot.title = element_text(vjust=1.5,size = 12,colour="purple"))
```

Figure 2. Tooth Length as a Function of Delivery Method and Dosage