Tooth Growth Data Analysis

authored by Gran Ville Lintao June 22, 2015

Overview

This document contains statistical analysis for the ToothGrowth dataset in R.
It is divided into 4 parts:
1. **Exploratory data analysis** summarizing an initial observation
2. Usage of **Confidence Intervals** for the mean difference of the two types of supplement
3. Usage of a Hypothesis Test - **Unequal Variance T Test** for the two types of supplement
4. **Assumptions and Conclusions** about the tests performed.

Exploratory Data Analysis

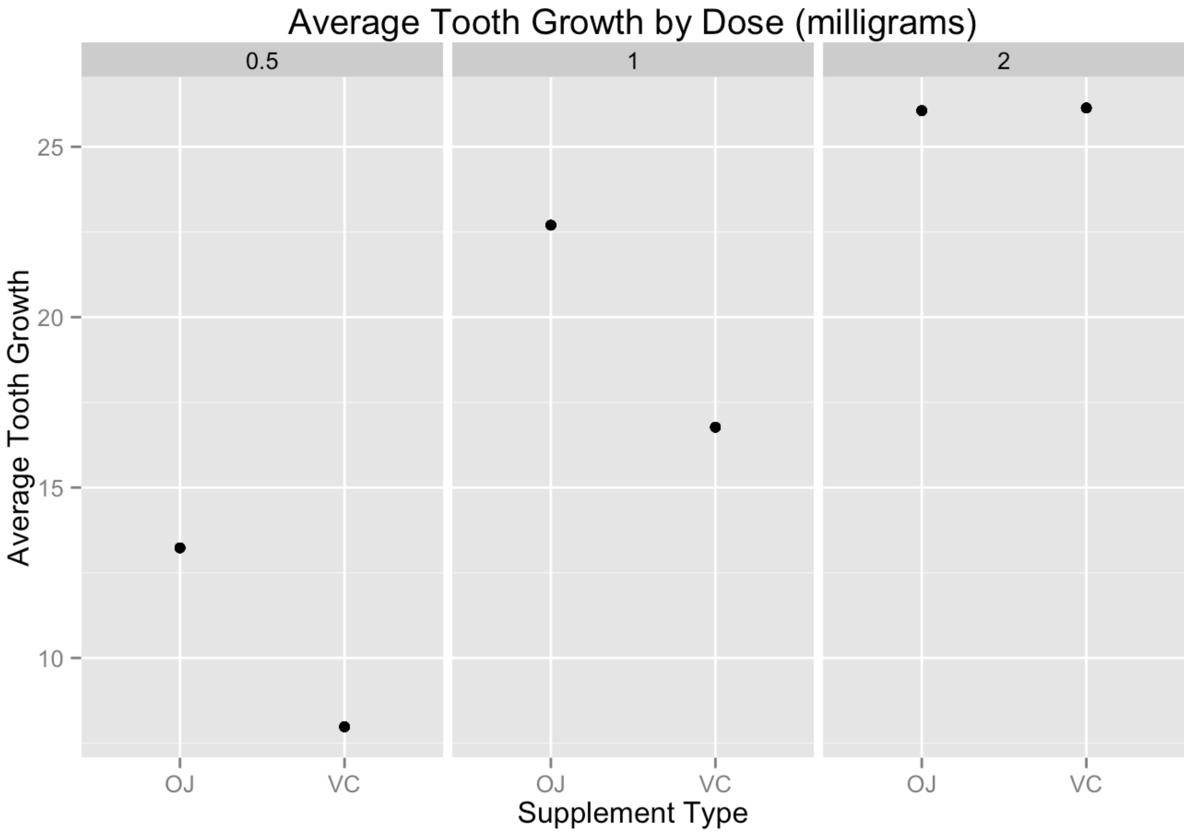
In this part we load the libraries needed, the data, and then using plyr, *generate a summary* by averaging the toothGrowth by supplement and dose:

```
library(ggplot2)
library(plyr)

data(ToothGrowth)
dataSum <- ddply(ToothGrowth, ~supp * dose, summarise,
                 aveGrowth=ave(len))
```

We then plot this summary and we can quickly see the positive correlation between dose and length of tooth growth on both doses. Furthermore, we can see that the supp **OJ** is correlated with **higher tooth growth** compared to **VC** on the doses **0.5 and 1.0**.

```
qplot(supp, aveGrowth, data=dataSum, facets=.~dose,
      main="Average Tooth Growth by Dose (milligrams)",
      ylab="Average Tooth Growth",
      xlab="Supplement Type")
```



Usage of Confidence Intervals

In this part, we use confidence intervals to compare the means between the doses VC and OJ. Note that these are independent groups.

First we take divide them in g1 and g2 for easier computing

```
g1 <- subset(ToothGrowth, supp=="VC")$len
g2 <- subset(ToothGrowth, supp=="OJ")$len
n1 <- length(g1); n2 <- length(g2)
```

Then we use can use the following computation for manually getting the confidence intervals

```
# confidence intervals option 1 manual computation practice :)
sp <- sqrt( ((n1-1) * sd(g1)^2 + (n2-1) * sd(g2)^2) / (n1+n2-2) ) # pooled s
d estimate
md <- mean(g2) - mean(g1) # mean difference
semd <- sp * sqrt(1/n1 + 1/n2) # std error of the mean difference
ciManual <- md + c(-1,1) * qt(.975,n1+n2-2) * semd
ciManual
```

```
## [1] -0.1670064  7.5670064
```

We can also use the t.test() function to check if our computation above is correct

```
# confidence intervals option 2 using t.test
ciTTest <- t.test(g2, g1, paired=FALSE, var.equal=TRUE)$conf
ciTTest
```

```
## [1] -0.1670064  7.5670064
## attr(,"conf.level")
## [1] 0.95
```

Here we can see that the manually computed **ciManual** : -0.1670064, 7.5670064 is the same with **ciTTest** : -0.1670064, 7.5670064

Unequal Variance T Test

Here we use an unequal variance T Test for comparing whether the means in the two supplements OJ and VC are equal under the null Hypothesis. The alternative hypothesis is that they're different.

```
uvTTest <- t.test(data=ToothGrowth, len ~ supp, paired=FALSE, var.equal=TRUE
)$statistic
uvTTest
```

```
##        t
## 1.915268
```

Here we can see that the estimated standard errors the difference in means from the hypothesized mean is : 1.9152683. Based on this we reject the null Hypothesis.

Assumptions and Conclusions

Conclusions:
1. Larger dosages is correlated with larger tooth growth
2. dose OJ has larger toothgrowth in lower doses 0.5, 1.0 compared to dose VC
3. We use an Indenpendent Group T Confidence Intervals due to assumption 1
4. The means in supplements OJ and VC are not equal - We reject the null hypothesis in favor of the alternative

Assumptions:
1. Groups OJ and VC are independent and not paired
2. We use an Unequal Variance T Test for comparing OJ and VC, which can be treated as ‘diets’