Coursera Statistical Inference: Course Project 1, Question 2

Ely Nazir

October 25, 2015

Introduction

Here we're going to analyze the ToothGrowth data in the R datasets package and report on this analysis.

The ToothGrowth data is 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).

A data frame with 60 observations on 3 variables.

[,1] len numeric Tooth length [,2] supp factor Supplement type (VC or OJ). [,3] dose numeric Dose in milligrams.

Analysis

1. Load the ToothGrowth data and perform some basic exploratory data analyses

We'll load the data into memory and look at the first few entries to see what the data look like, and to make sure it matches the provided description:

```
library(datasets)
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 ...
```

```
table(ToothGrowth$supp, ToothGrowth$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

We know from the definition of the data that `dose` is actually a predetermined, discrete factor and not a continuous number, so we'll make it a factor and run a summary:

```
ToothGrowth$dose <- factor(ToothGrowth$dose)
summary(ToothGrowth)
```

```
##       len          supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
```

```
##  Median :19.25              2  :20
##  Mean    :18.81
##  3rd Qu.:25.27
##  Max.    :33.90
```

We can see here that as described there are a total of 60 total records, which is 30 for each supplement, or 10 trials for each of the 6 supplement/dose combinations.

Let's break this down visually to explore the supplement methods and dosages independent effects on tooth growth:

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.1
```

```r
# Load and prepare Orange Juice data for plot
oj <- ToothGrowth[ToothGrowth$supp == "OJ",]
plot.oj <- ggplot(oj, aes(x=factor(dose),y=len,fill=factor(dose))) +
    geom_boxplot() +
    scale_x_discrete("Dose (mg)") +
    scale_y_continuous("Tooth Length") +
    ggtitle("Tooth Length by Dosage of Orange Juice")

# Load and prepare Vitamin C data for plot
vc <- ToothGrowth[ToothGrowth$supp == "VC",]
plot.vc <- ggplot(vc, aes(x=factor(dose),y=len,fill=factor(dose))) +
    geom_boxplot() +
    scale_x_discrete("Dose (mg)") +
    scale_y_continuous("Tooth Length") +
    ggtitle("Tooth Length by Dosage of Vitamin C")

plot.oj
```

Tooth Length by Dosage of Orange Juice

```
plot.vc
```
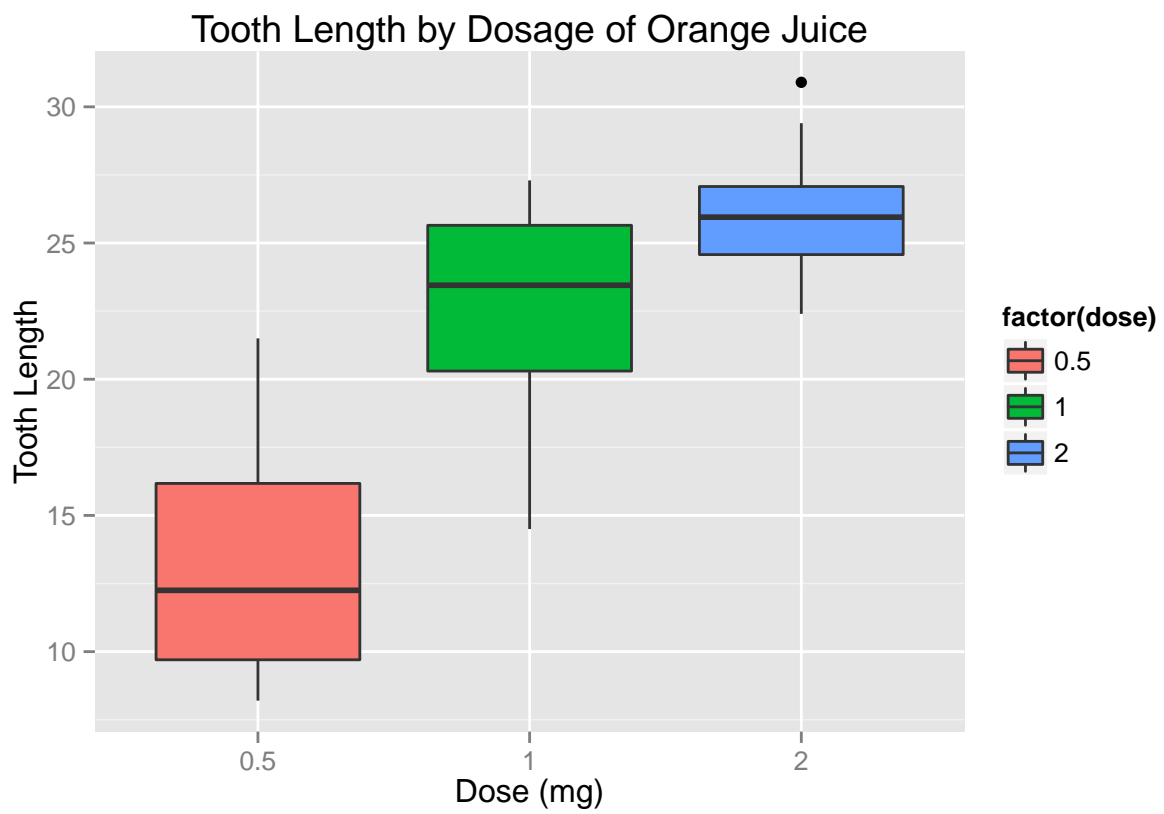
Tooth Length by Dosage of Vitamin C
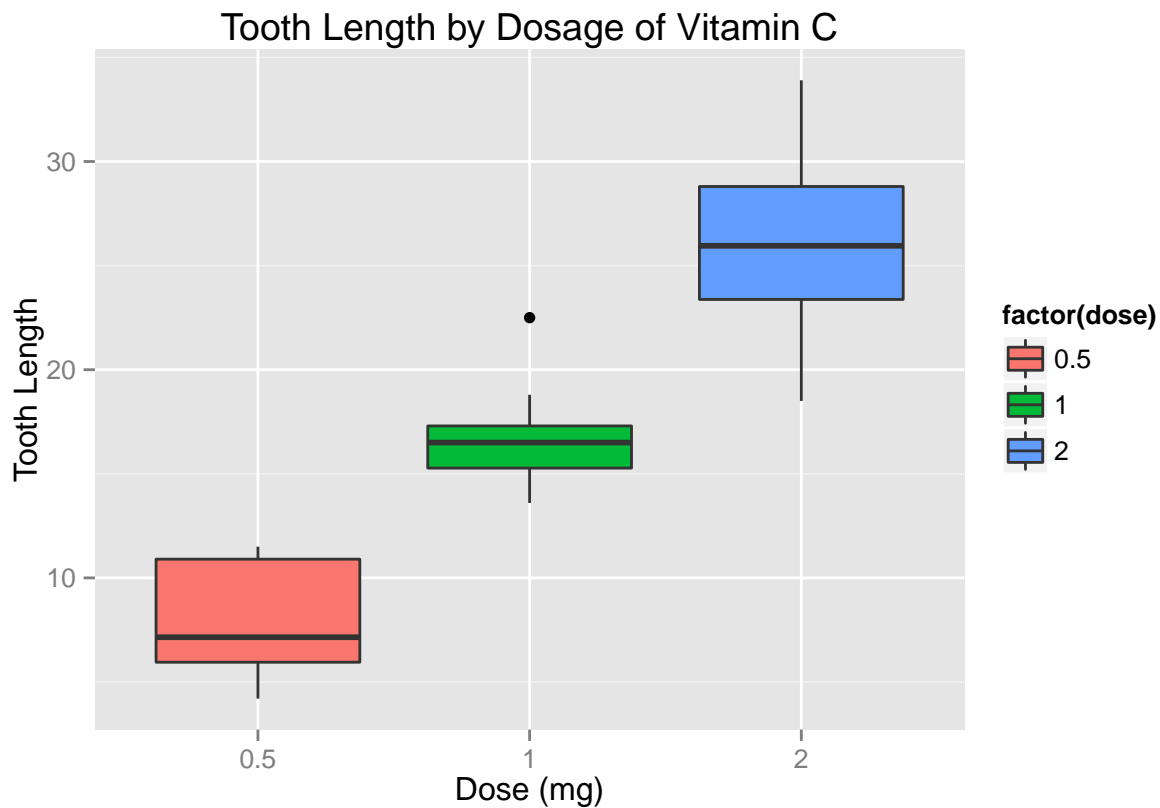
2. Provide a basic summary of the data.

The data matches the description of the dataset, which is always nice.

On first look the tooth lengths look to be substatinally higher for OJ in most cases, and that tooth length increases along with the dose of Vitamin C for both Ascorbic Acid and Orange Juice supplement methods.

Let's dive in to see whether our hypotheses here are valid.

3. Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose. (Use the techniques from class even if there's other approaches worth considering)

To test our ideas from the exploratory analysis we will use the `t.test` function here to perform a Student's t-Test on the data with the defaults of 95% confidence level and equal variances.

Our null hypothesis in all cases below is that the means of the two groups being tested are equal. The alternative hypothesis is that the observed difference in tooth lengths is statistically significant.

We will reject the null hypothesis in favor of the alternative hypothesis when we find a p-value of less than 0.05.

Effects of Supplement Type   First, we'll break the data up into subsets to help clarity in all the analysis below.

```
vc05 <- vc[vc$dose==0.5,]
vc10 <- vc[vc$dose==1.0,]
vc20 <- vc[vc$dose==2.0,]
oj05 <- oj[oj$dose==0.5,]
oj10 <- oj[oj$dose==1.0,]
oj20 <- oj[oj$dose==2.0,]
```

Now we'll test the effect of supplement type on tooth length while holding dosage constant at each level.

```
# Perform t test on vc vs oj at 0.5mg dose
t.05.supps.vc.and.oj <- t.test(len ~ supp, data=rbind(vc05,oj05), var.equal=FALSE)

# Perform t test on vc vs oj at 1.0mg dose
t.10.supps.vc.and.oj <- t.test(len ~ supp, data=rbind(vc10,oj10), var.equal=FALSE)

# Perform t test on vc vs oj at 2.0mg dose
t.20.supps.vc.and.oj <- t.test(len ~ supp, data=rbind(vc20,oj20), var.equal=FALSE)
```

Let's take a quick look at a summary of the results:

supplements compared / dose level / p-value / conf. int. (-) / conf. int. (+)
vc and oj / 0.5mg / 0.0063586 / 1.7190573 / 8.7809427
vc and oj / 1.0mg / 0.0010384 / 1.7190573 / 9.0578518
vc and oj / 2.0mg / 0.9638516 / 1.7190573 / 3.6380705

Effects of Dosage

Now we'll test the effect of dosages on tooth length:

```
# Perform t test on 0.5mg vs 1.0mg, within each supplement
t.vc.doses.05.and.10 <- t.test(len ~ dose, data=rbind(vc05,vc10), var.equal=TRUE)
t.oj.doses.05.and.10 <- t.test(len ~ dose, data=rbind(oj05,oj10), var.equal=TRUE)

# Perform t test on 1.0mg vs 2.0mg, within each supplement
t.vc.doses.10.and.20 <- t.test(len ~ dose, data=rbind(vc10,vc20), var.equal=TRUE)
t.oj.doses.10.and.20 <- t.test(len ~ dose, data=rbind(oj10,oj20), var.equal=TRUE)

# Perform t test on 0.5mg vs 2.0mg, within each supplement
t.vc.doses.05.and.20 <- t.test(len ~ dose, data=rbind(vc05,vc20), var.equal=TRUE)
t.oj.doses.05.and.20 <- t.test(len ~ dose, data=rbind(oj05,oj20), var.equal=TRUE)
```

The p-values and lower confidence intervals for each of these three cases follow here:

```
options("scipen"=100)
```

doses compared / supplement / p-value / conf. int. (-) / conf. int. (+)
0.5mg and 1.0mg / VC / 0.0000006 / -11.2643455 / -6.3156545
0.5mg and 1.0mg / OJ / 0.0000836 / -13.4108143 / -5.5291857
1.0mg and 2.0mg / VC / 0.000034 / -12.9689598 / -5.7710402
1.0mg and 2.0mg / OJ / 0.0373628 / -6.5005017 / -0.2194983
0.5mg and 2.0mg / VC / 0 / -21.8328433 / -14.4871567
0.5mg and 2.0mg / OJ / 0.0000003 / -16.2782226 / -9.3817774

4. State your conclusions and the assumptions needed for your conclusions.

Based on the above analysis, there is a definite dependence on the dosage level on tooth growth. p-values for all tests were less than the threshold and confidence intervals do not include zero, indicating that the increase in tooth growth when supplement dose is increased is significantly significant.

Furthermore, we see there is a correlation between the supplement type (Orange Juice vs Vitamin C) and tooth growth that disappears at a higher dose. For 0.5mg and 1.0mg doses, we see that the p-values are below our threshold and the confidence interval does not include zero, indicating that the increase in tooth growth that's seen with OJ vs VC is statistically significant. However, at 2.0mg this discrepancy vanishes and our null hypothesis at this level is accepted.