Statistical Inference Coursera Project - Part 2

Nick Horne

25 September 2017

Overview

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package. This data looks at the `Effect of Vitamin C on Tooth Growth in Guinea Pigs`

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

```
# Load the data
data(ToothGrowth)

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
```

```
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
```

2. Provide a basic summary of the data.

The data contains 3 variables `len` - The length of odontoblasts (cells responsible for tooth growth). `supp` - Orange juice (OJ) or ascorbic acid (VC) or `dose` - The dose levels of vitamin C (mg/day).

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

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

```
# Convert it to a factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

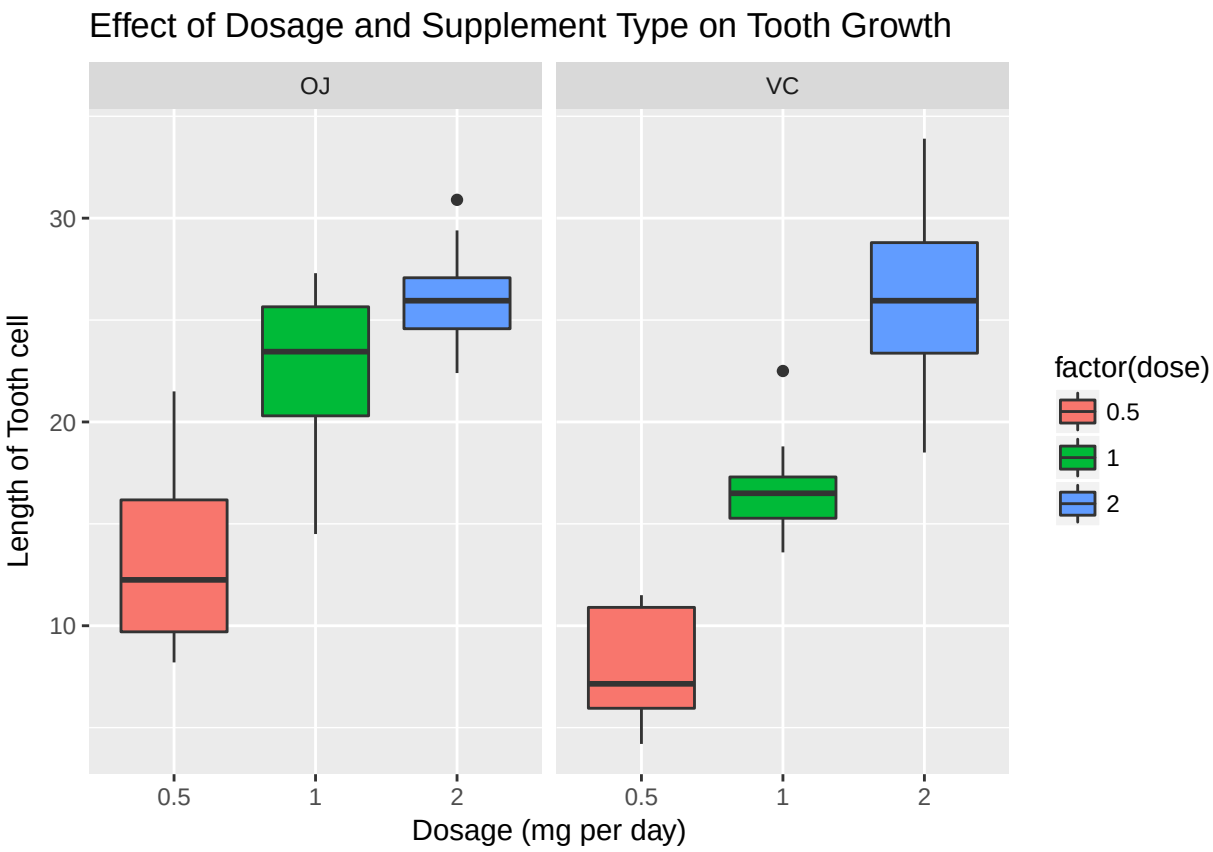
The dose levels are either 0.5, 1 or 2 (mg/day) and are equaliy distrubuted amongs the 60 observations. There are 30 observations for each suppliment.

```
aggregate(len~supp+dose, data=ToothGrowth, length)
```

```
##   supp dose len
## 1   OJ  0.5  10
## 2   VC  0.5  10
## 3   OJ    1  10
## 4   VC    1  10
## 5   OJ    2  10
## 6   VC    2  10
```

There are 10 observations for each combination of dose and suppliment.

```
plot <- ggplot(ToothGrowth,
               aes(x=factor(dose),y=len,fill=factor(dose)))
plot + geom_boxplot(notch=F) + facet_grid(.~supp) +
     xlab("Dosage (mg per day)") +
     ylab("Length of Tooth cell") +
     ggtitle("Effect of Dosage and Supplement Type on Tooth Growth")
```



The graph shows that both suppliments increase tooth cell length with dosage with the orange juice being more effective at lower dosages.

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

From the data it seems like the OJ suppliment has an increased effect but this tails off at higher levels. Lets look at it statistically and perform a t-test.

```r
suppTest <- t.test(ToothGrowth[ToothGrowth$supp=="OJ",]$len,
       ToothGrowth[ToothGrowth$supp=="VC",]$len,
       paired = FALSE,
       var.equal = FALSE)
suppTest
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth[ToothGrowth$supp == "OJ", ]$len and ToothGrowth[ToothGrowth$supp == "VC", ]$len
## 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 of x mean of y
##  20.66333  16.96333
```

Given the resulting p value 0.061 we cannot be certain that the supplimentery results are not due to chance.

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

Findings:

- Higher doses of vitamin C result in a longer total cell length regardless of the suppliment.
- Orange juice appears to be more effective at lower levels when compared to the ascorbic acid and the orange juice effectiveness seems to tail off at higher dose levels. However with the data provided we cannot prove this is the case statistically and not just down to chance.

These findings are based on the assumptions:

- The guinea pigs were randomly assigned and representative of the entite population