Statistical Inference Course Project - Part 2

Mike McFarren

November 27, 2016

Synopsis

Now in the second portion of the project, I'm going to analyze the ToothGrowth {datasets} in the R datasets package.

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

```
library(datasets)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
data("ToothGrowth")

data <- ToothGrowth

summary(data)
```

```
##       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
```

```
str(data)
```

```
## '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 ...
```

```
head(data)
```

```
##    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
```

The ToothGrowth dataset describes the effect of Vitamin C on tooth growth in guinea pigs. This dataset contained 60 observations on 3 variables:

- len (numeric): Length of odontoblasts (teeth) in millimeters
- supp (factor): Vitamin C supplement type (OJ=Orange Juice, VC=Ascorbic Acid)
- dose (numeric): Dose in milligrams/day (0.5, 1, or 2)

2. Provide a basic summary of the data.

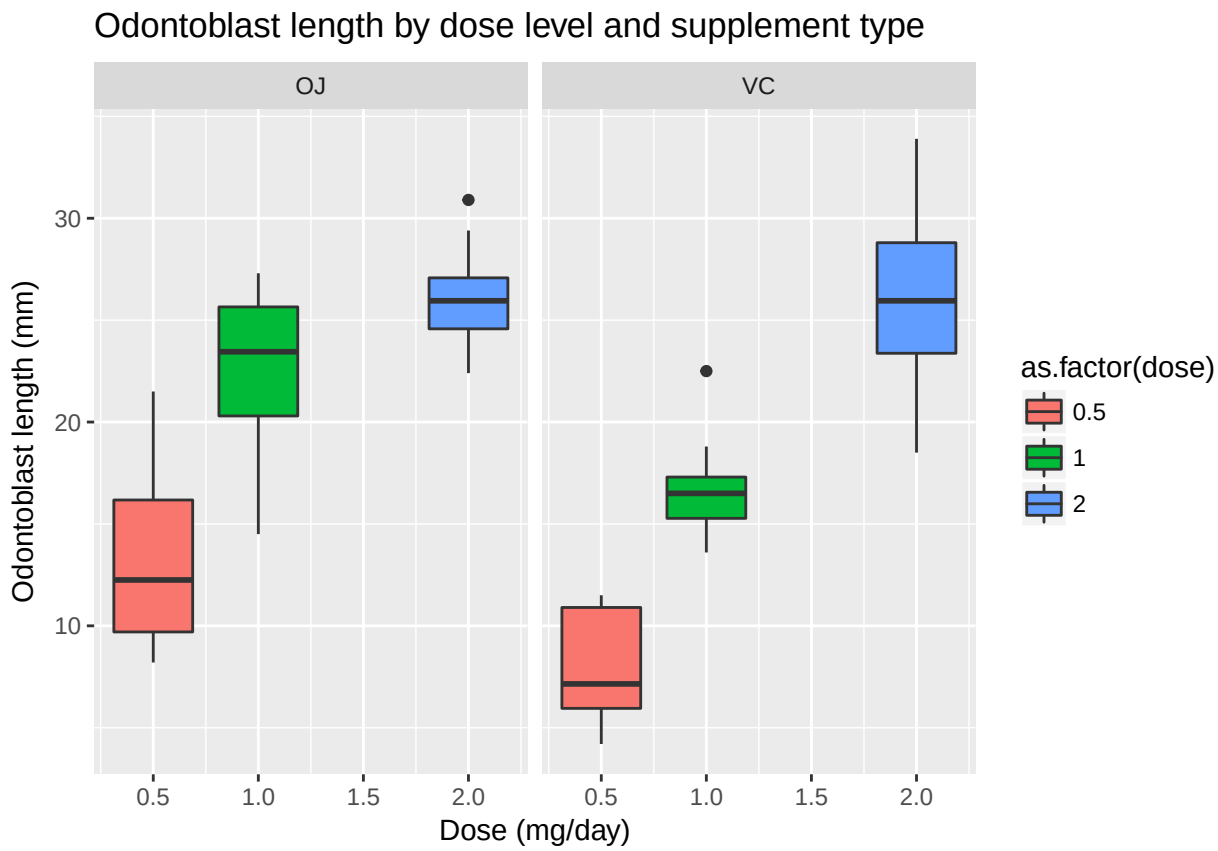
The average odontoblasts lengths from the sample set is 18.813, with a standard deviation of 7.649.

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

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

We also know that there are 10 guinea pigs in each of the subgroups, dose by supp.

```
ggplot(data, aes(dose, len)) +
  geom_boxplot(aes(fill = as.factor(dose))) +
  facet_grid(. ~ supp) +
  xlab('Dose (mg/day)') +
  ylab('Odontoblast length (mm)') +
  ggtitle('Odontoblast length by dose level and supplement type')
```



From this chart, we can see a positive correlation in both supplements between odontoblast length and Vitamin C doses.

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)

For the purpose of this assignment, I am assuming that the guinea pigs were randomly assigned, thus independence is guaranteed and the observation came from a nearly normal distribution. And since we are dealing with small sets of data (n <= 30), the skewness of the population could yield a distribution that is not normal, so I have chosen to use Gosset's T distribution for hypothesis testing and therefore T confidence intervals will be constructed.

```
t_diff_supp <- t.test(len ~ supp, ToothGrowth, var.equal = FALSE)
t_diff_supp
```

```
##
##  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 of 0.0606345 is larger than the significance value of 0.05, so we fail to reject the null hypothesis. As illustrated above, the confidence interval includes 0, so the data does not provide a statistically significant difference between the 2 groups.

```
test1 <- t.test(ToothGrowth$len[ToothGrowth$dose == 2], ToothGrowth$len[ToothGrowth$dose == 1])
test1
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 2] and ToothGrowth$len[ToothGrowth$dose == 1]
## t = 4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.733519 8.996481
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

```
test2 <- t.test(ToothGrowth$len[ToothGrowth$dose == 2], ToothGrowth$len[ToothGrowth$dose == 0.5])
test2
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 2] and ToothGrowth$len[ToothGrowth$dose == 0.5]
## t = 11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.83383 18.15617
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

```
test3 <- t.test(ToothGrowth$len[ToothGrowth$dose == 1], ToothGrowth$len[ToothGrowth$dose == 0.5])
test3
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 1] and ToothGrowth$len[ToothGrowth$dose == 0.5]
## t = 6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.276219 11.983781
## sample estimates:
## mean of x mean of y
##    19.735    10.605
```

Performing T-test permutations to cover all possible dose combinations, we can see that in all 3 cases the p-value is smaller than 0.05, which means we reject the null hypothesis for all 3 cases, and can conclude from this data that there is strong evidence that the odontoblast length of guinea pigs is, on average, different when compared to other dose levels.

For example, from these T-tests,we are 95% confident that the odontoblast length for guinea pigs that were administered 2 mg/day doses of Vitamin C is, on average, between 3.7335195 and 8.9964805 mm larger than those guinea pigs that were administered 1 mg/day.

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

The observed difference of odontoblast lengths, on average, across supplement types is statistically not different from 0. Our observations also led us to the finding that there is a positive correlation of odontoblast length and dose levels.