Assignment Part 2: Basic Inferential Data Analysis

Sam Edwardes

April 5, 2019

### Overview

The purpose of this document is to analyze tooth growth data in the R datasets package. From the R Documentation (?ToothGrowth):

*The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).*

*The dataframe has 60 observations on 3 variables:*

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

### Load data

# load required libraries  
library(ggplot2)  
library(dplyr)  
library(Hmisc)  
  
# load ToothGrowth data into a dataframe  
tg <- datasets::ToothGrowth

### Basic data summary

First, lets explore what the data looks like, and run some basic summary statistics.

describe(tg)

## tg   
##   
## 3 Variables 60 Observations  
## ---------------------------------------------------------------------------  
## len   
## n missing distinct Info Mean Gmd .05 .10   
## 60 0 43 0.999 18.81 8.839 6.37 8.11   
## .25 .50 .75 .90 .95   
## 13.07 19.25 25.27 27.30 29.57   
##   
## lowest : 4.2 5.2 5.8 6.4 7.0, highest: 29.4 29.5 30.9 32.5 33.9  
## ---------------------------------------------------------------------------  
## supp   
## n missing distinct   
## 60 0 2   
##   
## Value OJ VC  
## Frequency 30 30  
## Proportion 0.5 0.5  
## ---------------------------------------------------------------------------  
## dose   
## n missing distinct Info Mean Gmd   
## 60 0 3 0.889 1.167 0.678   
##   
## Value 0.5 1.0 2.0  
## Frequency 20 20 20  
## Proportion 0.333 0.333 0.333  
## ---------------------------------------------------------------------------

From analyzing each column, there are several insights gleaned:

* [len] column has the most variability, with values range from 4.2 up to 33.9
* [supp] column has only two variables, OJ and VC
* [does] column has three potential values, either 0.5, 1, or 2

It looks like we will want to compare tooth growth based on delivery method and/or dose.

oj\_mean <- filter(tg, supp == "OJ") %>% select(len) %>% summarise(mean(len))  
vc\_mean <- filter(tg, supp == "VC") %>% select(len) %>% summarise(mean(len))  
  
one\_half\_mean <- filter(tg, dose == 0.5) %>% select(len) %>% summarise(mean(len))  
one\_mean <- filter(tg, dose == 1.0) %>% select(len) %>% summarise(mean(len))  
two\_mean <- filter(tg, dose == 2.0) %>% select(len) %>% summarise(mean(len))  
  
# by delivery method  
pull(oj\_mean); pull(vc\_mean)

## [1] 20.66333

## [1] 16.96333

# by dose  
pull(one\_half\_mean); pull(one\_mean); pull(two\_mean)

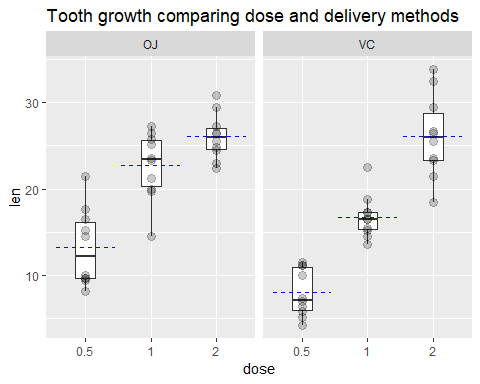
## [1] 10.605

## [1] 19.735

## [1] 26.1

When comparing tooth growth by delivery method, its clear orange juice results in more tooth growth. When comparing by dose, its clear that the higher dose results in the highest growth. Since we have multiple variables, it will be useful to plot and compare these by creating six groups (OJ: low, medium, and high dose; and VC: low, medium, and high dose).

# Plot the results by the six groups  
g3 <- ggplot(tg, aes(y = len, x = as.character(dose))) +  
 geom\_boxplot(outlier.shape = NA, width = 0.3) +  
 stat\_summary(aes(ymax = ..y.., ymin = ..y..),fun.y = mean, color='blue', geom="errorbar", linetype = "dashed") +   
 geom\_point(alpha = 0.2, size = 3) +  
 labs(title = "Tooth growth comparing dose and delivery methods", x = "dose") +  
 facet\_grid(.~supp)  
  
g3



# Show the mean for each group  
tg %>%  
 group\_by(supp, dose) %>%  
 summarise(mean = mean(len))

## # A tibble: 6 x 3  
## # Groups: supp [?]  
## supp dose mean  
## <fct> <dbl> <dbl>  
## 1 OJ 0.5 13.2   
## 2 OJ 1 22.7   
## 3 OJ 2 26.1   
## 4 VC 0.5 7.98  
## 5 VC 1 16.8   
## 6 VC 2 26.1

As the chart above demonstrates (note the dotted blue line represents the mean):

* it looks like at the highest dose, both delivery methods result in similar results
* at lower doses (0.5 and 1.0), orange juice delivers superior results.

### Confidence intervals

Now lets confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose to see if our initial observations are valid.

First, we will compare the delivery methods of orange juice vs. absorbic acid. The mean tooth growth for samples who received orange juice is 20.6633333. The mean tooth growth for samples who received absorbic acid is 16.9633333. The delta between these two means is 3.7.

t\_test\_result <- t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = tg)  
t\_test\_result

##   
## 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

As the results demonstrate, the 95% confidence interval contains 0. This means that although our samples showed orange juice resulted in more growth, it would not be unusual for use to see a delta of 0, or Orange Juice not resulting in more growth. So maybe Orange juice is not as powerful as we thought?

Lets try comparing now a high dose vs. a low dose, ignoring the delivery method. The mean tooth growth for a high dose is 26.1. The mean tooth growth for a low dose is 10.605. The delta between these two means is 15.495.

t\_test\_result <- t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = tg[tg$dose == 0.5 | tg$dose ==2,])  
t\_test\_result

##   
## Welch Two Sample t-test  
##   
## data: len by dose  
## 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:  
## -18.15617 -12.83383  
## sample estimates:  
## mean in group 0.5 mean in group 2   
## 10.605 26.100

This time, 0 does not fall in the 95% confidence interval. With this knowledge, we can say that it is very unlikely that a dose of 2 is less effective than a dose of 0.5.

### Conclusions

Our analysis showed that:

* The orange juice method delivered higher growth than absorbic acid, however the difference was not statistically significant.
* The higher the dose, the higher the observed tooth growth. When comparing a dose of 0.5 vs 2.0, the difference was statistically significant

From this analysis, we can conclude that under either delivery method, a dose of 2.0 will with a high degree of confidence result in tooth growth.