Coursera - Statistical Inference Project - Part 2

Introduction

Note: For this assignment all R-Code will be included in the Appendix.

For Part 2 of the assignment, basic analysis of the ToothGrowth data in R is to be performed. The ToothGrowth data consists of the following (obtained by using the ?ToothGrowth command in R):

```
Description

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.
```

Tasks to be addressed in this assignment:

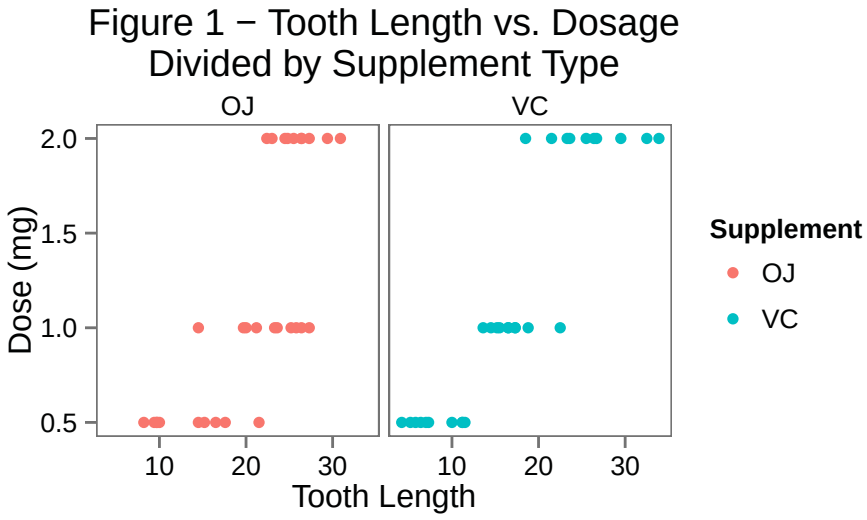
1. Load the ToothGrowth data and perform some basic exploratory data analyses

2. Provide a basic summary of the data.

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)

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

Exploratory Data Analysis:

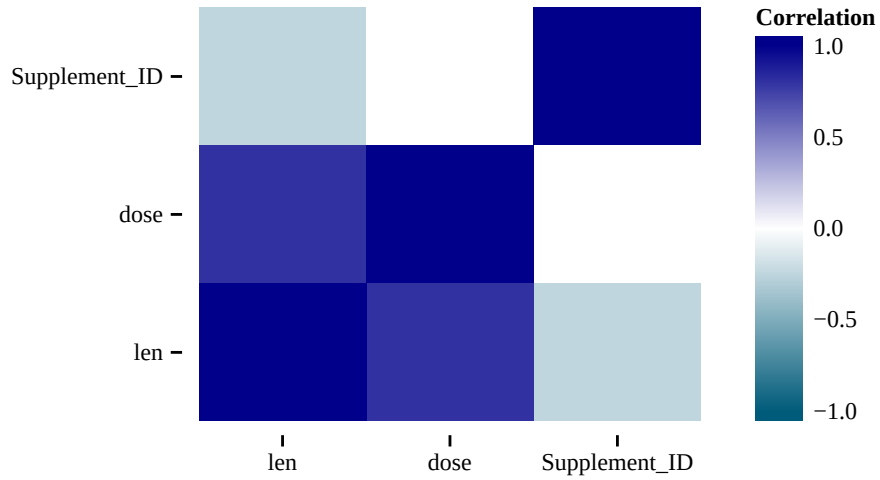


In *Figure 1*, tooth length was plotted against the dosage of Vitamin C (0.5mg, 1.0mg, 2.0mg) and divided by supplement type (orange juice = OJ, or ascorbic acid = VC). From this first plot, it appears that dosage is

having a significant impact on tooth length, with higher dosages resulting in seemingly longer teeth. The effect of the supplement type is unclear from this plot.

To further explore the data, R's *cor* (correlation) function was used to see if there is any correlation between tooth length and either dosage or supplement type. To this end, a heat map was generated (*Figure 2* - see below), showing a strong positive correlation between tooth length and Vitamin C dosage, but also a weak negative correlation between tooth length and supplement type (marked as Supplement_ID in *Figure 2*)

Figure 2 − Correlation Heatmap of Tooth Growth Data



Data summarization and analysis

For this data analysis, the effect of both Vitamin C dosage and supplementation type will be assessed with regards to the impact on *mean* tooth length. From the exploratory analysis above, it appears dosage is having a measurable effect on tooth length, while the impact of supplementation type is less clear.

As *mean* tooth length will be examined, first let us see some of the basic summary statistics from the ToothGrowth data set (called tooth_data below), specifically for the tooth length:

```
summary(tooth_data$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.2    13.1    19.2    18.8    25.3    33.9
```

As the overal sample size for this data set is small, the mean tooth length, 18.8, may not be a good indicator of the actual mean, or may not be very conducive in obtaining good confidence intervals for the sample means divided amongst dosage and supplementation types. To that end, a *bootstrapping* technique was applied to the data in order to differentiate the various sample means according to dosages and supplementation types.

This was done in several steps:

1. The data was subdivided into 5 separate data sets:

- tooth_OJ - Observations of only OJ supplementation type - across all dosage levels - 30 observations
- tooth_VC - Observations of only VC supplementation type - across all dosage levels - 30 observations
- tooth_05 - Observations of only 0.5mg dosages - across all supplementation types - 20 observations
- tooth_10 - Observations of only 1.0mg dosages - across all supplementation types - 20 observations

- tooth_20 - Observations of only 2.0mg dosages - across all supplementation types - 20 observations

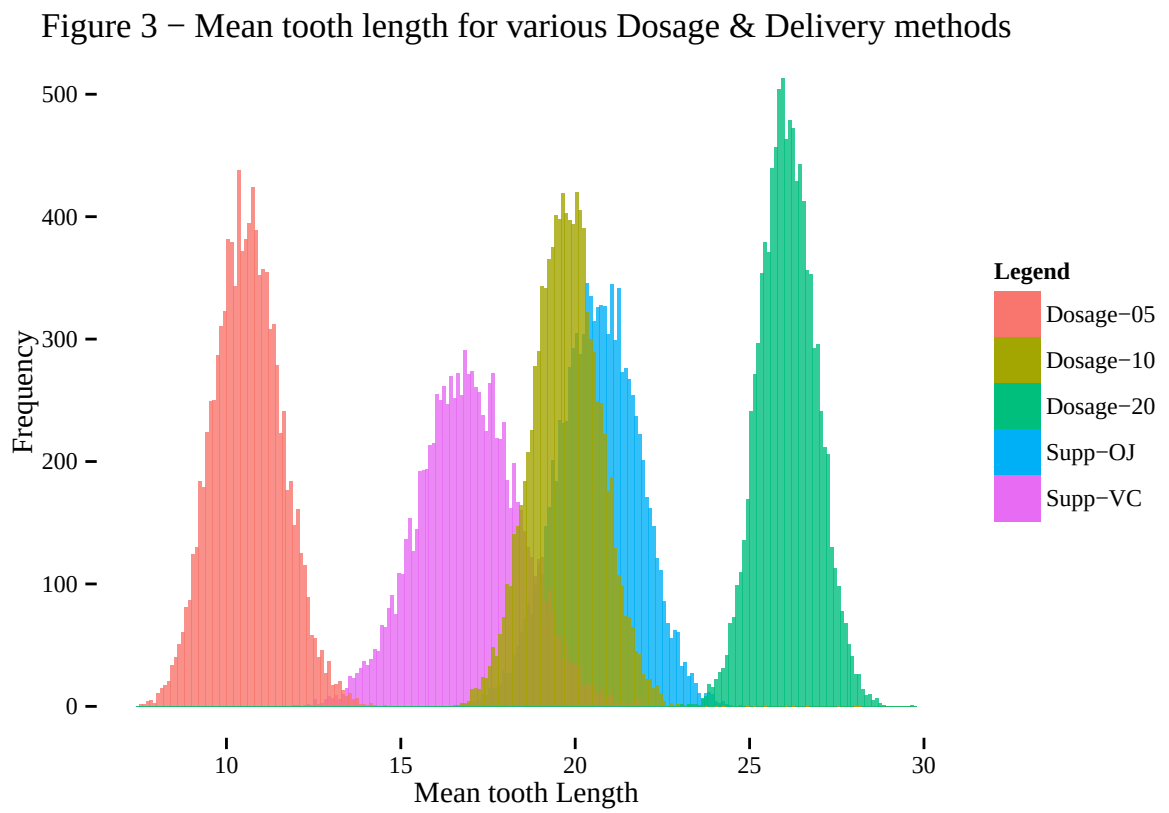
2. Five resample matrices were created - one for each of the data sets above. See below for an example:

```
set.seed(42)
resamples_OJ <- matrix(sample(tooth_OJ[,1],n_OJ * boots, replace = TRUE), boots, n_OJ)
means_OJ <- apply(resamples_OJ,1,mean)
```

In the example above, the dataset (*tooth_OJ*) was sampled with replacement 300000 times. This came from the number of rows in the dataset ($n\_OJ = 30$) and 10000 for the bootstrap variable (*boots*). This was then used to create a matrix of 10000 rows and 30 columns. A new dataset, was then created by taking the mean along all the rows of the *resamples_OJ* matrix. This dataset, *means_OJ*, consists of 10000 entries, all re-sampled means from the original dataset. Based on the Central Limit Theorem (CLT), the mean of this new data set should be able to approximate the population mean quite well, while at the same time having a *near* normal distribution.

A similar procedure was done for the other four datasets (*tooth_VC*, *tooth_05*, *tooth_10*, *tooth_20*). This should allow for the calculation of relevant confidence intervals for the mean across dosage levels and supplementation types.

3. All five *mean* distributions were then plotted on a histogram (*Figure 3*):



From *Figure 3*, a clear delineation is apparent amongst the *means* of the dosage types. There appears to be a clear relationship between *mean* tooth length and the dosage given, with Dosage-20 (corresponding to 2.0mg) having a significantly higher *mean* tooth length. *Figure 3* also shows a marked differentiation amongst the *means* controlled for supplementation type, with Supp-OJ (corresponding to orange juice) exhibiting a higher *mean* tooth length as compared to the Supp-VC (ascorbic acid). Amongst the supplementation types, there is significantly more overlap of the *means* distribution then found amongst the dosage types.

Confidence Intervals

From the distributions calculated and plotted above, confidence intervals for the *means* of the sample distributions can be calculated. These were done using the R *quantile* function, and tested for a 95 percent confidence interval (i.e. between 2.5 and 97.5 quantiles). See below for the confidence intervals, along with the means and standard deviations, by supplement type and dosage level:

```
quant_OJ; mean(means_OJ); sd(means_OJ)
```

```
##  2.5% 97.5%
## 18.34 22.91
```

```
## [1] 20.66
```

```
## [1] 1.169
```

```
quant_VC; mean(means_VC); sd(means_VC)
```

```
##  2.5% 97.5%
## 14.09 19.89
```

```
## [1] 16.96
```

```
## [1] 1.471
```

```
quant_05; mean(means_05); sd(means_05)
```

```
##  2.5% 97.5%
##  8.77 12.62
```

```
## [1] 10.62
```

```
## [1] 0.9857
```

```
quant_10; mean(means_10); sd(means_10)
```

```
##  2.5% 97.5%
## 17.84 21.64
```

```
## [1] 19.74
```

```
## [1] 0.9624
```

```
quant_20; mean(means_20); sd(means_20)
```

```
##  2.5% 97.5%
## 24.50 27.71
```

```
## [1] 26.09
```

```
## [1] 0.8158
```

t-test

We will set the NULL Hypothesis for the t-test to be performed as "True difference in mean is equal to 0".

Per *Figure 3*, there appears to be little need to perform a test on the distributions with regards to differences in dosage levels, as the *means* of those distributions were fairly wide ranging within relatively strict standard deviations and no overlap in the calculated confidence intervals.

As there was some overlap between the distributions for supplementation types, a t-test will be performed on those distributions. 4 t-tests will be performed, with *var.equal* and *paired* going from TRUE to FALSE.

NOTE: The value "paired = TRUE" doesn't actually make much sense in this scenario as there was no overlap between guinea pigs that received the OJ vs VC. It is being included solely for completeness, but in all 4 instances below, the p-values are so low that the NULL Hypothesis is rejected in all instances.

```
t.test(means_OJ,means_VC, var.equal = FALSE, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  means_OJ and means_VC
## t = 196.9, df = 19031, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.663 3.737
## sample estimates:
## mean of x mean of y
##     20.66     16.96
```

```
t.test(means_OJ,means_VC, var.equal = TRUE, paired = FALSE)
```

```
##
##  Two Sample t-test
##
## data:  means_OJ and means_VC
## t = 196.9, df = 19998, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.663 3.737
## sample estimates:
## mean of x mean of y
##     20.66     16.96
```

```
t.test(means_OJ,means_VC, var.equal = FALSE, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  means_OJ and means_VC
```

```
## t = 335.3, df = 9999, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.678 3.721
## sample estimates:
## mean of the differences
##                     3.7
```

```
t.test(means_OJ,means_VC, var.equal = TRUE, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  means_OJ and means_VC
## t = 335.3, df = 9999, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.678 3.721
## sample estimates:
## mean of the differences
##                     3.7
```

From the t-tests above, the NULL Hypothesis - "True difference in mean is equal to 0" is rejected in all cases.

Conclusion

For this analysis, a bootstrap method was employed in order to separate out the contribution of dosage levels and supplementation type to tooth growth. Specifically *mean* tooth length was examined. Several underlying assumptions were made, including that the tooth growth amongst guinea pigs does not vary so wildly that the bootstrap method and the CLT would no longer apply. In order to account for unequal variances, multiple t-tests were run on the bootstrapped samples, and in each case there was a clear delineation of effect within dosage levels (higher dosage levels resulted in greater *mean* tooth length), and in supplementation type (OJ appearing to result in greater tooth length).

APPENDIX - R Code

```
library(knitr)
library(ggplot2)
library(ggthemes)
library(plyr)
library(reshape2)
library(scales)

data(ToothGrowth)

boots <- 10000
tooth_data <- ToothGrowth
tooth_data <- ddply(tooth_data, .(supp), mutate, Supplement_ID = as.numeric(supp))
toothCor <- cor(tooth_data[,c(1,3,4)])
toothMelt <- melt(toothCor, varnames = c("x","y"), value.name = "Correlation")
toothMelt <- toothMelt[order(toothMelt$Correlation),]

summary(tooth_data$len)


ggplot(tooth_data, aes(x = len, y = dose)) + geom_point(aes(color = supp )) +
        facet_grid(. ~ supp) +
        xlab("Tooth Length") + ylab("Dose (mg)") +
        ggtitle("Figure 1 - Tooth Length vs. Dosage\nDivided by Supplement Type") +
        scale_colour_discrete(name = "Supplement") +
        theme_few()

ggplot(toothMelt, aes(x=x, y=y)) + geom_tile(aes(fill=Correlation)) +
        scale_fill_gradient2(low=muted("lightblue"), mid = "white", high="darkblue",
                             guide = guide_colorbar(ticks = FALSE,
                                                    barheight = 10),
                             limits = c(-1,1)) +  labs(x=NULL, y=NULL) +
        ggtitle("Figure 2 - Correlation Heatmap of Tooth Growth Data") + theme_tufte()

#################################################################
##  Bootstrapping
#################################################################

tooth_OJ <- tooth_data[tooth_data[,2] == "OJ",]
tooth_VC <- tooth_data[tooth_data[,2] == "VC",]

tooth_05 <- tooth_data[tooth_data[,3] == 0.5,]
tooth_10 <- tooth_data[tooth_data[,3] == 1.0,]
tooth_20 <- tooth_data[tooth_data[,3] == 2.0,]

n_OJ <- nrow(tooth_OJ)
n_VC <- nrow(tooth_VC)
n_05 <- nrow(tooth_05)
n_10 <- nrow(tooth_10)
n_20 <- nrow(tooth_20)

set.seed(42)
```

```r
resamples_OJ <- matrix(sample(tooth_OJ[,1],n_OJ * boots, replace = TRUE), boots, n_OJ)
means_OJ <- apply(resamples_OJ,1,mean)
quant_OJ <- quantile(means_OJ, c(0.025,0.975))

set.seed(42)
resamples_VC <- matrix(sample(tooth_VC[,1],n_VC * boots, replace = TRUE), boots, n_VC)
means_VC <- apply(resamples_VC,1,mean)
quant_VC <- quantile(means_VC, c(0.025,0.975))

set.seed(42)
resamples_05 <- matrix(sample(tooth_05[,1],n_05 * boots, replace = TRUE), boots, n_05)
means_05 <- apply(resamples_05,1,mean)
quant_05 <- quantile(means_05, c(0.025,0.975))


set.seed(42)
resamples_10 <- matrix(sample(tooth_10[,1],n_10 * boots, replace = TRUE), boots, n_10)
means_10 <- apply(resamples_10,1,mean)
quant_10 <- quantile(means_10, c(0.025,0.975))

set.seed(42)
resamples_20 <- matrix(sample(tooth_20[,1],n_20 * boots, replace = TRUE), boots, n_20)
means_20 <- apply(resamples_20,1,mean)
quant_20 <- quantile(means_20, c(0.025,0.975))

################################################
#Histograms
################################################

## OJ vs VC vs Dosages
ggplot() + geom_histogram(data = data.frame(means_OJ = means_OJ),
                          aes(x=means_OJ, fill = "Supp-OJ"), alpha = 0.8, binwidth = 0.1 ) +
        geom_histogram(data = data.frame(means_VC = means_VC),
                       aes(x=means_VC,fill = "Supp-VC"),
                       alpha = 0.8, binwidth = 0.1) +
        geom_histogram(data = data.frame(means_05 = means_05),
                       aes(x=means_05,fill = "Dosage-05"),
                       alpha = 0.8, binwidth = 0.1) +
        geom_histogram(data = data.frame(means_10 = means_10),aes(x=means_10,
                                fill = "Dosage-10"),
                       alpha = 0.8, binwidth = 0.1) +
        geom_histogram(data = data.frame(means_20 = means_20),aes(x=means_20,
                                         fill = "Dosage-20"),
                       alpha = 0.8, binwidth = 0.1) +
        xlab("Mean tooth Length") + ylab("Frequency") + scale_fill_discrete("Legend") +
        ggtitle("Figure 3 - Mean tooth length for various Dosage & Delivery methods")  +
        theme(legend.title=element_blank()) + theme_tufte()


t.test(means_OJ,means_VC, var.equal = FALSE, paired = FALSE)
t.test(means_OJ,means_VC, var.equal = TRUE, paired = FALSE)
t.test(means_OJ,means_VC, var.equal = FALSE, paired = TRUE)
t.test(means_OJ,means_VC, var.equal = TRUE, paired = TRUE)
```