BIOS 4120: Lab 13

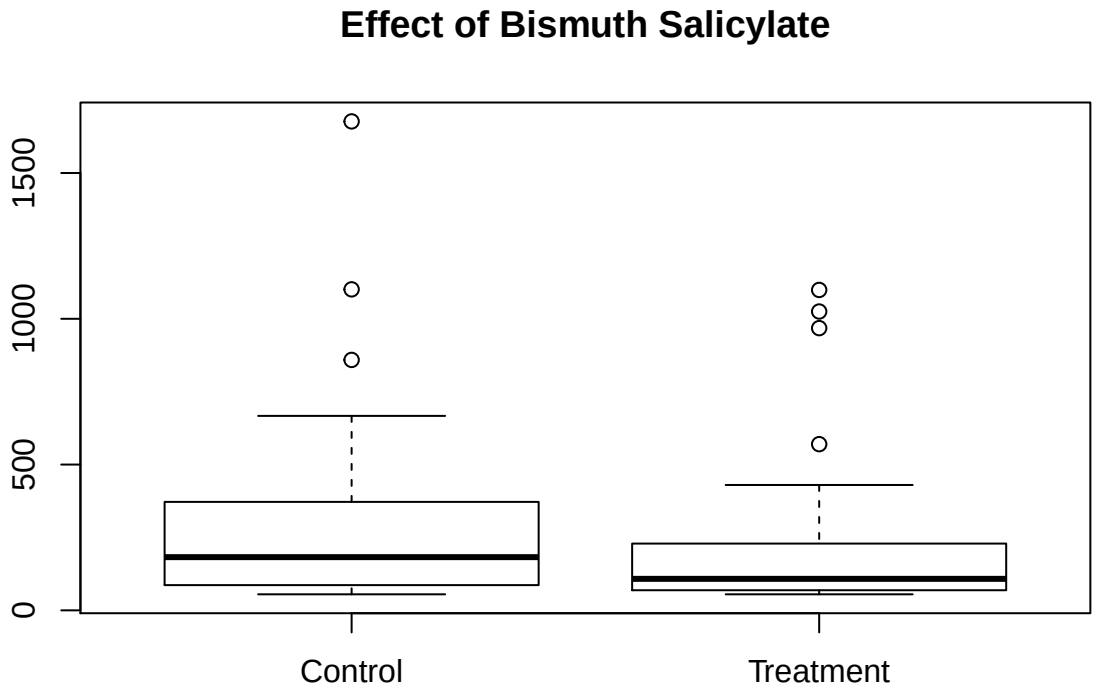
April 17-18, 2018

Last week in lab, we began analyzing the Infant Diarrhea study. In this lab, we will further analyze that data set using what we now know about outliers, transforming data, and non-parametric testing procedures.

Examining the data

To start, let's examine the distribution of the Infant Diarrhea data by group. Notice that the data are right-skewed with several outliers.

```
diarrhea <- read.delim("http://myweb.uiowa.edu/pbreheny/data/diarrhea.txt")
boxplot(diarrhea$Stool~diarrhea$Group, main = "Effect of Bismuth Salicylate")
```

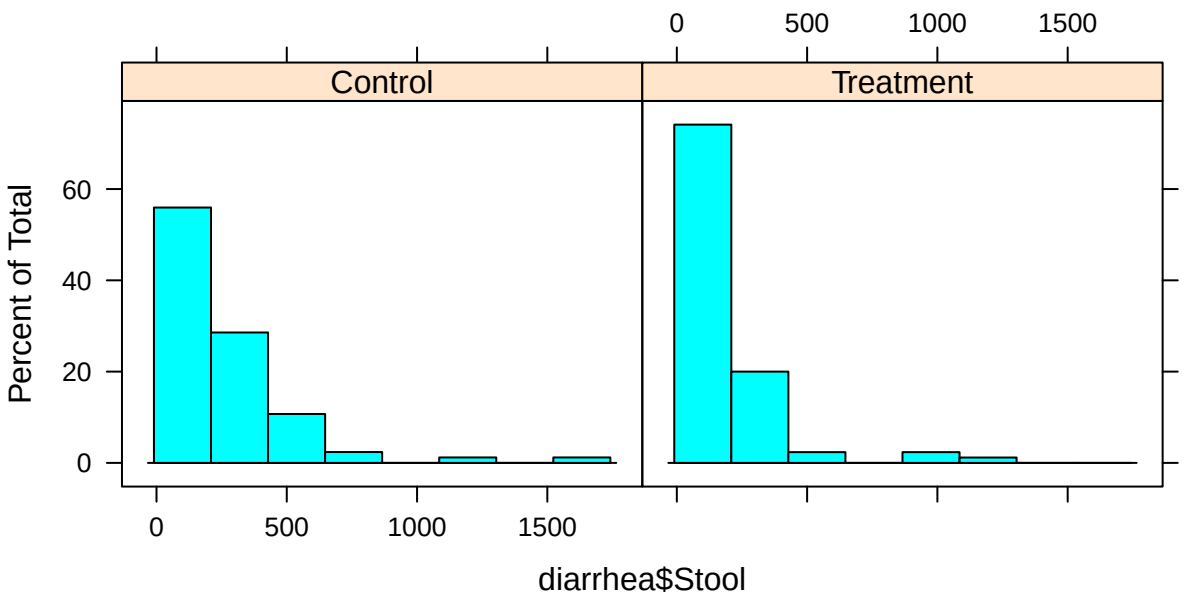


The mean and standard error are heavily influenced by outliers, therefore the two-sample t-test may be inadequate for analyzing this data.

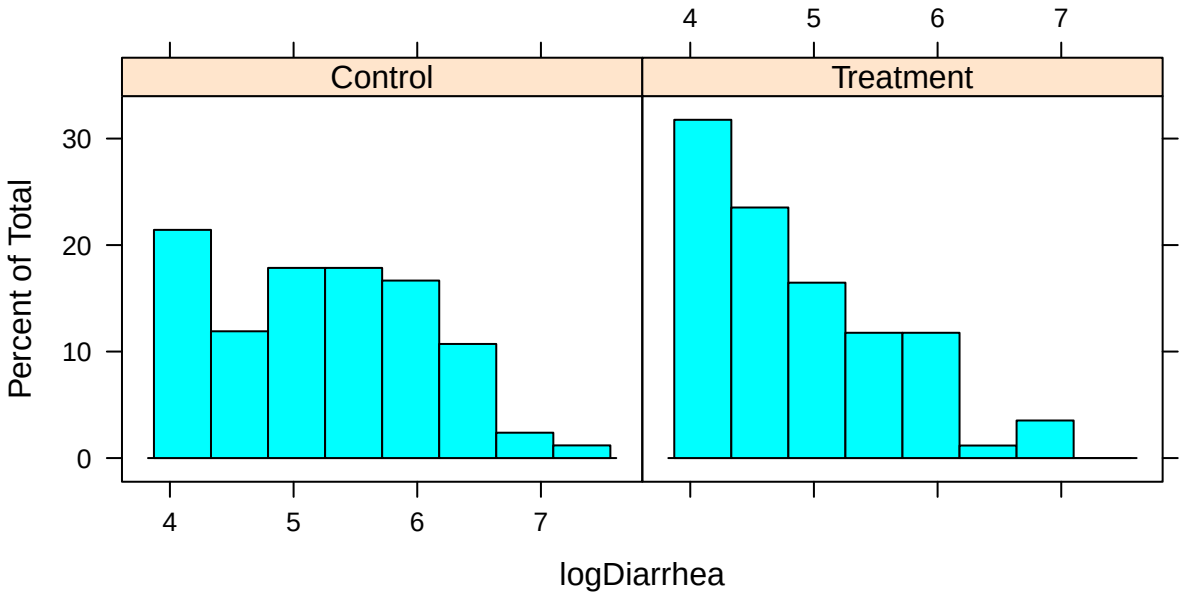
Transformations

Recall the distribution of the Odds Ratio, and how it was right-skewed. We 'fixed' this by transforming it to a new statistic (Log Odds Ratio) which is normally distributed. The same idea can be used for right-skewed data. (Note: If you've switched computers throughout the semester, you will have to install the lattice package prior to running the code.)

```r
logDiarrhea <- log(diarrhea$Stool)
library(lattice)

# Original, for comparison:
histogram(~diarrhea$Stool | diarrhea$Group)
```




```r
# Log-transformed
histogram(~logDiarrhea | diarrhea$Group)
```



We can see that the distribution of logDiarrhea is still right-skewed; however, it is not as severe as before. Now we can perform a two-sample t-test and achieve a more powerful result. Compare this with the t-test with the original data.

```
# Original
t.test(diarrhea$Stool~diarrhea$Group,var.equal=TRUE)

# log-transformed
t.test(logDiarrhea~diarrhea$Group,var.equal=TRUE)
```

Note: The confidence bounds provided are on the log scale. In order to obtain a more interpretable interval, we need to exponentiate them.

Exponentiating the Confidence Interval

Using our previous t-test, we can extract the confidence interval using "conf.int" and store the results as a variable before exponentiating.

```
logCI <- t.test(logDiarrhea~diarrhea$Group,var.equal=TRUE)$conf.int
exp(logCI)
```

Note: When we're working on the log scale, we're now thinking about the ratio between the two groups, since log(a)-log(b) = log(a/b).

By taking the difference of the log means, we are actually calculating the log ratio of the two groups. We can exponentiate this result for interpretaion.

```
estimate <- 5.2124-4.8706
exp(estimate)
```

How would you interpret this estimate?

Non-parametric tests

When the normality assumption is violated, another way to analyze the data is with a non-parametric test. These tests do not require a distributional assumption and are robust to the presence of outliers. These 'rank-based methods' are a powerful way to analyze data when distributional assumptions are questionable, and particularly effective in the presence of outliers.

- Two-sample studies: Mann-Whitney U Test also known as the Wilcoxon Rank Sum Test
- One-sample (or paired) studies: Wilcoxon Signed-Rank Test
- Both continuous: Spearman's Rank Correlation

Wilcoxon Rank Sum Test

Refer back to the Infant Diarrhea study. Instead of transforming the data or discarding outliers, we can use the Wilcoxon Rank Sum Test to test whether the treatment and control groups have different stool output values. Ranking the data minimizes the impact of outliers, and removes assumptions about the underlying distribution of the data.

```
# Rank-Sum Test
wilcox.test(diarrhea$Stool~diarrhea$Group)
```

How would you interpret this result?

Wilcoxon Signed-Rank Test

If the data we are analyzing is **paired**, the signed-rank test is a non-parametric procedure that takes in to account the relationship between the two groups. Let's revisit the familiar paired data set from the Oatbran study.

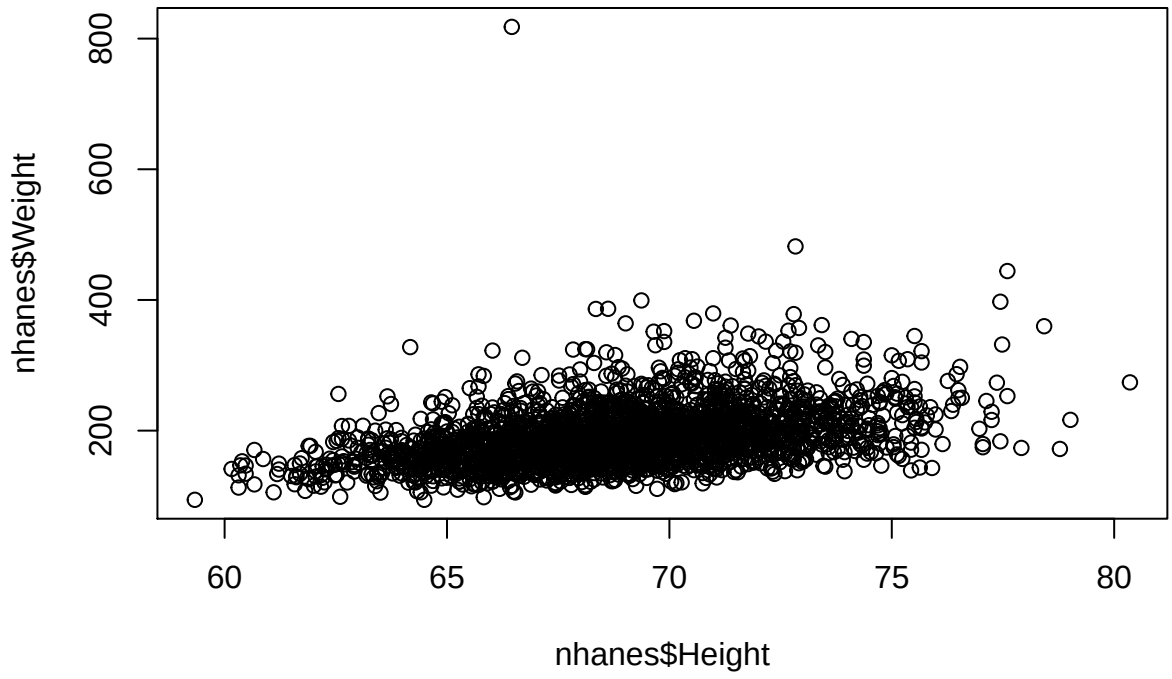
```
# Signed Rank Test
oatbran <- read.delim("http://myweb.uiowa.edu/pbreheny/data/oatbran.txt")
wilcox.test(oatbran$CornFlakes, oatbran$OatBran, paired=TRUE)

# For comparison, in case you don't remember:
t.test(oatbran$CornFlakes, oatbran$OatBran, paired=TRUE)
```

Describe the difference between the two results. What conclusions can you draw?

Spearman's Rank Correlation

Another non-parametric option is Spearman's Rank Correlation for continous data. This test ranks the values of both variables in the data set before computing a correlation coefficient. Let's revisit the nhanes data for men. Remember that the two variables of height and weight are continuous. When we plot the data, we can see a clear outlier that is over 800 pounds.

```
nhanes <- read.delim("http://myweb.uiowa.edu/pbreheny/data/nhanes-am.txt")
plot(nhanes$Weight~nhanes$Height)
```



Let's see how our results would have been influenced by this outlier using Spearman's Rank Correlation.

```r
#Spearman's Rank Correlation
cor.test(nhanes$Height, nhanes$Weight, method = "spearman")

#For comparison, here's what we got using Pearson's correlation:
cor.test(nhanes$Height, nhanes$Weight, method = "pearson")
```

How much of an influence did the outlier have?

Parametric advantages: When the assumptions hold, the parametric tests are more powerful and construction of the confidence intervals is straightforward.

Non-parametric advantages: There are minimal assumptions, and they are more powerful when parametric assumptions are invalid.

In summary, when you have continuous data, don't automatically use a t-test. Look at the data, and if it's skewed or contains large outliers, consider a transformation or a non-parametric option.