Lab 13

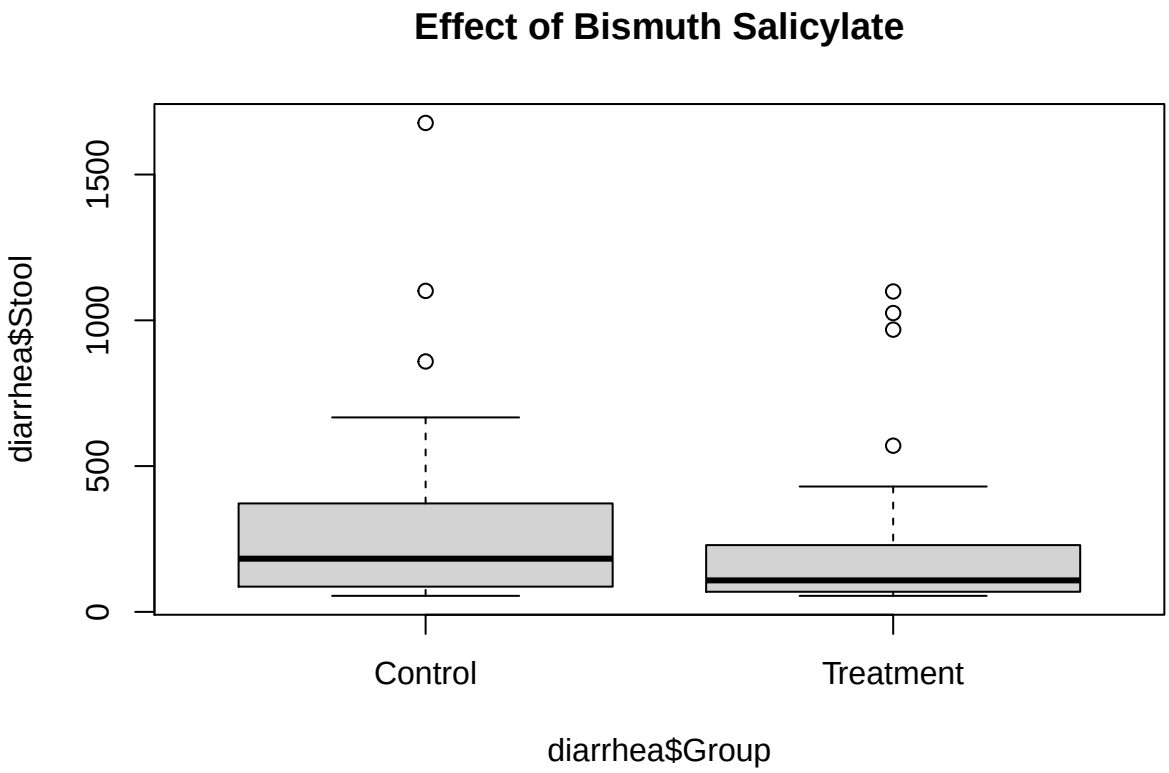
April 28th, 2020

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")
```

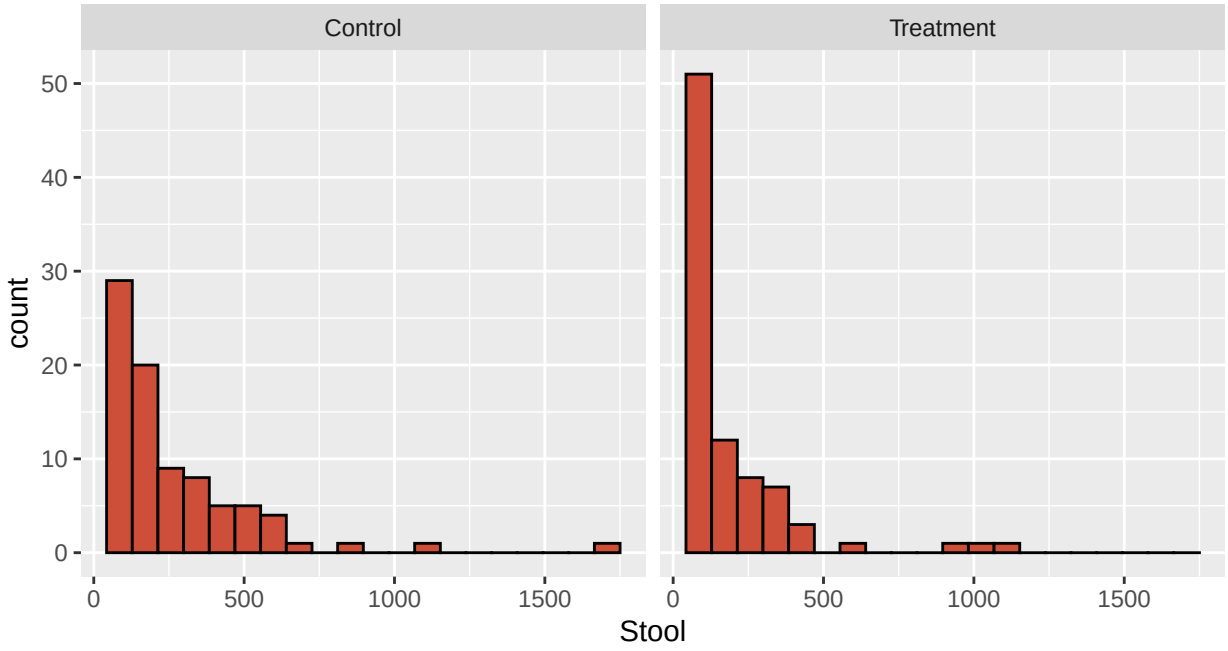


The mean and standard error are heavily influenced by outliers, therefore the two-sample t-test may be inadequate for analyzing this data.

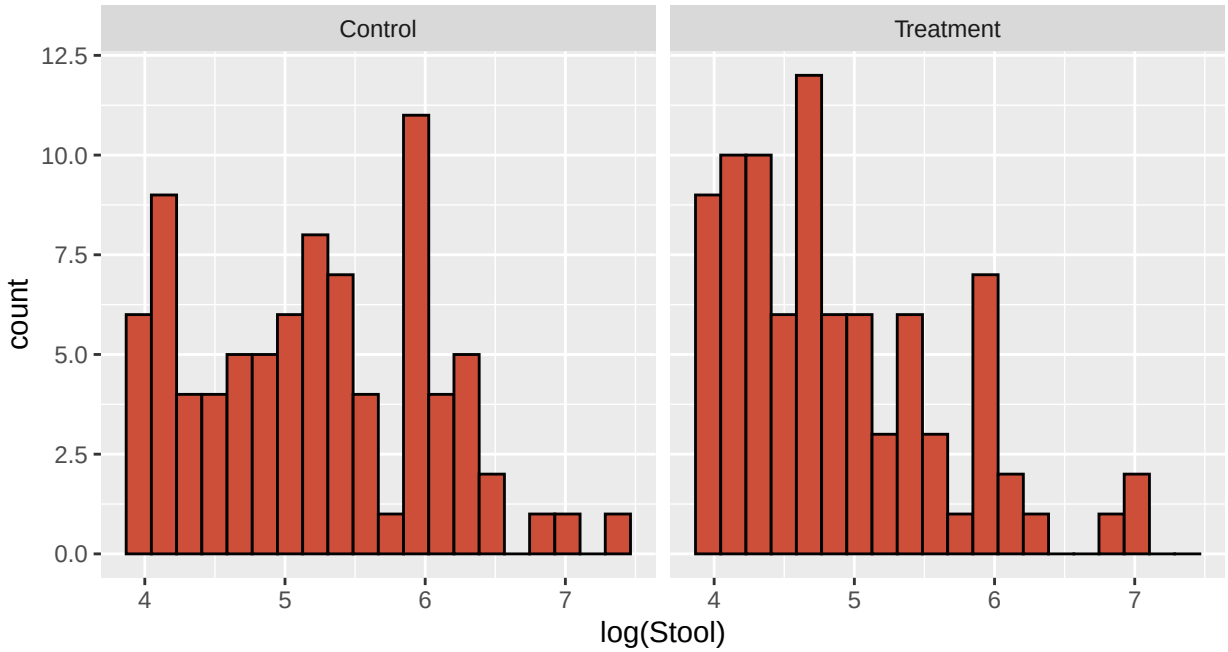
Transformations

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(ggplot2)

#Original
ggplot(diarrhea, aes(x = Stool)) + geom_histogram(fill = "tomato3", color = "black", bins = 20) + facet_
```




```r
#Log-transformed
ggplot(diarrhea, aes(x = log(Stool))) + geom_histogram(fill = "tomato3", color = "black", bins = 20) + t
```



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.

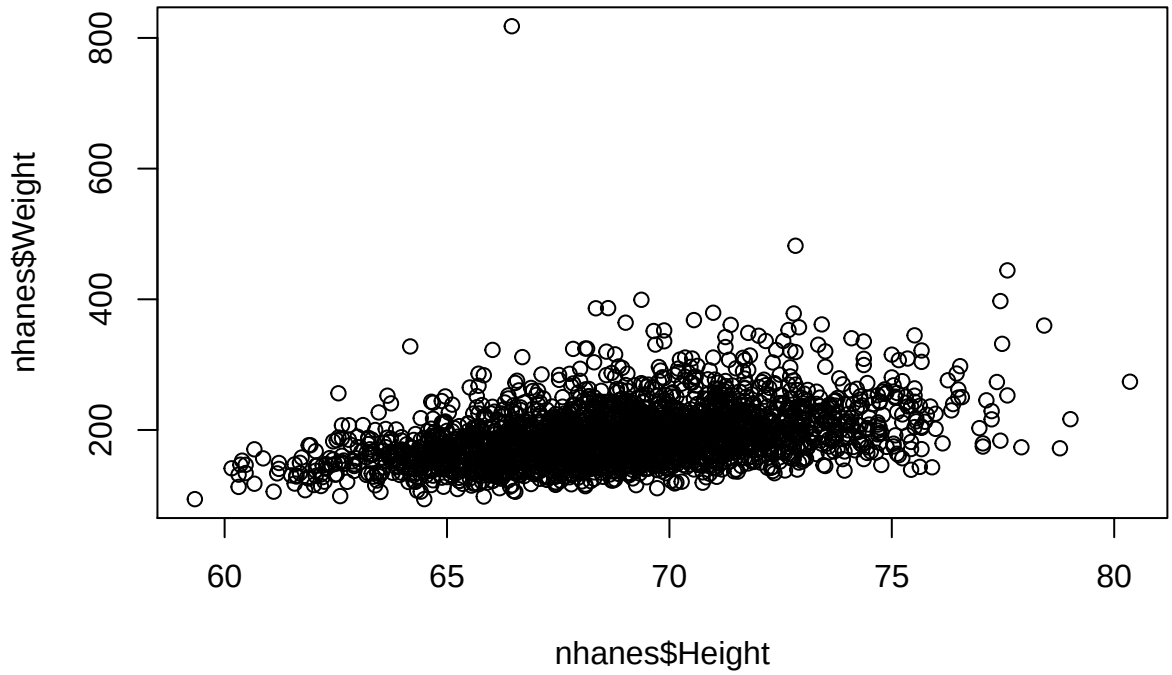
```
# Signed Rank Test
oatbran <- read.delim("http://myweb.uiowa.edu/pbreheny/data/oatbran.txt")
wilcox.test(oatbran$CornFlakes, oatbran$OatBran, paired=TRUE)

# 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.

```
#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.

Quiz Review

Two-sample categorical data

Chi-square Be able to create a contingency table with the outcome as the columns and the groups as the rows. Know how to compute a Chi-squared test. The formula for the chi-squared statistic is the sum of $\frac{(observed-expected)^2}{expected}$. Then find the value on a chi-squared table and **find the area to the right** (always take the complement of the probability you find on the table to compute p-value for observing something as extreme or more extreme than expected).

Fisher's exact test: Know Fisher's tests the same hypothesis as a chi-squared test, but is an **exact test**, not an approximation (chi-square). Especially useful when any expected cell count is below 5, and is necessary when the expected counts are lower than 1.

Odds Ratio Remember from a contingency table this is ad/bc. It is important you know how to interpret the odds. If the odds ratio > 1, for example 1.5, then we would say "The odds for (group 1) experiencing (outcome"Y") is 1.5 times the odds of (group 2) experiencing (outcome"Y")."

If the odds ratio is < 1, for example 0.6, we would say "The odds for (group 1) experiencing (outcome"Y") is 0.6 times the odds of (group 2) experiencing (outcome"Y"). OR"The odds for (group 1) experiencing (outcome "Y") are 40% lower than the odds of (group 2) experiencing (outcome "Y")." The CI for an odds ratio is exp(log(OR) +- Z*sqrt(1/a + 1/b + 1/c + 1/d)), where Z for a 95% CI is 1.96.

Two-sample continuous data

Recall, this test is used to determine whether the means of two groups from and unpaired dataset are significantly different. We can either use Welch's (computer only) or Student's test (can compute by hand). Know the difference between the two and why it would be appropriate to use one test over the other.

Review methods for handling outliers and poor skews. We can log transform the data, use a Mann-Whitney rank sum test, or use a permutation test. Know the difference between parametric and non-parametric.

Different study designs

Understand the differences between retrospective, prospective, and cross-sectional studies.

Example Questions

Example 1

A study examined 793 individuals who were in bike accidents. It was found that of 147 riders that wore helmets, 17 of them had a head injury. Of the rest of the bikers who did not wear helmets, 428 did not get a head injury. You may do this by hand or use R.

1) Make a contingency table for the data
2) Without running any tests, does there appear to be a benefit to wearing a helmet? (hint: odds ratio)
3) Make a 95% CI for this odds ratio
4) What are the expected counts for the contingency table?
5) Calculate a p-value using the chi-squared statistic and interpret.

Example 2

A study compared the miles per gallon of American cars (sample 1) to Japanese cars (sample 2). The sample size for American cars was 249 with a sample mean of 20.145 and sample standard deviation of 6.415. Japanese cars had a sample size of 79, sample mean of 30.481, and sample standard deviation of 6.108. (pooled standard deviation is 6.343)

1) If we assume data is normal what test do we run? What else might we consider to determine what test is most appropriate?
2) Conduct a t-test comparing the two group means and interpret the results
3) Further analysis shows that two of the American cars in the sample were getting less than 5 miles per gallon. How might this affect the test results? How might you remedy this issue?

Example 3

The Predators (a hockey team) just reached the second round of playoffs. I was curious if the Predators experienced any benefit to playing at home this season, so I gathered the data on how many goals they scored each game and whether they were home or away (regular season only). In home games they scored an average of 3.098 goals in 41 games with a sd of 1.841. In away games (41) they scored 2.237 with a sd of 1.43. The pooled sd is 1.650.

1) Which test would you use to examine the association between location and goals scored?
2) It seems they did better at home could this be a difference explained by chance alone?