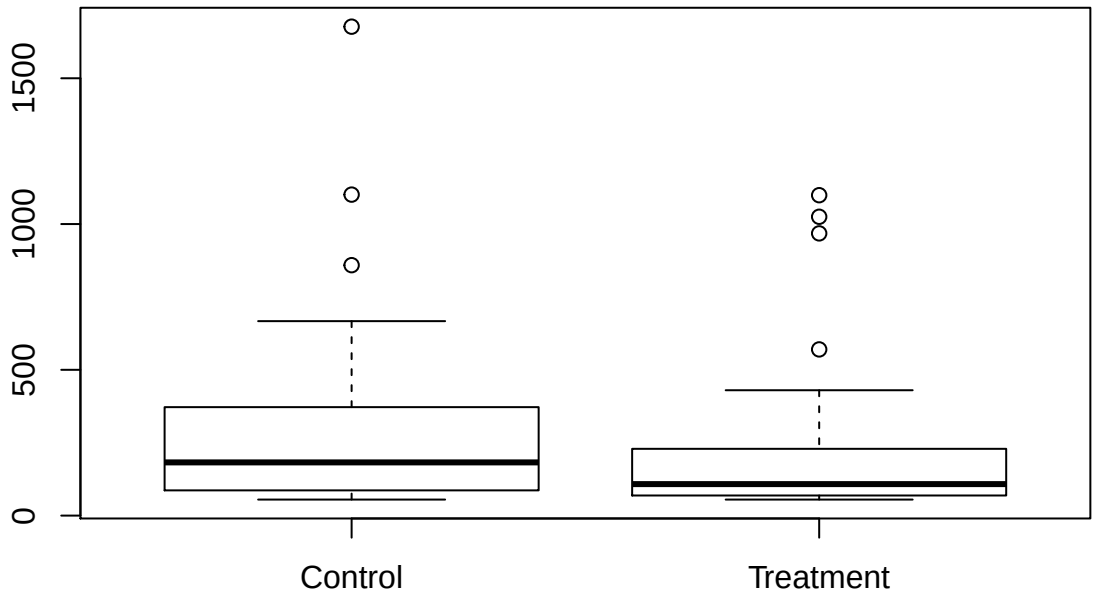
Lab 13: Transformations and nonparametric tests

Last week in lab, we began analyzing the Infant Diarrhea study. In lab #13 we will further analyze that data set using what we now know about outliers, transforming data, and non-parametric testing procedures.

Examining the data

We glossed over this bit last week, but now that we have ways to deal with it, we'll pay more attention. We begin by examining the distribution of the Infant Diarrhea data, stratified by group. Notice the right-skewness and the outliers.

```
babyPoop <- read.delim("http://myweb.uiowa.edu/pbreheny/data/diarrhea.txt")
boxplot(babyPoop$Stool~babyPoop$Group)
```



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

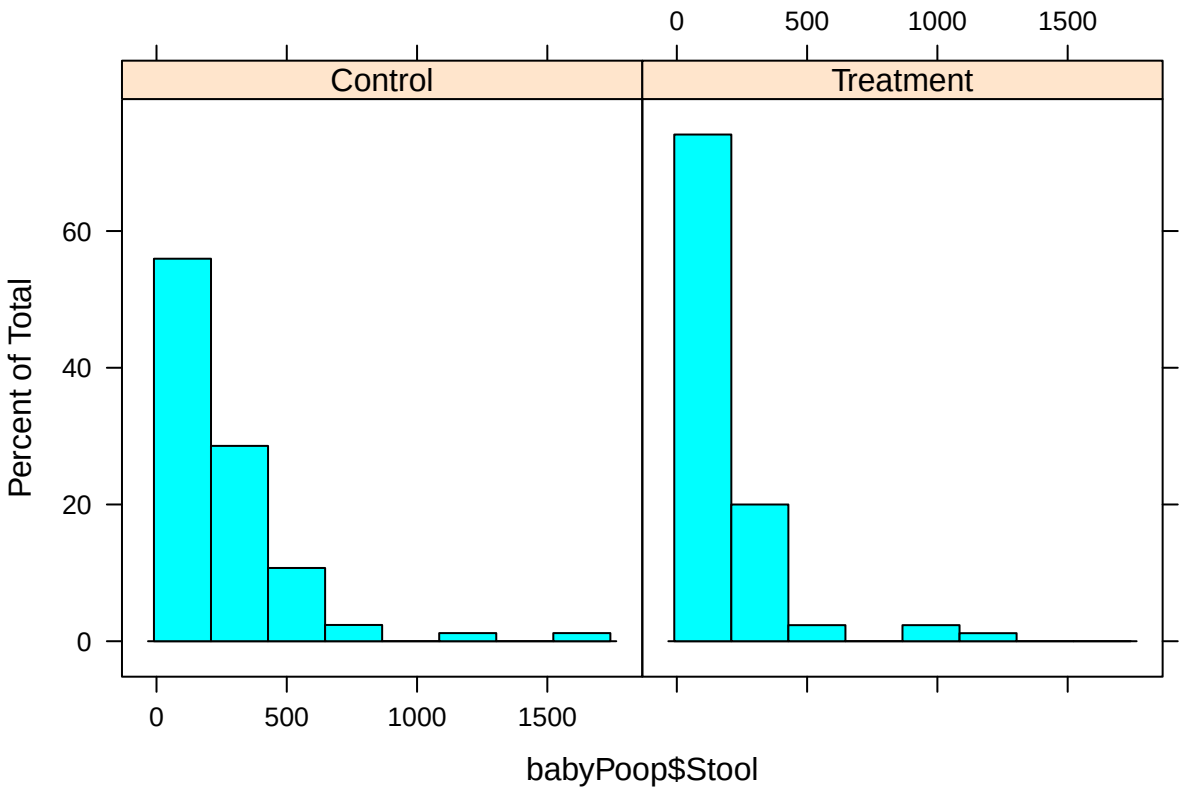
Pros and cons of "throwing out" outliers:
The severe outliers have a large impact on the two-sample t-test. It might not be desirable to have the analysis so highly affected by such a small percentage of the data. Other measures could be taken to analyze both groups. (see second half of lab)
Think about the dataset. Could there be any justification to just exclude the outliers?

Transformations

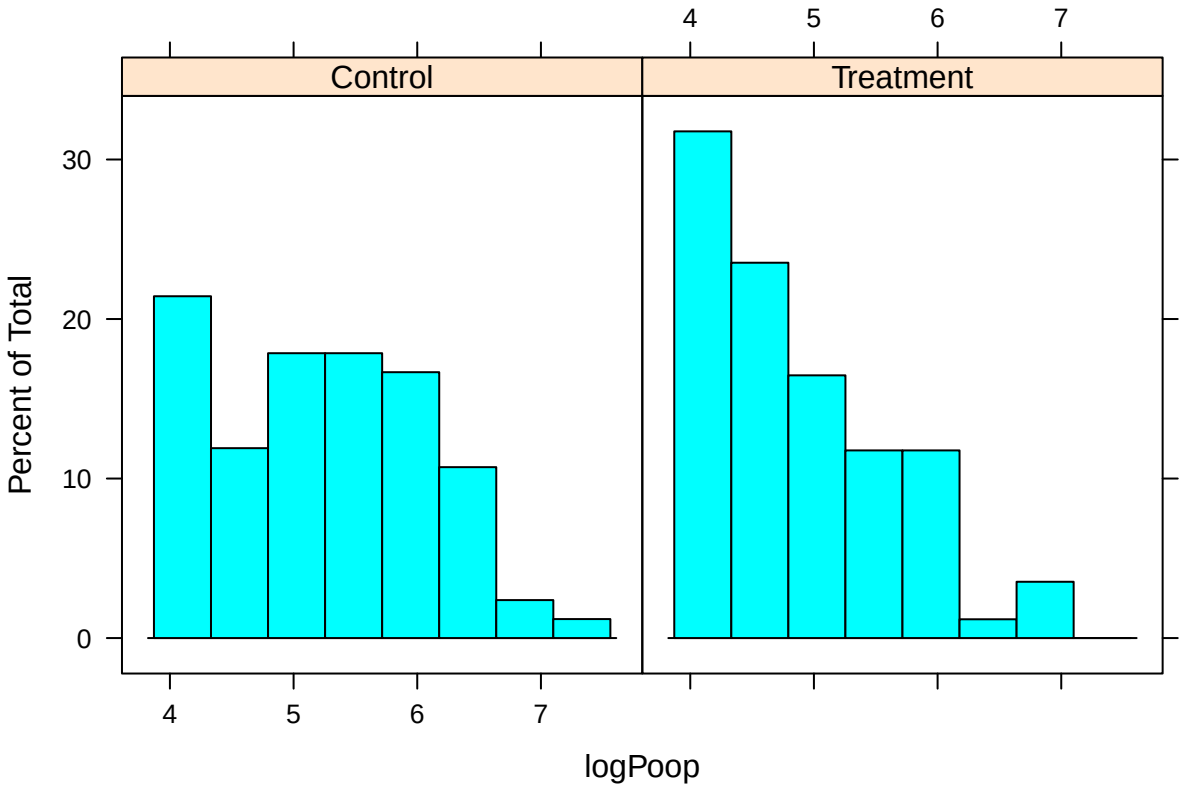
Recall the distribution of the Odds Ratio, and how it was right-skewed. We 'fixed' this skewness by transforming it to a new statistic (Log Odds Ratio) which is normally distributed. The same idea can be used for right-skewed data.

```
logPoop <- log(babyPoop$Stool)
library(lattice)

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




```
# Log-transformed
histogram(~logPoop | babyPoop$Group)
```



We can see there is still some skewness in the distribution of logStool; however, it is not as severe as before. We can do a two-sample t-test and achieve a more powerful result. Compare to the t-test with the original data.

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

```
##
##  Two Sample t-test
##
## data:  Stool by Group
## t = 2.245, df = 167, p-value = 0.02608
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    9.457362 147.396700
## sample estimates:
##   mean in group Control mean in group Treatment
##                260.2976                181.8706
```

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

```
##
##  Two Sample t-test
##
```

```
## data:  logPoop by babyPoop$Group
## t = 2.82, df = 167, p-value = 0.005383
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1025092 0.5810823
## sample estimates:
##   mean in group Control mean in group Treatment
##                5.212370                4.870574
```

Also note the confidence limits provided are on the log scale. In order to obtain a more interpretable interval, we need to exponentiate them.

```
logTest <- t.test(logPoop~babyPoop$Group,var.equal=TRUE)
exp(logTest$conf.int)
```

```
## [1] 1.107948 1.787973
## attr(,"conf.level")
## [1] 0.95
```

```
estimate <- logTest$estimate[1]-logTest$estimate[2]
names(estimate) <- NULL
exp(estimate)
```

```
## [1] 1.407473
```

Note that 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).
The point estimate for the ratio is 1.41, as in: infants in the control group have 1.41 times more diarrhea than the treatment group.

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 / Wilcoxon Rank Sum Test (same thing, two names)
One-sample (or paired) studies: Wilcoxon Signed-Rank Test
Both continuous: Spearman Correlation

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 values.

```
# Rank-Sum Test
with(babyPoop, wilcox.test(Stool~Group))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Stool by Group
## W = 4452, p-value = 0.005573
## alternative hypothesis: true location shift is not equal to 0
```

Ranking the data minimizes the impact of outliers, and allows us to not make assumptions on the underlying distribution of the data.

If the data we are analyzing is matched/paired, the signed-rank test is a non-parametric procedure that takes in to account the relatedness between the two groups. The signed-rank test is analogous to a paired t-test.

```
# Signed Rank Test
# Note: Bringing in the Oatbran study, because it's paired and familiar.
oatbran <- read.delim("http://myweb.uiowa.edu/pbreheny/data/oatbran.txt")
with(oatbran, wilcox.test(CornFlakes, OatBran, paired=TRUE))
```

```
##
##  Wilcoxon signed rank test
##
## data:  CornFlakes and OatBran
## V = 93, p-value = 0.008545
## alternative hypothesis: true location shift is not equal to 0
```

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

```
##
##  Paired t-test
##
## data:  CornFlakes and OatBran
## t = 3.3444, df = 13, p-value = 0.005278
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1284606 0.5972537
## sample estimates:
## mean of the differences
##               0.3628571
```

Parametric advantages: more powerful when distribution assumptions hold, and are straightforward with construction of confidence intervals

Nonparametric advantages: Minimal assumptions, more powerful when parametric assumptions are invalid.

The bottom line of this lab (literally): when you have continuous data, don't blindly apply a t-test. Look at the data, and if it's skewed or contains large outliers, consider a transformation or a rank-based analysis.