BIOS 4120: Introduction to Biostatistics
Breheny

Lab #13

In the previous 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.

1. Infant diarrhea study

We will begin by examining the distribution of the Infant Diarrhea data, stratified by their group. Pay particular attention to the right-skewness and the outliers.

```
boxplot(Stool ~ Group, col = "gray")
```

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

Discuss the pros and cons of 'throwing out' the outliers.

- The severe outliers have a large impact on the two-sample t-test. This might not be desirable to have your analysis so highly affected by such a small percentage of your data. Other measures could be taken to analyze both groups.

- Think about the data set we just analyzed. Can it be justified to discard the outliers?

2. Transformations

Remember 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) where it is normally distributed. The same idea can be used for right-skewed data.

```
logStool <- log(Stool)
require(lattice)
histogram(~logStool | 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.

```
t.test(logStool ~ Group, var.equal = TRUE)
t.test(Stool ~ Group, var.equal = TRUE)
```

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

```
exp(5.212-4.871)
exp(0.103)
exp(0.581)
```

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

3. 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/Wilcoxon Rank Sum Test.
- One sample studies: Wilcoxon Signed-Rank Test.
- Both variables 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.

```
wilcox.test(Stool ~ Group)
```

We get a p-value of 0.006. What do we conclude? Compare to the two previous two-sample t-tests.

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

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.

4. Wilcoxon Signed Rank Test

If the data you 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.

In R, you can do the Wilcoxon Signed Rank test using the wilcox.test function with the argument paired = TRUE.

As an example we will use the oat bran and corn flakes cholesterol data that we have previously looked at in several homework assignments.

```
wilcox.test(CornFlakes, OatBran, paired = TRUE)
```

Recall that the paired t-test resulted in a p-value of 0.005, compare this to results from the above approach.

The bottom line of this lab: 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.