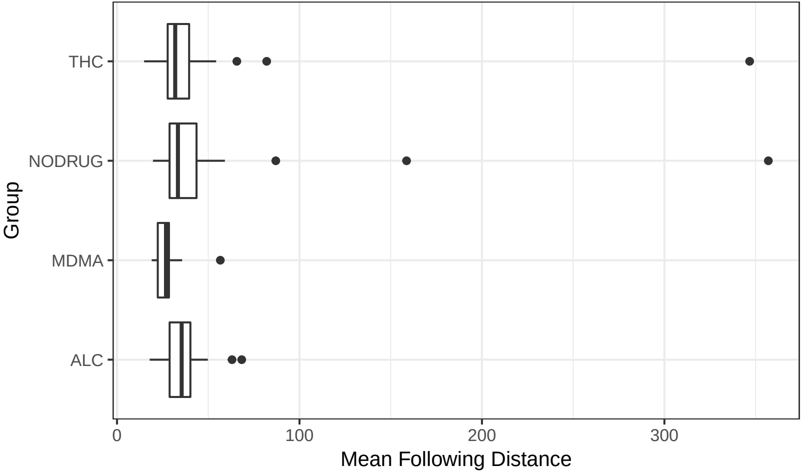
Transformations and Outliers

Ryan Miller

Outliers

- We've previously analyzed the tailgating data when learning about the Bonferroni adjustment
- In that analysis I neglected to show several large outliers



Practice - Outliers

With your group, load the Tailgating Data into Minitab (the variable "D" contains each subject's average following distance), then:

1. Use Minitab to evaluate the difference mean following distances for the MDMA and THC groups using a two-sample *t*-test (Hint: perform the test using summary statistics)
2. Manually delete the outlier in the THC group and repeat the test
3. How does the *p*-value change after deleting the outlier?

Outliers

- With the outlier included the p-value of the t-test is 0.09
- If the outlier is deleted, the p-value is 0.03
- It is tempting to remove the outlier, imagine your team spent hundreds of hours on this study...
  - But *should* the outlier be discarded?
- Selectively discarding data raises major ethical questions
  - p-values calculated when data is selectively discarded are at best questionable and at worst meaningless
  - Unfortunately these situations occur regularly and can be impossible for outsiders discover

Is it Ever Okay to Remove Outliers?

- Discarding recorded data should be approached with caution, but sometimes there are valid reasons to remove outliers:
  - Recording/measurement errors (a pulse of 0, or a "teen" with an age of 155)
  - Or, in the tailgating study, the outliers could have been individuals who weren't taking the study seriously
  - In any of these cases those values don't belong in the analysis and should be excluded
- When outliers are real data points, it is better to alter the analysis approach instead of manipulating the raw data
- Sometimes, outliers are the most interesting and important components of the data
  - A famous example involves NASA's monitoring of the Earth's ozone layer

Ozone, Outliers, and the Nimbus-7

- In the mid 1980's a large hole in the ozone layer above Antarctica was discovered, garnering worldwide attention
- Since the early 1970's, NASA had been monitoring the Earth's atmosphere using data collected by the satellite Nimbus-7
  - However, this data collection seemed to have completely missed the ozone hole! Or did it...
- Technology in the 1970s was prone to measurement errors, leading scientists build data processing programs that automatically discarded certain unusual observations as errors
- During the controversy of the 1980's, scientists revisited the Nimbus-7 raw data (including what was automatically being discarded)
  - Evidence of the ozone hole existed nearly a decade earlier, but in the data that was being automatically excluded!

How to Analyze Data with Outliers

- As demonstrated in the drug use and tailgating example, outliers can severely reduce the power of many statistical tests
- There are two popular approaches to analyzing data with outliers:
  - **Transform** the variable of interest so that its distribution is more normal
  - Use a **non-parametric test** (ie: test the difference in medians)
- We'll first explore the former approach, specifically using **log-transformation**

Logarithms

- Statisticians use "log" to refer to what most call the *natural logarithm*:

$$log(X) = T \leftrightarrow X = e^T$$

- A key property of logarithms is that differences on the log-scale correspond with ratios on the original scale *after exponentiating*:

$$log(X) - log(Y) = log(X/Y)$$

$$e^{log(X/Y)} = X/Y$$

Logarithms - Example

- For the tailgating study, we can perform a log-transformation using a Minitab formula creating new columns containing "log(following distance)"
- For the THC and MDMA groups, the means of this new variable are 3.54 and 3.28 respectively
  - The difference in means on the log-scale is 0.26
  - $exp(0.26) = 1.30 =$ mean following distance of THC group is *30% higher* than the mean in the MDMA group
- This concept applies to confidence intervals too:
  - The 95% CI on the log-scale is (0.05, 0.47), which we can exponentiate to (1.05, 1.60)
  - So we can be 95% confident the mean following distance is somewhere between 5% and 60% higher for THC users in the population these data represent

Logarithms - Additional Details

- On a technical note, $\sum log(x_i)/n \neq log(\sum x_i/n)$; the exponentiated mean of the log-transformed data is actually the *geometric mean*
    - So 1.30 (in the last example) was actually the ratio of geometric means, not the ratio of arithmetic means
    - This is a technical detail which I mention for completeness, it is not an important distinction practically speaking
    - the big picture take-away is that analyzing the log-transformed data allows us to measure *relative changes* across groups (after the transformation is undone via exponentiation)

Example - Applying the Log-transformation

Using the Tailgating Data:

1. Use a Minitab formula to create a new variable: “LogDistance”, check that it matches the existing variable “LD”
2. Construct the 95% confidence interval for the mean *relative increase* in following distance of No Drug and THC users
3. Perform a two-sample t-test using the log-transformed data for No Drug and THC groups, compare the results with a two-sample t-test on the untransformed data

Example - Solution

1. Not shown
2. The 95% CI on the log scale is (-0.151, 0.318), exponentiating the interval yields (0.86, 1.37) which it's plausible that the mean following distance in the "no drug" group could be anywhere from 14% shorter to 37% longer than the THC group
3. The test statistic on the log scale is 0.71 and the *p*-value is 0.478, on the original scale the test statistic is 0.39 and the *p*-value is 0.70.

The test is much more powerful on the log-transformed data, though neither test indicates a statistically significant difference in the average following distance of these two groups.

Remarks

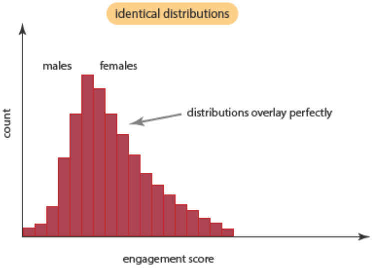
- There are many transformations that statisticians sometimes apply to non-normally distributed data
  - The log-transformation is popular because it retains interpretability (we can use exponentiation make relative comparisons)
- **Non-parametric** tests are a completely different alternative to transforming the data
  - In the slides that follow I will *briefly* introduce a couple of non-parametric analogs to the one-sample and two-sample t-tests
  - You *will not be responsible* for understanding the details of these tests, but you should be aware of when they might be used (and you should consider them for your final project)

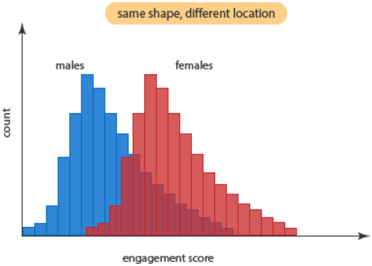
Wilcoxon Signed-Rank test (one-sample test)

- The **Wilcoxon Signed-Rank test** is a non-parametric analog to the one-sample t-test (single mean)
  - It is most often used to test whether the *median difference* in a paired design is zero
- Formally, the test specifies $H_0 : m = m_0$, or the median is some theoretical median
  - It proceeds by ranking the data-points (1:N) based upon how far they are from m_0
  - Next signs are given to these ranks based upon whether the data-point was above m_0 (+ sign) or below m_0 (- sign)
  - Under the null hypothesis, the sum of the signed-ranks is expected to be zero, so we can use this sum to derive a null distribution and a p-value (something we won't cover)

Mann-Whitney U-test (two-sample test)

- The **Mann-Whitney U-test** is commonly used as a non-parametric analog to the two-sample t-test (difference in means)
  - It tests whether the location of one distribution is *shifted* relative to another
  - In doing so it makes no assumptions regarding the shape of the distributions (they could both be skewed, have outliers, etc.)





Mann-Whitney U-test (two-sample test)

- Formally, the Mann-Whitney U-test specifies $H_0 : \text{dist}(X_1) = \text{dist}(X_2)$ and $H_A : \text{dist}(X) \neq \text{dist}(Y)$
  - It proceeds by ranking each data-point, regardless of group, from smallest to largest (1:N)
  - These ranks are summed within each group, yielding the quantities R_1 and R_2
  - R_1 and R_2, along with n_1 and n_2 are used to construct the U-statistic
- An exact test or a z-test can be performed using U (something we won't cover)

Example - Non-parametric Tests in Minitab

1. Using the tailgating data, use a Mann-Whitney U-test to evaluate the difference in following distances of the No Drug and THC groups. How do the results of this test compare with the *p*-value of the *t*-test on the log-transformed data (0.48), and the *p*-value of the *t*-test on the un-transformed data (0.70)?
2. Using the wetsuits data, create a new column "difference" and then use the Wilcoxon Signed-rank test to evaluate whether swim velocity when wearing a wetsuit differs from swim velocity without a wetsuit. How do the results of this test compare with the *p*-value of the paired *t*-test (0.000)?

Example - Solution

1. The p-value of the Mann-Whitney test is 0.43, a p-value that is very similar to the log-transformed result. This illustrates how a non-parametric test or a log-transformation can both be effective strategies for data with skew and/or outliers.
2. The p-value of the Wilcoxon Signed-Rank test is 0.003, when the assumptions of parametric tests (such as the paired t-test) are satisfied that test is generally the more powerful than its non-parametric analogs.

Conclusion

Right now you should...

1. Understand the concerns involved with excluding outliers from a statistical analysis
2. Recognize situations where removing outliers is a appropriate
3. Understand how to apply *and interpret* log-transformations
4. Be aware of non-parametric tests as an alternatives to the one-sample and two-sample t-test

If you'd like another perspective on this topics read Ch 15 of "Introduction to the Practice of Statistics" available here: http://bcs.whfreeman.com/webpub/statistics/ips9e/9781319013387/companionchapters/companionchapter15.pdf