Outliers and Transformations

Ryan Miller

Data with Multiple Groups

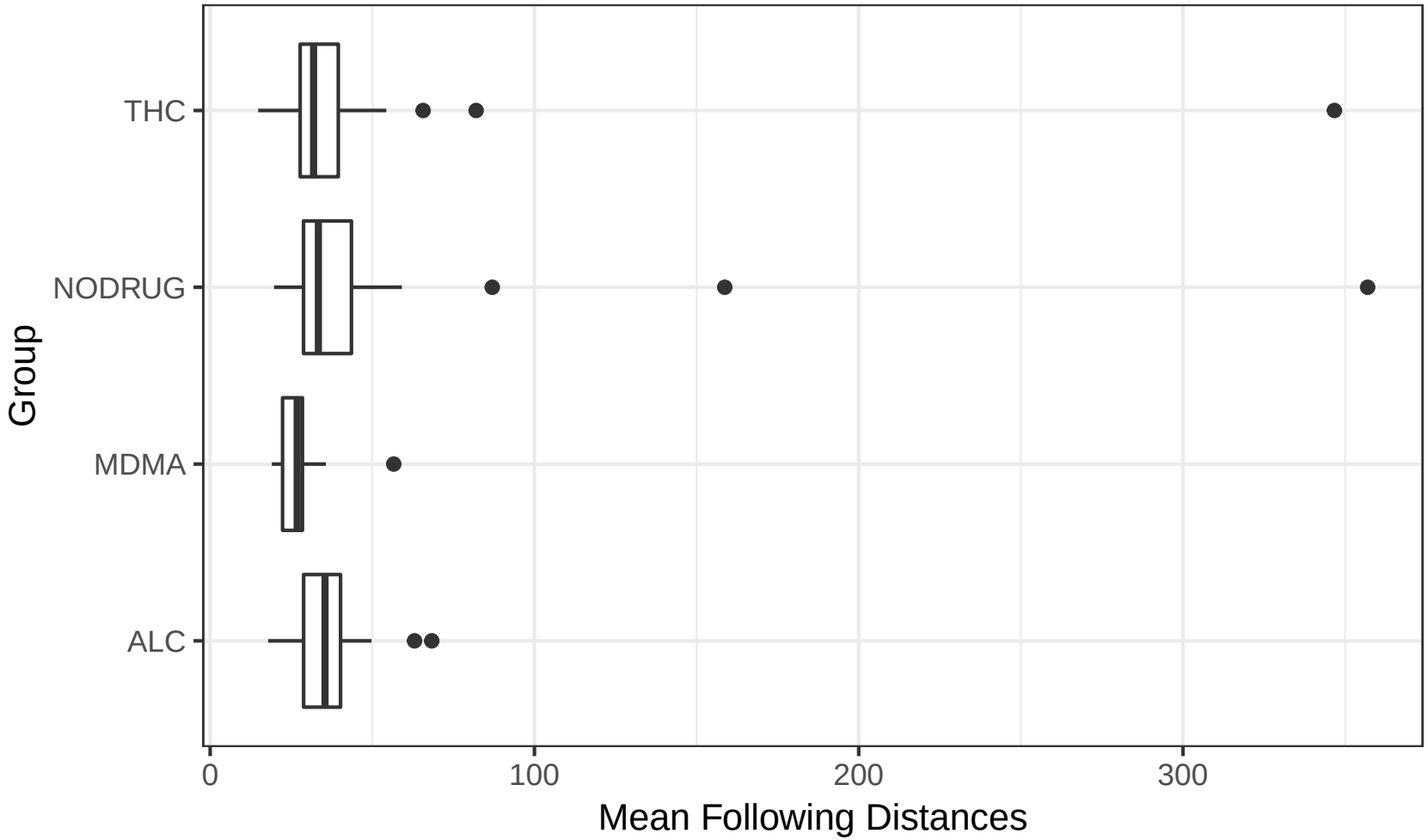
- The Chi-Squared test allowed us to avoid relying upon many differences in proportions tests when analyzing a categorical explanatory variable with more than two categories
- We will now shift our attention to a similar setting where:
  - We still have a categorical explanatory variable with more than two categories
  - The outcome variable is quantitative (rather than categorical like it was for the Chi-Squared test)
  - Before doing so, we will take a closer look at some difficulties that frequently arise when analyzing quantitative data

Drug Use and Tailgating

- A study conducted at the National Advanced Driving Simulator (NADS) aimed to link drug use with risky behavior in other areas (driving)
- In a driving simulator, subjects were told to follow a lead vehicle that was programmed to vary its speed unpredictably
  - As the lead vehicle erratically changed speed, more cautious drivers follow at a larger distance, while riskier drivers tailgate the vehicle
- The outcome variable was the mean following distance of each participant during the simulation
- The explanatory variable was the participant's drug use group: Alcohol, MDMA, THC, or no drugs used
  - Participants who used multiple drugs were classified according to the "hardest" drug they used (MDMA $>$ THC $>$ Alcohol)

Drug Use and Tailgating

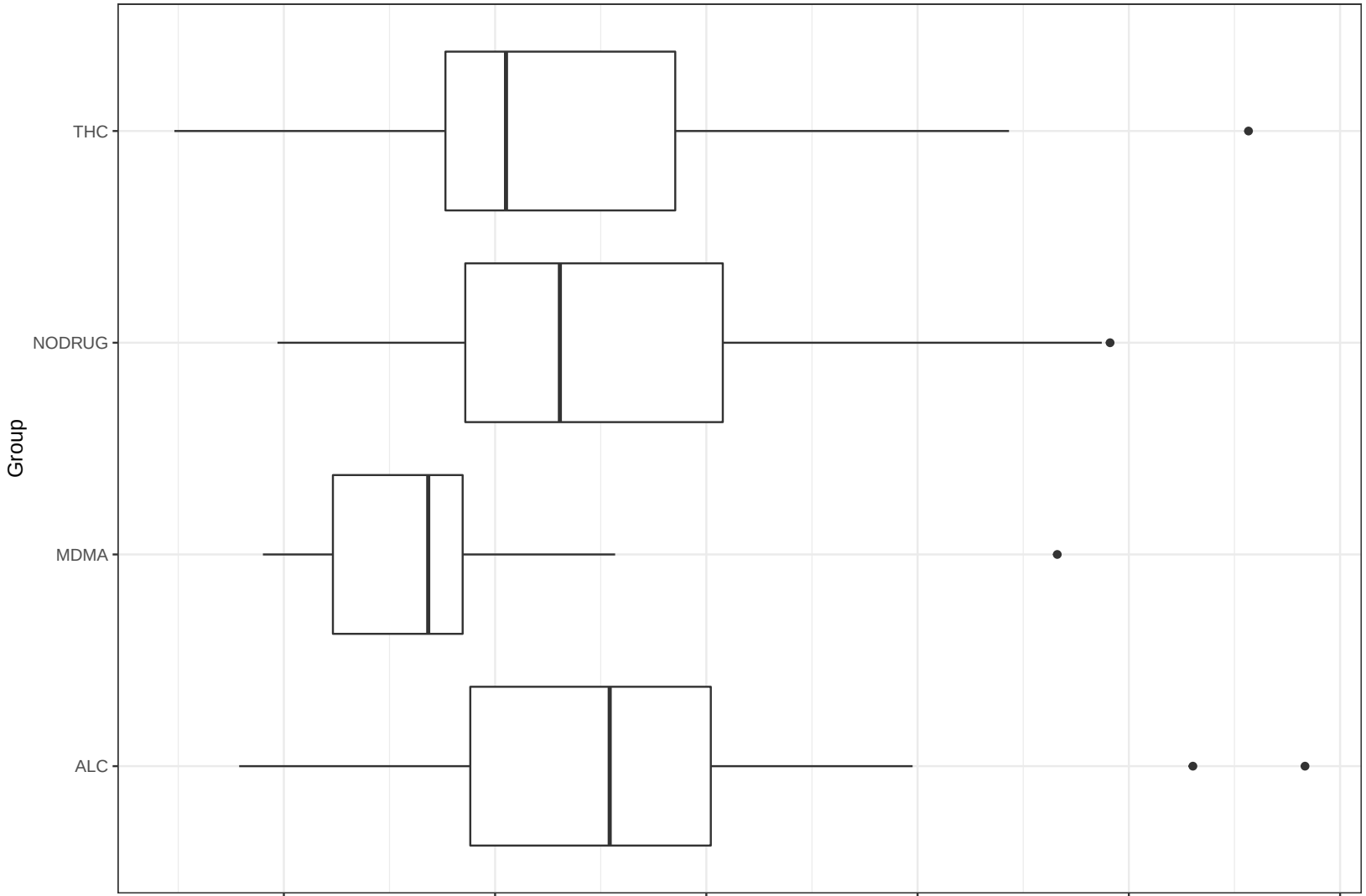
The plot below shows each individual's average following distance:



What can you conclude?

Drug Use and Tailgating

The large outliers in the "NODRUG" and "THC" groups make it difficult to see any trends. Here's what the data look like without those outliers:



Outliers

- Outliers can influence the results of approaches that rely on normality (such as the t-test!)
  - But how big of an impact do they make?
- With your group, load the Tailgating Data into Minitab (The variable "D" contains each subject's average following distance)
  1. Compare the mean following distance in the MDMA and THC groups using a two-sample t-test (Hint: do the test using summary statistics)
  2. Manually delete the outlier in the THC group and repeat the test
  3. How do the results of these two tests compare?

Outliers

- With the outlier included, the p-value of the t-test is 0.09, but if the outlier is deleted, the p-value is 0.03
- There is a temptation to remove the outlier, imagine you invested hundreds of hours on a study leading to an unconvincing p-value of 0.09
  - But *should* the outlier be discarded?
- Selectively choosing which data should be kept and which should be excluded raises ethical questions
  - The 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

What to do with Outliers

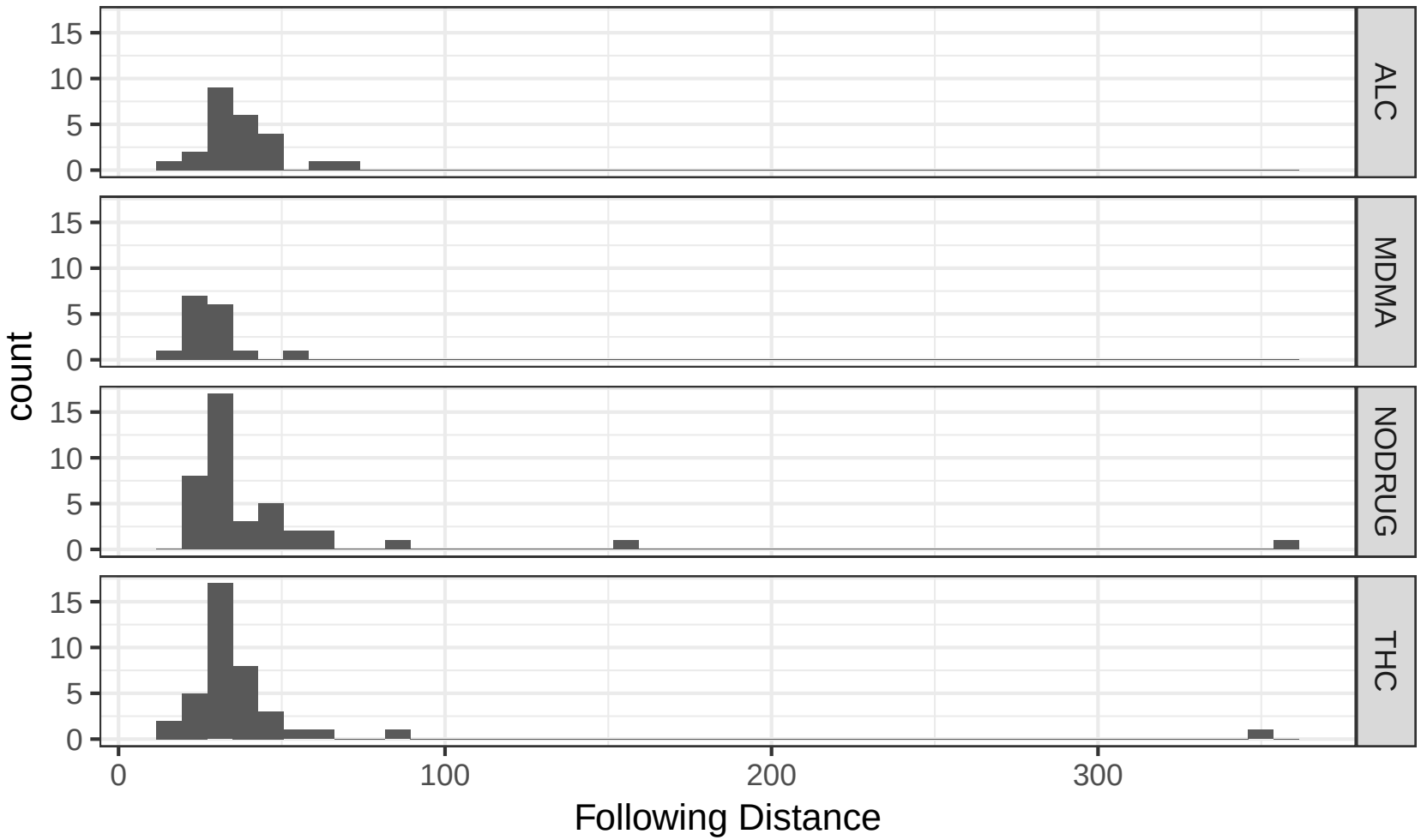
- Sometimes there are good reasons to remove outliers:
  - The outliers could be artifacts of recording/measurement errors (a pulse of 0, or a "teen" with age 155)
  - Or, in the tailgating study, the outliers could have been individuals who weren't taking the study seriously
  - In either case, those values don't belong in the analysis and should be excluded
- But when outliers are real data points, it is better to alter the analysis approach instead of manipulating the raw data
- Sometimes, outliers be the most interesting and important components of the data
  - A famous example illustrating the downsides of excluding real data outliers involves NASA's monitoring of the Earth's ozone layer

Nimbus-7 and Ozone Outliers

- 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 monitoring seemed to have completely missed the ozone hole
- The Nimbus-7's data was processed automatically in a way that discarded certain unexpected 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

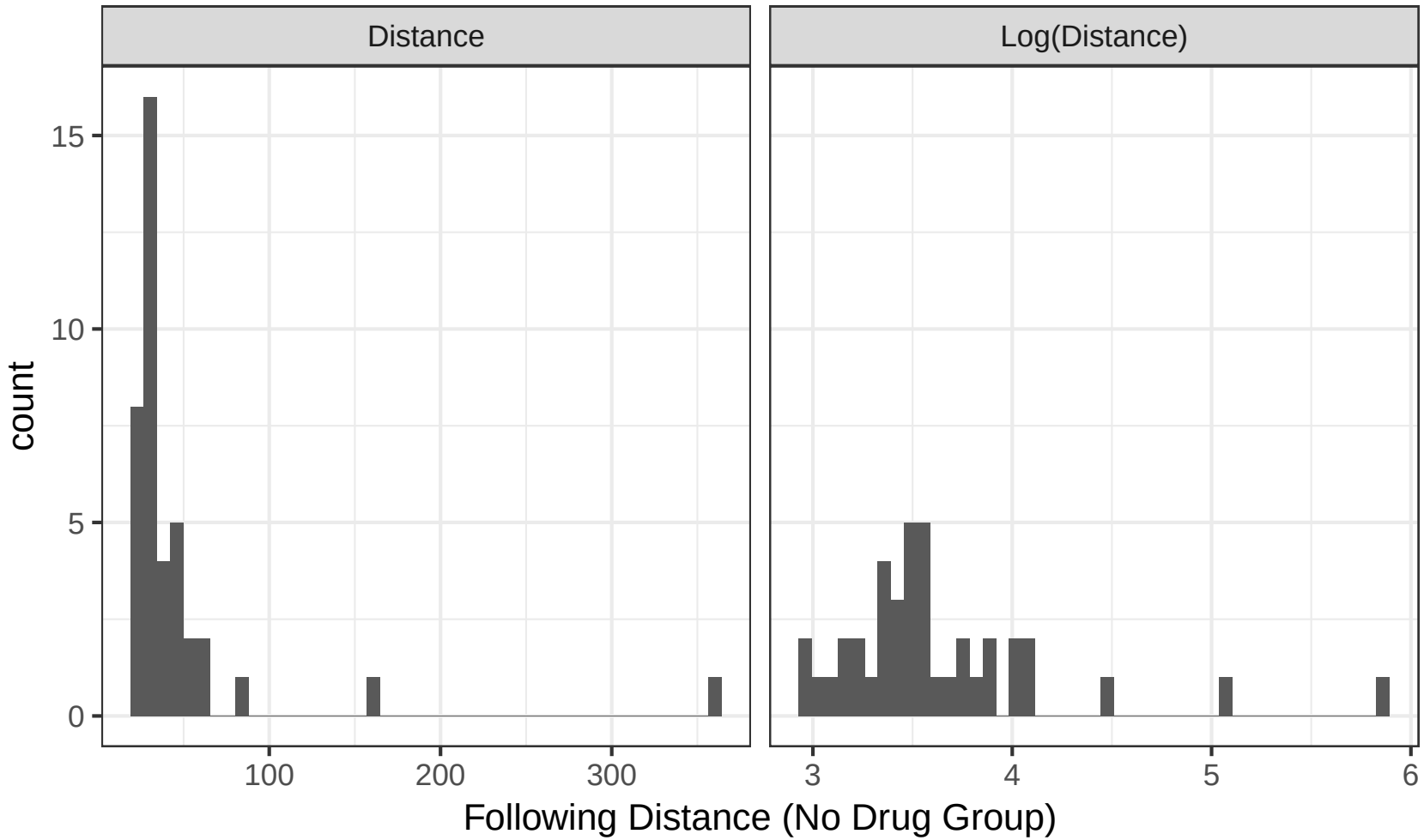
Transforming the Data

Assuming the outliers in the tailgating study are real and should be included, we have a problem with right skew:



Transforming the Data

A very common approach to analyzing right-skewed data is to apply a **log transformation**



Note: Statisticians use "log" to mean the *natural logarithm*

Transforming the Data

- After transforming these data, the normality assumption of the t-test is much more reasonable
- Because our outcome is now log(Distance), the way we interpret the t-test needs to change
- As an example we will use a t-test to compare following distances in the No Drug and THC groups
  - The observed sample statistic of interest is mean difference in log distances, which is 0.084
  - Differences on the log scale are *transformed ratios* on the original scale:

$$log(A) - log(B) = log(A/B)$$

  - Undoing the log transformation by exponentiating provides the relative change in group means:

$$exp(log(A/B)) = A/B$$

Transforming the Data

- For the tailgating study, $exp(0.084) = 1.09$
  - This tells us that mean following distance of No Drug group was 9% higher than the THC group
- On a technical note, $\sum log(x_i)/n \neq log(\sum x_i/n)$; so the exponentiated mean of the log-transformed data is actually the *geometric mean*
  - So 1.09 is actually the ratio of geometric means, rather than the ratio of arithmetic means, which is 1.11 for these data
  - This is a technical detail which I mention for completeness, it is not an important distinction in any real sense
  - the big picture take-away is that analyzing the log-transformed data provides relative changes across groups (after the transformation is undone)

Transforming the Data - Example

- A possible advantage of using log-transformed data is that we can construct confidence intervals for the relative changes across groups
- To do this, we simply calculate a confidence interval in the usual way using the log-transformed data, then exponentiate the end points

Practice: With your group:

1. Create a new variable: "LogDistance" in Minitab, 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

Transforming the Data - Example (solution)

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 no drug group's mean following distance 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 some interpretability (it describes changes on a relative scale)
- **non-parametric** tests are an alternative to transforming the data
  - We won't cover non-parametric methods in this class, but they are a topic to be aware of

Conclusion

These notes are a pre-cursor to Ch 8 of the textbook. Right now you should be able to...

1. Understand the role of outliers in statistical procedures
2. Know when it is appropriate to discard outliers and when it isn't
3. Know how to apply a log-transformation to alieve concerns regarding outliers and right skew
4. Understand how to analyze log-transformed data (relative interpretations)

I encourage you to read Ch 8 of the book and its examples.