Quantitative Data with Multiple Groups - ANOVA, Outliers, and Transformations

Ryan Miller

Data with Multiple Groups

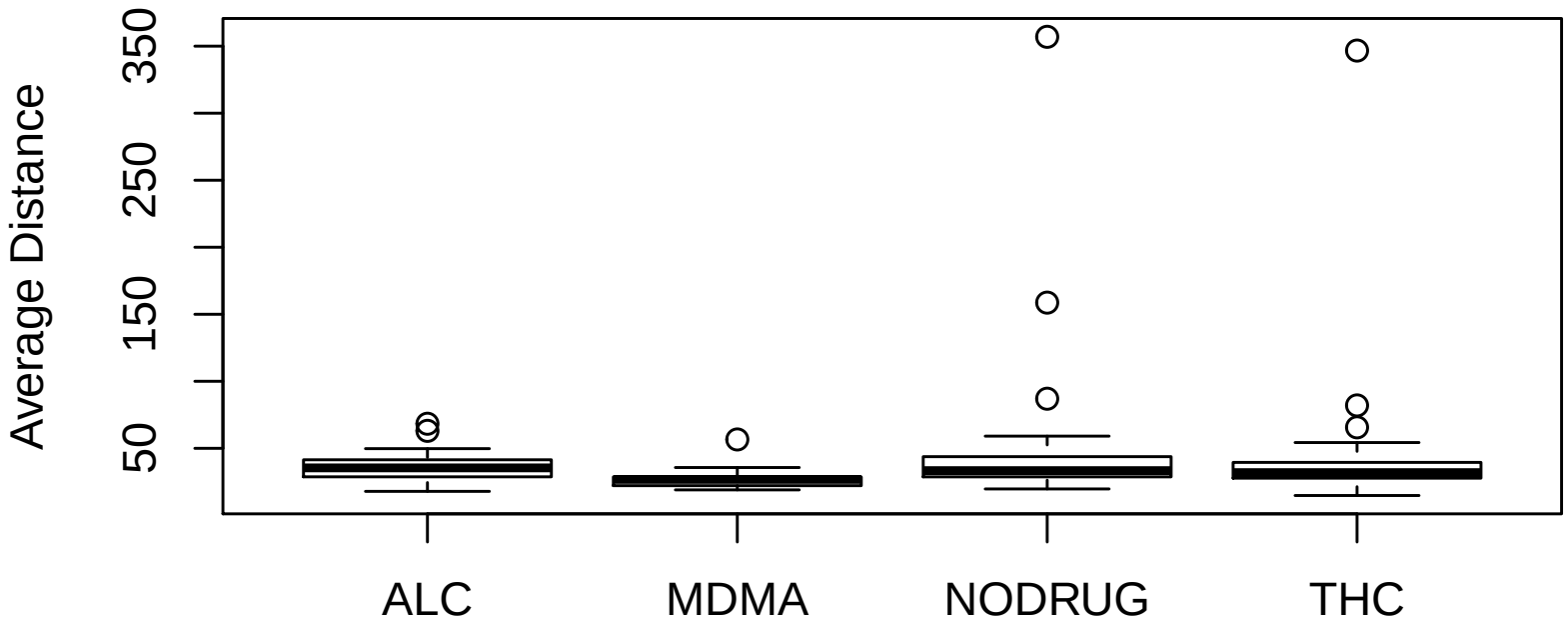
- A while ago, we learned how to analyze a single proportion and differences in proportions across two groups
- Later, we used Chi-Square tests to accommodate categorical variables with more than two groups (imagine a two-frequency table with more than 2 rows)
- Now we will learn how to analyze quantitative data with more than two groups
    - In doing so, we will also explore how to address some real data challenges (outliers and skew) that we've yet to talk much about

Drug Use and Tailgating

- Our example involves a driving study done at the National Advanced Driving Simulator (NADS), which attempted to link drug use with risky behavior in other areas (such as 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 tend to tailgate the vehicle
- The study's outcome variable was the average following distance of each participant
- The study's 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

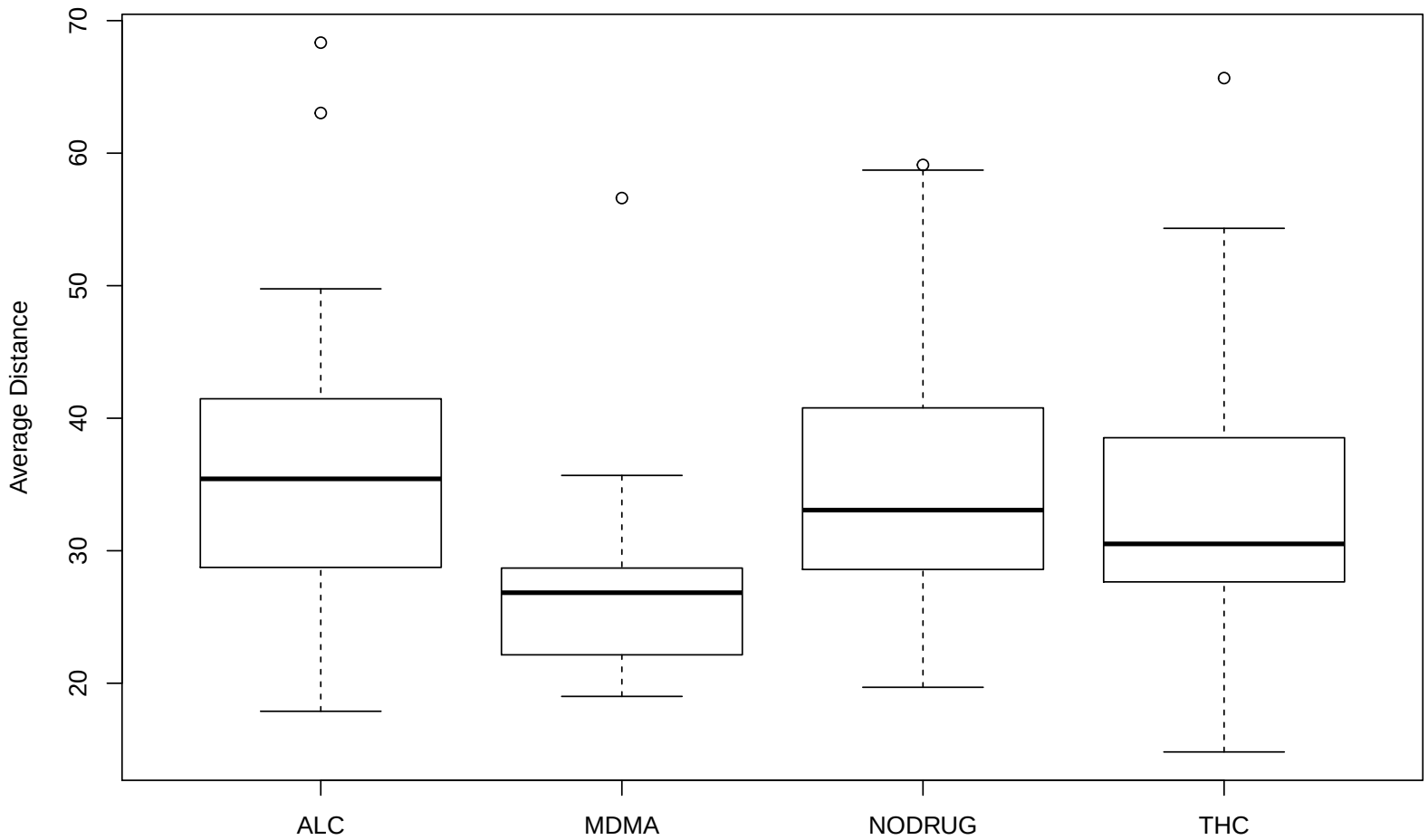
The plot below shows the average following distances across groups:



What can you conclude?

Drug Use and Tailgating

It's difficult to see what is going on in the data due to large positive outliers in the "NODRUG" and "THC" groups. 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 data ("Tailgating" on p-web) 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 (You may want to subset the rows of the data)
  2. Manually delete the outlier in the THC group and repeat the test
  3. How do the results of these tests compare?

Outliers

- When the outlier included, the p-value of the t-test is 0.09, when the outlier is deleted, the p-value becomes 0.03
- It is tempting to throw away the outlier, imagine you spend hundreds of hours on a study and got 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, this happens quite often and is nearly impossible for outside observers discover

What to do with Outliers

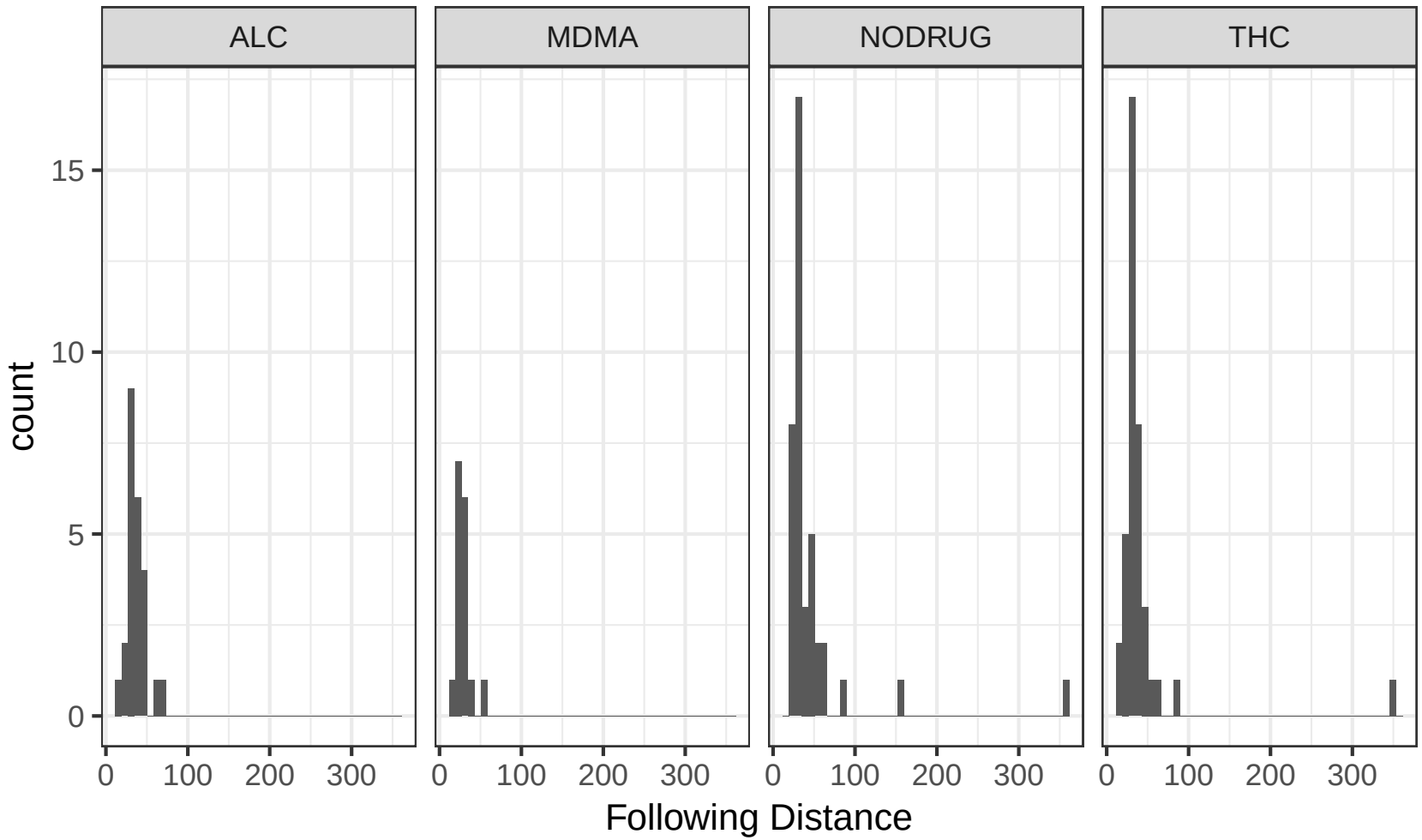
- Sometimes there are good reasons to throw away outliers
  - Certain outliers could be artifacts of recording or measurement errors (for example a subject with a pulse of 0 or an age of 155)
  - In the tailgating study, the outliers could have been individuals who weren't taking the study seriously
  - In either case, those values don't belong in the analysis and should be excluded
- When the outliers are real data points, it is better to alter the analysis approach than it is to manipulate the raw data
  - These outliers can sometimes be the most interesting and important aspects of the data
  - A famous example showcasing 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 was discovered in the ozone layer above Antarctica, garnering worldwide attention
- Since the early 1970's, NASA had been monitoring the Earth's atmosphere using data from Nimbus-7, a satellite which measured atmospheric conditions
  - The monitoring seemed to have missed the ozone hole
- The Nimbus-7's data was processed automatically in a way that discarded certain unexpected observations as errors
- After the controversy in the 1980's, scientists revisited the Nimbus-7 raw data (including what was automatically being discarded)
  - They found evidence of the ozone hole existed nearly a decade earlier
  - The evidence was in the outliers, which were programmed to be automatically excluded

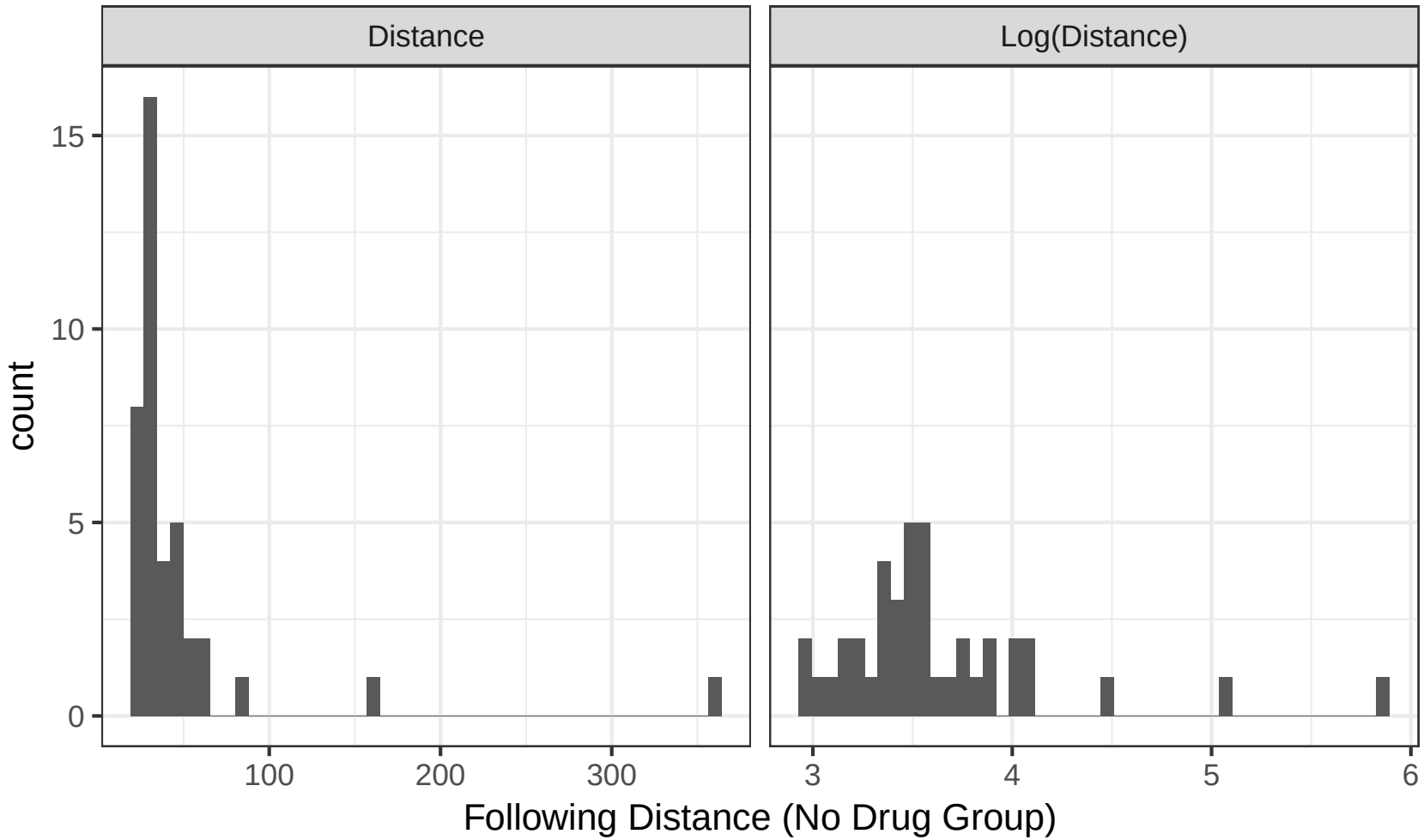
Transforming the Data

Assuming the outliers in the tailgating study are real and should be included, we still 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), how we interpret the t-test's results needs to change
- As an example we will use a t-test to compare following distance in the No Drug and THC groups
  - The observed sample statistic of interest is mean log distance, which is 0.084
  - Differences on the log scale are transformed ratios on the original scale:

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

  - 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, $exp(0.084) = 1.09$
  - So the mean following distance of No Drug group was 9% higher than the THC group
- Technically, $\sum log(x_i)/n \neq log(\sum x_i/n)$; so the exponentiated mean of the log-transformed data is actually the *geometric mean*
  - This means that 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 that is worth knowing for the sake of 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

Transforming the Data - Example

- An advantage of analyzing log-transformed data is that we can construct confidence intervals for the relative changes across groups
- To do this, we simple calculate a confidence interval the usual way on the log-scale and 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 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% larger 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, although neither test indicates a statistically significant difference in the average following distance of these two groups.

Comparing Multiple Groups

- Comparing the No Drug and THC groups is interesting, but there are 4 groups in the tailgating data
- We could make 6 pairwise comparisons, but this will increase the overall **type I error rate** of the study
- A key property of the p-value is that rejecting H_0 using a threshold of $\alpha = 0.05$ means the test has a 5% chance of making a type I error when the null hypothesis is true
  - If we perform 6 hypothesis tests in our analysis of an experiment we've now got 6 opportunities to make a type I error
  - The probability of making *at least one* type I error in the experiment is now much larger 5%
- There are two ways to analyze the data while limiting the experiment's overall type I error rate to 5%
  - We can adjust the significance threshold for the 6 pairwise tests
  - We can do a single joint test assessing the association between group and following distance

The Bonferroni Adjustment

- We will look at both of these approaches, starting with adjusting the significance threshold
- Suppose the null hypothesis is true for all 6 pairwise tests in the tailgating study (and the tests are independent); using $alpha = 0.05$:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05)^6 = 26.5\% \end{aligned}$$

- This calculation suggests a simple correction to significance threshold: $\alpha^* = \alpha/h$, where h is the number of hypothesis tests being performed

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05/6)^6 \approx 5\% \end{aligned}$$

The Bonferroni Correction

- Modifying the significance threshold to $\alpha^* = \alpha/h$ when conducting h hypothesis tests is known as the **Bonferroni correction**
  - This adjustment controls the **family error rate** (the experiment's overall type I error rate) at α
  - In some cases Minitab will provide Bonferroni corrected confidence intervals, which are created by using α^* to adjust the critical value (z^* or t^*) used to construct the interval
- The Bonferroni correction tends to be conservative; in most cases it actually controls the family error rate below α
- There are many other, more powerful, approaches to family error rate control that we won't discuss. The details of these approaches is far less important than understanding what they achieve

The Bonferroni Correction - Example

- With your group, perform the 6 pairwise t-tests on comparing the log following distance of each group and answer the following:
    1. How many differences are significant using an unadjusted significance threshold of $\alpha = 0.05$? Also, describe the expected family error rate using this threshold.
    2. How many differences are significant using the bonferroni corrected significance threshold? Also, describe the expected family error rate using this threshold.
    3. How does using the bonferroni correction impact statistical power?

The Bonferroni Correction - Example (solution)

1. ALC vs NODRUG, p-value $= 0.5102$
2. ALC vs MDMA, p-value $= 0.00417$
3. ALC vs THC, p-value $= 0.8959$
4. THC vs NODRUG, p-value $= 0.4782$
5. THC vs MDMA, p-value $= 0.01383$
6. MDMA vs NODRUG, p-value $= 0.00216$

The bonferroni corrected significance threshold when conducting 6 hypothesis tests is $0.05/6 = 0.0083$; so we conclude that the following distances are different for ALC and MDMA groups, as well as the MDMA and NODRUG groups.

ANalysis Of VAriance (ANOVA)

- Using the Bonferroni correction, we can analyze the tailgating study data using 6 pairwise t-tests without increasing our chances of making a type I error
  - Unfortunately, this decreases the statistical power of the experiment
- Another option is to use a single test of the hypothesis:

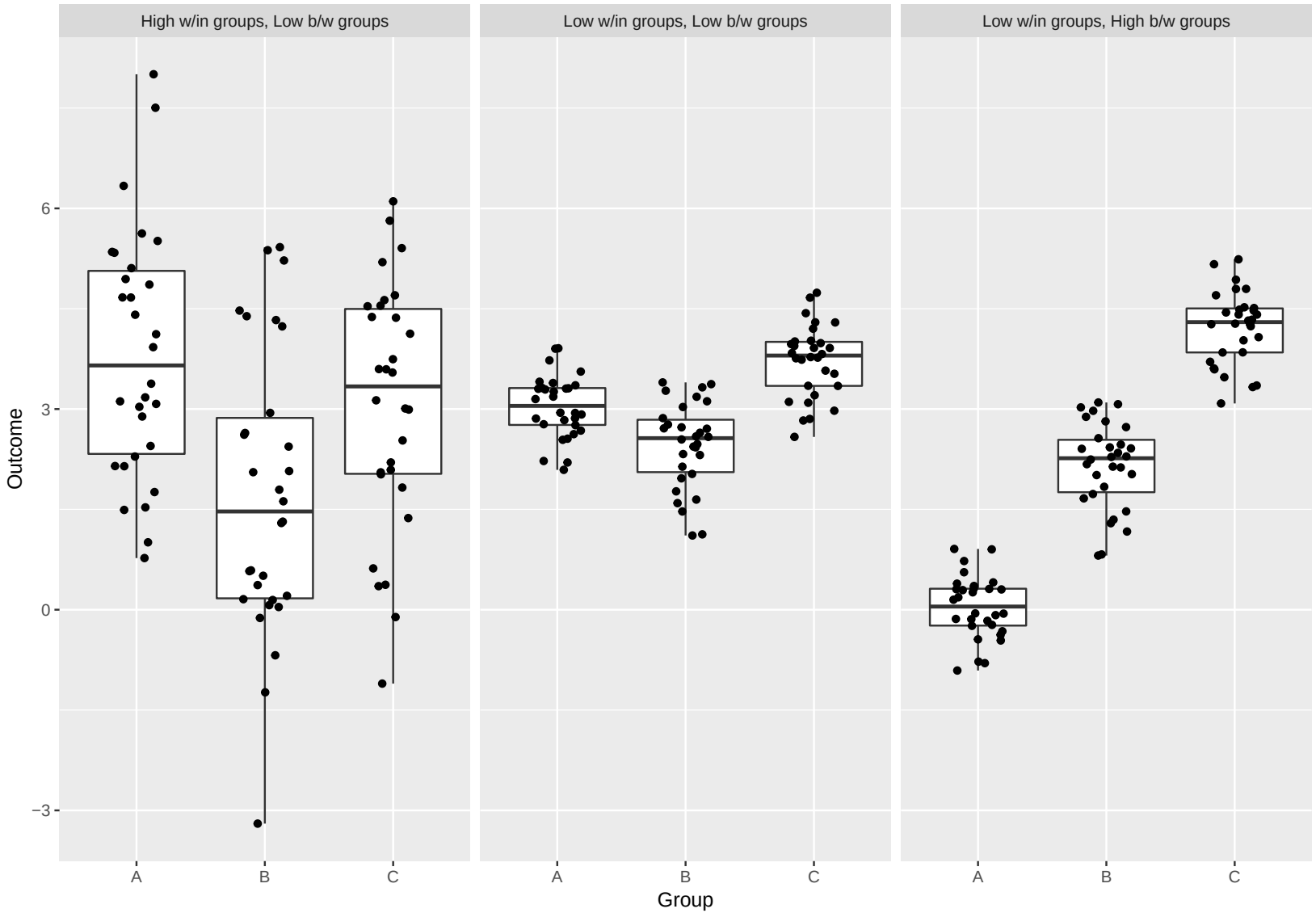
$$H_0 : \mu_{ND} = \mu_{THC} = \mu_{ALC} = \mu_{MDMA}$$

- The test we will use is: "Analysis of Variance" or ANOVA

Partitioning Variability

- The idea behind ANOVA is to split the total variability in the data into two pieces
  - The variability between groups
  - The variability within groups
- If the grouping variable is clearly associated with the quantitative outcome, we expect there to be much more variability between groups than there is within groups

Partitioning Variability



How can we Measure these Sources of Variability?

- To understand how ANOVA works, we need to briefly discuss *statistical modeling*
- A model is a simplified characterization of how the world works
  - The goal of a model is to explain the variability in a certain outcome variable (ie: a child's adult height can be predicted using their age, their parents height, etc.)
  - Generally, it is impossible for a model to perfectly explain of the variability in the outcome (no model can predict exactly how tall every single child will be)
  - But some models are better than others (they are capable of explaining more of the uncertainty in height)

Modeling Distance in the Tailgating Data

- Returning to the tailgating study, the simplest model is one where a single mean tailgating distance is the best prediction for everyone in the study (this model implies all variability about that mean is unexplainable)
  - This simplest possible model is sometimes called the *null model*
- Under this model, every subject deviates from the mean, $\bar{y}$, by a **residual** denoted:

$$r_i = y_i - \bar{y}$$

- We can summarize the size of the model's residuals using a **sum of squares**:

$$SST = \sum_i r_i^2 \text{ for the null model}$$

- We call this SST (sum of squares total) because this is the largest possible sum of squares of any model

Modeling Distance in the Tailgating Data

- This sum of squares measures variability, it is quantifies how much variability in the outcome variable isn't explained by the null model
  - The exact value of SST doesn't tell us much by itself, but it does provide a good baseline for comparison
- With the tailgating data, we can consider a more complex alternative model where each group had its own unique mean
- Under this model, the residuals look like:

$$r_i = y_i - \bar{y}_i$$

- $\bar{y}_i$ is now the mean for the group of which subject i is a member
- The sum of squares describing this model's residuals is denoted SSE (sum of squares error), in reference to the accuracy of this alternative model's predictions

Explained Variability (R-squared)

- R^2, the coefficient of variation, of the alternative model can be expressed using these two sums of squares:

$$R^2 = \frac{SST - SSE}{SST}$$

- For the tailgating data, $R^2 = 0.055$, so the model where each group gets its own mean explains 5.5% of the variability in tailgating distance (verify this yourself in Minitab)
- Any more complicated model will always have a lower SSE (this concept is called overfitting)
  - What we really want to determine whether SSE, the sum of squares for the alternative model, is lower more than would be expected to see by random chance when adding that complexity

Modeling and ANOVA

- The special type of model we've been considering, where a single categorical variable is used to predict a quantitative outcome, is actually an analysis of variance model (ANOVA)
- ANOVA uses the test statistic:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{\text{Std. Error}}$$

- d_1 and d_0 refer to the number of parameters in the model being considered and the null model, in the tailgating example $d_0 = 1$ (the single overall mean) and $d_1 = 4$ (each group's mean)
- The F statistic can be interpreted as the standardized drop in the sum of squares per parameter included in the alternative model

What is the Standard Error?

- Previously we've seen that standard errors tend to look like a measure of variability divided by the sample size
- In this setting:

$$\text{Std. Error} = \frac{SSE}{n - d_1}$$

- This is the sum of squares of the alternative model divided by its *degrees of freedom*, $df = n - d_1$
- Using this standard error, the F statistic can be expressed:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{SSE/(n - d_1)}$$

Connecting the F-test to Variability

- SST is the sum of squares for the null model, this model predicts each y_i using the overall mean $\bar{y}$
  - $SST = \sum_i r_i^2$ where $r_i = y_i - \bar{y}$
  - SST describes total variability in y
- SSE is the sum of squares for the alternative model, this model predicts each y_i using a group-specific mean $\bar{y}_i$
  - $SSE = \sum_i r_i^2$ where $r_i = y_i - \bar{y}_i$
  - SSE describes the variability that remains after accounting for group
- By subtraction, we can determine how much variability is being explained by the parameters included in the alternative model:

$$SST = SSE + SSG$$

- SSG, the sum of squares groups, denotes the amount of variability explained by using "group"

Connecting the F-test to Variability

- Using SSG, we can express the F-statistic as:

$$F = \frac{SSG/(d_1 - d_0)}{SSE/(n - d_1)}$$

- These sums of squares divided by their degrees of freedom are often called **mean squares**, this allows for a simpler looking F statistic:

$$F = \frac{MSG}{MSE}$$

- MSG is the mean square of groups, MSE is the mean square of error

The ANOVA Table

- Calculating sums of squares and mean squares by hand is extremely tedious and something we won't spend time doing in this class
- However, you will be expected to read and understand a common piece of software output known as the **ANOVA table**
- The general form of these tables is shown below:

Source	*df*	Sum Sq.	Mean Sq.	*F*-statistic	*p*-value
"Group"	$d_1 - d_0$	SSG	MSG	MSG/MSE	Use $F_{d_1-d_0, n-d_1}$
Error	$n - d_1$	SSE	MSE		
Total	$n - d_0$	SST			

- In the typical ANOVA setting:
  - $d_0 = 1$, the null model has one parameter, a single overall mean
  - $d_1 = k$, the alternative model has k parameters, a mean for each group

The ANOVA Table - Example

With your group, complete the following ANOVA table (assuming $d_0 = 1$):

Source	*df*	Sum Sq.	Mean Sq.	*F*-statistic	*p*-value
"Group"	4	200	?	?	?
Error	?	440	?		
Total	59	?			

Additionally, make a sketch of what the boxplots for these data might look like (disregarding units)

The ANOVA Table - Example (solution)

Here $k = 5$ and $n = 60$, so:

Source	*df*	Sum Sq.	Mean Sq.	*F*-statistic	*p*-value
"Group"	4	200	50	6.25	0.0003
Error	55	440	8		
Total	59	640			

- The *p*-value is found using the right-tail area beyond 6.25 of an F distribution with (4, 55) degrees of freedom
- The boxplots corresponding to this table will show high variability between groups and low variability within groups

ANOVA - Example

With your group, analyze the "Tailgating" in Minitab using ANOVA (Stat -> ANOVA -> One-Way), be sure to report:

1. Your null and alternative hypotheses
2. Your test statistic
3. Your p-value and a one sentence conclusion

ANOVA - Example (solution)

1. $H_0 : \mu_{ND} = \mu_{THC} = \mu_{ALC} = \mu_{MDMA}$
2. F $= 2.23$
3. The p-value here is 0.088. There is borderline evidence that drug use is predictive of following distance, it appears that the MDMA group is most different, with shorter following distances (on the log-scale)

Inference for Means after ANOVA

- The results of an ANOVA test only tell us whether or not a difference in group means exists, not the specific groups that are different
- The next step when the ANOVA test is statistically significant is to investigate which groups differ

CI for μ_i	$\bar{x}_i \pm t^* \sqrt{MSE/n_i}$
CI for $\mu_i - \mu_j$	$(\bar{x}_i - \bar{x}_j) \pm t^* \sqrt{MSE * (1/n_i + 1/n_j)}$
Test of $H_0 : \mu_i = \mu_j$	$t_{\text{test}} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE * (1/n_i + 1/n_j)}}$

- Each procedure/test uses t-distributions with $n - d_1$ degrees of freedom $(n - k)$

Inference for Means after ANOVA - Example

Practice: With your group, conduct a follow up analysis of the "tailgating" data using ANOVA and answering the following questions:

1. Which groups are most different?
2. Which groups are least different?
3. Construct and interpret the confidence interval for the difference between the "NODRUG" and "MDMA" groups (remember the variable "LD" is on the log scale)

Inference for Means after ANOVA - Example (solution)

1. This can be determined by comparing means, NODRUG and MDMA are the most different
2. THC and ALC are the least different
3. $(3.63 - 3.28) \pm t^* \sqrt{0.212 * (1/40 + 1/16)}$, $df = 115$, so $t^* = 1.98$ and the interval on the log-scale is: (0.08, 0.62), by exponentiating the endpoints we get: (1.08, 1.86).

We conclude that the mean following distance of the NODRUG group is between 8% and 86% greater than the mean following distance of the MDMA group.

More on ANOVA and Modeling

- In ANOVA, we use a single categorical variable to predict a quantitative variable
  - If using that categorical variable improves prediction beyond what could be attributed to random chance, the ANOVA test will be statistically significant
- Statistical modeling is an extremely broad topic, it is so vast that you'd likely need several courses to cover it thoroughly
- Nevertheless, for our next topic we will return to regression
  - ANOVA actually is a special case of regression modeling where the predictor variable is a categorical
  - Our goal will be to build and understand *multiple regression* models that involve several predictor variables that can be categorical or quantitative

Conclusion

These notes cover Ch 8 of the textbook. Right now you should be able to...

1. Understand how ANOVA testing relates to statistical modeling
2. Understand the partitioning of variability in ANOVA
3. Make conclusions using ANOVA table
4. Fill out an incomplete ANOVA table
5. Conduct the appropriate follow-up analyses after ANOVA

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