Analysis of Variance (ANOVA)

Ryan Miller

Analysis Of Variance (ANOVA)

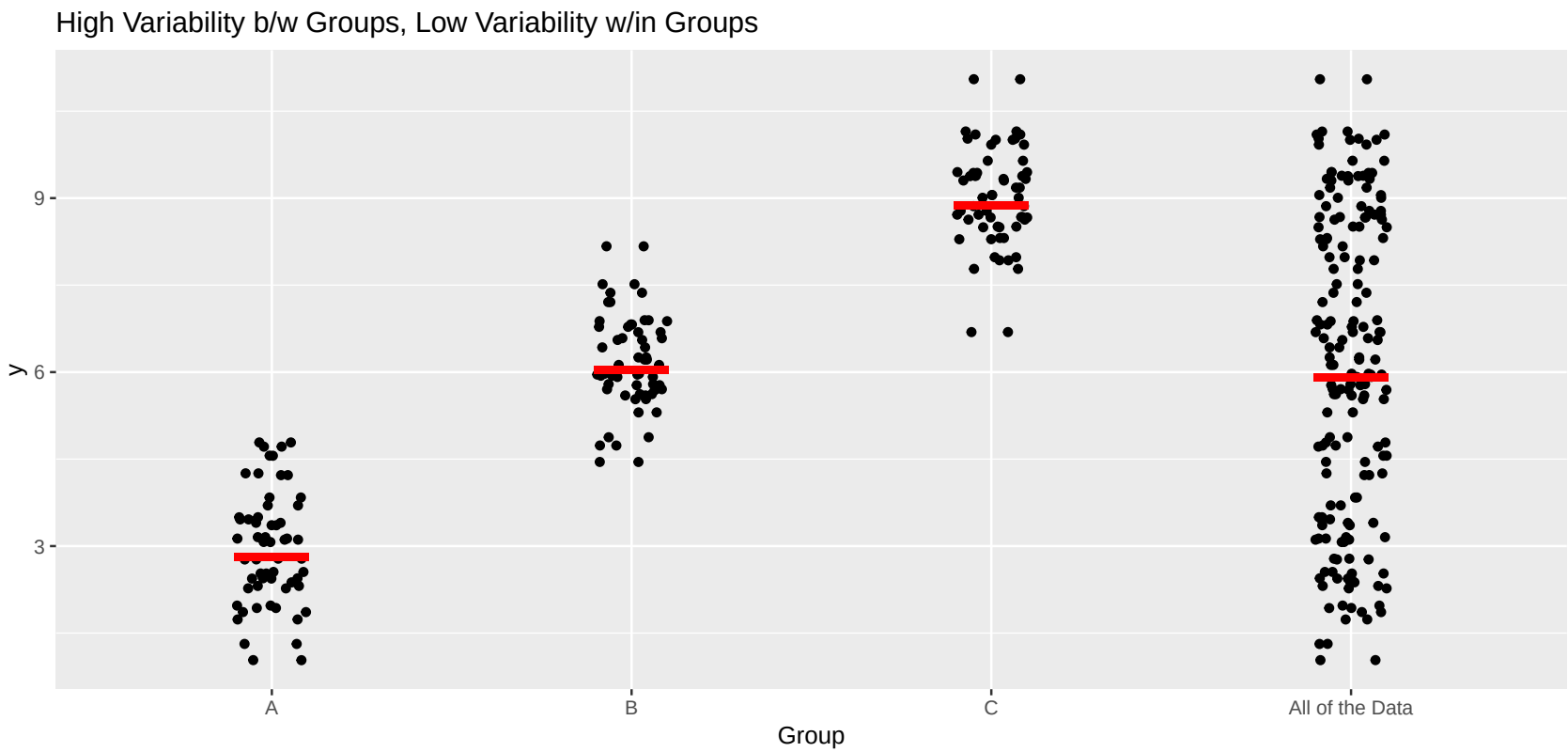
- Last time, we saw how to handle the skew and outliers of the Tailgating Data
  - However, we were limited to comparisons involving only 2 of the 4 groups
  - We could do 6 different tests (1 for each possible pairing), but what problem would this approach have?
- A more sophisticated approach is a single test of the hypothesis:

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

- This situation, where the *explanatory variable is categorical* (with more than two groups) and the *outcome variable is quantitative*, is handled using "ANalysis Of VAriance" or ANOVA

Partitioning Variability

- The ANOVA works by to splitting the *total variability* in the outcome variable into two parts
  - The variability between groups
  - The variability within groups



Modeling

- To more formally evaluate "variability", we need to learn about *statistical modeling*
- A model is a simplified characterization of reality
  - We might model a child's adult height as a function of their age, current height, etc.
- The goal of a model is to *explain variability* in an outcome variable
  - A model explains variability if its predictions are "better" than guessing
- The model above involves multiple variables, which can get pretty complex
  - We'll start by looking at the *simplest possible model*

The "Null Model"

- In the tailgating data, the average mean following distance was $\bar{y} = 41$ feet, which is our best estimate of μ the *population mean* following distance
- If there are no useful explanatory variables, a reasonable model for any individual's following distance is:

$$y_i = \mu + \epsilon_i$$

- ϵ_i is an unexplainable deviation of that individual from that mean
- This model suggests the prediction: $\hat{y}_i = \bar{y}$
  - The expected (predicted) following distance for any individual is just the overall average

Summarizing the Null Model

- Under the null model (or any model), each subject deviates from their prediction by a **residual**:

$$\begin{aligned} r_i &= \hat{y}_i - y_i \text{ (Definition of a residual)} \\ &= \bar{y} - y_i \text{ (Residuals for the null model)} \end{aligned}$$

- We can *summarize* how close the null model is to the truth 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 it is the largest possible sum of squares (of any justifiable model)

Modeling Distance in the Tailgating Data

- The null model makes the same prediction for everyone
- An *alternative model* suggests different predictions for each drug group (indexed by k):

$$\text{This alternative model is: } y_i = \mu_k + \epsilon_i$$
$$\text{suggesting predictions: } \hat{y}_i = \bar{y}_k$$

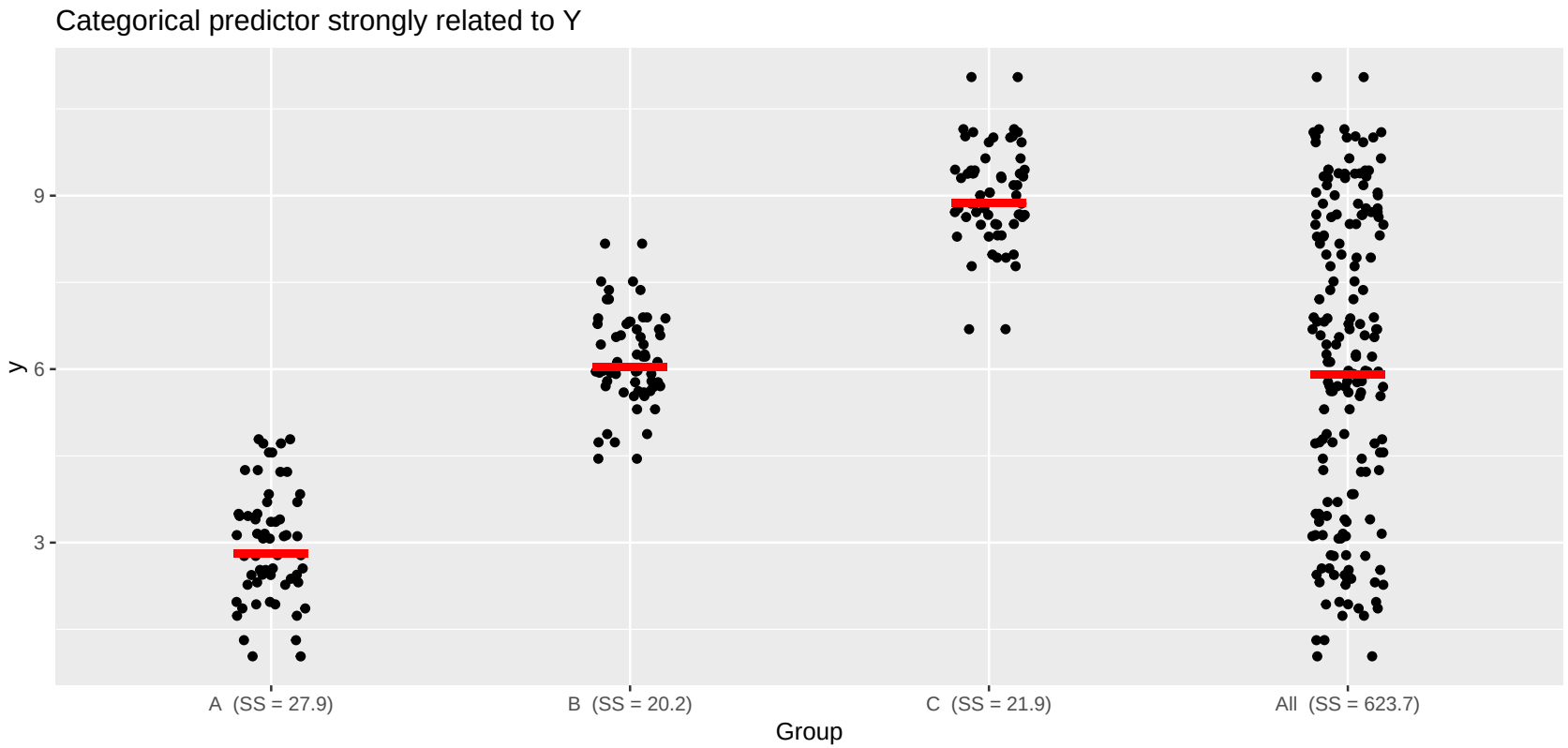
- This alternative model can also be summarized using a **sum of squares**:

$$SSE = \sum_i r_i^2 \text{ for the alternative model}$$

- We call this SSE because it summarizes the errors made by the model we seek to evaluate

SSE versus SST

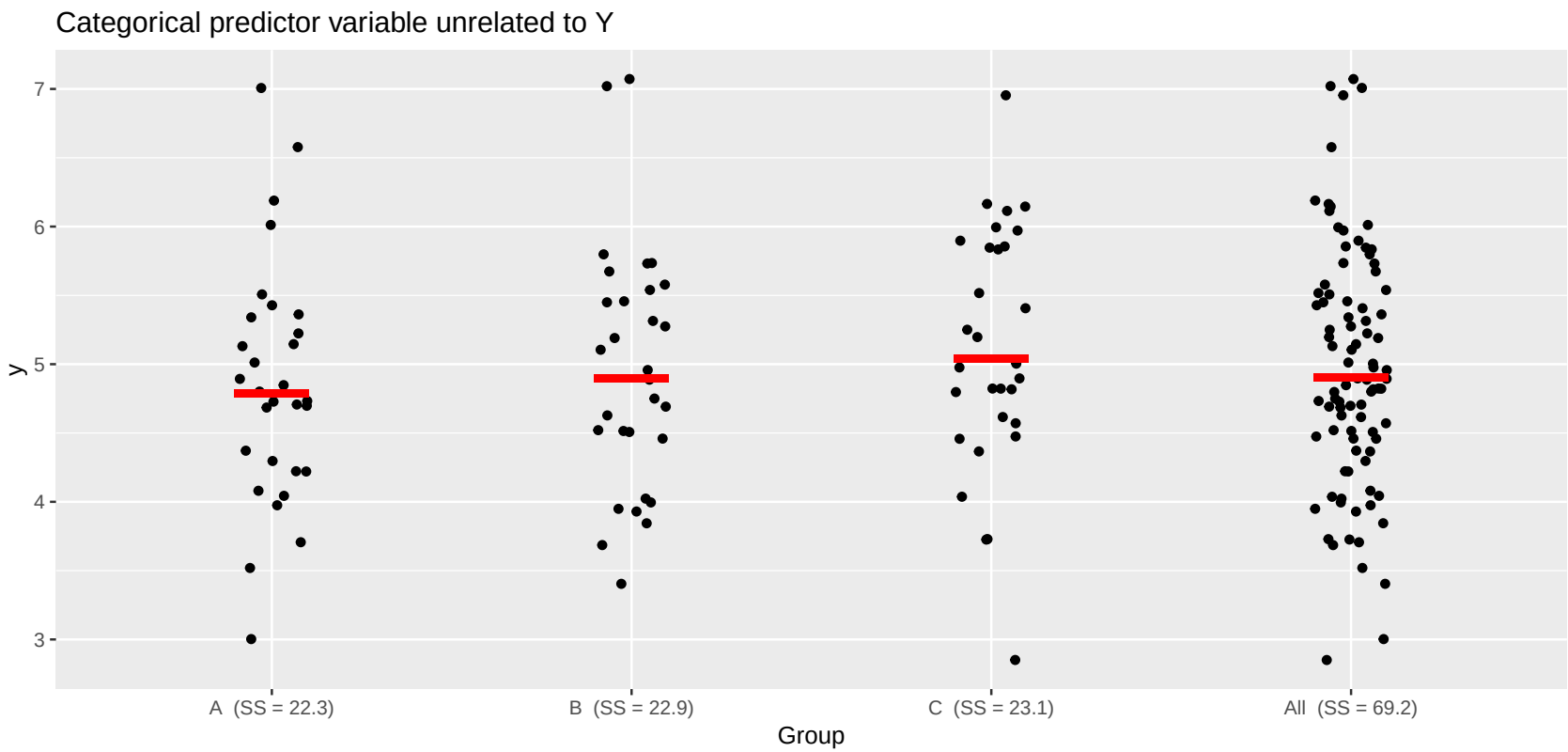
- If the alternative model is *superior* to the null model (ie: the group means really are *different* at the *population level*), *SSE* will be *much smaller* than *SST*



Note: SSE is the total of each group's SS (ie: 27.9 + 20.2 + 21.9), SST is the SS for all of the data (ie: 623.7)

SSE versus SST

- If the grouping variable is *not associated* with Y (ie: the *group means* are identical at the *population level*), SSE will still be *somewhat smaller* than SST



An Important Question

- A lower sum of squares for the alternative model implies the population level means are different
- But if SSE will always be lower than SST, how should we decide if we should believe alternative model?

Hint: When considering $H_0 : p = 0.5$ how do you decide whether to believe $p \neq 0.5$? Would seeing a sample with $\hat{p} = 0.52$ be sufficient?

Evaluating the Role of Random Chance

- Because SSE will *always* be less than SST, we should be asking:
  - *"Does the grouping variable improve model fit beyond what might be expected due to random chance?"*
- ANOVA answers this question using the test statistic:

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

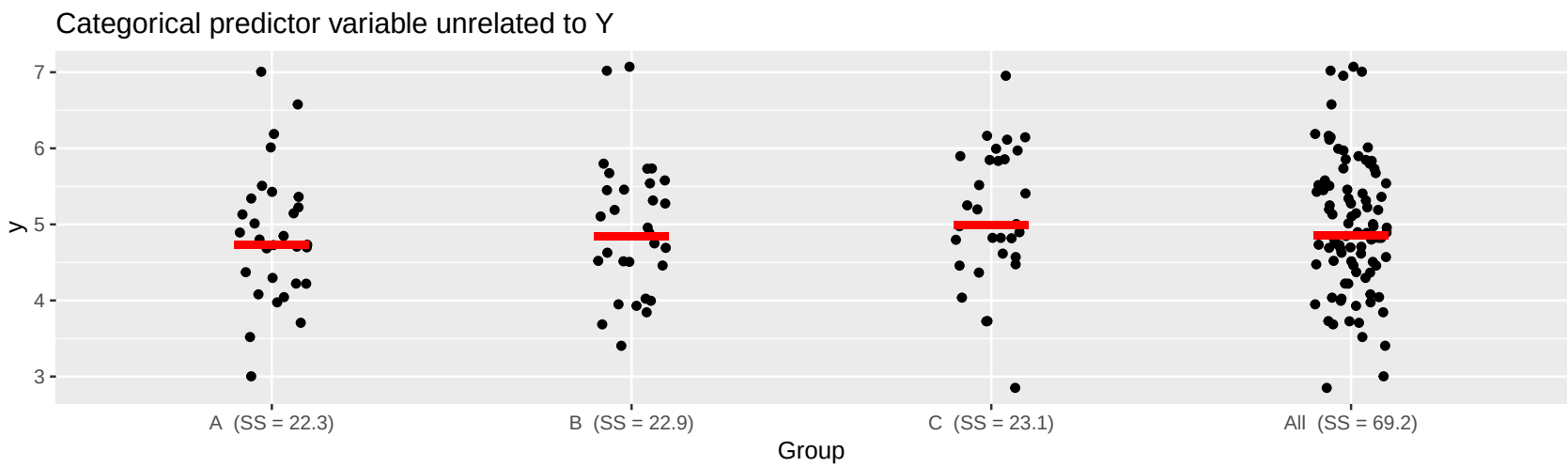
- d_1 and d_0 refer to the number of parameters in the model being considered and the null model, in the tailgating example $d_0 = 1$ (the single overall mean) and $d_1 = 4$ (each group's mean)
- The F statistic can be interpreted as the *standardized drop* in the sum of squares *per additional parameter*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

The F-Distribution

- Under the null hypothesis (ie: the null model is true), this F-statistic follows an F-distribution that depends upon two different degrees of freedom (df) parameters
  - The *numerator df* is $d_1 - d_0$
  - The *denominator df* is $n - d_1$
- We can use StatKey to view various F-distribution curves

The F-Distribution - Practice

- For the data displayed below, assuming the standard error is 2.53, calculate the F statistic comparing an alternative model the uses 3 group means (A, B, and C) against the null model (using the overall mean)
- For these data, $n = 90$, use this to locate your F-statistic on the appropriate distribution. Explain what the "Right Tail" area beyond your F-statistic describes.



The F-Distribution - Solution

1. $SSE = 22.3 + 22.9 + 23.1 = 68.3$, $SST = 69.2$, $d_1 = 3$, $d_0 = 1$, and $SE = 2.53$; so the F-statistic is given by: $\frac{(69.2-68.3)/(3-1)}{2.53} = 0.18$
2. The area to the right of this statistic is 0.836 on the F(2,87) distribution, indicating there is an 0.836 chance of seeing data like ours *if the null model were true*

What is the Standard Error?

- We've seen that standard errors tend to look like a measure of variability divided by the sample size
- In the ANOVA 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:

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

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 which group a data points belongs to
- 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 the "group" variable

Simplifying the F-statistic

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

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

- Sums of squares divided by their degrees of freedom are often called **mean squares**, they allow 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 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 application:
  - $d_0 = 1$, the null model has one parameter, a single overall mean
  - $d_1 = k$, the alternative model has k parameters, a different mean for each group

The ANOVA Table - Example #1

With your group, complete the following ANOVA table (assuming this is a typical ANOVA test where $d_0 = 1$):

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

Additionally, roughly a sketch of what a set of boxplots for these data (broken down by group) might look like (disregarding the units)

The ANOVA Table - Example #1 (solution)

In this example $d_0 = k = 5$ 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 corresponding boxplots should show high variability between groups and low variability within groups

ANOVA - Example #2

With your group, analyze the Tailgating Data (using log(distance) or LD as your outcome variable) in Minitab with 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 #2 (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 which groups are different
- After a statistically significant ANOVA test we should further investigate which groups differ
- In Minitab this is done using **Tukey's honest significant difference (HSD) test** (sometimes called Tukey's range test)
  - Tukey's HSD naturally controls the type I error rate for all possible pairwise comparisons (so we avoid the problem of doing multiple tests)
- Our textbook provides an alternate set of formulas for *post-hoc testing*, you won't be held responsible for those formulas

Inference for Means after ANOVA - Example

Practice: With your group, conduct a follow up analysis of the Tailgating Data using ANOVA and Tukey's HSD (click "comparisons" in the ANOVA menu) and answer the following questions:

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

Inference for Means after ANOVA - Example (solution)

1. The NODRUG and MDMA groups are the most different, but the difference just misses statistical significance ($p = 0.055$)
2. The THC and ALC groups are the least different
3. On the log-scale the interval is (-0.005, 0.705), after exponentiation we get: (0.995, 2.024).

We conclude that the mean following distance of the NODRUG group is between 0.5% shorter and 102.4% 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 outcome 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 using a single categorical predictor variable
  - Our goal will be to build and understand *multiple regression* models that involve *several predictor variables* which can be categorical or quantitative

Conclusion

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

1. Know the situations where ANOVA can be used
2. Understand the concepts of Null and Alternative Models
3. Know how to fill out an incomplete ANOVA table
4. Understand how to interpret an ANOVA table
5. Conduct appropriate follow-up analyses after ANOVA

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