ANOVA...and more!

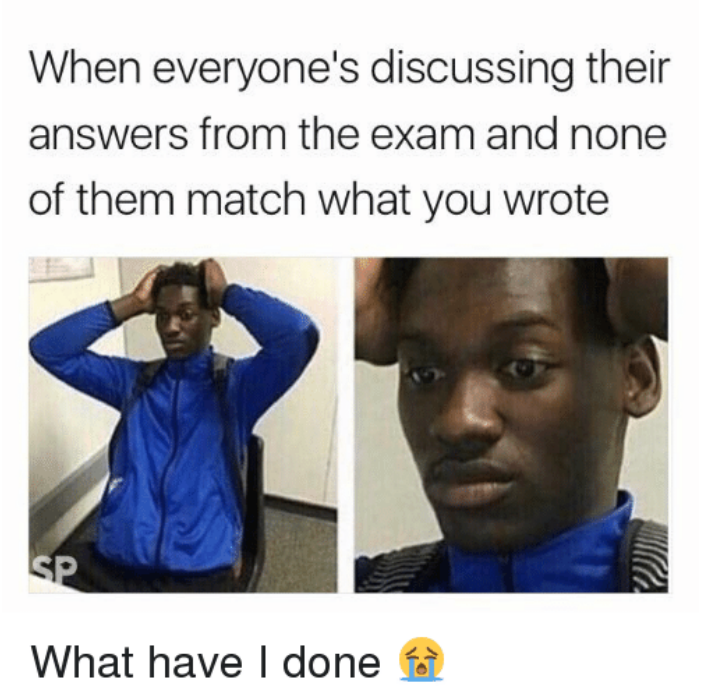
Javier E. Flores

April 17, 2019

Introduction

In our last exam, we were introduced to data collected for the purposes of comparing the foam index of espressos brewed three different ways.

When asked how we might test for differences in foam index across the three groups, with our current toolbox, our only option would be to perform tests for each pairwise comparison (for a total of three tests).



Introduction

One issue with this approach, which becomes increasingly apparent when you consider larger numbers of groups, is performing multiple comparisons.

> With only three groups, testing each pairwise comparison makes for a total of $\binom{3}{2} = 3$ tests.
>
> If instead we were comparing across six groups, testing each pairwise comparison would require we perform $\binom{6}{2} = 15$ tests!

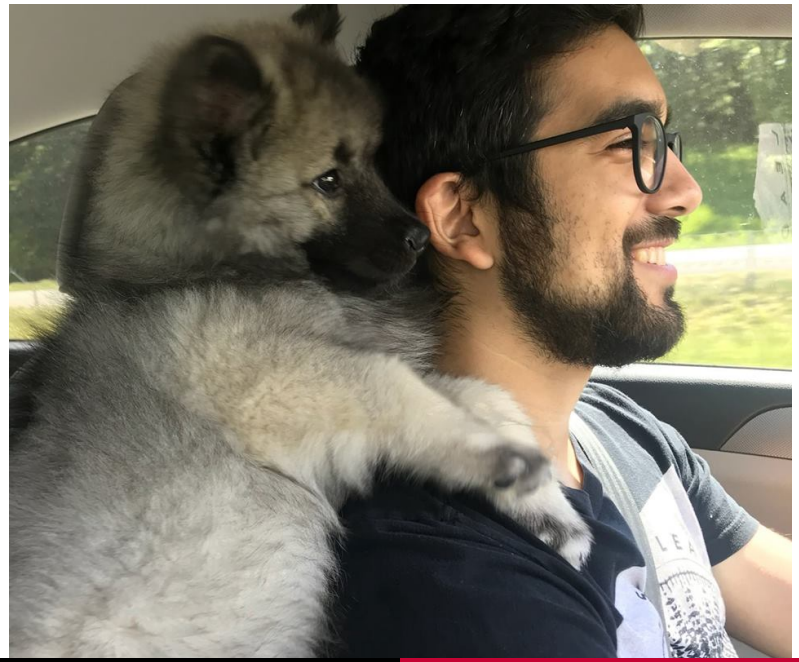
An alternative, more efficient, approach in assessing differences in a quantitative measure across several categories of a categorical variable is **ANOVA**, or **Analysis of Variance**.

Introduction

In order to avoid further revisiting of any second exam traumas, we'll abandon the espresso dataset as a motivating example.

Instead, we'll use data collected from the National Advanced Driving Simulator (NADS) for a study investigating the link between drug use and risky driving behavior.

Figure 1: An Example of Risky Driving Behavior



NADS Study

In the NADS study, participants were first asked to describe their drug use and were subsequently categorized into one of four groups: Alcohol, MDMA, THC, or no drug use.

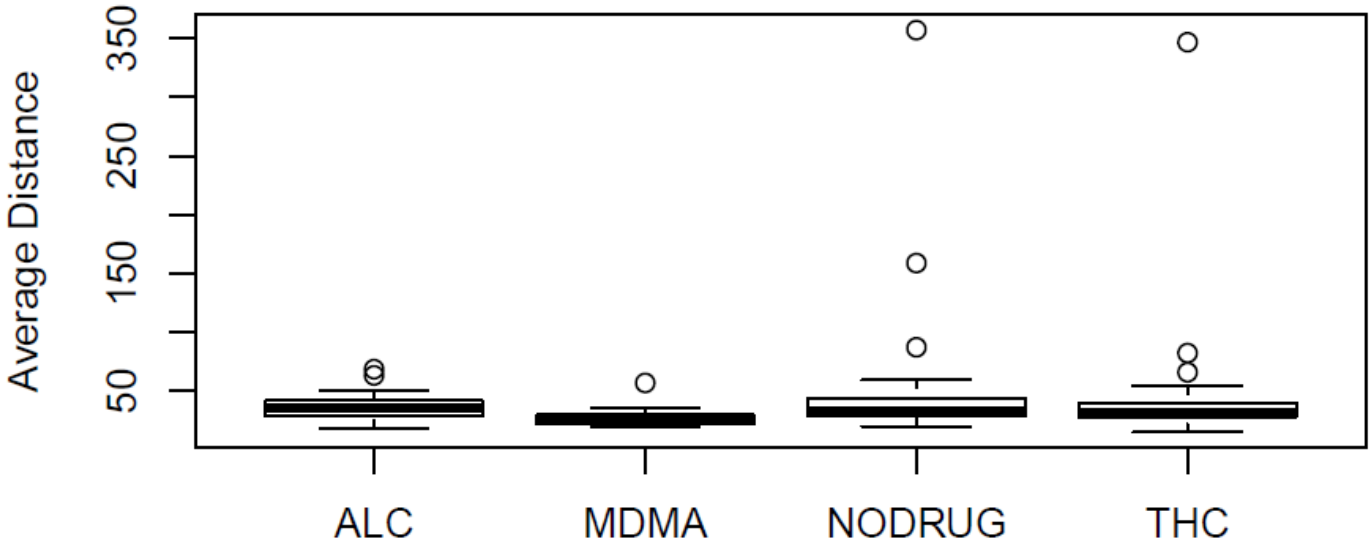
Subjects who described using multiple drugs were classified according to the "hardest" drug they used (i.e. MDMA $>$ THC $>$ Alcohol).

Using a driving simulator, subjects were then asked to follow a lead vehicle that was programmed to vary its speed unpredictably.

The average following distance of each driver was recorded, and drivers were considered more cautious if they followed at larger distances.

NADS Study

Shown below are boxplots of the average following distances across each group:

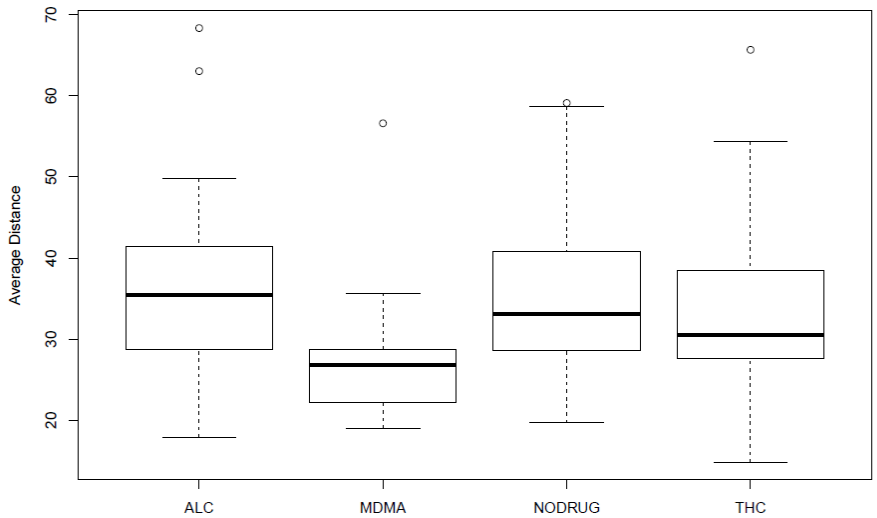


Question: Can you infer anything about the relationship between drug use and risky driving? Is this plot informative?

Outliers

The previous boxplot makes it difficult to assess any differences in following distance across the groups.

This is largely due to the distortion created by the extreme outliers in the no drug use and THC categories. If we remove these outliers, we obtain the following figure:



Question: Now can you infer anything about the relationship between drug use and risky driving? Is this plot informative?

Outliers

With the outliers removed, the plot becomes much more informative. We see that most groups generally appear the same, with MDMA being a possible exception.

Additionally, rather than the extreme skewness initially observed, we see that the data within each group is roughly symmetric and normally distributed.

By visually comparing both graphs, it is clear that outliers can have a substantial effect on how data are interpreted.

Furthermore, outliers can influence the results of approaches (e.g. t-tests) that rely on distributional assumptions (e.g. normality assumptions).

Impact of Outliers

With your groups, load the Tailgating dataset into Minitab and answer the following questions. Note that the variable "D" contains each subject's average following distance.

1) Is there a difference in average following distance between the MDMA and THC groups? Use a two-sample t-test.
2) Manually delete the outlying observation in the THC group (the subject whose average following distance is around 350), and repeat the test.
3) How do the two test results compare?

Impact of Outliers

Performing a two-sample t-test with the outlier included yields a p-value of 0.09. Assuming a significance level of 5%, we would fail to reject the null hypothesis.

However, if we remove the outlier, the same test yields a p-value of 0.03. In this instance, we would reject the null hypothesis...so which result should be used?

While it may be tempting to use the second set of results (with the outlier removed), it is important to first consider whether there is some justification for removal beyond the sole fact that the data point is an outlier.

Impact of Outliers

Sufficient justification for removal would include:

- The outlier is the result of recording or measurement error (e.g. pulse rate of 0 or age of 155).
- The subject was not taking the study seriously or was not complying with study protocol.

Without justification, we are essentially selectively choosing which data should be kept and which should be discarded.

This data "cherry-picking" will, at best, lead to severely biased inference, and, at worst, lead to a completely meaningless set of results.

Dealing with Outliers

So what do we do if we can't justifiably remove an outlier, and it is, in fact, a real data point?

In these instances, rather than manipulate the raw data in some way, we should think of alternate analytic approaches that might accommodate outliers.

In some cases, outliers can provide the greatest insight towards answering a question.

To (briefly) illustrate this point, consider the famous 1980 NASA ozone fiasco.

Nimbus-7 and the Ozone

In the mid 1980's, a large hole in the ozone layer above Antarctica was discovered, garnering worldwide attention.

Given the size of the hole, it was surprising that NASA, who had been monitoring the Earth's atmosphere using the Nimbus-7 satellite since the early 70's, had only then discovered the hole.

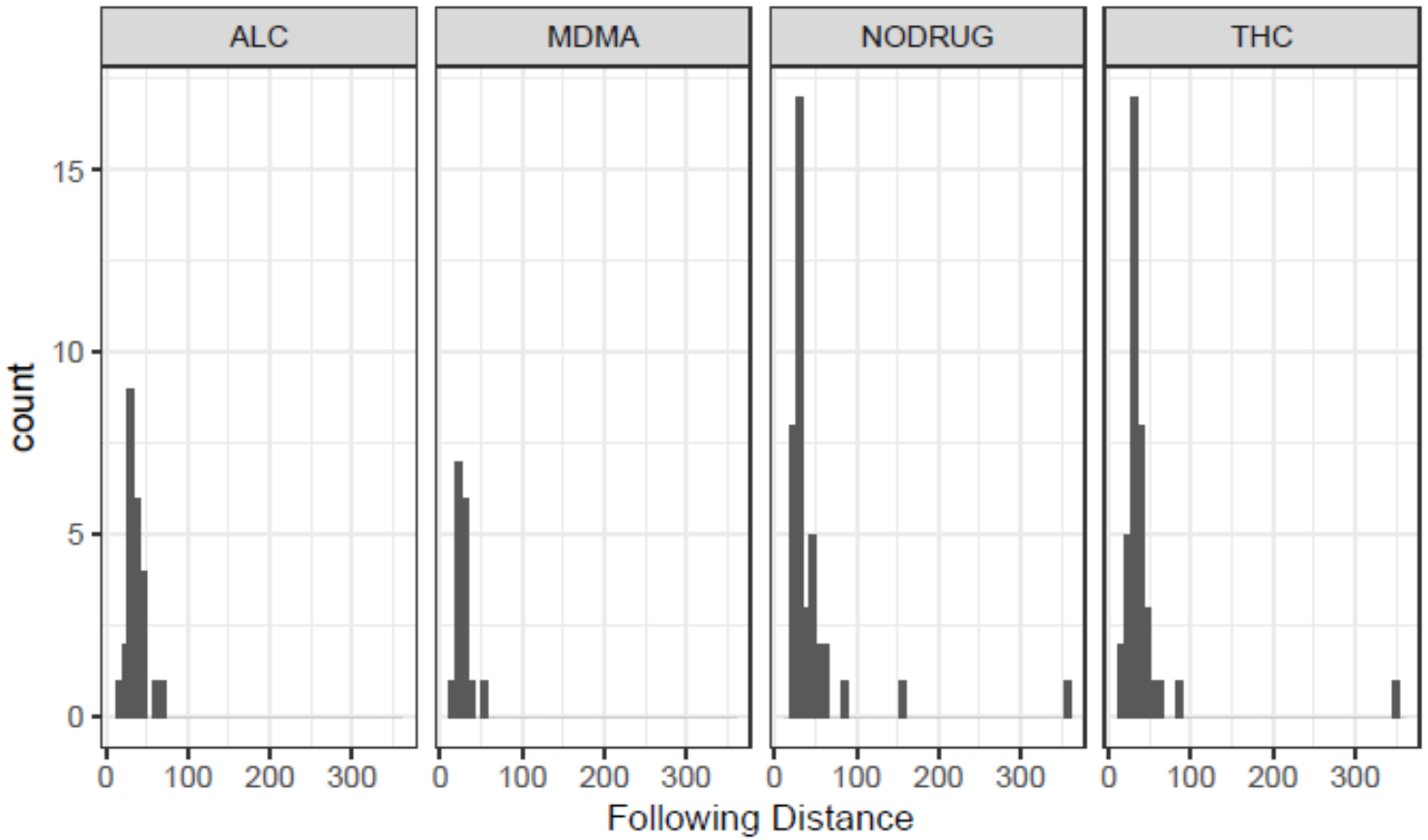
After a bit of headscratching, NASA discovered that the Nimbus-7 was programmed to automatically discard certain unexpected observations as errors (i.e. outliers).

Further inspection of the raw Nimbus-7 data (including the discarded outliers) showed that the Antarctic ozone hole had been around for nearly a decade!

Lesson: Don't discard outliers without sufficient reason!

Back to the NADS Study...

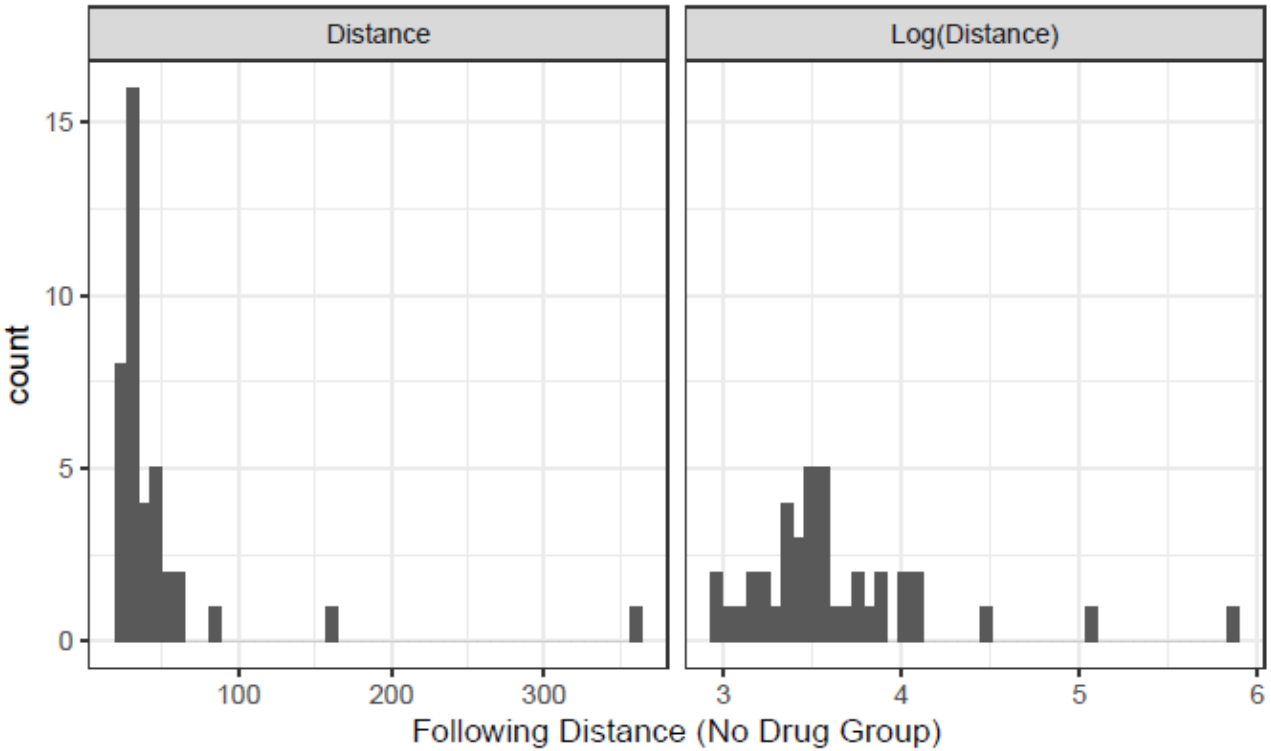
If we were to assume that, in our NADS Study, the outliers are "real" data points and there is no justification for their removal, our data are severely right-skewed:



Log Transformations

One way that we alter our analytic approach to address this skewness is to apply a **log transformation** to our outcome data.

In doing so, rather than looking at group differences in average following distance, we are assessing the group differences in the average natural logarithm of following distance.



Log Transformations

While log transformations aren't perfect solutions, they can often greatly attenuate the degree of skewness in your data.

As a result, assuming normality for the sake of a t-test becomes a much more reasonable thing to do.

However since, in this example, we are now working with the $\log(\textit{Distance})$ as opposed to the *Distance*, the way we interpret our t-test is different.

If, for example, we wanted to compare distances in the No Drug and THC groups we would compute the difference in mean $\log(\textit{Distance})$ and obtain 0.084.

This difference (on a log scale) is equivalent to a ratio in the untransformed scale:

$$\log(A) - \log(B) = \log(A/B)$$
$$\implies \quad \exp(\log(A/B)) = A/B$$

Log Transformations

If we were to exponentiate the difference in mean log(*Distance*) (i.e. 0.084), we would obtain: $\exp(0.084) = 1.09$.

This value, 1.9, is the ratio of average following distance in the No Drug and THC groups. From this ratio, we can say that the mean following distance of the No Drug group was 9% greater than the THC group average following distance.

To be a bit technical, this ratio is actually a ratio of the *geometric* means rather than the *arithmetic* means.

Despite this technicality, the bottom line is that analyzing log-transformed data provides an idea of the relative change across groups.

Log Transformations

Aside from sometimes allowing us to make normality assumptions more comfortably, log transformations allow us to construct confidence intervals for the relative changes across groups.

To do so, we first calculate a confidence interval in the usual way using the log-transformed data and then exponentiate the end points:

CI for difference in $\log(Distance)$:

$$\overline{\log(D_1)} - \overline{\log(D_2)} \pm t_{crit}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (EP_1, EP_2)$$

CI for relative change in *Distance*: $(\exp(EP_1), \exp(EP_2))$

Practice

With your groups create a new variable labeled "LogDistance" in Minitab by taking the log of each value under "D". Check that this new variable matches the existing variable "LD".

Construct a 95% confidence interval for the mean relative increase in following distance of No Drug and THC users.

Perform a two-sample t-test using the log-transformed data for No Drug and THC groups and compare the results with a two-sample t-test on the untransformed data.

Solution

We first obtain a 95% confidence interval for the difference in mean log(*Distance*) between the two groups:

$$\overline{\log(D_{ND})} - \overline{\log(D_{THC})} \pm t_{crit}\sqrt{\frac{s^2_{ND}}{n_{ND}} + \frac{s^2_{THC}}{n_{THC}}} = (-0.151, 0.318)$$

Exponentiating the endpoints of this interval, we obtain the 95% CI for the mean relative increase in following distance of No Drug and THC users:

$$(\exp(-0.151), \exp(0.318)) = (0.86, 1.37)$$

The above result indicates that the No Drug following distance plausibly ranges from 14% shorter to 37% longer than the THC following distance.

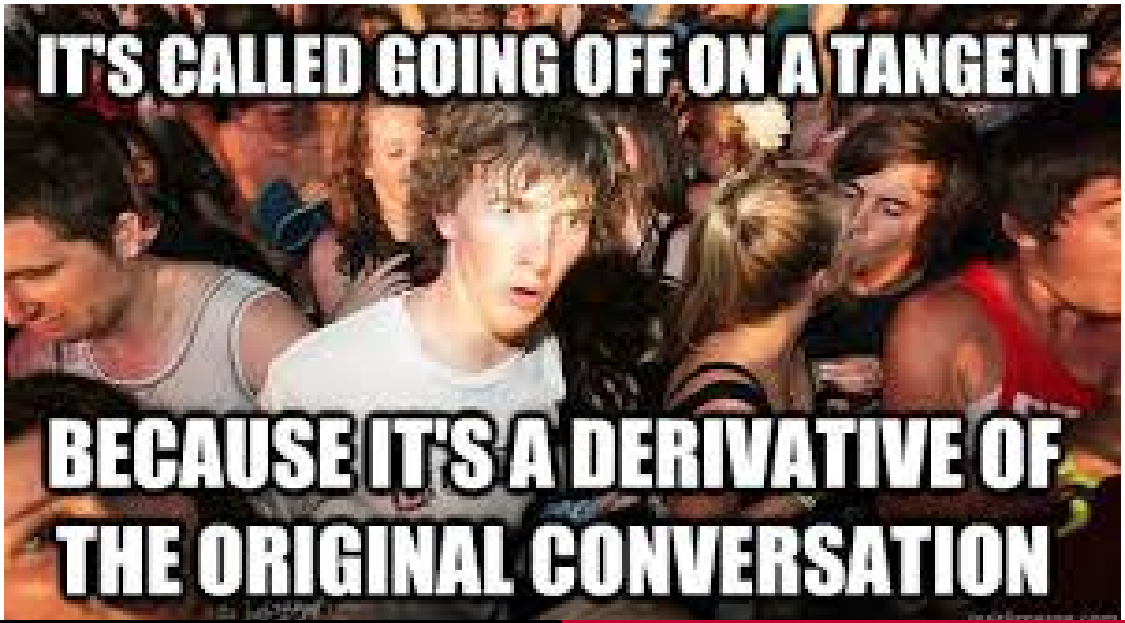
The test statistic on the log scale is 0.71 with a p-value of 0.478, and, on the original scale, the test statistic is 0.39 with a p-value of 0.70.

While both would lead us to fail to reject the null hypothesis, the test on the log-scale is more powerful since it better meets the t-test's normality assumptions.

Weren't We Supposed to Talk About ANOVA?

We began this lecture by teasing an analytic approach called ANOVA, and subsequently discussed the tangential topics of outliers and log-transformations.

With these important concepts behind us, we'll next speak broadly about making comparisons across multiple groups, and then lead into a discussion of statistical modeling and ANOVA.



Comparing Multiple Groups

As I mentioned at the start of the lecture, we technically have the capability of finding differences across multiple groups.

To do so, we would need to perform a series of pairwise comparisons (i.e. two sample t-tests), and compile the results of all tests to gain some overall understanding.

The problem with this approach is that often times several pairwise comparisons need to be made, making it difficult to interpret results holistically and, perhaps more importantly, leading to an increased (overall) type I error.

In order to address the type I error rate issue, we have two options:

1) We can adjust the significance threshold for each individual pairwise comparison.
2) We can perform a single joint test assessing the association between group and following distance (i.e. ANOVA).

Adjusting α

Suppose that, for the NADS study, we test all pairwise comparisons at $\alpha = 0.05$ for a total of six tests.

If we assume that each test was (probabilistically) independent of one another and that for each comparison, the null was actually true, the probability of committing at least one type I error would be:

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

The above calculation suggests a simple correction to the significance level of each test: $\alpha^* = \alpha/h$, where h is the total number of tests being performed. This correction is known as the **Bonferroni Correction**.

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

The Bonferroni Correction

The Bonferroni correction, or adjustment, controls the **family-wise**, or **experiment-wise**, type I error rate at α.

This adjustment tends to be conservative in that it usually controls the family-wise error rate below α, rather than at exactly α.

Aside from the Bonferroni correction, there are a slew of other, more powerful, approaches to family error rate control. However the Bonferroni adjustment is arguably the most intuitive and easiest to implement.

In some cases Minitab will even provide Bonferroni corrected confidence intervals, which are created by using α^* to adjust the critical value (z_{crit} or t_{crit}) used to construct the interval.

Practice

With your groups, perform the six pairwise t-tests comparing the log(*Distance*) between each group. Answer the following questions:

1) How many differences are significant using a non-adjusted α of 0.05? Do you expect the experiment-wise error rate to be higher or lower than the significance level for each individual test?
2) How many differences are significant using a Bonferroni-adjusted α? Do you expect the experiment-wise error rate to be higher or lower than the significance level for each individual test?
3) Does using the bonferroni correction affect the power of each test? If so, in what way? If not, why not?

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

Using the unadjusted threshold, there are three significant results: 2), 5), and 6). However, we expect our overall type I error rate to be higher than 0.05.

Using the Bonferroni-adjusted threshold ($0.05/6 = 0.0083$), we find only two significant results: 2) and 6).

We would then conclude that the following distances are different between the ALC and MDMA groups, as well as between the MDMA and NODRUG groups.

What About Power?

While using a Bonferroni adjustment effectively controls our overall type I error rate, it comes at the cost of statistical power.

Applying a Bonferroni adjustment decreases the significance level, α, of each individual test. As a result, the amount of evidence necessary to reject each null hypothesis increases, making the number of rejections fewer than if using an unadjusted α.

With fewer rejections, we can then expect a decrease in power (and increase in type II error rate).

If we wanted to avoid this tradeoff and still control our type I error rate, we could turn to option (2) and use a single, joint test (i.e. ANOVA) of the hypothesis:

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

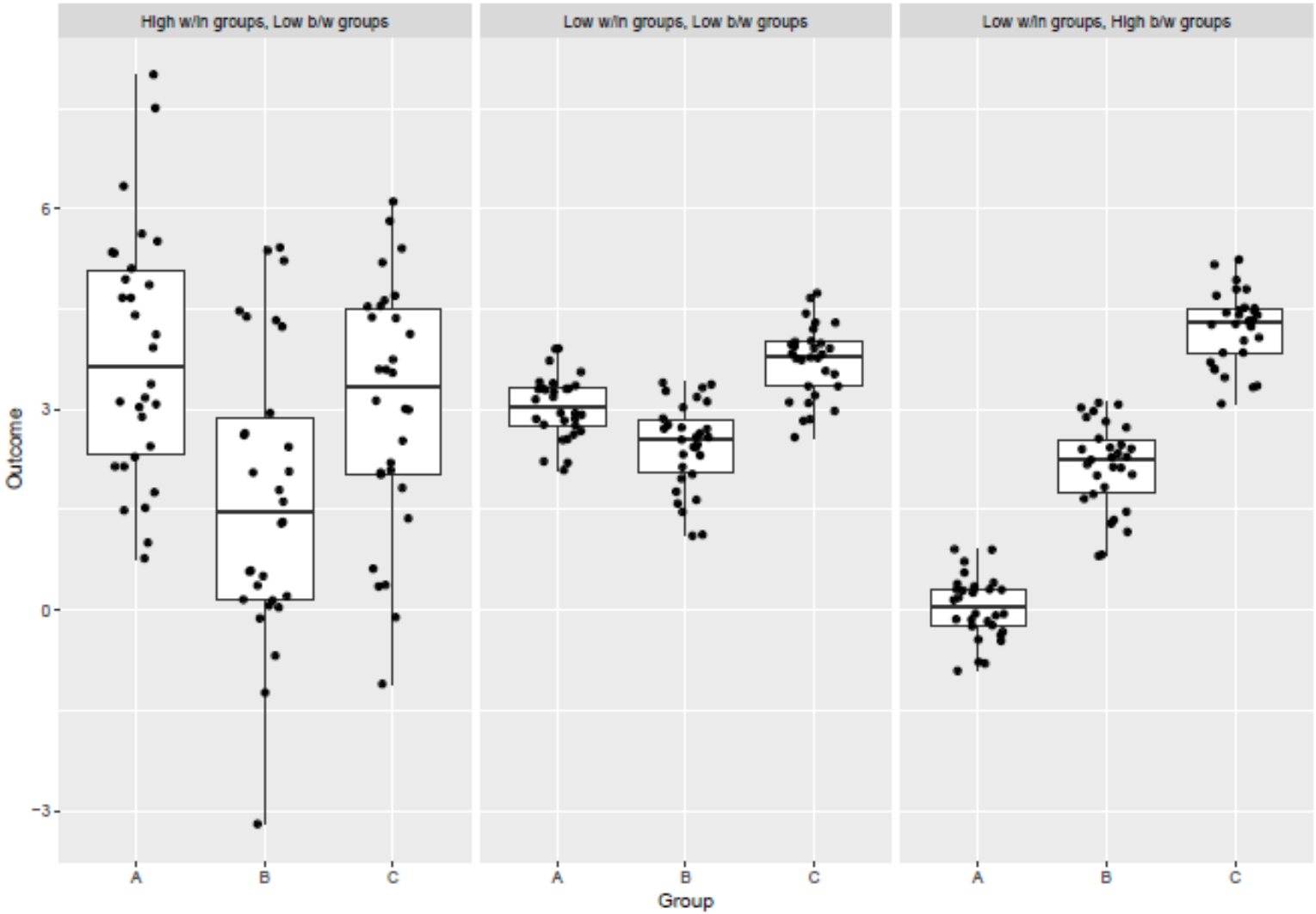
ANOVA

As the name implies (i.e. Analysis of Variance), in ANOVA we analyze the total observed variability in our outcome data in order to appropriately partition it between two sources:

1) Variability arising due to differences between groups (i.e. between-group variability)
2) Variability arising due to differences within groups (i.e. within-group variability)

Intuitively, if there was more between-group variability than within-group variability, it would be reasonable to conclude that there are significant group-level differences in our outcome.

Partitioning Variability



Statistical Modeling

To better frame the idea of partitioning variability, it is important that we first discuss **statistical modeling**.

A model is a simplified characterization of a certain process or relationship. Good models are those which, while maintaining some degree of simplicity, accurately describe or explain the phenomenon of interest.

Statistical models are models in every sense of the word, but their "quality" is measured by their ability to "explain" the variability in a certain outcome variable.

It is generally impossible for models to explain all of the variability in an outcome, but some models are better than others in this endeavor.

As an example, consider the following two models:

1) Using the height of a child's parents to predict the child's adult height
2) Using the child's weight at birth to predict the child's adult height.

Clearly, we would expect the first model to do a better job in predicting, or explaining the variability in, the child's adult height.

Statistical Models

The simplest statistical model, often referred to as the **null model**, is one which posits that all outcome variability is "unexplainable" and should therefore be modeled using a single mean.

In the case of the NADS study, the null model would correspond to using the mean following distance (across all groups) as the prediction for everyone in the study.

More complex statistical models are those which use one or more *explanatory* variables to explain the variability in an outcome of interest.

For the NADS study, an example of a more complex statistical model would involve using the mean following distance for a study group (NODRUG, ALC, THC, or MDMA) as the prediction for individuals in that group.

Total Variability

The quality of a model is measured by its ability to "explain" the variability in a certain outcome variable.

We often summarize the total variability in an outcome by the **Total Sum of Squares (SST)**:

$$SST = \sum_i (y_i - \bar{y})^2$$

You may recognize that this is simply the sum of the **residuals**, $r_i = y_i - \bar{y}$.

We can then assess the quality (or fit) of a model by the share of SST it explains. The greater share, or proportion, of SST that is accounted for by a model, the better that that model is.

Sum of Squares Error

Since models generally are not perfect, it is often of interest to quantify the amount by which their predictions fall short.

We often use the **Sum of Squares Error**, or **SSE**, as a measure of model error. The larger the SSE, the worse a model is.

The way in which the SSE is computed is dependent on the model that is fit.

As an example, for the NADS study, the SSE for the null model is computed:

$$\sum_i (y_i - \bar{y})^2$$

In contrast, if we wanted to use the study groups to model our outcome, the SSE would be computed:

$$\sum_i (y_i - \bar{y}_i)^2$$

where $\bar{y}_i$ is the appropriate group-specific mean.

Model Fit

Using both the SSE and SST, we can quantify the fit of a model (i.e. proportion of explained variability) through the **coefficient of determination**, R^2:

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

Under the null model, $SST = SSE$, and so R^2 is 0.

For the tailgating data, under the model where each group gets its own mean, $R^2 = 0.055$ indicating that the model explains 5.5% of the total variability.

Increasing the complexity of a model will always lower the SSE and increase R^2. Despite this, care must be taken not to **overfit** your model by adding unnecessary complexity.

Rather, we should test whether introducing complexity lowers the SSE by more than what we'd expect to see by random chance.

ANOVA

ANOVA is simply a special type of statistical model in which a single categorical variable is used to predict a quantitative outcome.

In order to test whether the drop in SSE resulting from including a single categorical variable is greater than what we'd expect to see by random chance, we use the test statistic:

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

where d_1 and d_0 refer to the number of parameters in the model being considered (i.e. the model using a single categorical variable as a predictor) and the null model.

For the NADS study, $d_0 = 1$ (the single overall mean) and $d_1 = 4$ (each group's mean).

ANOVA

For tests we've learned in the past that involve quantitative variables, we've seen that the standard errors involve some measure of variability divided by the sample size (e.g. $\frac{s}{\sqrt{n}}$).

In the ANOVA setting:

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

Using this standard error, the F statistic can be expressed:

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

F-Test and Variability

SST is the total sum of squares which quantifies the total variability in the outcome, y. $SST = \sum_i (y_i - \bar{y})^2$.

SSE is the outcome variability that remains unexplained after implementing your model. Under the null model, SSE = SST. Under the ANOVA model, $SSE = \sum_i (y_i - \bar{y}_i)^2$.

By subtraction, we can determine how much variability is being explained by the inclusion of our categorical variable:

$$SSG = SST - SSE$$

where SSG, the **Sum of Squares Groups**, quantifies the amount of variability explained using the categorical variable groups in our model.

F-Test and Variability

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

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

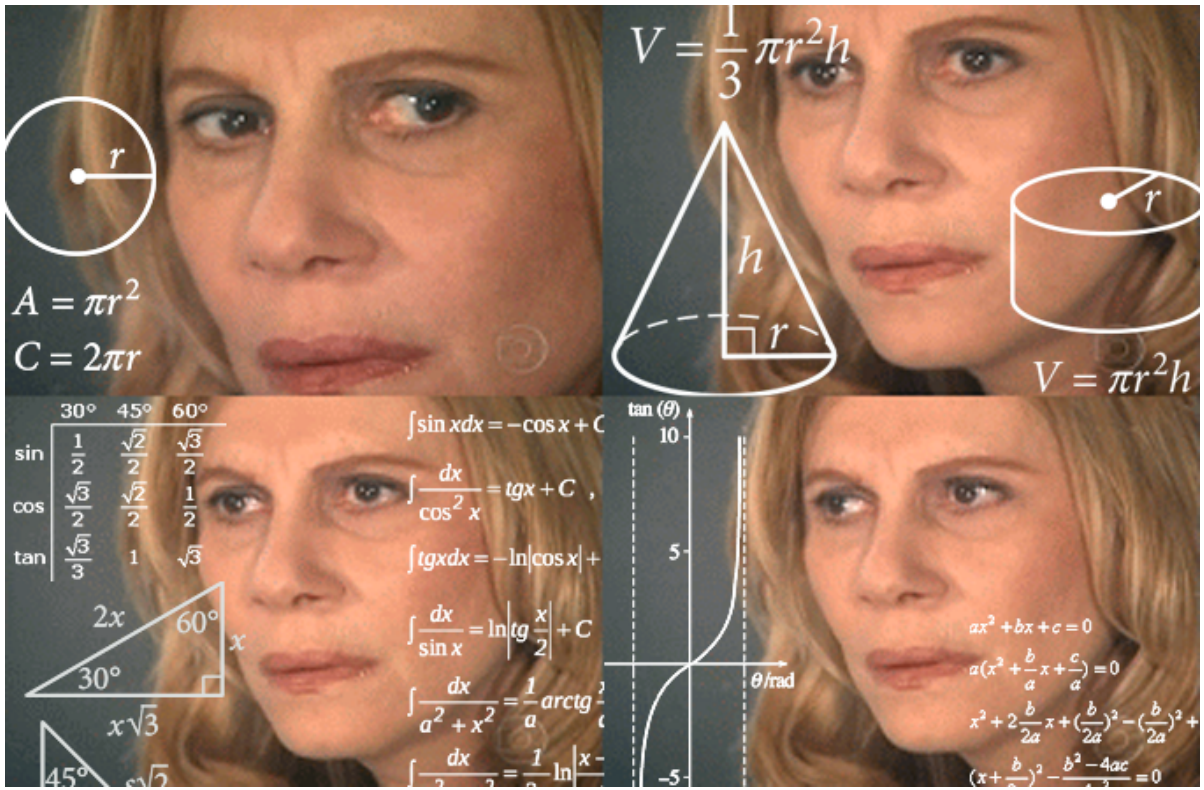
Sums of squares divided by their degrees of freedom are often called **mean squares**, and allows us to express the F-statistic as:

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

where MSG is the mean square of groups and MSE is the mean square of errors.

Math Overload

At this point in the lecture, and with all the math thrown at you all, I'm seeing a lot of this:



Math Overload

This is completely understandable, and the good news is that I will not expect you to memorize these formulas or compute any of these things by hand.

However you will be expected to understand and interpret an **ANOVA table**, which is a common piece of software output for ANOVA.

Then general form of these tables is shown below:

Source	DF	SS	MS	F-Value	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			

$d_0 = 1$, 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.

Practice

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

Source	DF	SS	MS	F-Value	P-Value
Group	4	200	?	?	?
Error	?	440	?		
Total	59	?			

Solution

In this example, $d_1 = k = 5$ and $n = 60$ so:

Source	DF	SS	MS	F-Value	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.

Minitab Practice

With your groups, use Minitab to analyze the Tailgating data with ANOVA (Stat $->$ ANOVA $->$ One-way). Use log(*Distance*) as your outcome variable. Be sure to report:

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

Solution

1) $H_0 : \mu_{ND} = \mu_{THC} = \mu_{ALC} = \mu_{MDMA}$; $H_A : \mu_i \neq \mu_j$ for at least one pair.

2) $F = 2.23$

3) Using a $F_{3,115}$ distribution, we obtain a p-value of 0.088. There is borderline evidence that drug use is predictive of following distance. It appears that the MDMA group is most different in that the group generally has shorter following distances than the rest.

Inference After ANOVA

The results of an ANOVA test only tells us whether a difference across groups exists, and not which specific groups are different.

Because of this, ANOVA is often followed by a few pairwise comparisons in order to investigate which groups differ.

In Minitab we can do this using **Tukey's honest significant difference (HSD) test** (sometimes called Tukey's range test).

Similar to the Bonferroni adjustment, Tukey's HSD controls the family type I error rate for all possible pairwise comparisons (so we don't need to worry about an inflated overall type I error rate).

Practice

With your groups, use Minitab to conduct a follow up analysis of the Tailgating data using ANOVA and Tukey's HSD (click "comparisons" in the ANOVA menu). Answer the following:

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 that the variable "LD" is on the log scale).

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), and after exponentiating becomes (0.995, 2.024).

   We then 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.

The ANOVA test will be statistically significant only if the categorical variable improves prediction beyond what could be attributed to random chance.

The ANOVA model is just one type of model in the vast array of statistical models. If we were to cover all statistical models, you'd be halfway towards getting a PhD in statistics!

We'll instead restrict the last of our lectures to regression models, of which ANOVA is a special case.

Aside from ANOVA, regression models include the simple regression models seen earlier in the course (i.e. single predictor) and multiple regression models with several predictors.

Wrap-Up

Right now, you should...

- Understand how ANOVA testing is related to statistical modeling
- Understand the partitioning of variability in ANOVA
- Construct and use an ANOVA table to draw conclusions
- Conduct appropriate follow-up analyses after ANOVA

These notes cover sections 8.1-8.2 of the textbook. Please read through these sections and their examples along with any links provided in this lecture.