Multiple samples: Modeling and ANOVA

Patrick Breheny

April 28

Multiple group studies

- In the latter half of this course, we have discussed the analysis of data coming from various kinds of studies – we started out with one-sample studies and then moved on to two-sample studies
- This week, we will discuss the analysis of data in which three or more groups/samples are present
- For example, in the tailgating study, we compared illegal drug users with non-illegal drug users
- However, there were really four groups: individuals who use marijuana, individuals who use MDMA (ecstasy), individuals who drink alcohol, and drug-free individuals

Comparing multiple groups

- One possible approach would be to simply make all 6 pairwise comparisons separately
- As we have seen, however, making multiple comparisons increases the Type I error rate unless we make some sort of correction
- In the special case of multiple groups, however, we can take another approach: to simultaneously test the equality of all the groups – i.e., use the entire collection of data to carry out a single test
- To do this, however, we will need a different approach than the ones we have used so far in this course: we will need to build a *statistical model*

The philosophy of statistical models

- There are unexplained phenomena that occur all around us, every day: Why do some die while others live? Why does one treatment work better on some, and a different treatment for others? Why do some tailgate the car in front of them while others follow at safer distances?
- Try as hard as we may, we will never understand any of these things in their entirety; nature is far too complicated to ever understand perfectly
- There will always be variability that we cannot explain
- The best we can hope to do is to develop an oversimplified version of how the world works that explains some of that variability

The philosophy of statistical models (cont'd)

- This oversimplified version of how the world works is called a *model*
- The point of a model is not to accurately represent exactly what is going on in nature; that would be impossible
- The point is to develop a model that will help us to understand, to predict, and to make decisions in the presence of this uncertainty – and some models are better at this than others
- The philosophy of a statistical model is summarized in a famous quote by the statistician George Box: "All models are wrong, but some are useful"

Residuals

- What makes one model better than another is the amount of variability it is capable of explaining
- Let's return to our tailgating study: the simplest model is that there is one mean tailgating distance for everyone and that everything else is inexplicable variability
- Using this model, we would calculate the mean tailgating distance for our sample
- Each observation y_i will deviate from this mean by some amount r_i: $r_i = y_i - \bar{y}$
- The values r_i are called the *residuals* of the model

Residual sum of squares

- We can summarize the size of the residuals by calculating the *residual sum of squares*:

$$\text{RSS} = \sum_i r_i^2$$

- The residual sum of squares is a measure of the *unexplained variability* that a model leaves behind
- For example, the residual sum of squares for the simple model of the tailgating data is $(-23.1)^2 + (-2.1)^2 + \ldots = 230,116.1$
- Note that residual sum of squares doesn't mean much by itself, because it depends on the sample size and the scale of the outcome, but it has meaning when compared to other models applied to the same data

A more complex model

- A more complex model for the tailgating data would be that each group has its own unique mean
- Using this model, we would have to calculate separate means for each group, and then compare each observation to the mean of its own group to calculate the residuals
- The residual sum of squares for this more complex model is $(-18.9)^2 + (2.1)^2 + \ldots = 225,126.8$

Explained variability

- We can quantify how good our model is at explaining the variability we see with a quantity known as the *explained variance* or *coefficient of multiple determination*
- Letting RSS_0 and RSS_1 denote the residual sums of squares from the null model and the more complex model, the percentage of variance explained by the model is:

$$\begin{aligned} R^2 &= \frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_0} \\ &= \frac{230,116.1 - 225,126.8}{230,116.1} \\ &= 0.022 \end{aligned}$$

- In words, our model explains 2.2% of the variability in tailgating distance

Complex models always fit better

- The more complex model has a lower residual sum of squares; it must be a better model then, right?
- Not necessarily; the more complex model will always have a lower residual sum of squares
- The reason is that, even if the population means are exactly the same for the four groups, the sample means will be slightly different
- Thus, a more complex model that allows the modeled means in each group to be different will always fit the observed data better
- But that doesn't mean it would explain the variability of future observations any better (this concept is called *overfitting*)

ANOVA

- The real question is whether this reduction in the residual sum of squares is larger than what you would expect by chance alone
- This type of model – one where we have several different groups and are interested in whether the groups have different means – is called an *analysis of variance* model, or *ANOVA* for short
- The meaning of the name is historical, as this was the first type of model to hit on the idea of looking at explained variability (variance) to test hypotheses
- Today, however, many different types of models use this same idea to conduct hypothesis tests

Parameters

- To answer the question of whether the reduction in RSS is significant, we need to keep track of the number of *parameters* in a model
- For example, the null model had one parameter: the common mean
- In contrast, the more complex model had four parameters: the separate means of the four groups
- Let's let d denote the number of parameters, so $d_0 = 1$ and $d_1 = 4$
- Of particular importance is the fact that the more complex model has $4 - 1 = 3$ additional parameters compared to the null model

Variation explained per parameter

- In the tailgating example, the four-mean model explained an additional $4,989.3\ \mathrm{m}^2$ of variation
- However, since adding parameters will always lower the RSS, a better way of stating the difference between the models is that the more complicated model explained an additional $1,663.1\ \mathrm{m}^2$ of variation per parameter added
- However, this still isn't quite satisfactory – if we measured following distance in feet instead of meters, we'd get a different number
- To finish standardizing, we need to take the standard deviation into account

Estimating variability

- As in Student's t test, if we're willing to assume that the variability is the same for all observations, then we can estimate the standard deviation (or *variance*, which is the standard deviation squared) by pooling the residuals from the more complex model:

$$\hat{\sigma}^2 = \frac{\sum_i r_i^2}{n - d_1}$$

- Again, as in Student's t test, it is worth keeping in mind that this is only an estimate, and thus has some uncertainty to it
- In this case, the pooled standard deviation is based on $n - d_1$ pieces of information (degrees of freedom)
- For the tailgating data, $\hat{\sigma}^2 = 225,126.8/(119 - 4) = 1957.6$

The F test

- Thus, we finally arrive at the ANOVA test statistic:

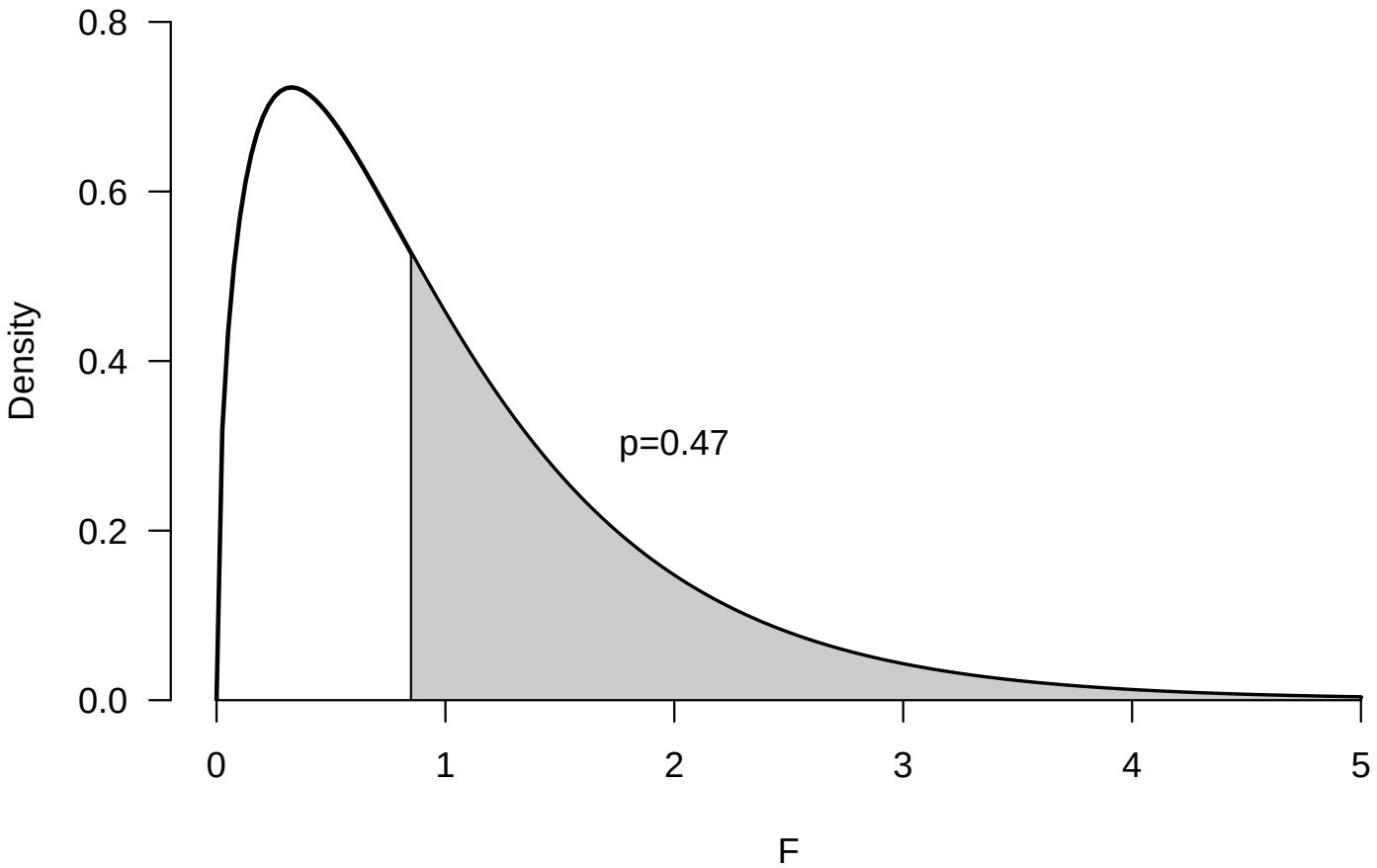
$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(d_1 - d_0)}{\hat{\sigma}^2};$$

  the test statistic is usually denoted F after Ronald Fisher, who first developed the idea of ANOVA
- For the tailgating data,

$$F = \frac{1663.1}{1957.6} = 0.85$$

- To determine significance, we would have to compare this number to a new curve called the F *distribution*

F distribution (with $df = 3, 115$)



Outliers

- Recall, however, that this data had large outliers
- If we rank the data and then perform an ANOVA, we get a different picture of how strong the relationship is between drug use and tailgating behavior:

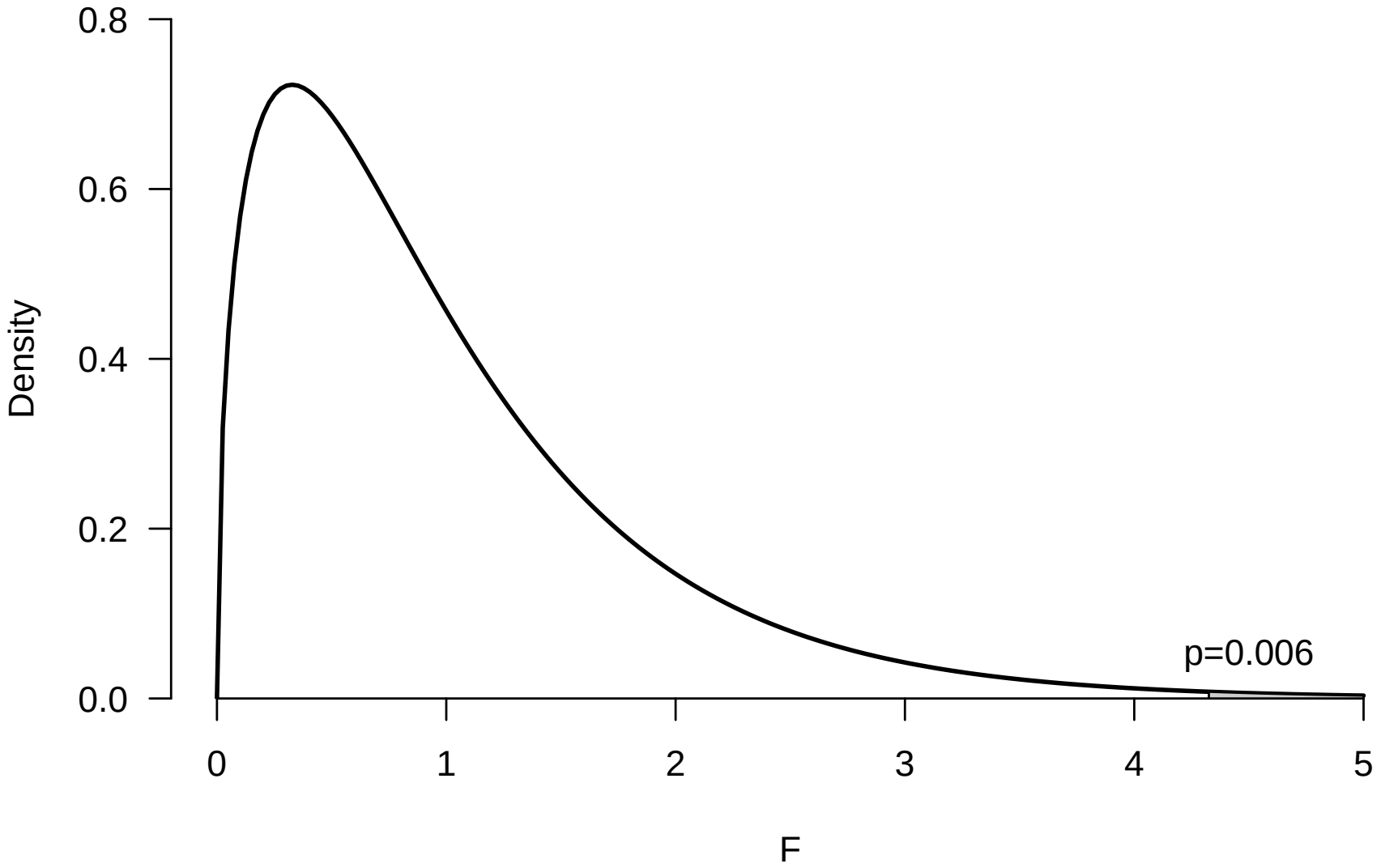
$$\text{RSS}_0 = 140,420$$
$$\text{RSS}_1 = 126,182$$

- Now, our model explains 10.1% of the variability in following distance:

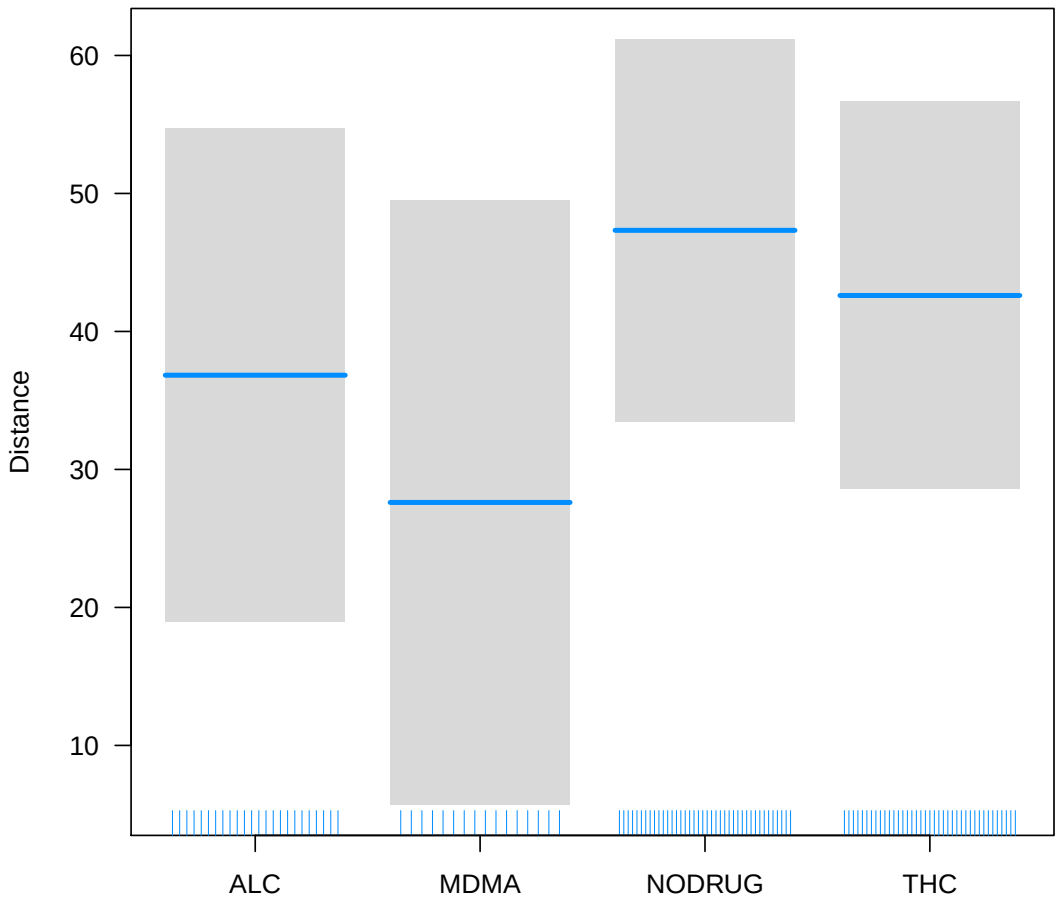
$$\frac{140,420 - 126,182}{140,420} = .101$$

- Furthermore, our F statistic is 4.3

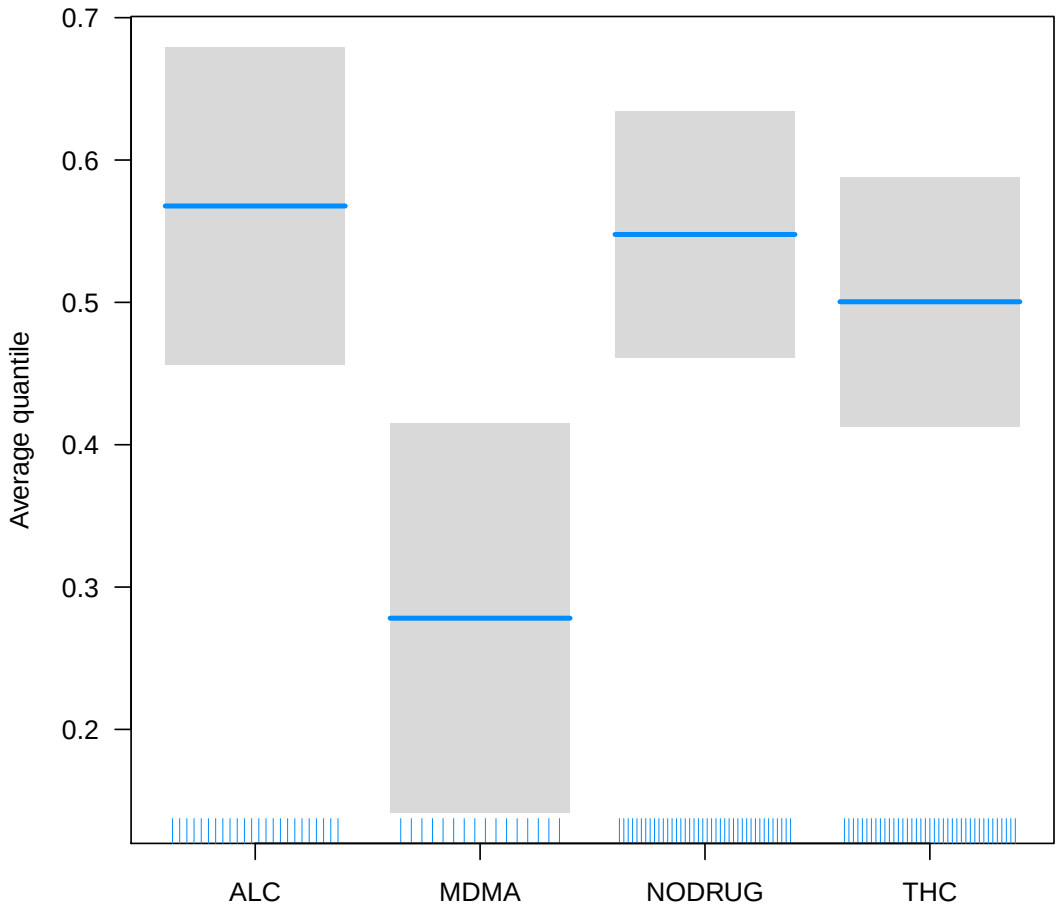
F test: ranked data



Means and CIs: Original



Means and CIs: Ranked



ANOVA for two groups?

- We have seen that ANOVA models can be used to test whether or not three or more groups have the same mean
- Could we have used models to carry out two-group comparisons?
- Of course; however, comparing the amount of variability explained by a two-mean vs. a one-mean model produces exactly the same test as Student's t-test

Other uses for statistical models

- Statistical models have uses far beyond comparing multiple groups, such as adjusting for the effects of confounding variables, predicting future outcomes, and studying the relationships between multiple variables
- Statistical modeling is a huge topic, and we are just skimming the surface today
- Statistical models are the focus of the next course in this sequence, BIOS 5120

Summary

- Statistical models attempt to explain variation; the ability of a model to do so is measured by the coefficient of multiple determination (R^2)
- More complex models always explain more variation in the data than simpler models do, even if the simpler model is correct
- To test whether this increase in explained variation is significant, we can use an F test
- This idea can be used to carry out a single test of whether three or more groups have the same mean and avoid the need for multiple comparison adjustment; this test is known as an ANOVA