Multiple comparisons and ANOVA

Patrick Breheny

April 19
So far in this class, I’ve painted a picture of research in which investigators set out with one specific hypothesis in mind, collect a random sample, then perform a hypothesis test.

Real life is a lot messier.

Investigators often test dozens of hypotheses, and don’t always decide on those hypotheses before they have looked at their data.

Hypothesis tests and $p$-values are much harder to interpret when multiple comparisons have been made.
As an example, suppose we see five cases of a certain type of cancer in the same neighborhood.

Suppose also that the probability of seeing a single case in neighborhood this size is 1 in 10.

If the cases arose independently (our null hypothesis), then the probability of seeing three cases in the neighborhood in a single year is \( \left( \frac{1}{10} \right)^5 = 0.00001 \).

This looks like pretty convincing evidence that chance alone is an unlikely explanation for the outbreak, and that we should look for a common cause.

This type of scenario occurs all the time, and suspicion is usually cast on a local industry and their waste disposal practices, which may be contaminating the air, ground, or water.
But there are a lot of neighborhoods and a lot of types of cancer

Suppose we were to carry out such a hypothesis test for 100,000 different neighborhoods and 100 different types of cancer

Then we would expect $(100,000)(100)(.00001) = 100$ of these tests to have $p$-values below .00001 just by random chance

As a result, further investigations by epidemiologists and other public health officials rarely succeed in finding a common cause

The lesson: if you keep testing null hypotheses, sooner or later, you’ll find significant differences regardless of whether or not one exists
If an investigator begins with a clear set of hypotheses in mind, however, and these hypotheses are independent, then there are methods for carrying out tests while adjusting for multiple comparisons.

For example, consider a study done at the National Institutes of Health to find genes associated with breast cancer.

They looked at 3,226 genes, carrying out a two-sample $t$-test for each gene to see if the expression level of the gene differed between women with breast cancer and healthy controls (i.e., they got 3,226 $p$-values).
If we accepted $p < .05$ as convincing evidence, what is the probability that we would make at least one mistake?

\[
P(\text{At least one error}) = 1 - P(\text{All correct}) \\
\approx 1 - .95^{3.226} \\
\approx 1
\]

If we want to keep our overall probability of making a type I error at 5%, we need to require $p$ to be much lower.
The Bonferroni correction

- Instead of testing each individual hypothesis at $\alpha = .05$, we would have to compare our $p$-values to a new, lower value $\alpha^*$, where

$$\alpha^* = \frac{\alpha}{h}$$

where $h$ is the number of hypothesis tests that we are conducting (this approach is called the Bonferroni correction).

- For the breast cancer study, $\alpha^* = .000015$

- Note that it is still possible to find significant evidence of a gene-cancer association, but much more evidence is needed to overcome the multiple testing.
Another way to adjust for multiple hypothesis tests is the *false discovery rate*.

Instead of trying to control the overall probability of a type I error, the false discovery rate controls the proportion of significant findings that are type I errors.

If a cutoff of $\alpha$ for the individual hypothesis tests results in $s$ significant findings, then the false discovery rate is:

$$FDR = \frac{h\alpha}{s}$$
So for example, in the breast cancer study, $p < 0.01$ for 207 of the hypothesis tests.

By chance, we would have expected $3226(0.01) = 32.26$ significant findings by chance alone.

Thus, the false discovery rate for this $p$-value cutoff is

$$FDR = \frac{32.26}{207} = 15.6\%$$

We can expect roughly 15.6% of these 207 genes to be spurious results, linked to breast cancer only by chance variability.
Breast cancer study: Visual idea of FDR
Breast cancer study: $FDR$ vs. $\alpha$
The issue of multiple testing comes up a lot – for example,

- Subgroup analyses: separately analyses of the subjects by sex or by age group or patients with severe disease/mild disease
- Multiple outcomes: we might collect data on whether the patients died, how long the patients were in the intensive care unit, how long they required mechanical ventilation, how many days they required treatment with vasopressors, etc.
- Multiple risk factors for a single outcome
A different kind of multiple comparisons issue arises when there is only one outcome, but there are multiple groups present in the study.

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.
The problem with multiple $t$-tests

- We talked about how to analyze one-sample studies and two-sample studies; how do we test for significant differences in a four-sample study?
- We could carry out 6 different $t$/Mann-Whitney tests (one for each two-group comparison), but as we have seen, this will increase our type I error rate (unless we correct for it)
- Instead, it is desirable to have a method for testing the single hypothesis that the mean of all four groups is the same
- To do this, however, we will need a different approach to than the ones we have used so far in this course: we will need to build a statistical model
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.”
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.
We can summarize the size of the residuals by calculating the *residual sum of squares*:

\[ RSS = \sum_{i} r_i^2 \]

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 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 $\text{RSS}_0$ and $\text{RSS}_1$ denote the residual sums of squares from the null model and the more complex model, the percentage of variance explained by the model is:

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

- In words, our model explains 2.2% of the variability in tailgating distance.
Complex models always fit better

- Still, 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*)
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.
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 $d_1 = 4$. 
The $F$ test

- Testing for whether the decrease in $\text{RSS}$ is significant depends on a quantity similar to the explained variance, but with some corrections for the sample size and number of parameters:

\[
F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(d_1 - d_0)}{\text{RSS}_1/(n - d_1)}
\]

- For the tailgating data, $F = 0.85$

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

$p = 0.47$
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:

\[
\begin{align*}
RSS_0 &= 140,420 \\
RSS_1 &= 126,182
\end{align*}
\]

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

![Graph showing ranked distances for ALC, MDMA, NODRUG, and THC]
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 barely skimming the surface today.
- Statistical models are the focus of the next course in this sequence, Biostatistics II (CPH 630).