FES 524: Natural Resources Data Analysis

Reading 3.2: Multiple comparisons and power

Contents

1 Multiple comparisons **1**
- 1.1 Type I and Type II errors . . . 1
- 1.2 Beyond Type I and Type II errors . . . 2
- 1.3 When to use an adjustment . . . 2
- 1.4 Options for adjustments in `emmeans` . . . 3

2 Inconclusive results **3**
- 2.1 Issue of sample size . . . 4
- 2.2 Issue of variation . . . 4
- 2.3 Limitations . . . 4

3 Power analysis **4**
- 3.1 Basic approach . . . 5
- 3.2 Why do a power analysis? . . . 5
- 3.3 Simulations . . . 6

1 Multiple comparisons

This week we will do a quick review of the issue of multiple comparisons (i.e., multiple hypothesis tests). Multiple comparisons is most often considered to be an issue when doing many comparisons using the same response variable, although sometimes folks consider comparisons from related response variables. All the comparisons we consider to be related in some way may be referred to as the family of comparisons.

The issue of multiple comparisons comes up when people are worried about the *familywise* Type I error rate; the Type I error rate for the whole family of comparisons.

1.1 Type I and Type II errors

Type I errors are when we falsely reject a "true" null hypothesis. When we consider Type I errors, we are often worried about saying something is going on when there really isn't. The more comparisons we do, the more likely it is we will make a mistake, assuming all null hypotheses we are testing are "true". The terms *false positive* or *false discovery* can also be used for this kind of mistake.

You will see a lot of focus on Type I errors in statistics classes. However, this is not the only kind of error you can make nor is it unquestionably the most important kind of error to worry about. Type II errors, while less discussed, may also be important. Type II errors occur when we fail to reject a false null hypothesis. These are issues of false negatives, where we conclude that the data are not inconsistent with a hypothesis of nothing going on, but there really is. Which mistake worse? That depends on the situation. For example, a false negative could be terrible in a cancer screening because we want to catch the disease in its early stages, while a false positive could be terrible in a DNA test at a crime scene because we could put an innocent person behind bars.

It's important to recognize that Type I and Type II errors are about long run behavior, and have little to do with p-values calculated from an individual study. They can only be calculated if we do the exact same study

many times, taking a different sample in each study iteration. In addition, and more important, they can only be calculated if we already know the true answer. To calculate a Type I error rate, we must assume there really is no difference among means or no relationship between variables (i.e., we assume the null hypothesis is *actually* true). To calculate a Type II error rate, we must assume there is really a difference among, for example, group means, and we would need to know exactly how large that true difference is (i.e., we need to specify an effect size).

While the concepts of Type I and Type II errors can be useful when in the planning stages of a study, they are not useful when looking at observed results. First, they can help lead us into dichotomous thinking, where we say there absolutely was or wasn't an effect only based on statistical measures. Second, observed results from a single study are not long run averages; Type I and Type II errors aren't meaningful when looking at the observed results from a single analysis, they are theoretical properties of statistical testing procedures under different conditions. They provide a means of comparing testing procedures on theoretical bases, but that's really about it.

1.2 Beyond Type I and Type II errors

Instead of thinking about Type I/Type II errors in a purely statistical sense, we can use them as a proxy for thinking about our study results and whether we are most concerned with overstating or understating the observed study results.

We need to think about how likely it is that whatever we are studying should show some effect, such as a difference in means or a non-zero slope. The probability that what you see is a "false discovery" is really based on an underlying truth we don't know and the unrealistic assumption that we can repeat the study over and over again under the same conditions. Given that, we need to think about what we believe that underlying truth to be based on expert knowledge when thinking about how to frame our results.

What kind of mistake are you willing to make in your own research? Are you more concerned with being too conservative and reporting little effect when there really was a large effect? This would be a situation where your confidence interval is overly wide and so you are being conservative since a wide confidence interval has many plausible values in it for the true effect. Or are you more concerned about overstating the results, making a strong statement of an effect that is actually spurious? This would be a situation where your confidence intervals are too narrow, and you overstate the precision of the results. Be prepared to discuss this in class and justify your choices in your own work to your committee, collaborators, and paper reviewers.

1.3 When to use an adjustment

Given the issues with Type I and Type II errors, do we need to adjust p-values and / or confidence intervals for multiple comparisons? The answer, like so many in statistics, is a solid "maybe". It depends on the goals of your research and your answers to some of the questions I listed above. It may also depend on your field. For good or ill, some fields always use certain multiple comparison adjustments regardless of any of the other issues we will consider.

The amount of planning that went into the comparisons is something to consider when deciding about whether you will use an adjustment or not. A few, carefully planned comparisons you defined before you collected any data is considered by many to be a different situation than when you are doing all possible pairwise comparisons across many groups. With preplanned comparisons you may be able to justify not using a multiple comparisons adjustment because you planned things so carefully using *a priori* knowledge. If you have a strong idea that there will be an effect based on a current scientific theory (i.e., reducing competition increases growth), it may be easier to justify not doing an adjustment. When doing all pairwise comparisons, though, you may be more concerned about overstating the results and so decide to use an adjustment for multiple comparisons. For example, what is more compelling?:

1. I defined three contrasts I was interested in before the study because strong differences would support my hypothesis, while only two strong difference would support an alternative hypothesis. Say I found strong effects and the p-values from the contrasts suggest the data are not particularly compatible with the null hypotheses of no differences.

2. I looked at 100 pairwise comparisons and found 3 with strong differences for which the p-values from the test against no difference were small.

The type of research you are doing could also affect your decision to make an adjustment or not. Confirmatory research is usually more strict about being careful not to overstate results than exploratory research. However, multiple comparisons can be a useful tool in exploratory research as well. For example, investigators doing exploratory genetics research tend to use multiple comparison adjustments as a way to weed out some of the many many (on the order of thousands to tens of thousands) potentially real associations between genetic loci and phenotypes, if nothing more than to focus their future efforts exploring impacts of genetics variation at specific loci.

What the results are going to be used for may affect whether we want to be more conservative (wider CI) or less conservative (narrower CI). This is based on scientific expertise and not statistics, which is one reason why I can't give you a blanket rule on when we should or shouldn't use multiple comparisons adjustments.

Whatever you choose to do, you should state which adjustment you used and why. This indicates you thought about the problem and made a decision as an expert in your field. Below are some examples of stating why adjustments were or were not used so you can see the kind of language you might use in assignments and in your thesis.

1.4 Options for adjustments in `emmeans`

We are using package `emmeans` in this class for doing comparisons of means across groups. This package comes with several built-in adjustments for multiple comparisons. You can see more information on these adjustments in the documentation for `summary.emmGrid()`. Also review your notes from previous statistics classes if you can't remember when some of these are used.

Below are the emmeans built-in adjustments below. The stars indicate adjustments that are extremely conservative for large families of comparisons. Unless these are standard in your field or you are very concerned about overstating results, you might want to avoid these adjustments.

- Tukey's honestly significant difference (HSD) - For all pairwise comparisons.
- Scheffe's method
- Sidak correction*
- Bonferroni correction*
- Dunnett's method - For all comparisons against a control
- Multivariate-t correction - Useful for adjusting confidence intervals
- No correction

2 Inconclusive results

Unfortunately, there are going to be situations where the results of a study are inconclusive. The thinning example from week 2 is one such situation. In Figure **??**, we can see that the differences between the moderate and heavy thinning versus the light thinning were in the expected direction but were smaller than the practically important difference. In addition, the confidence intervals are quite wide, encompassing both practically important differences and differences in the "wrong" direction.

There is nothing we can do about this after the fact. Investigators need to be up front that little was learned, which can be difficult to admit, especially when journals prefer to publish studies with strong effects and publications carry such weight on a CV (this biases the perception of what we think we know in the scientific community, though).

2.1 Issue of sample size

Having wide confidence intervals is a common problem when an analysis ends up with inconclusive results. The size of the confidence interval is affected by the standard errors, which are affected by the sample size.

As a reminder, here is the calculation of the standard error:

$$SE = \sqrt{\frac{s^2}{n}}$$

where s^2 is the sample variance. What happens to the standard error as n gets large?

Given the same standard deviation (SD), a larger sample will reduce the standard error and so the size of confidence intervals. While this is true for any finite standard deviation, there is one **very important** nuance to this. Sample size is not the only thing that affects the standard error. You can collect infinitely many samples from a study in which your protocol or variable of interest is confounded with another variable, such as the blocking factor discussed in reading 3.1. If both are in the model, the width of the CI will remain infinitely large! Thus, increasing your sample size is a good use of your time and resources *if* you have a solid experimental / sampling design. Sometimes, however, you can get a bigger "bang for your buck" in terms of smaller standard errors by improving the experimental / sampling design instead of simply sampling till you drop (the brute force option).

2.2 Issue of variation

Unexplained variation is another important factor that can cause wide confidence intervals. This is variation that is not explained by other variables in the model or controlled for by the study design.

What happens to the standard error as the SD gets small, given a fixed sample size?

The affect of unexplained variation on the results is why we spent time talking about sources of variation. If we can control the variation via study design (e.g., blocking) or explain it through covariates, we can get a more precise answer without having to sample more units.

2.3 Limitations

You will be asked to discuss study limitations on assignments and in your final project.

One such limitation for inconclusive results like the one above could be related to a wide confidence interval. If you believe the study should have chosen a larger sample size, be specific on how that would help with conclusions in a future study.

The amount of unexplained variation is another common limitation for studies with inconclusive results. If so, you could discuss other factors that you think investigators should control or collect data on in future studies.

The representativeness of the sample to the population of interest could be a limitation you want to discuss. For example, you could question how representative a very small sample really is of the population. The smaller the sample, the greater an effect "odd" observation will have on your results, despite being a valid observation from the population of interest.

Scope of inference will commonly come up in your limitations section even if results were "conclusive". I find this especially true for observational studies done in a single year or some other short time frame. We know as ecologists that conditions in nature can vary drastically from year to year.

3 Power analysis

Statistical power is defined mathematically as 1 minus the probability of a Type II error. In words, power is the probability that we will reject the null hypothesis when it is indeed false, given some fixed alpha level.

This is a statistical term that requires theoretical long-run averages and the assumption that we know the truth; remember to avoid dichotomous thinking in *observed* results as discussed above.

I almost hesitate to include the following information because I think it risks instilling too much dichotomous thinking. Read below with the caveat in your mind that Type I and Type II errors only arise *because we are thinking in a dichotomous way.* While this is valid in the theoretical context, most of what we do en ecology shouldn't be boiled down to "discoveries" and "lack-thereof". However, power analyses can be useful during the study design phase, so the general ingredients that are needed for power analysis are outlined below.

3.1 Basic approach

3.1.1 Define an effect size

Power analysis requires you to define an effect size that you think is realistic and would indicate something practically important in your field. Generally speaking, the larger the effect size, the fewer samples you need to detect a signal.

3.1.2 Define the expected variance around that effect size.

Again, this must be a realistic estimate of the variance. The smaller the variance, the fewer samples you need to get a precise estimate of the effect size.

It is often nice to define the effect size and expected variance may come from previous research. But beware the winner's curse! These values should not be solely based on estimates from small pilot studies. Studies with very low power can lead to exaggerated effect sizes while severely underestimating the true variance. The short blog post by Andrew Gelman you are also reading this week hits on this same idea. If you are interested in more information on the winner's curse, see the Button et al. 2013 paper "Power failure: why small sample size undermines the reliability of neuroscience".

3.1.3 Choose a minimum power your study should have

This is where you think about the seriousness of making a Type II error in what you are studying. The higher the power you choose (i.e., the lower the Type II error rate), the more samples you will need.

3.1.4 Choose an alpha level

This is where you think about the seriousness of making a Type I error. You shouldn't simply default to choosing 0.05 as your cut-off but seriously consider what it would mean to your research to make a Type I error as well as how likely you think it is that you will see an effect. The larger the alpha value the fewer samples you will need to detect "an effect" (again, potentially nonsensical dichotomous thinking here).

Unlike after you have observed the results, considering Type I and Type II error rates and how serious each is in your research can be helpful during the design phase.

Once you have all of these pieces of information you can calculate the number of samples you need per group (for simple study designs). If working with categorical explanatory variables, we usually make the simplifying assumption that we will have the same number of observations in every group when doing a power analysis. Indeed, balanced designs are more powerful than unbalanced designs.

3.2 Why do a power analysis?

There are a lot of positives that come out of doing a power analysis.

Investigators must define the effect of interest. This means they are forced to consider the size of the effect they think would be important, which is a big improvement over only considering the value from the null hypothesis.

Since values are needed for both an effect size and a reasonable estimate of variance around that effect, investigators must review the previous research that is similar to the proposed research with a critical eye. For example, they need to consider how "good" estimates are that are coming out of previous research.

Analyses are based on a lot of assumptions, and a power analysis forces investigators to state those assumptions. This includes the alpha value they are willing to consider as well as a value for power.

For simple analyses, such as a t-test or a basic ANOVA, you may find simple power analysis calculators online, like this one. For more complicated study designs or models like mixed models or generalized linear models you will more likely need to use a simulation-based approach.

3.3 Simulations

Simulations are a powerful approach to exploring design trade offs and understanding how a model works in less-than-ideal circumstances (such as if an assumption hasn't been met). Simulations usually involve a fair amount of coding, which makes them difficult to fit into most classes that are set up like this one. We unfortunately will not have enough time to learn to do simulations in this class.

You can see some blog posts on how to do get started doing simulations in R for different kinds of linear models here: https://aosmith.rbind.io/tags/#simulation-list.