11.5: Introduction to Pairwise Comparisons

Any time you run an ANOVA with more than two groups and end up with a significant effect (reject the null hypothesis), the first thing you'll want to do is see if the mean differences are in the direction that you predicted in your research hypothesis. This means that you'll have to find out which groups are actually different from one another; remember, the ANOVA only tells you that at least one mean is different from at least one other mean. In our job applicant example, our null hypothesis was that all three types of degrees (None, Related, and Unrelated) have the same average test score. But if you think about it, the null hypothesis is actually claiming *three* different things all at once here. Specifically, it claims that:

- $\overline{X}_N = \overline{X}_R$
- $\overline{X}_N = \overline{X}_U$
- $\overline{X}_R = \overline{X}_U$

If any *one* of those three claims is false, then the null hypothesis for the whole ANOVA is also false. So, now that we've rejected our null hypothesis, we're thinking that *at least* one of those things isn't true. But which ones? All three of these propositions are of interest; that's why the research hypothesis predicts how each pair of group means relates to one another. When faced with this situation, its usually helps to look at the data,. For instance, if we look at the plots in Figure 11.4.2, it's tempting to conclude that having a Related degree is better than having No degree, but it's not quite clear if having a Related degree results in a significantly higher average test score than having an Unrelated degree. If we want to get a clearer answer about this, it might help to run some tests.

Running "pairwise" t-tests

How might we go about solving our problem? Given that we've got three separate pairs of means ($\overline{X}_N$ versus $\overline{X}_R$; $\overline{X}_N$ versus $\overline{X}_U$; $\overline{X}_R$ versus $\overline{X}_U$) to compare, what we could do is run three separate t-tests and see what happens. If we go on to do this for all possible pairs of variables, we can look to see which (if any) pairs of groups are significantly different to each other. This "lots of t-tests idea" isn't a bad strategy, though as we'll see later on there are some problems with it. However, our current problem is that it's a *pain* to calculate all of these t-tests by hand. You might be asking if statistical software would do this, and the answer is yes! But still, if your experiment has 10 groups, then you would have to run 45 t-tests.

There's Always a "But..."

In the previous section it was hinted that there's a problem with just running lots and lots of t-tests. The concern is that when running these analyses, what we're doing is going on a "fishing expedition": we're running lots and lots of tests without much theoretical guidance, in the hope that some of them come up significant. This kind of theory-free search for group differences is referred to as post hoc analysis ("post hoc" being Latin for "after this").

It's okay to run post hoc analyses, but a lot of care is required. For instance, running a t-test for each pair of means is actually pretty dangerous: each *individual* t-test is designed to have a 5% Type I error rate (i.e., α=.05), and I ran three of these tests so now I have about 15% chance of rejecting the null hypothesis when it is really true (Type I Error). Imagine what would have happened if my ANOVA involved 10 different groups, and I had decided to run 45 "post hoc" t-tests to try to find out which ones were significantly different from each other! You'd expect 2 or 3 of them to come up significant *by chance alone*. As we saw in the chapter when we first learned about inferential statistics, the central organizing principle behind null hypothesis testing is that we seek to control our Type I Error rate, but now that I'm running lots of t-tests at once, in order to determine the source of my ANOVA results, my actual Type I Error rate across this whole set of tests has gotten completely out of control.

The usual solution to this problem is to introduce an adjustment to the p-value, which aims to control the total error rate across the set of tests. There are different ways to do this adjustment. We'll discuss some common analyses in this section, but you should be aware that there are many other methods out there.

Corrections of p-values with Multiple Comparisons

These first two post-hoc analysis focus on calculating a probability based on the raw p-values from analyses conducted by statistical software, although these calculations can easily be done by hand.

Bonferroni Corrections

The simplest of these adjustments is called the Bonferroni correction, and it's very very simple indeed. Suppose that my post hoc analysis consists of "m" separate tests (in which "m" is the number of pairs of means you need to compare), and I want to ensure that the total probability of making *any* Type I errors at all is a specific alpha (α), such as 0.05. If so, then the Bonferroni correction just says "multiply all your raw p-values by m". If we let p denote the original p-value, and let p_B be the corrected value, then the Bonferroni correction tells that:

$$p_{Bonferroni} = m \times p$$

If you're using the Bonferroni correction, you would reject the null hypothesis if your Bonferroni probability is smaller than the alpha $(p_{Bonferroni} < \alpha)$.

Example 11.5.1

What would this look like for our job applicant scenario if the raw p-value was .004 for the difference between No Degree and a Related Degree?

Solution

$$p_{Bonferroni} = m \times p$$

We are making three comparisons ($\overline{X}_N$ versus $\overline{X}_R$; $\overline{X}_N$ versus $\overline{X}_U$; $\overline{X}_R$ versus $\overline{X}_U$), so m = 3.

$$p_{Bonferroni} = 3 \times 0.004$$

$$p_{Bonferroni} = 0.012$$

Because our Bonferroni probability (p_B) is smaller than our typical alpha $(\alpha)(0.012 < 0.05)$, we reject the null hypothesis that this set of pairs (the one with a raw p-value of .004)

The logic behind this correction is very straightforward. We're doing m different tests; so if we arrange it so that each test has a Type I error rate of at most α/m, then the *total* Type I error rate across these tests cannot be larger than α. That's pretty simple, so much so that in the original paper, the author writes:

> *The method given here is so simple and so general that I am sure it must have been used before this. I do not find it, however, so can only conclude that perhaps its very simplicity has kept statisticians from realizing that it is a very good method in some situations (pp 52-53 Dunn 1961)*

Holm Corrections

Although the Bonferroni correction is the simplest adjustment out there, it's not usually the best one to use. One method that is often used instead is the ***Holm correction*** (Holm, 1979). The idea behind the Holm correction is to pretend that you're doing the tests sequentially; starting with the smallest raw p-value and moving onto the largest one. For the j-th largest of the p-values, the adjustment is *either*

$$p'_j = j \times p_j$$

(i.e., the biggest p-value remains unchanged, the second biggest p-value is doubled, the third biggest p-value is tripled, and so on), *or*

$$p'_j = p'_{j+1}$$

whichever one is *larger*. This might sound a little confusing, so let's go through it a little more slowly. Here's what the Holm correction does. First, you sort all of your p-values in order, from smallest to largest. For the smallest p-value all you do is multiply it by m, and you're done. However, for all the other ones it's a two-stage process. For instance, when you move to the second smallest p value, you first multiply it by m−1. If this produces a number that is bigger than the adjusted p-value that you got last time, then you keep it. But if it's smaller than the last one, then you copy the last p-value. To illustrate how this works, consider the table below, which shows the calculations of a Holm correction for a collection of five p-values:

Table 11.5.1-Holm Calculations and p-values

raw p	rank j (m)	p×j	Holm p
.001	5	.005	.005
.005	4	.020	.020
.019	3	.057	.057
.022	2	.044	.057
.103	1	.103	.103

Hopefully that makes things clear.

Although it's a little harder to calculate, the Holm correction has some very nice properties: it's more powerful than Bonferroni (i.e., it has a lower Type II error rate), but – counterintuitive as it might seem – it has the *same* Type I error rate. As a consequence, in practice there's never any reason to use the simpler Bonferroni correction, since it is always outperformed by the slightly more elaborate Holm correction.

Those are the types of corrections to the p-values that can be done to make sure that you don't accidently commit a Type I Error. Next, we'll cover post-hoc analyses that find a critical value for the differences between each pair of means.

Reference

Dunn, H. L. (1961). *High-level wellness: A collection of twenty-nine short talks on different aspect os the them "High Level Wellness for Man and Society"*. Arlington, VA: Beatty.

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)