Lecture 14

Corrections for Multiple Comparisons

In this lecture, we will discuss making many comparisons at once. We will discuss what the problem with making too many comparisons is mathematically and then we will discuss possible corrections to the problem, including Fisher's philosophy that there is no problem if the ANOVA was significant. The data we will use for this lecture come from Carnegie Mellon's StatLib Data and Story collection: http://lib.stat.cmu.edu/DASL/Stories/cuckoo.html. The data are the length of Cuckoo eggs found in the nests of other species.

Mathematics of conducting many tests at once

Suppose you conduct n hypothesis tests, each at the α significance level. What that means is that for each test you have probability α of making a specific mistake: namely, the mistake of rejecting the null hypothesis of no difference when the null hypothesis is actually true. In the most extreme case when the n tests are independent of each other, your family-wide error rate (the probability of making at least one mistake when all n null hypotheses are true) is then:

$$P(\text{at least one mistake in } n \text{ tests}) = 1 - P(\text{no mistake in } n \text{ tests}) = 1 - (1-\alpha)^n \approx n\alpha$$

Thus, your type I error grows approximately linearly with the number of tests you conduct when α is small compared to $1/n$. If your goal is to have the family-wide error rate be α so that your probability of making any mistake is α, then you need to adjust the significance level at which you run each individual test.

The Bonferroni Correction

The Bonferroni correction simply says that if you conduct n hypothesis tests at one time, conduct each one at a significance level of α/n and then your family-wide error rate will be no more than α. The correction is simple, straightforward, and works well under many circumstances. It can be applied to pair-wise comparisons after an ANOVA, but it is generally too strong of a correction in those circumstances because the pair-wise comparisons are not all independent of each other.

Fisher's ANOVA protected method

Sir Ronald Fisher had a different opinion, at least about doing multiple comparisons after an ANOVA. His argument was as follows: if the ANVOA is significant then you know that there are differences between the populations and you are free to explore each of those possible differences at the α significance level. This is called an ANOVA-protected multiple comparison - you make your multiple comparisons only if an ANOVA indicates that there are differences, but then you make comparisons at an uncorrected 5% level.

Tukey's method

One problem with the Bonferroni method is that it works properly in the case where you are conducting independent tests, but pairwise comparisons after an ANOVA are NOT independent of each other - if you know something about the comparisons between populations 1 and 2 and between populations 2 and 3 then you have information about the comparison between populations 1 and 3. Tukey's method is based on the expected range for the comparisons made after an ANOVA:

$$\frac{\bar{Y}\text{max} - \bar{Y}\text{min}}{\text{S.E.}} = \text{Q}$$

Q is called the studentized range distribution.

The confidence interval for the difference between two population means with Tukey's correction is

$$\bar{Y}_i - \bar{Y}_j \pm \frac{Q(1-\alpha, I, n-I)}{\sqrt{2}} s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Where I is the number of populations being compared in the ANOVA, n is the total sample size from all the populations, s is the pooled standard deviation from the ANOVA, and n_i and n_j are the individual sample sizes from population i and j respectively. Kramer contributed an adjustment for unbalanced designs so the method is sometimes called Tukey-Kramer.

Dunnett's method and Hsu's method

Dunnett's method specifies one population to be the control and it compares the other populations to this one control population. This can be useful, especially in medical applications, when not all pairwise comparisons are relevant.

Hsu's method compares all populations to the "best" population where the best population is one that has the smallest (or largest) mean. You specify which extreme is best. This too is useful in medical applications - you have 5 different drugs and you want the one that reduces the viral count the most. What's relevant is whether you can detect a difference from the best drug - if not, the difference might be due to chance and the drugs might be equally effective, but if you can, then why waste time on a drug that is not as good as the "best" one.

False Discovery Rate

Statisticians are inherently conservative creatures and the methods they invent to control for making any mistake are so stringent that the techniques above are not always useful to scientists who are interested in making discoveries. A significant and insightful contribution to the myriad of methods correcting for multiple tests came from Benjamini and Hochberg in 1995 who turned the question around and instead of asking to control the probability of making any mistake instead

developed a method to control the expected proportion of mistakes made among all "discoveries." Their method is useful in most data mining and microarray analysis situations where most tests are insignificant, but you are conducting so many tests that there is still a large number of discoveries to be make. (So 1000 tests, 100 discoveries possible, for instance.) Their procedure is simply the following: collect the p-value for each test conducted separately. Put those p-values in order from smallest to largest ($p_{(1)} < p_{(2)} < \cdots < p_{(n)}$) Find the largest k so that

$$\frac{n}{k} p_{(k)} < \alpha$$

and reject all tests corresponding to the p-values equal to and smaller than $p_{(k)}$. Then roughly your expected number of false discoveries among all your discoveries has proportion α.

Minitab

In Minitab, you can make pairwise comparisons while conducting an ANOVA. In the one-way ANOVA dialog box, click on "comparisons" and select the method you want to use. If you want to use a Bonferroni correction (not recommended after an ANOVA because the tests are not independent) then just fill in the correct significance level (α/number of tests) you want into Fisher's correction box.

References and Readings

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

[2] E. P. Chance. The truth about the cuckoo. *Country Life*, 1940.

Exercises for Lecture 14

1. –

2. –