About Multiple Comparison (or Pairwise Comparison) Analyses

If your research design has only two conditions, the omnibus-F test will be sufficient to test your research hypothesis (but be sure to check if the direction of the mean difference agrees with your research hypothesis). However, if you have three or more conditions of the qualitative grouping variable, rejecting the H0: tells you that the condition means are not all the same but doesn't tell you what the pattern of the mean differences is. For that you need to perform additional statistical analyses, one kind of which is called "multiple pair-wise comparisons". "Pairwise" means that each comparison looks at the difference between the means of a pair of design conditions. "Multiple" reminds us that there will be at least three pairwise comparisons, in order to obtain a complete description of the pattern of mean differences among the IV conditions.

How many pairwise comparisons? That depends upon the number of research conditions. The formula for the number of independent pairwise comparisons is k(k-1)/2, where k is the number of conditions. If we had three conditions, this would work out as 3(3-1)/2 = 3, and these pairwise comparisons would be Gap 1 vs.Gap 2, Gap 1 vs. Gap 3, and Gap 2 vs. Grp3. Notice that the reference is to "independent" pairwise comparisons. This is because comparing Gap 1 vs. Gap 2 is the same as comparing Gap 2 vs. Gap 1, so we do only one of them.

Although pairwise comparisons are a useful way to fully describe the pattern of mean differences (and so, to test a research hypothesis), performing multiple analyses also creates a problem for us. When we reject a H0: because the obtained (calculated) value is larger than the critical value (looked up from the table), we know that we might be making a Type I statistical decision error (rejecting the H0: based on this sample when that H0: is really true in the population of interest). We even know the probability of making that Type I error -- the value of p. If we reject the H0: based on a critical value looked up in the table based on p= .05, we know that we have a 5% chance of committing a Type I error when we reject the H0:. When we make multiple comparisons using the p=.05 criterion, we have a chance of making a Type I error on each comparison. Also, the more comparisons we make, the more likely we are to make at least one Type I error. The chance of committing at least one Type I error for a set of comparisons can be estimated as c * .05, where c is the number of comparisons made. For our example with three conditions, we can make three comparisons, this would be 3 * .05 = .15, or a 15% chance of making at least one Type I error.

There are two opposing approaches to what we should do about this increase in the possibility of making Type I error when we make multiple comparisons. One view is that Type I errors are a real problem, and we should "protect" ourselves from them by using more strict criteria for rejecting the H0: when we do multiple comparisons. We can easily do this by setting the p-value for rejecting H0: for each pairwise comparisons to some value lower than p=.05, say p=.01. The second view reminds us that if we adjust the p-value to reduce the Type I error rate we are consequently raising the probability of a Type II error (retaining the H0: based on this sample when the H0: is false in the population of interest), and so should test each pairwise comparisons at p = .05. This second view also reminds us that if we set more strict standards for rejecting H0: of pairwise comparisons, we might end up with the troublesome result that we reject the omnibus-F (and so, know there is are least one difference among the condition means) but then find no mean differences based on the more strict pairwise comparisons.

The first of these views emphasizes protection from Type I error. It is often labeled as "conservative", in that it requires extra evidence (lower p-value) to reject H0: for pairwise comparisons. The second view emphasizes protection from Type II error. It is often referred to as emphasizing "sensitivity", because it uses the α=.05 criterion for each pairwise comparison. Researchers often differ (quite loudly) about which of these approaches has greater merit, usually based on whether they are more concerned about "missing effects" (making Type II errors -- these folks usually favor sensitive pairwise testing) or "claiming to find effects that aren't really there" (making Type I errors -- these folks usually favor conservative pairwise testing). Because of these differences of opinion, below is one for completing "sensitive" pairwise comparisons, called the Least Significant Difference (LSD) procedure, and one for completing "conservative" pairwise comparisons, called the Honestly Significant Difference (HSD) procedure.