
Editor's Corner: Too Much of a Good Thing

This editorial is dedicated to my friend and colleague, Tony Carpenter, who served as Editor and then Editor Emeritus for the Journal of Studies on Alcohol. He was a marvelous teacher who dedicated a significant portion of his life to helping students and investigators see the forest and not get lost in the trees. I would like to think that he would have found some of these thoughts amusing, and perhaps even useful.

I RECENTLY HAD the opportunity to reread a comment published in 1998 in the *British Medical Journal* (Perneger, 1998) that makes some points I think are well worth heeding. The topic was the Bonferroni correction, or statistical adjustment, used to mitigate the impact of carrying out multiple statistical evaluations on the same data. I will return to the specific article later, but I think there are some broader issues that might be highlighted. These comments might rest under the subheading of the need to use common sense when applying statistical rules.

It is important to emphasize at the outset that I do not know enough about statistical formulas to be a heretic. I have been trained as a clinician and clinical researcher and need help from my colleagues for any but the more straightforward statistical tests. I am also a teacher and enjoy helping young scientists develop academic careers, which gives me the opportunity to watch their usual approach to statistics as they begin to apply classroom-based rules to the real world of science.

It is quite common for one of my junior colleagues to tell me that something is “meaningful” or “not worth noting” based solely on the p value. For example, not too long ago a student comparing spouses of alcoholics with controls concluded that there were no meaningful differences based primarily on the absence of a significant p value for one item, while ignoring the pattern of consistent group differences that approached statistical significance on items that related directly to the original hypothesis. This person began with the mind set that, regardless of the size of the sample involved, if the differences did not reach $p < .05$, there was nothing of importance likely to be there. He missed an important pattern of results.

I try to preach a somewhat different approach. My advice is to first analyze the data without paying much heed to statistical values in order to determine if patterns emerge from the numbers. To use a simple example, it is worth-

while asking if there is a consistent manner in which two groups differ in a way that either fits with or conflicts with the original hypothesis. It is also relevant to ask if the pattern of relationships among variables in a series of regression analyses maintains similar correlations or beta weights in various subsets of the data as are observed for the entire population. In this instance, a beta weight might be exactly the same for a subgroup as was observed in the entire sample, but it may be no longer statistically significant because of the smaller number of subjects evaluated. It would be unwise to interpret these results as demonstrating that the general conclusions have no meaning to the particular subgroup evaluated.

After observing patterns of findings, it is then important to pay attention to the results of the tests of significance. These can help confirm one's observations of patterns and can help determine whether subsequent steps are required (such as increasing the number of subjects). However, the statistical values alone should not dictate one's interpretations. Even findings that are not statistically significant can be of great potential importance.

Of course, if the sample is large enough or if groups differ on additional relevant characteristics, large and highly statistically significant group differences can be observed, but by themselves they do not necessarily make the data meaningful. It takes judgment and some level of experience to evaluate whether the statistically significant differences carry potentially important implications. Here, another statistic, the effect size, can also be a helpful guide. This possible “overinterpretation” of data can also be made by experienced investigators who are attempting to draw conclusions outside of their usual area of expertise. For instance, an investigator with only limited clinical experience finds significant differences between two clinical groups without recognizing that the differential was almost inevitable based on the way in which the groups were structured. In this case, statistical results can help guide the researcher, but on their own do not dictate the potential importance of the results. To human beings involved in research projects, this should be reassuring as one corollary of this premise is that meaningful decisions are not likely to be made by computers or statistical formulas alone.

That brings me back to the particular article by Perneger (1998). The Bonferroni correction attempts to control for a Type I error where the researcher concludes that something is present when it is not. The potential problem occurs when, for example, two groups are compared on multiple variables, with the assumption that each statistic should now be considered in light of a p value that becomes more and more demanding with each test used. The author points out, however, that the correction was originally developed to help deal with decisions regarding highly repetitive situations where, for example, a manufacturer looking for imperfections in a product coming off an assembly line must consider the number of times a sample was taken from that line in order to appropriately determine whether the product was truly defective. Thus, the Bonferroni adjustments are best applied to problems in interpreting repeated decisions of the same type from the same sample, but problems may arise in interpreting results from a research experiment or clinical situation that does not involve using exactly the same test over and over again. The correction has much less meaning, for example, when one looks at outcomes of an intervention and evaluates survival rate, quality of life scores and complication rates for the same subjects. Perneger also reminds us that the too rigid application of Bonferroni adjustments increases the probability of rejecting an important finding even though it is meaningful (a Type II error).

The purpose of this note is to help remind investigators, especially those relatively new to science, that even the most useful statistical rule must be applied with common

sense. There are, indeed, potentially important problems that can result from carrying out multiple statistical tests in a given dataset. In some instances (for example, where the test might be looked at as small variations in approaches asking the same question of the dataset), Bonferroni adjustments are important. In others, it is sufficient to warn the reader that the investigator recognizes that multiple evaluations are carried out and that the initial univariate analyses are only a step toward more sophisticated multivariate approaches where the significance level might be more appropriately determined, although patterns might not be optimally highlighted by this approach. In any event, as noted by Perneger, the essential step is to accurately describe the analyses that were carried out and to remind the reader of potential problems in interpretation of the data.

To paraphrase the main point of this editorial, the determination of whether a specific result is meaningful requires much more than carrying out statistical tests. These are important tools to help the researcher and clinician place findings into appropriate perspective, but, when used rigidly and without appropriate thoughtfulness, they can be too much of a good thing.

Reference

- PERNEGER, T.V. What's wrong with Bonferroni adjustments. *Brit. Med. J.* **316**: 1236-1238, 1998.

Marc A. Schuckit
Editor