This Week's Citation Classic®

Tests of the equality of means fall into just two classes: a t test or an omnibus F test. The relation between these two classes is illustrated, and the relation between several minor variations of the t test is shown. [The *SCI®* and the *SSCI®* indicate that this paper has been cited in over 145 publications.]

———◆●◆———

Origins of Multiple Comparisons of Means

Paul A. Games
Department of Human Development
and Family Studies
College of Health and Human Development
Pennsylvania State University
University Park, PA 16802

July 30, 1990

In my graduate statistics course, I had established the algebraic identity of the common t statistic for two independent groups, used by Fisher as the Least Significant Difference, and the Tukey test variously referred to as the Honestly Significant Difference or Wholly Significant Difference. The statistic was the same, with the Tukey test just using a larger critical value, hence being less powerful, but with a lower risk of Type I error. A Monte Carlo article in an excellent journal[1] reported results as described above, but misinterpreted the results as if they were based on two entirely different statistics. This encouraged me to prepare my proof for publication since it would remove the confusion.

As I worked on this paper, I tried extending the algebraic logic to other tests and concluded that the Dunnett test and the Newman-Keuls test also were special cases of t differing only in critical values. The largest thrill of exploration was that I was able to relate the region of retention of the above t tests to the region of retention of the omnibus F test. The negative feature of this exploration was that the paper was getting longer, and I worried about where to publish it.

In exploring the literature further, I encountered the book by R.G. Miller, Jr.,[2] and discovered that he

had covered some of the same topics, so that some of my paper was a rediscovery. Fortunately, Miller's book was oriented toward mathematical statisticians and had a higher level of abstraction than would be understood by behavioral scientists with limited statistical backgrounds. However, it was reassuring to note that my conclusions agreed with Miller's. Thus, I continued my efforts and eventually sent in a long manuscript to the *American Educational Research Journal (AERJ)*. This was the longest manuscript I have ever published in a journal (as contrasted with a book), and in printed form it took 35 pages with seven figures.

Fortunately, the initial reviews were favorable, with most supporting the length as needed to provide comprehensive coverage, and the editor was willing to tolerate the length. The galley proofs were frustrating since they did not include any of the Greek letters or tildes, and these had to be entered by hand. Also, I never had a chance to make any corrections when the typesetters messed up some of these characters. However, I provided an errata sheet with the original reprints and made hand corrections for later photocopy reprints.

The article was well received and shared the Palmer Johnson Award as one of the two best articles in the 1971 issues of the *AERJ*. I am grateful that the journal was willing to accept such a long article with so many figures, since I have had other experiences where all the reviewers were positive, but the editor said he was not going to publish because of length and figures. The article probably is cited often because it removed the confusion between the various statistics that may be used in multiple comparisons on means. Till that time different texts gave different recommendations and did not recognize the similarity of the different tests based on the t statistic. The article is considerably shorter, less abstract, and less comprehensive than Miller's book,[2] which also included topics like multiple regression.

Since then, I've tried to extend the robustness of some of the procedures,[3] and Miller has produced a second edition.[4] Hochberg and Tamhane[5] have provided a more recent book that is very comprehensive, but it includes some techniques that I believe are best ignored.

1. Petrinovich L F & Hardyck C D. Error rates for multiple comparison methods: some evidence concerning the frequency of erroneous conclusions. *Psychol. Bull.* 71:43-54, 1969. (Cited 75 times.)
2. Miller R G. *Simultaneous statistical inference.* New York: McGraw-Hill, 1966. (Cited 1,625 times.)
3. Games P A & Howell J F. Pairwise multiple comparisons procedures with unequal n's and variances: a Monte Carlo study. *J. Educ. Stat.* 1:113-25, 1976. (Cited 40 times.)
4. Miller R G. *Simultaneous statistical inference.* New York: Springer-Verlag, 1981. (Cited 545 times.)
5. Hochberg Y & Tamhane A C. *Multiple comparison procedures.* New York: Wiley, 1987. (Cited 20 times.)