Statistical Science
2011, Vol. 26, No. 4, 604–607
DOI: 10.1214/11-STS356C
Main article DOI: 10.1214/11-STS356

Discussion of "Multiple Testing for Exploratory Research" by J. J. Goeman and A. Solari

Peter H. Westfall

1. INITIAL COMMENTS

Closure-based multiple testing procedures for controlling the familywise error rate (FWER) have been around for decades, but they have not been well understood, and hence have been under-appreciated and under-utilized. Goeman and Solari (GS) provide a service by highlighting important practical features of closure. Using elegant notation for closure-based methods, they develop a handy book-keeping tool for presenting additional results of closed testing that are available when non-consonant testing methods are used, and they prove its validity.

In their Figure 1, GS provide the confidence set $\tau(\{2,3\}) \in \{0,1\}$, where $\tau(\{2,3\})$ is the number of true nulls in the set $\{H_2, H_3\}$. In doing so, GS highlight a not-so-well known fact about closure: inferences for the additional $(2^n - 1) - n$ composite hypotheses H_I are available "free of charge" whenever one performs closed testing for the original n elementary hypotheses H_i. This follows from the fact that "the closure of the closure is the closure;" that is, that no new hypotheses are generated when the set of $2^n - 1$ intersection hypotheses is treated as the set of elementary hypotheses. Hence, in GS's Figure 1, the significance of $H_{\{2,3\}}$ can be stated with full FWER control over the set of $2^3 - 1 = 7$ hypotheses, and the conclusion $\tau(\{2,3\}) \le 1$ follows immediately. Again, GS provide a service in reminding statisticians (or in teaching those who have not heard about it in the first place) of this nice feature of closure.

GS's paper also implicitly explains the following paradox: while closure is based on composite hypotheses, it is not true that more powerful composite tests lead to more powerful closure-based multiple tests. When considering only the elementary hypotheses, Bonferroni (or MaxT) types of composite tests, which are usually thought to be the least powerful of the class of composite testing methods (e.g., Nakagawa, 2004), tend to give higher power for closure-based multiple tests (Romano, Shaikh and Wolf, 2011). However, when the goal is to establish how many true effects there might be among a collection of hypotheses, GS suggest indeed that more powerful composite tests lead to more powerful multiple tests.

Peter Westfall is Paul Whitfield Horn Professor of Statistics, Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, Texas 79409-2101, USA (e-mail: peter.westfall@ttu.edu).

The Fisher combination test is a useful choice of composite test, as noted by GS. But it is worth pointing out how bad this test can be compared to the Bonferroni test, when both are used via closure for testing elementary hypotheses. Consider analyzing a version (available from the author) of the classic dataset reported by Golub at al. (1999), testing 7,129 genes for association with either acute myeloid or acute lymphoblastic leukemia, using 7,129 two-sample t-tests. The closed Fisher combination method is and has been available in PROC MULTTEST of SAS/STAT with the $O(n^2)$ shortcut since release 8.1 of SAS in 2000; this software computes closure-based adjusted p-values (defined below) to assess significance of elementary hypotheses. Despite the fact that the Fisher combination test is liberal with correlated data, the *smallest* adjusted p-value using the closed Fisher combination test is 1.000 (rounded), hence none of the 7,129 tests are significant at any reasonable nominal FWER level. On the other hand, 37 of the 7,129 genes have adjusted p-values less than the nominal 0.05 FWER level when using closed Bonferroni (or Holm, 1979) tests; the smallest adjusted p-value is 1.7×10^{-6} and is therefore extremely significant, even after multiplicity adjustment.

I have some other comments/critiques about the paper that fall into the following categories: (i) the assumption of free combinations and its consequences, (ii) use of adjusted p-values rather than rigid nominal thresholds, (iii) computational shortcuts, and (iv) permutation testing.