Statistical Science
2011, Vol. 26, No. 4, 601–603
DOI: 10.1214/11-STS356B
Main article DOI: 10.1214/11-STS356

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

Nicolai Meinshausen

I want to congratulate the authors on this thought-provoking and important paper on multiple testing in exploratory settings.

Standard Multiple Testing procedures can appear very mechanistic. Hypotheses are ordered by increasing p-value. Given a Type I error criterion, the Multiple Testing procedure selects a cut-off in this list. Simply working down the list of hypotheses in order of their p-values is perhaps suboptimal for exploratory analysis as a lot of information is lost in this way and important discoveries might be missed. Some previous work has addressed this issue by changing the ranking of the hypotheses. To highlight only three examples: Tibshirani and Wasserman (2006) devised a method to borrow strength across highly correlated test statistics in microarray experiments. Storey (2007) proposed an "optimal discovery" procedure that again leads to a different ranking of variables than the ranking implied by the marginal p-values. One of the authors also proposed a very powerful way of incorporating known network structure into the testing procedure [Goeman and Mansmann, 2008].

The proposed approach to exploratory multiple testing is more radical, though, than changing the cut-off or changing the ranking of hypotheses. Instead of the perhaps rather dull task of selecting a cut-off in a list of ordered hypotheses, the researcher can reject for follow-up analysis any set of hypotheses he or she regards as interesting, using all the information at hand. The method then returns a lower bound on the number of false null hypotheses (true discoveries) in this set. Since the bound is valid simultaneously across all sets, an exploratory approach does not invalidate the error bound.

I think this method will be very important and useful in many fields as it allows a flexible exploration of possibly interesting sets of hypotheses, while at the same time protecting the practitioner against too many false rejections (or at least managing expectations about the number of true discoveries one can hope to make).

There is a price to be paid for the simultaneous nature of the bound, though. I have some doubts (hopefully unfounded) about the applicability to large-scale testing situations as they arise, for example, in genomics or astronomy for two reasons: computational complexity and statistical power.

It is obvious and also acknowledged by the authors that the proposed procedure without shortcuts will be impractical for even just a few dozen hypotheses. The computational complexity is simply too high. An example is shown in Figure 1 for a genomics regression example with less than one hundred observations. The proposed method takes already more than half a minute for 12 predictor variables on a standard computer with a 3 GHz CPU and the supplied `cherry` R-package and the complexity seems to be (super-)exponential in the number of hypotheses, as one would expect. The proposed shortcuts are not applicable in all settings. If they are applicable, they seem to be very effective in reducing the computational complexity, making large-scale testing feasible. Figure 1 shows that even testing situations with $> 10^6$ tests are handled in about a second or less.

Maybe more worrying, the statistical power of the method deteriorates with an increasing number of hypotheses. This is due to the simultaneous nature of the bound on the number of correctly rejected hypotheses among all possible sets of hypotheses. I compared the power for a simple setting, in which there are m independent p-values p_i with $i = 1, \ldots, m$ with distribution $p_i \sim U([0, c_i])$ and $c_i = 1$ if $i > 10$ and $c_i = 0.1/m$ if $i \leq 10$ (there are hence 10 false null hypotheses). If rejecting all hypotheses, the lower bound for the number of correctly rejected hypotheses is shown as a function of m in Figure 1, along with the bound for the same quantity proposed by Meinshausen and Rice (2006). The proposed approach works very well up to a few dozen hypotheses. If the number of hypotheses

Nicolai Meinshausen is University Lecturer, Department of Statistics, University of Oxford, UK (e-mail: meinshausen@stats.ox.ac.uk).