Sankhyā : The Indian Journal of Statistics
2008, Volume 70, Part 2, pp. 169-176

Discussion: On Methods Controlling the False Discovery Rate

Joseph P. Romano
Stanford University, USA
Azeem M. Shaikh
University of Chicago, USA
Michael Wolf
University of Zurich, Switzerland

1 Introduction

It is a pleasure to acknowledge another insightful article by Sarkar. By developing clever expressions for the FDP, FDR, and FNR, he manages to prove fundamental properties of stepdown and stepup methods. It is particularly important that the theory is sufficiently developed so as to apply to what Sarkar calls adaptive BH methods. Here, the goal is to improve upon the Benjamini Hochberg procedure by incorporating a data-dependent estimate of the number of true null hypotheses. Theoretical justification of such methods is vital and Sarkar's analysis is useful for this purpose.

A perhaps more ambitious task is to develop methods which implicitly or explicitly estimate the joint dependence structure of the test statistics (or p-values). The focus of our discussion is to show how resampling methods can be used to construct stepdown procedures which control the FDR or other general measures of error control. The main benefit of using the bootstrap or subsampling is the ability to estimate the joint distribution of the test statistics, and thereby offer the potential of improving upon methods based on the marginal distributions of test statistics. The procedure below is a generalization of one we developed for FDR control, and the utility of the bootstrap is that it can apply to essentially arbitrary measures of error control, such as the pairwise FDR of Sarkar, the k-FWER, or the tail probabilities of the false discovery proportion. However, it is important to note that the justification of such methods is asymptotic.

2 Setup and Notation

Our setup is as follows. Suppose data $X = (X_1, \ldots, X_n)$ is available from some probability distribution $P \in \Omega$. A hypothesis H may be viewed as a subset ω of Ω. In this paper we consider the problem of simultaneously testing null hypotheses $H_i : P \in \omega_i, i = 1, \ldots, s$ on the basis of X. The alternative hypotheses are understood to be $H_i' : P \notin \omega_i, i = 1, \ldots, s$.

We assume that test statistics $T_{n,i}, i = 1, \ldots, s$ are available for testing $H_i, i = 1, \ldots, s$. Large values of $T_{n,i}$ are understood to indicate evidence against H_i. Let

$$T_{n,(1)} \leq \cdots \leq T_{n,(s)}$$

denote the ordered test statistics (from smallest to largest) and let $H_{(1)}$, $\ldots$,$H_{(s)}$ denote the corresponding null hypotheses. A stepdown multiple testing procedure rejects $H_{(s)}, \ldots, H_{(s-j^*)}$, where j^* is the largest integer j that satisfies

$$T_{n,(s)} \geq c_s, \ldots, T_{n,(s-j)} \geq c_{s-j} \ ;$$

if no such j exits, the procedure does not reject any null hypotheses. The problem is how to construct the c_i so as to control the given measure of error control.

Denote by $I(P)$ the set of indices corresponding to true null hypotheses; that is,

$$I(P) = \{1 \leq i \leq s : P \in \omega_i\} \ . \tag{2.1}$$

For a given multiple testing procedure, let F denote the number of false rejections and let R denote the total number of rejections. Our goal is to construct a stepdown procedure so that

$$\limsup_{n \to \infty} E_P[g(F, R)] \leq \alpha \tag{2.2}$$

for all $P \in \Omega$, where α is some fixed value, not necessarily in $(0, 1)$. Some choices of g are $F/R \cdot I\{R > 0\}$, $I\{F \geq k\}$, $I\{F/R \cdot I\{R > 0\} > \lambda\}$ and Sarkar's pairwise FDR defined by $[F(F-1)]/[R(R-1)] \cdot I\{R > 1\}$.

3 Motivation for Method

In order to motivate the method, first note that for any stepdown procedure based on critical values $c_1, \ldots c_s$ we have that

$$\begin{aligned} E_P[g(F,R)] &= \sum_{1\le r\le s} E_P[g(F,r)|R=r]P\{R=r\} \\ &= \sum_{1\le r\le s} E[g(F,r)|R=r]P\{T_{n,(s)} \ge c_s, \ldots, \\ &\qquad T_{n,(s-r+1)} \ge c_{s-r+1}, T_{n,(s-r)} < c_{s-r}\} \ , \end{aligned}$$

where the event $T_{n,s-r} < c_{s-r}$ is understood to be vacuously true when $r = s$. Let $s_0 = |I(P)|$ and assume without loss of generality that $I(P) = \{1, \ldots, s_0\}$. Under weak assumptions, one can show that all false hypotheses will be rejected with probability tending to one. For the time being, assume that this is the case. Let $T_{n,r:t}$ denote the rth ordered of the t test statistics $T_{n,1}, \ldots, T_{n,t}$, ordered from smallest to largest (here and throughout). In particular, when $t = s_0$, $T_{n,r:s_0}$ denotes the rth ordered of the test statistics corresponding to the true hypotheses. Then, with probability approaching one, we have that

$$\begin{aligned} E_P[g(F,R)] = \sum_{s-s_0+1\le r\le s} & g(r-s+s_0,r)P\{T_{n,s_0:s_0} \ge c_{s_0}, \ldots, \\ & T_{n,s-r+1:s_0} \ge c_{s-r+1}, T_{n,s-r:s_0} < c_{s-r}\} \ , \end{aligned} \tag{3.1}$$

where the event $T_{n,s-r:s_0} < c_{s-r}$ is again understood to be vacuously true when $r = s$.

Our goal is to ensure that (3.1) is bounded above by α for any P, at least asymptotically. To this end, first consider any P such that $s_0 = |I(P)| = 1$. Then, (3.1) reduces to

$$g(1,s)P\{T_{n,1:1} \ge c_1\} \ . \tag{3.2}$$

A suitable choice of c_1 is thus the smallest value for which (3.2) is bounded above by α; that is,

$$c_1 = \inf\{x \in \mathbb{R} : P\{T_{n,1:1} \ge x\} \le \alpha/g(1,s)\} \ .$$

Note that if $\alpha/g(1,s) \ge 1$, then c_1 so defined is equal to $-\infty$. Having determined c_1, now consider any P such that $s_0 = 2$. Then, (3.1) is simply

$$g(1,s-1)P\{T_{n,2:2} \ge c_2, T_{n,1:2} < c_1\} + g(2,s)P\{T_{n,2:2} \ge c_2, T_{n,1:2} \ge c_1\} \ . \tag{3.3}$$

A suitable choice of c_2 is therefore the smallest value for which (3.3) is bounded above by α.

In general, having determined $c_1, \ldots, c_{j-1}$, the jth critical value may be determined by considering P such that $s_0 = j$. In this case, (3.1) is simply

$$\sum_{s-j+1 \leq r \leq s} g(r-s+j, r) P\{T_{n,j:j} \geq c_j, \ldots, T_{n,s-r+1:j} \geq c_{s-r+1}, T_{n,s-r:j} < c_{s-r}\} \,. \tag{3.4}$$

An appropriate choice of c_j is thus the smallest value for which (3.4) is bounded above by α.

Of course, the above choice of critical values is infeasible, since it depends on the unknown P through the distribution of the test statistics. We therefore focus on feasible constructions of critical values based on resampling.

4 A Bootstrap Approach

We now specialize our framework to the case in which interest focuses on a parameter vector

$$\theta(P) = (\theta_1(P), \ldots, \theta_s(P)) \,.$$

The null hypotheses may be one-sided, in which case

$$H_j : \theta_j \leq \theta_{0,j} \quad \text{vs.} \quad H_j' : \theta_j > \theta_{0,j} \tag{4.1}$$

or the null hypotheses may be two-sided, in which case

$$H_j : \theta_j = \theta_{0,j} \quad \text{vs.} \quad H_j' : \theta_j \neq \theta_{0,j} \,. \tag{4.2}$$

Test statistics will be based on an estimate $\hat{\theta}_n = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,s})$ of $\theta(P)$ computed using the data X. We will consider 'studentized' test statistics

$$T_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j})/\hat{\sigma}_{n,j} \tag{4.3}$$

for the one-sided case (4.1) or

$$T_{n,j} = \sqrt{n}|\hat{\theta}_{n,j} - \theta_{0,j}|/\hat{\sigma}_{n,j} \tag{4.4}$$

for the two-sided case (4.2). Note that $\hat{\sigma}_{n,j}$ may either be identically equal to 1 or an estimate of the standard deviation of $\sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j})$. This is done

to keep the notation compact; the latter is preferable from our point of view but may not always be available in practice.

Recall that the construction of critical values in the preceding section was infeasible because of its dependence on the unknown P. For the bootstrap construction, we therefore simply replace the unknown P with a suitable estimate $\hat{P}_n$. To this end, let $X^* = (X_1^*, \ldots, X_n^*)$ be distributed according to $\hat{P}_n$ and denote by $T_{n,j}^*, j = 1, \ldots, s$ test statistics computed from X^*. For example, if $T_{n,j}$ is defined by (4.3) or (4.4), then

$$T_{n,j}^* = \sqrt{n}(\hat{\theta}_{n,j}^* - \theta_j(\hat{P}_n))/\hat{\sigma}_{n,j}^* \tag{4.5}$$

or

$$T_{n,j}^* = \sqrt{n}|\hat{\theta}_{n,j}^* - \theta_j(\hat{P}_n)|/\hat{\sigma}_{n,j}^* \ , \tag{4.6}$$

respectively, where $\hat{\theta}_{n,j}^*$ is an estimate of θ_j computed from X^* and $\hat{\sigma}_{n,j}^*$ is either identically equal to 1 or an estimate of the standard deviation of $\sqrt{n}(\hat{\theta}_{n,j}^* - \theta_j(\hat{P}_n))$ computed from X^*. For the validity of this approach, we require that the distribution of $T_{n,j}^*$ provides a good approximation to the distribution of $T_{n,j}$ whenever the corresponding null hypothesis H_j is true, but, unlike Westfall and Young (1993), we do not require subset pivotality. The exact choice of $\hat{P}_n$ will, of course, depend on the nature of the data. If the data $X = (X_1, \ldots, X_n)$ are i.i.d., then a suitable choice of $\hat{P}_n$ is the empirical distribution, as in Efron (1979). If, on the other hand, the data constitute a time series, then $\hat{P}_n$ should be estimated using a suitable time series bootstrap method; see Lahiri (2003) for details.

Given a choice of $\hat{P}_n$, define the critical values recursively as follows: having determined $\hat{c}_{n,1}, \ldots, \hat{c}_{n,j-1}$, compute $\hat{c}_{n,j}$ according to the rule

$$\begin{aligned}\hat{c}_{n,j} = \inf\{c \in \mathbb{R} : \sum_{s-j+1 \le r \le s} g(r-s+j, r)\hat{P}_n\{T_{n,j:j}^* \ge c, \\ \ldots, T_{n,s-r+1:j}^* \ge \hat{c}_{n,s-r+1}, T_{n,s-r:j}^* < \hat{c}_{n,s-r}\} \le \alpha\} \ .\end{aligned} \tag{4.7}$$

It is important to be clear about the meaning of the notation $T_{n,r:t}^*$, with $r \le t$, in (4.7). By analogy with the "real" world, it should denote the rth ordered of the observations corresponding to the first t true null hypotheses. However, the ordering of the true null hypotheses in the bootstrap world is not $1, 2, \ldots, s$, but it is instead determined by the ordering $H_{(1)}, \ldots, H_{(s)}$ from the real world. So if the permutation $\{k_1, \ldots, k_s\}$ of $\{1, \ldots, s\}$ is defined such that $H_{k_1} = H_{(1)}, \ldots, H_{k_s} = H_{(s)}$, then $T_{n,r:t}^*$ is the rth ordered of the observations $T_{n,k_1}^*, \ldots, T_{n,k_t}^*$.

We now provide conditions under which the stepdown procedure with critical values defined by (4.7) satisfies (2.2). The following result applies to the case of two-sided null hypotheses, but the one-sided case can be handled using a similar argument. In order to state the result, we will require some further notation. For $K \subseteq \{1, \ldots, s\}$, let $J_{n,K}(P)$ denote the joint distribution of

$$(\sqrt{n}(\hat{\theta}_{n,j} - \theta_j(P))/\hat{\sigma}_{n,j} : j \in K) \ .$$

It will also be useful to define the quantile function corresponding to a c.d.f. $G(\cdot)$ on $\mathbb{R}$ as $G^{-1}(\alpha) = \inf\{x \in \mathbb{R} : G(x) \geq \alpha\}$.

THEOREM 4.1. *Consider the problem of testing the null hypotheses $H_i, i = 1, \ldots, s$ given by (4.2) using test statistics $T_{n,i}, i = 1, \ldots, s$ defined by (4.4). Suppose that $J_{n,\{1,\ldots,s\}}(P)$ converges weakly to a limit law $J_{\{1,\ldots,s\}}(P)$, so that $J_{n,I(P)}(P)$ converges weakly to a limit law $J_{I(P)}(P)$. Suppose further that $J_{I(P)}(P)$*

(i) has continuous one-dimensinal marginal distributions;

(ii) has connected support, which is denoted by supp$(J_{I(P)}(P))$;

(iii) is exchangeable.

Also, assume

$$\hat{\sigma}_{n,j} \xrightarrow{P} \sigma_j(P) \ ,$$

where $\sigma_j(P) > 0$ is nonrandom. Let $\hat{P}_n$ be an estimate of P such that

$$\rho(J_{n,\{1,\ldots,s\}}(P), J_{n,\{1,\ldots,s\}}(\hat{P}_n)) \xrightarrow{P} 0 \ , \tag{4.8}$$

where ρ is any metric metrizing weak convergence in $\mathbb{R}^s$. Then, the stepdown method with critical values defined by (4.7) satisfies

$$\limsup_{n\to\infty} E_P[g(F, R)] \leq \alpha \ .$$

The proof of the Theorem closely follows the special case of FDR control in Romano et. al. (2008), and generalizes quite easily. The power of computer-intensive methods is evident as the bootstrap adapts easily to different choices of g.

Note that, in the definitions of $T^*_{n,j}$ given by (4.5) or (4.6) used in our bootstrap method to generate the critical values, one can typically replace $\theta_j(\hat{P}_n)$ by $\hat{\theta}_{n,j}$; but, see Remark 5.5 of Romano et. al. (2008).

An alternative approach can be based on subsampling, which avoids the exchangeability assumption; see Romano et. al. (2008). The bootstrap approach performed better in simulations.

5 Concluding Remarks

We have developed a bootstrap method which provides asymptotic control of the false discovery rate, or other generalized error rates. Asymptotic validity of the bootstrap holds under fairly weak assumptions, but we require an exchangeability assumption for the joint limiting distribution of the test statistics corresponding to true null hypotheses. However, simulations support the use of the bootstrap method under a wide range of dependence. Even under independence, our bootstrap method is competitive with that of Benjamini et. al. (2006), and outperforms it under dependence. While the approach is a generalization of one we developed for FDR control, for other measures of error control, other constructions may be preferable, such as those in Romano and Wolf (2007). The method described above requires calculation of s critical values, which may be prohibitive.

The bootstrap method succeeds in generalizing Troendle (2000) to allow for non-normality. However, it would be useful to also consider an asymptotic framework where the number of hypotheses is large relative to the sample size. Furthermore, it would be useful to know that limit result is uniform in P. Future work will address this.

There is clear tradeoff between methods based on marginal p-values and those based on resampling which attempt to account for the entire joint distribution of the test statistics. While exact finite-sample results like those in Sarkar are ideal, it must be pointed out that these methods may only serve as approximations if the marginal p-values are based on large-sample appoximations. Ultimately, the choice of methods for a particular application will heavily depend on the number of hypotheses and the sample size, and further work is needed to choose among competing methods. Given that there exists a growing number of competing methods, the situation is begging for guidance by some kind of optimality theory. Perhaps Sarkar's expressions for FNR can be used towards the construction of procedures with smallest FNR subject to the constraint of error control.

Acknowledgements. The research of the first author has been supported by the National Science Foundation grant DMS-0707085. The research of the second author has been supported by the National Science Foundation

grant DMS-0820310. The research of the third author has been supported by the University Research Priority Program "Finance and Financial Markets", University of Zurich, and by the Spanish Ministry of Science and Technology and FEDER, grant MTM2006-05650.

References

Benjmini, Y., Krieger, A.M. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, **93**, 491–507.

Efron, B. (1979). Botstrap methods. Another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Lahiri, S.N. (2003). *Resampling Methods for Dependent Data.*, Springer, New York.

Romano, J.P., Shaikh, A.M. and Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling, (with discussion), *Test*, **17**, 417–442.

Romano, J.P. and Wolf, M. Control of generalized error rates in multiple testing. *Ann. Statist.*, **35**, 1378–1408.

Troendle, J.F. (2000). Stepwise normal theory test procedures controlling the false discovery rate. *J. Stat. Plan. Inf.* **84**, 139-158.

Westfall, P.H. and Young, S.S. (1993). *Resampling-Based Multiple Theory Testing: Examples and Methods for P-value Adjustment.*, John Wiley, New York.

Joseph P. Romano
Departments of Statistics and Economics
Stanford University , USA
E-mail: romano@stanford.edu

Azeem M. Shaikh
Department of Economics
University of Chicago, USA
E-mail: amshaikh@uchicago.edu

Michael Wolf
Institute for Empirical Research in Economics
University of Zurich
E-mail: mwolf@iew.uzh.ch

Paper received October 2008; revised January 2009.