ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Dariusz Parys**

THE MODIFICATION OF STEPWISE MULTIPLE PROCEDURE

Abstract: In this paper we discuss stepdown methods that control the familywise error rate in finite samples. Such methods proceed stagewise by testing an intersection hypothesis without regard to hypotheses previously rejected. However, one cannot always achieve strong control in such a simple manner. By understanding the limitations of this approach in finite samples, we can then see why an asymptotic approach will be valid under fairly weak assumptions. It turns out that a simple monotonicity condition for theoretical critical values allows for some immediate results.

Key words: multiple testing, familywise error rate, stepdown procedure.

I. INTRODUCTION

Suppose data X generated from some unknown probability distribution P. In anticipation of asymmetric results, we may write $X = X^{(n)}$, where n typically refers to the sample size. A model assumes that P belongs to a certain family of probability distributions Ω, though we make no rigid requirements for Ω. Indeed, Ω may be a nonparametric model, a parametric model, or a semiparametric model.

Consider the problem of simultaneously testing a hypothesis H_j against $H_j^{'}$ for $j = 1, ..., k$. Of course, a hypothesis H_j can be viewed as a subset, ω_j, of Ω, in which case the hypothesis H_j is equivalent to $P \in \omega_j$ and $H_j^{'}$ is equivalent to $P \notin \omega_j$. For any subset $K \subset \{1,...,k\}$, let $H_K = \bigcap_{j \in K} H_j$ be the hypothesis that $P = \bigcap_{j \in K} \omega_j$.

Suppose that a test of the individual hypothesis H_j is based on a test statistic $T_{n,j}$, with large values indicating evidence against the H_j. For an individual hypothesis, numerous approaches exist to approximate a critical value, such as

* Ph. D., Chair of Statistical Methods, University of Łódź.

those based on classical likelihood theory, bootstrap tests, Edgeworth expansions, permutation tests, etc. The main problem addressed in the present work is to construct a procedure that controls the familywise error rate (FEW). Recall that the familywise error rate is the probability of rejecting at least one true null hypothesis. More specifically, if P is the true probability mechanism, let $I = I(P) \subset \{1,..., k\}$ denote the indices of the set of true hypotheses; that is, $i \in I$ if and only $P \in \omega_i$. The FWE is the probability under P that any H_i with $i \in I$ is rejected. To show its dependence on P, we may write FEW = FWE_P. We require that any procedure satisfy that the familywise error rate to no bigger than α (at least asymptotically). Furthermore, this constraint must hold for all possible configurations of true and null hypotheses; that is, demand strong control of the FEW. A procedure that only controls the FEW when all k null hypotheses are true is said to have weak control of the FEW. As remarked by Dudoit et. al. (2002), this distinction is often ignored.

For any subset K of $\{1,\ldots, k\}$, let $c_{n,K}(\alpha, P)$ denote an α-quantile of the distribution of $\max_{j\in K} T_{n,j}$ under P. Concretely,

$$c_{n,K}(\alpha, P) = \inf\{x : P\{\max_{j\in K} T_{n,j} \le x\} \ge \alpha\}. \tag{1}$$

For testing the intersection hypothesis H_K, it is only required to approximate a critical value for $P \in \bigcap_{j\in K} \omega_j$. Because there may be many such P, we define

$$c_{n,K}(\alpha - 1) = \sup\{c_{n,K}(1-\alpha, P) : P \in \bigcap_{j\in K} \omega_j\}. \tag{2}$$

At this point, we acknowledge that calculating these constants may be formidable in some problems (which is why we later turn to approximate or asymptotic methods).

Let

$$T_{n,r_1} \ge T_{n,n_2} \ge \ldots \ge T_{n,r_k} \tag{3}$$

denote the observed ordered test statistics, and let $H_{r_1}, H_{r_2},\ldots, H_{r_k}$ be the corresponding hypotheses.

II. STEPDOWN PROCEDURES

Stepdown procedures begin by testing the joint null hypothesis $H_{\{1,\ldots,k\}}$ that all hypotheses are true. This hypothesis is rejected if T_{n,r_1} is large. If it is not large, accept all hypotheses; otherwise, reject the hypothesis corresponding to the largest test statistic. Once a hypothesis is rejected, remove it and test the remaining hypotheses by rejecting for large values of the maximum of the remaining test statistics, and so on. Thus, at any step, one tests an intersection hypothesis, and an ideal situation would be to proceed at any step without regard to previous rejections (or not having to consider conditioning on the past). Because the Holm procedure works in this way, one might hope that one can generally test the intersection hypothesis at any step without regard to hypotheses previously rejected. Forgetting about whether or not such an approach generally yields strong control for the time being, we consider the following conceptual algorithm, which proceeds in stages by testing intersection hypotheses.

Algorithm 2.1 (Idealized Stepdown Method)

1. Let $K_1 = \{1,\ldots,\}$. If $T_{n,r_1} \le c_{n,K_1}(1-\alpha)$, then accept all hypotheses and stop; otherwise, reject H_{r_1} and continue.
2. Let K_2 be the indices of the hypotheses not previously rejected. If $T_{n,r_2} \le c_{n,K_2}(1-\alpha)$, then accept all remaining hypotheses and stop; otherwise, reject H_{r_2} and continue.

⋮

j. Let K_j be the indices of the hypotheses not previously rejected. If $T_{n,r_j} \le c_{n,K_j}(1-\alpha)$, then accept all remaining hypotheses and stop; otherwise, reject H_{r_j} and continue.

⋮

k. If $T_{n,k} \le c_{n,K_k}(1-\alpha)$, then accept H_{r_k}; otherwise, reject H_{r_k}.

The above algorithm is an idealization for two reasons: the critical values may be impossible to compute and, without restriction, there is no general reason why such a stepwise approach strongly controls the FWE. The determination of conditions where the algorithm leads to strong control will help us understand the limitations of a stepdown approach as well as understand how such a general approach can at least work approximately in large samples. First, we present an example to show that some condition is required to exhibit strong control.

Example 2.1 Suppose $T_{n,1}$ and $T_{n,2}$ are independent and normally distributed, with $T_{n,1} \sim N(\theta_1,(1+\theta_2)^{2p})$ and $T_{n,2} \sim N(\theta_2,(1+\theta_2)^{-2p})$, where $\theta_1 \geq 0$ and $\theta_2 \geq 0$. (The index n plays no role here, but we retain it for consistent notation). Here, p is a suitable positive constant, chosen to be large. Also, let $\Phi(\cdot)$ denote the standard normal cumulative distribution function. The hypothesis H_i specifies $\theta_i = 0$ while $H_i^{'}$ specifies $\theta_i \geq 0$. Therefore, the first step of Algorithm 2.1 is to reject the overall joint hypothesis $\theta_1 = \theta_2 = 0$ for large values of $\max(T_{n,1},T_{n,2})$ when $T_{n,1}$ and $T_{n,2}$ are i.i.d. $N(0, 1)$. Specifically, accept both hypotheses if

$$\max(T_{n,1},T_{n,2}) \leq c(1-\alpha) \equiv \Phi^{-1}(\sqrt{1-\alpha});$$

otherwise, reject the hypothesis corresponding to the larger $T_{n,i}$. Such a procedure exhibits weak control but not strong control. For example, the probability of rejecting the H_1 at the first step when $\theta_1 = 0$ and $\theta_2 = c(1-\alpha)/2$ satisfies

$$P_{0,\theta_2}\{T_{n,1} > c(1-\alpha), T_{n,1} > T_{n,2}\} \to 1/2$$

as $p \to \infty$. So, if $\alpha < 1/2$, for some large enough but fixed p, the probability of incorrectly declaring H_1 to be false is greater than α. Incidentally, this also provides an example of a single-step procedure which exhibits weak control but not strong control. (Single-step procedures are those where hypotheses are rejected on the basis of a single critical value; see Westfall and Young (1993).)

Therefore, in order to prove strong control, soma condition is required. Consider the following monotonicity assumption: for $I \subset K$,

$$c_{n,K}(1-\alpha) \geq c_{n,I}(1-\alpha). \tag{4}$$

The condition (4) can be expected to hold in many situations because the left hand side is based on computing the $1-\alpha$ quantile of the maximum of $|K|$ variables, while the right hand side is based on the maximum of $|I| \leq |K|$ variables (though one must be careful and realize that the quantiles are computed under possibly different P, which is why soma condition is required). Romano i Wolf (2005) proved the following theorem:

Theorem 2.1 *Let P denote the true distribution generating the data.*
(i) Assume for any K containing I (P),

$$c_{n,K}(1-\alpha) \geq c_{n,I(P)}(1-\alpha). \tag{5}$$

Then, the probability that Algorithm 2.1 rejects any $i \in I(P)$ *is* $\leq \alpha$; *that is,* $FWE_p \leq \alpha$.

(ii) Strong control persists if, in Algorithm 2.1, the critical constants $c_{n,K}(1-\alpha)$ *are replaced by* $d_{n,K_j}(1-\alpha)$ *which satisfy*

$$d_{n,K_j}(1-\alpha) \geq c_{n,K_j}(1-\alpha) \tag{6}$$

(iii) Moreover, the condition (5) may be removed if the $d_{n,K_j}(1-\alpha)$ *satisfy*

$$d_{n,K}(1-\alpha) \geq d_{n,I(P)}(1-\alpha) \tag{7}$$

for any $K \supset I(P)$.

Remark 2.1 Under weak assumptions, one can show the sup over P of the probability that Algorithm 2.1 rejects any $i \in I(P)$ is equal to a. It then follows that the critical values cannot be made smaller, in hopes of increasing the ability to detect false hypotheses, without violating the strong control of the FWE. (However, this does not negate the possibility of smaller random critical values, as long as they are not smaller with probability one.)

Example 2.2 Assumptions stronger than (5) have Been used. Suppose, for example, that for every subset K C f 1, ..., k}, there exists a distribution P_K which satisfies

$$c_{n,K}(1-\alpha,P) \leq c_{n,K}(1-\alpha,P_K) \tag{8}$$

for all P such that $I(P) \supset K$. Such a P_K may be referred to being least favorable among distributions P such that $P \in \bigcap_{j \in K} \omega_j$. (For example, if H_j corresponds to a parameter $\theta_j \leq 0$, then intuition suggests a least favorable configuration should correspond to $\theta_j = 0$.)

In addition, assume the subset pivotality condition of Westfall and Young (1993); that is, assume there exists a Po with $I(P_0) = \{1,...,k\}$ such that the joint

distribution of $\{T_{n,i} : i \in I(P_K)$ under P_K is the same as the distribution of $\{T_{n,i} : i \in I(P_n)$ under P_0. This condition says the (joint) distribution of the test statistics used for testing the hypotheses H_i, $i \in I(P_K)$ is unaffected by the truth or falsehood of the remaining hypotheses (and therefore we assume all hypotheses are true by calculating the distribution of the maximum under Po). It follows that, in step j of Algorithm 2.1,

$$c_{n,K}(1-\alpha) = c_{n,K_j}(1-\alpha, P_{K_j}) = c_{n,K_j}(1-\alpha, P_0) = c_{n,K_j}(1-\alpha); \tag{9}$$

the outer equalities in (9) follow by the assumption (8) and the middle equality follows by the subset pivotality condition. Therefore, in Algorithm 2.1, we can replace $c_{n,K_j}(1-\alpha)$ by $c_{n,K_j}(1-\alpha, P_0)$, which in principle is known because it is the $1-\alpha$ quantile of the distribution of $\max(T_{n,i} : i \in K_j)$ under P_0, and P_0 is some fixed (least favorable) distribution. At the very least, this quantile may be simulated.

The asymptotic behavior of stepwise procedures is considered in Finner and Roters (1998), and they recognize the importance of monotonicity for the validity of stepwise procedures. However, they also suppose the existence of a single least favorable Po for all configurations of true hypotheses, which then guarantees monotonicity of critical values for stepdown procedures. As previously seen, such assumptions do not hold generally.

Example 2.3 To exhibit an example where condition (5) holds, but subset pivotality does not, suppose that $T_{n,1}$ and $T_{n,2}$ are independent, normally distributed, with $T_{n,1} \sim N(\theta_1, 1/(1+\theta_2^2))$ and $T_{n,2} \sim N(\theta_2, 1/(1+\theta_1^2))$. The hypothesis H_i specifies $\theta_i = 0$ while the alternative H_i' specifies $\theta_i > 0$. Then, it is easy to check that, with $K_1 = \{1, 2\}$,

$$c_{n,K_1}(1-\alpha) = \Phi^{-1}(\sqrt{1-\alpha}) > \Phi^{-1}(1-\alpha) = c_{n,\{i\}}(1-\alpha).$$

Therefore, (5) holds, but subset pivotality fails.

Example 2.4 Suppose $-T_{n,i} \equiv \hat{p}_{n,i}$ is a p-value for testing H_i; that is, assume the distribution of $\hat{p}_{n,i}$ is Uniform on (0, 1) when H_i is true. Note that this assumption is much weaker than subset pivotality (if $k > 1$) because we are only making an assumption about the one-dimensional marginal distribution of the p-value statistic. Furthermore, we may assume the weaker condition

$$P\{\hat{p}_{n,i} \le x\} \le x$$

for any $x \in (0,1)$ and any $P \in \omega_i$. If $I(P) \supset K$, the usual argument using the Bonferroni inequality yields

$$c_{n,K}(1-\alpha, P) \le -\alpha / |K|,$$

which is independent of P, and so

$$c_{n,K}(1-\alpha) \le -\alpha / |K|, \tag{10}$$

It is easy to construct joint distributions for which this is attained, and so we have equality here if the family Ω is so large that it includes all possible joint distributions for the p-values. In such case, we have equality in (10) and so the condition (5) is satisfied. Of course, even if the model is not so large, this procedure has strong control. Simply, let $d_{n,K}(1-\alpha) = -\alpha / |K|$, and strong control follows by Theorem 2.1(iii).

Part (iii) of Theorem 2.1 points toward a more general method that has strong control even when (5) is violated, and that can be much less conservative than the Holm procedure.

Corollary 2.1 *Let*

$$c^*_{n,K_j}(1-\alpha) = \max\{c_{n,K}(1-\alpha): K \in K_j\}. \tag{11}$$

Then, if you replace $c_{n,K_j}(1-\alpha)$ *by* $c^*_{n,K_j}(1-\alpha)$ *in Algorithm 2.1, strong control holds.*

Corollary 2.1 is simply the closure principle of Marcus et al. (1976); also see Hommel (1986) and Theorem 4.1 of Hochberg and Tamhane (1987). Thus, in order to have a valid stepdown procedure, one must not only consider the critical value $c_{n,K}(1-\alpha)$ when testing an intersection hypothesis H_K, one must also compute all $c_{n,I}(1-\alpha)$ for $I \subset K$.

REFERENCES

Dudoit A., Shaffer J., Boldric J. (2002), *Multiple hypothesis testing in microarray experiments*. Technical report, Division of Biostatistics, U.C. Berkeley.

Finner H., Roters M. (1998), *Asymptotic comparison of step-down and step-up multiple test procedures based on exchangeable test statistics*, "Annals of Statistics6", 2: 505–524.

Hochberg Y., Tamhane A. (1987), *Multiple Comparison Procedures* Wiley, New York.

Holm S. (1979), *A simple sequentially rejective multiple test procedure*, "Scandinavian Journal of Statistics", 6: 65–70.

Marcus R., Teritz E., Gabriek K. (1976), *On closed testing procedures with special reference to ordered analysis of variance,* "Biometrica" 63: 655–660.

Romano I. P., Wolf M. (2005), *Stepwise multiple testing as formalized data snooping,* "Econometrica", 73, 1237–1282

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

Dariusz Parys

MODYFIKACJA KROCZĄCEJ WSTĘPUJĄCEJ PROCEDURY TESTOWANIA WIELOKROTNEGO

Procedury kroczące w porównaniach wielokrotnych często nie są w stanie zachować silnej kontroli nad błędem rodziny (tzw. familywise errors rate FWE). Prezentujemy tutaj ogólną metodę wnioskowania wielokrotnego opartego na krokach zstępujących i na jej tle proponujemy metodę wykorzystując modyfikację stałych krytycznych, które lepiej sprawują kontrolę nad FWE dla prób skończonych.