Statistics and Probability Letters 113 (2016) 38–40

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Efficient computation of adjusted *p*-values for resampling-based stepdown multiple testing☆

Joseph P. Romano [a], Michael Wolf [b,*]

[a] *Departments of Statistics and Economics, Stanford University, Stanford, CA 94305, USA*

[b] *Department of Economics, University of Zurich, 8032 Zurich, Switzerland*

ARTICLE INFO

Article history:
Received 19 February 2016
Accepted 20 February 2016
Available online 2 March 2016

Keywords:
Adjusted *p*-values
Multiple testing
Resampling
Stepdown procedure

ABSTRACT

There has been a recent interest in reporting *p*-values adjusted for the resampling-based stepdown multiple testing procedures proposed in Romano and Wolf (2005a,b). The original papers only describe how to carry out multiple testing at a fixed significance level. Computing adjusted *p*-values instead in an efficient manner is not entirely trivial. Therefore, this paper fills an apparent gap by detailing such an algorithm.

1. Introduction

Romano and Wolf (2005a,b) propose resampling-based stepdown multiple testing procedures to control the familywise error rate (FWE); also see Romano et al. (2008, Section 3). The procedures as described are designed to be carried out at a fixed significance level α. Therefore, the result of applying such a procedure to a set of data will be a 'list' of binary decisions concerning the individual null hypotheses under study: reject or do not reject a given null hypothesis at the chosen significance level α.

In a series of recent papers, however, there has been an interest in computing adjusted *p*-values instead.[1] That is, for each null hypothesis under study, compute a corresponding *p*-value adjusted for stepdown multiple testing proposed in Romano and Wolf (2005a,b). Examples of such papers include Heckman et al. (2010), Hein et al. (2010), Campbell et al. (2014), Gertler et al. (2014) and Dobbie and Fryer (2015). Unfortunately, the descriptions in these papers of how to compute the adjusted *p*-values are often unclear or even missing altogether.

In principle, for a given individual hypothesis, an adjusted *p*-value can be obtained by 'trial and error' as the smallest significance level α at which the hypothesis can be rejected by the stepdown multiple testing procedure. But clearly this way of computing adjusted *p*-values would be rather cumbersome. Instead, it is desirable to have an efficient (or streamlined) algorithm for computing adjusted *p*-values. This paper details such an algorithm.

Of course, algorithms for computing *p*-values adjusted for multiple testing have been described before; for example, see Westfall and Young (1993) and the various references to earlier work listed in Section 1.3 of that book. But the contribution of this paper is to describe an algorithm that is custom-tailored to the stepdown multiple testing procedures proposed in Romano and Wolf (2005a,b), which will make it easier for practitioners to understand and implement this algorithm.

☆ We thank Henning Müller for helpful comments.

* Corresponding author.
E-mail addresses: romano@stanford.edu (J.P. Romano), michael.wolf@econ.uzh.ch (M. Wolf).

[1] Such adjusted *p*-values are sometimes also called multiplicity-adjusted *p*-values.

http://dx.doi.org/10.1016/j.spl.2016.02.012
0167-7152/

2. Notation and unadjusted p-values

We now give a stylized, high-level description of the multiple testing problem under study. The details – such as the construction of test statistics and sufficient conditions for (asymptotic) validity of the proposed stepdown procedures – depend on the context; see Romano and Wolf (2005a,b) and Romano et al. (2008, Section 3).

There are S individual hypothesis testing problems:

$$H_s \quad \text{vs.} \quad H_s' \quad \text{for } s = 1, \ldots, S,$$

where H_s denotes a null hypothesis and H_s' denotes an alternative hypothesis. The corresponding test statistics are denoted by $t_1, \ldots, t_S$. They are designed in a way such that large values are indicative of the alternative. (In particular, for two-sided testing problems, the test statistics would usually be based on absolute values.)

Stepdown multiple testing procedures are generally based on a set of null resampling test statistics $\boldsymbol{t}^{*,m} := (t_1^{*,m}, \ldots, t_S^{*,m})$, for $m = 1, \ldots, M$, where M denotes the number of resampling repetitions. Depending on context, the resampling can be carried out by a bootstrap method, a permutation method, or a randomization method. Details for the bootstrap method can be found in Romano and Wolf (2005a, Section 4.2), Romano and Wolf (2005b), and Romano et al. (2008, Section 4.3). Details for the permutation and randomization methods can be found in Romano and Wolf (2005a, Section 3.2).

Following Davison and Hinkley (1997, Chapter 4), an unadjusted (or marginal) p-value for H_s, denoted by $\hat{p}_s$, can be defined as

$$\hat{p}_s := \frac{\#\{t_s^{*,m} \geqslant t_s\} + 1}{M + 1}. \tag{2.1}$$

Note that this definition of unadjusted p-values is not unique. For example, some people instead use the definition

$$\hat{p}_s := \frac{\#\{t_s^{*,m} \geqslant t_s\}}{M}. \tag{2.2}$$

Clearly, when M is reasonably large (such as $M = 1000$), the difference between (2.1) and (2.2) is not practically relevant.

3. Stepdown multiple testing at fixed significance level

It will be convenient to first describe the generic stepdown multiple testing procedure that controls the FWE at fixed significance level α in the stylized notation of this paper. In this way, the algorithm to compute the adjusted p-values in the next section will be easier to understand.

The hypotheses are relabeled in descending order of the observed test statistics. More specifically, let $\{r_1, r_2, \ldots, r_S\}$ denote a permutation of $\{1, 2, \ldots, S\}$ that satisfies $t_{r_1} \geqslant t_{r_2} \geqslant \cdots \geqslant t_{r_S}$. In this way, H_{r_1} is the 'most significant' hypothesis and H_{r_S} is the 'least significant' hypothesis.

Let $\max_{t,j}^{*,m}$ denote the largest value of the vector $(t_{r_j}^{*,m}, \ldots, t_{r_S}^{*,m})$, that is,

$$\max_{t,j}^{*,m} := \max\{t_{r_j}^{*,m}, \ldots, t_{r_S}^{*,m}\} \quad \text{for } j = 1, \ldots, S \text{ and } m = 1, \ldots, M.$$

Furthermore, let $\hat{c}(1-\alpha, j)$ denote an empirical $1-\alpha$ quantile of the collection $\{\max_{t,j}^{*,m}\}_{m=1}^{M}$. (There is no unique definition of an empirical quantile.[2] But as long as M is reasonably large, the differences are not practically relevant.)

The algorithm for the stepdown multiple testing procedure at significance level α is as follows.

Algorithm 3.1 (*Stepdown Multiple Testing at Significance Level α*)**.**

1. For $s = 1, \ldots, S$, reject H_{r_s} iff $t_{r_s} > \hat{c}(1-\alpha, 1)$.
2. Denote by R_1 the number of hypotheses rejected. If $R_1 = 0$, stop; otherwise let $j = 2$.
3. For $s = R_{j-1} + 1, \ldots, S$, reject H_{r_s} iff $t_{r_s} > \hat{c}(1-\alpha, R_{j-1} + 1)$.
4. (a) If no further hypotheses are rejected, stop.
   (b) Otherwise, denote by R_j the number of all hypotheses rejected so far and, afterwards, let $j := j + 1$. Then return to step 3.

Remark 3.1 (*Alternative Description*)**.** It is easy to see that H_{r_s} will be rejected at level α by Algorithm 3.1 if and only if

$$t_{r_j} > \hat{c}(1-\alpha, j) \quad \text{for all } j = 1, \ldots, s.$$

Therefore, the set of hypotheses rejected at level α is given by the collection $\{H_{r_1}, \ldots, H_{r_n}\}$, where n is the largest integer in the set $\{1, \ldots, S\}$ such that $t_{r_j} > \hat{c}(1-\alpha, j)$ for all $j = 1, \ldots, n$. If no such n exists, then no hypothesis is rejected. ■

[2] For example, the statistical software R offers nine different versions of empirical quantiles in its function `quantile`. Our recommendation would be to simply use the default version.

4. Adjusting *p*-Values for Stepdown multiple testing

We denote the adjusted *p*-value for hypothesis H_s by $\hat{p}_s^{\text{adj}}$. The following algorithm describes how these adjusted *p*-values can be computed in an efficient manner.

Algorithm 4.1 (*Computation of p-Values Adjusted for Stepdown Multiple Testing*)**.**

1. Define
$$\hat{p}_{r_1}^{\text{adj}} := \frac{\#\{\max_{t,1}^{*,m} \geqslant t_{r_1}\} + 1}{M+1}.$$
2. For $s = 2, \ldots, S$,
   (a) first let
$$\hat{p}_{r_s}^{\text{initial}} := \frac{\#\{\max_{t,s}^{*,m} \geqslant t_{r_s}\} + 1}{M+1},$$
   (b) then enforce monotonicity by defining
$$\hat{p}_{r_s}^{\text{adj}} := \max\{\hat{p}_{r_s}^{\text{initial}}, \hat{p}_{r_{s-1}}^{\text{adj}}\}.$$

Remark 4.1 (*Enforcing Monotonicity*)**.** Step 2(b) in Algorithm 4.1 is essential. Without it, the adjusted *p*-values for the hypotheses $H_{r_2}, \ldots, H_{r_S}$ would generally be too optimistic (in the sense of providing evidence against the null). This fact is easiest to see by considering H_{r_S}. Without step 2(b), it would hold that $\hat{p}_{r_S}^{\text{adj}} = \hat{p}_{r_S}$, so that the adjusted *p*-value would be equal to the unadjusted *p*-value. ■

It is straightforward to see that the adjusted *p*-values are correct in the sense that, as long as *M* is reasonably large, H_s will be rejected at fixed level α by Algorithm 3.1 for all practical purposes if and only if the adjusted *p*-value for H_s computed by Algorithm 4.1 satisfies $\hat{p}_s^{\text{adj}} \leqslant \alpha$. The addition of "for all practical purposes" to this statement is due to the fact that, as previously mentioned, there exists a unique definition neither for the empirical quantiles $\hat{c}(1-\alpha, j)$ used in Algorithm 3.1 nor for the resampling-based *p*-values used in Algorithm 4.1. But as long as *M* is reasonably large (such as $M = 1000$), violations of the if-and-only-if statement could not occur before the third decimal place of α, which is not practically relevant.

Acknowledgment

Research was supported by NSF Grant DMS-1307973.

References

Campbell, F., Conti, G., Heckman, J.J., Moon, S.H., Pinto, R., Pungello, E., Pan, Y., 2014. Early childhood investmens substantially boost adult health. Science 343, 1478–1485.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge.
Dobbie, W.S., Fryer, R.G., 2015. The medium-term impacts of high-achieving charter schools. J. Polit. Econ. 123 (5), 985–1037.
Gertler, P., Heckman, J.J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S.M., Grantham-McGregor, S., 2014. Labor market returns to an early childhood stimulation intervention in Jamaica. Science 344, 998–1001.
Heckman, J.J., Moon, S.H., Pinto, R., Savelyev, P., Yavitz, A., 2010. Analyzing social experiments as implemented: A reexamination of the evidence from the High Scope Perry Preschool Program. Quant. Econ. 1 (1), 1–46.
Hein, G., Silani, G., Preuschoff, K., Batson, C.D., Singer, T., 2010. Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. Neuron 68, 149–160.
Romano, J.P., Shaikh, A.M., Wolf, M., 2008. Formalized data snooping based on generalized error rates. Econometric Theory 24 (2), 404–447.
Romano, J.P., Wolf, M., 2005a. Exact and approximate stepdown methods for multiple hypothesis testing. J. Amer. Statist. Assoc. 100 (469), 94–108.
Romano, J.P., Wolf, M., 2005b. Stepwise multiple testing as formalized data snooping. Econometrica 73 (4), 1237–1282.
Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment. John Wiley, New York.