Statistics and Probability Letters 82 (2012) 1507–1514

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

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

A sharp upper bound for the expected number of false rejections

Alexander Y. Gordon*

University of North Carolina at Charlotte, Department of Mathematics and Statistics, 9201 University City Blvd, Charlotte, NC 28223, United States

ARTICLE INFO

Article history:
Received 25 May 2011
Received in revised form 7 March 2012
Accepted 7 March 2012
Available online 14 March 2012

MSC:
primary 62J15
secondary 62G10

Keywords:
Multiple testing procedure
Monotone procedure
Per-family error rate
Step-down procedure
Step-up procedure

ABSTRACT

We consider the class of monotone multiple testing procedures (monotone MTPs). It includes, among others, traditional step-down (Holm type) and step-up (Benjamini–Hochberg type) MTPs, as well as their generalization – step-up-down procedures (Tamhane et al., 1998). Our main result – the All-or-Nothing Theorem – allows us to explicitly calculate, for each MTP in those classes, its per-family error rate – the exact level at which the procedure controls the expected number of false rejections under general and unknown dependence structure of the individual tests. As an illustration, we show that, for any monotone step-down procedure (where the term "step-down" is understood in the most general sense), the ratio of its per-family error rate and its familywise error rate (the exact level at which the procedure controls the probability of one or more false rejections) does not exceed 4 if the denominator is less than 1.

1. Introduction

The traditional control of the *familywise error rate* (the probability of one of more false rejections, abbreviated as FWER) (Tukey, 1953) becomes impractical in those applications of multiple hypothesis testing where the number of hypotheses is large (e.g., in microarray data analysis). Tukey (1953) also introduced another measure of type I error occurrence—the *per-family error rate* (PFER), which equals the expected number of false rejections. In those testing situations, where thousands of hypotheses are tested simultaneously, the control of the FWER (that is, the requirement that with a probability close to 1 no true hypothesis be falsely rejected) is no longer desirable, because it severely reduces chances to detect false hypotheses. In these situations the PFER appears to be a natural "heir" of the traditional FWER.

The present work is motivated by the following question: given a multiple testing procedure (MTP) $\mathcal{M}$, what is the exact level at which it controls the PFER? This number is an important characteristic of the procedure's safety against "false discoveries".

The MTPs considered below are of the most common type: such a procedure uses as input the observed p-values p_i associated with the hypotheses H_i being tested, and its output is the list of rejected hypotheses (or equivalently, the list of indices i of the p-values declared *$\mathcal{M}$-significant*). We assume, furthermore, that the procedure $\mathcal{M}$ is *symmetric* (which p-values will be declared $\mathcal{M}$-significant does not depend on the order in which they are listed), *cutting* (the $\mathcal{M}$-significant

---

Abbreviations: MTP, multiple testing procedure; TSD, threshold step-down; TSU, threshold step-up; TSUD, threshold step-up-down; PFER, per-family error rate; FWER, familywise error rate.

* Tel.: +1 704 687 4576; fax: +1 704 687 6415.
E-mail addresses: aygordon@uncc.edu, aygordon@yahoo.com.

0167-7152/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2012.03.008

p-values, if any, are smaller than the $\mathcal{M}$-insignificant ones, if any), and *monotone*: reduction in some or all p-values can only increase the number of rejections.

The object of our study is the exact level at which a given procedure $\mathcal{M}$ controls the PFER under a general and unknown dependence structure of the p-values. Roughly speaking, this is the expected number of false rejections (falsely rejected true hypotheses) for the least favorable joint distribution of the p-values.

The main result of this work is the All-or-Nothing Theorem, which states that, under the additional assumption that all the hypotheses H_i are true, such a least favorable distribution can be found among those distributions that have the following property: given a random vector with such distribution, the procedure almost surely rejects either all hypotheses or none. This result allows us to explicitly calculate the exact level of control of the PFER (both with and without the above assumption) for the commonly used classes of stepwise procedures.

The remaining part of the paper is organized as follows. In the following Section 2, we present the necessary definitions and notation that will be used throughout the paper. In Section 3, we formulate the main result – the All-or-Nothing Theorem – and prove it. This result pertains to the weak control of the PFER (all hypotheses are supposed to be true). In Section 4, we derive a theorem pertaining to the strong control of the PFER (no restrictions on the joint distribution of the p-values). In Section 5, we apply the previously obtained results to stepwise procedures. In Section 6, we consider an illustrative example: a comparison of the exact levels at which a given MTP controls the PFER and the FWER. Obviously, their ratio is $\leq m$, where m is the number of hypotheses being tested. The results of the present work and certain earlier results imply that for a monotone step-down procedure (where the term "step-down" is understood in the most general sense) this ratio, rather surprisingly, does not exceed 4 if the denominator is less than 1.

2. Basic notions

2.1. Uninformed multiple testing procedures

A multiple testing procedure (MTP) is a decision rule that, based on randomly generated data, selects for rejection a subset of the given set of hypotheses about the probability distribution from which the data are drawn.

We assume that there are in total m hypotheses $H_1, H_2, \ldots, H_m$, and associated with them are p-values $P_1, P_2, \ldots, P_m$. The p-value P_i is a random variable (determined by the data) such that:

(i) $0 \leq P_i \leq 1$;

(ii) if the hypothesis H_i is true, then

$$\mathrm{pr}\{P_i \leq x\} \leq x \quad \text{for all } x\ (0 \leq x \leq 1). \tag{1}$$

The p-value P_i measures the strength of the evidence against the hypothesis H_i provided by the data: the smaller P_i, the stronger the evidence. For a more detailed discussion of p-values, see, for example, Lehmann and Romano (2005, pp. 63–64).

From now on we assume that the hypothesis H_i is true if and only if (1) holds.

A *marginal-p-value-only-based*, or *uninformed*, multiple testing procedure (in the sequel — just *multiple testing procedure*, or *MTP*) is a Borel measurable mapping $\mathcal{M}: I^m \to 2^{\mathbf{N}_m}$ from the unit cube $I^m = [0, 1]^m$ to the set $2^{\mathbf{N}_m}$ of all subsets of $\mathbf{N}_m = \{1, 2, \ldots, m\}$. Applying $\mathcal{M}$ to a vector $\mathbf{p} = (p_1, \ldots, p_m)$ of observed p-values (p-*vector*), we obtain a subset $\mathcal{M}(\mathbf{p})$ of $\mathbf{N}_m$; the inclusion $i \in \mathcal{M}(\mathbf{p})$ means that, *given p-values* $p_1, \ldots, p_m$, $\mathcal{M}$ *rejects the hypothesis* H_i, or equivalently, *the ith p-value is* $\mathcal{M}$*-significant*. Otherwise, the hypothesis H_i is *retained* (not rejected), and the ith p-value is $\mathcal{M}$*-insignificant*.

A multiple testing procedure $\mathcal{M}$ is *symmetric* if for any $\mathbf{p} \in I^m$ and any one-to-one mapping (permutation) $\sigma: \mathbf{N}_m \to \mathbf{N}_m$, denoting by $\sigma(\mathbf{p})$ such $\mathbf{p}' \in I^m$ that $p'_{\sigma(i)} = p_i$ for all i, we have $\mathcal{M}(\sigma(\mathbf{p})) = \sigma(\mathcal{M}(\mathbf{p}))$. This means that if we arbitrarily permute the hypotheses (and their observed p-values), then the procedure $\mathcal{M}$ will reject the same hypotheses as before. Every vector $\mathbf{p} \in I^m$ is a permutation of a (unique) vector $\mathbf{t} \in \mathrm{Simp}^m$, where Simp^m is the m-dimensional simplex

$$\mathrm{Simp}^m = \{\mathbf{t} = (t_1, \ldots, t_m) \in \mathbf{R}^m : 0 \leq t_1 \leq \cdots \leq t_m \leq 1\};$$

therefore, a symmetric MTP is determined uniquely once it is defined on Simp^m.

A procedure $\mathcal{M}$ is *cutting* if, whenever $i \in \mathcal{M}(\mathbf{p})$ and $j \notin \mathcal{M}(\mathbf{p})$, we have $p_i < p_j$, that is, for any p-vector its $\mathcal{M}$-significant components (if any) are smaller than its $\mathcal{M}$-insignificant components (if any).

All MTPs considered below are assumed to be symmetric and cutting. We denote the set of all such procedures by Proc^m, where m is the number of hypotheses being tested.

Remark 1. Note that if $\mathcal{M} \in \mathrm{Proc}^m$, $\mathbf{p} \in I^m$ and $p_i = p_j$ for some i, j $(i \neq j)$, then either both p_i and p_j are $\mathcal{M}$-significant or both are $\mathcal{M}$-insignificant.

Comparison of procedures

Following Liu (1996), we say that a multiple testing procedure $\mathcal{M}'$ *dominates* a procedure $\mathcal{M}$ if for all $\mathbf{p} \in I^m$ we have $\mathcal{M}'(\mathbf{p}) \supset \mathcal{M}(\mathbf{p})$, i.e., $\mathcal{M}'$ rejects all hypotheses H_i rejected by $\mathcal{M}$ (and maybe some others); in this case we write $\mathcal{M}' \succeq \mathcal{M}$.