Published in final edited form as:

Stat Probab Lett. 2008 September 15; 78(13): 1878–1884. doi:10.1016/j.spl.2008.01.055.

Optimality of the Holm procedure among general step-down multiple testing procedures

Alexander Y. Gordon[†] and Peter Salzman^{‡,§}

[†]Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA; E-mail: aygordon@uncc.edu

[‡]Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Ave, Rochester, NY 14642, USA; E-mail: psalzman@bst.rochester.edu

Abstract

We study the class of general *step-down multiple testing procedures*, which contains the usually considered procedures determined by a nondecreasing sequence of thresholds (we call them *threshold step-down*, or *TSD, procedures*) as a parametric subclass. We show that all procedures in this class satisfying the natural condition of *monotonicity* and controlling the family-wise error rate (FWER) at a prescribed level are dominated by one of them – the classical Holm procedure. This generalizes an earlier result pertaining to the subclass of TSD procedures (Lehmann and Romano, *Testing Statistical Hypotheses*, 3rd ed., 2005). We also derive a relation between the levels at which a monotone step-down procedure controls the FWER and the generalized FWER (the probability of k or more false rejections).

Keywords

multiple testing procedure; p -value; step-down procedure; monotone procedure; family-wise error rate; generalized family-wise error rate; Holm procedure

1 Introduction

Multiple testing procedures (MTPs) have many applications. Among major applications are those in biology and medicine, such as DNA microarray experiments, clinical trials with multiple endpoints, functional neuroimaging. Their needs have largely motivated the intensive development of the theory and methods of multiple hypothesis testing in the last decade. For a brief introduction to MTPs in the context of microarray experiments, see Dudoit et al. (2003). More information can be found in Hochberg and Tamhane (1987), Shaffer (1995), Hsu (1996), and Benjamini et al. (2004).

The problem of multiple testing arises when we want to simultaneously test several, say m , hypotheses about the probability distribution from which our data are drawn. We assume that

© 2008 Elsevier B.V. All rights reserved.

Correspondence to: Alexander Y. Gordon.

[§]Research of this author was supported by grant K99LM009477 from the National Library of Medicine

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

associated with each hypothesis H_i is a p -value, P_i , which is a random variable, such that $0 \leq P_i \leq 1$ and, if H_i is true,

$$\text{pr}\{P_i \leq x\} \leq x, \quad 0 \leq x \leq 1. \quad (1)$$

The observed p -value measures the strength of the evidence against the corresponding hypothesis, typically provided by an appropriate test statistic: the smaller the p -value, the stronger the evidence. For a more detailed discussion of p -values, see, e.g., Lehmann and Romano (2005b, pp. 63–64).

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

In the rest of the Introduction we will outline the notions necessary for our main result, as well as the result itself. Rigorous formulations will be given in Section 2 and Section 3.

A *multiple testing procedure*, say $\mathcal{M}$, is a decision rule that, based on the vector of observed p -values (p -vector), selects some hypotheses for rejection; the corresponding p -values are said to be $\mathcal{M}$ -*significant*, or just *significant*, if this causes no ambiguity.¹

We assume that an MTP has two basic properties: it is *cutting* (a significant p -value cannot exceed an insignificant one) and *symmetric* (permuting components of a p -vector does not affect which of them will be found significant).

Our main result pertains to monotone step-down procedures. The natural condition of *monotonicity* means that if some p -values are replaced by smaller ones, the number of rejected hypotheses can only increase. The fact that $\mathcal{M}$ is a *step-down* MTP essentially means: whether or not a p -value is $\mathcal{M}$ -significant depends only on this and smaller p -values.

We consider those monotone step-down procedures for which the probability of rejecting at least one true hypothesis (the *family-wise error rate*, or *FWER*) does not exceed a chosen level α ($0 < \alpha < 1$), whatever the joint distribution of the p -values. Our main result states that all those MTPs are *dominated* by one of them – the classical Holm procedure. Domination means that whenever such a procedure $\mathcal{M}$ rejects some hypotheses, the Holm procedure, based on the same p -values, rejects those hypotheses too.

A particular case of this result pertaining to the algorithmically defined class of what we call *threshold step-down (TSD)* procedures (see Subsection 2.3 below) was obtained earlier by Lehmann and Romano (2005b, Chapter 9).

The remaining part of the paper is organized as follows. In Section 2 we give necessary definitions, mainly following Gordon (2007a) and (Gordon 2007b) (with some changes in terminology). In Section 3 the main result is rigorously formulated and proven. In Section 4 we apply it to the study of generalized family-wise error rates of a monotone step-down procedure.

2 Multiple testing procedures and the FWER

2.1 Multiple testing procedures

A *p -value-only-based multiple testing procedure* (in the sequel – *multiple testing procedure*, or *MTP*) is a Borel measurable mapping $\mathcal{M}: I^m \rightarrow 2^{K_m}$ from the unit cube $I^m = [0, 1]^m$ to the

¹We call MTPs of this type *p -value-only-based* (see Subsection 2.1). There are other types of procedures in the literature, see, e.g., Westfall and Young (1993).

set 2^{K_m} of all subsets of $K_m = \{1, 2, \dots, m\}$. Applying $\mathcal{M}$ to a vector $\mathbf{p} = (p_1, \dots, p_m)$ of observed p -values (p -vector), we obtain a subset $\mathcal{M}(\mathbf{p})$ of K_m ; the inclusion $i \in \mathcal{M}(\mathbf{p})$ means that, given p -values $p_1, \dots, p_m$, $\mathcal{M}$ rejects hypothesis H_i , or equivalently, the i th p -value is $\mathcal{M}$ -significant.

Example 2.1—The classical Bonferroni procedure, which we denote by $Bon f^\alpha$ ($0 < \alpha < 1$), is defined by $Bon f^\alpha(\mathbf{p}) := \{i \in K_m : p_i \leq \alpha/m\}$, $\mathbf{p} \in I^m$.

A multiple testing procedure $\mathcal{M}$ is *symmetric*, if for any p -vector $\mathbf{p}$ and any one-to-one mapping (permutation) $\sigma: K_m \rightarrow K_m$, denoting by $\sigma(\mathbf{p})$ such p -vector $\mathbf{p}'$ that $p'_{\sigma(i)} = p_i$ for all i , we have

$$\mathcal{M}(\sigma(\mathbf{p})) = \sigma(\mathcal{M}(\mathbf{p})). \quad (2)$$

Informally speaking, this means that if we list the hypotheses (and their p -values) in a different order, the procedure $\mathcal{M}$ will reject the same hypotheses as before.

Procedure $\mathcal{M}$ is *cutting*, if it satisfies the following condition: whenever, given p -vector $\mathbf{p}$, p_i is $\mathcal{M}$ -significant and p_j is not, we have $p_i \leq p_j$. In fact, in this case we even have $p_i < p_j$, provided $\mathcal{M}$ is symmetric; indeed, symmetry implies that if $p_i = p_j$, then the corresponding hypotheses H_i and H_j are either both rejected or both accepted. (To see this, use (2) choosing the permutation σ that transposes i and j .)

All MTPs considered below are assumed to be symmetric and cutting.

Procedure $\mathcal{M}$ determines a function $s_{\mathcal{M}}: I^m \rightarrow \mathbf{L}_m = \{0, 1, 2, \dots, m\}$ defined as follows: $s_{\mathcal{M}}(\mathbf{p})$ is the number of rejected hypotheses, given p -vector $\mathbf{p}$ (i.e., $s_{\mathcal{M}}(\mathbf{p}) = \text{Card } \mathcal{M}(\mathbf{p})$). The function $s_{\mathcal{M}}(\cdot)$ is symmetric (permuting components of the p -vector $\mathbf{p}$ does not affect $s_{\mathcal{M}}(\mathbf{p})$, which follows from (2)) and, in turn, allows to uniquely recover procedure $\mathcal{M}$: given $\mathbf{p}$, $\mathcal{M}$ rejects hypotheses with the $s_{\mathcal{M}}(\mathbf{p})$ smallest p -values. Note that the function $s_{\mathcal{M}}(\cdot)$, in view of its symmetry, is completely determined by its values on the simplex

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

the same is true for $\mathcal{M}$. The restriction of $s_{\mathcal{M}}(\cdot)$ to Simp^m can be any Borel measurable function $s(\cdot)$ on Simp^m with the following property:

$$\text{if } \mathbf{t} \in \text{Simp}^m \text{ and } r := s(\mathbf{t}) \text{ satisfies } 1 \leq r \leq m-1, \text{ then } t_r < t_{r+1}.$$

This property is equivalent to the fact that declaring the first $s(\mathbf{t})$ components of $\mathbf{t} \in \text{Simp}^m$ significant will not result in inequalities between significant and insignificant components.

For a more detailed discussion, see Gordon (2007a).

2.2 Comparison of procedures

Let $\mathcal{M}$ and $\mathcal{M}'$ be two multiple testing procedures. Following Liu (1996), we say that procedure $\mathcal{M}'$ *dominates* procedure $\mathcal{M}$, if for any p -vector $\mathbf{p} = (p_1, p_2, \dots, p_m)$ 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 $\mathcal{M}' \geq \mathcal{M}$.

It is easy to see that $\mathcal{M}' \geq \mathcal{M}$ if and only if $s_{\mathcal{M}'}(\mathbf{t}) \geq s_{\mathcal{M}}(\mathbf{t})$ for all $\mathbf{t} \in \text{Simp}^m$.

Note that the partial order $\geq$ between MTPs is *not linear*: both relations $\mathcal{M}' \geq \mathcal{M}$ and $\mathcal{M} \geq \mathcal{M}'$ may be false. Therefore, given an arbitrary class of MTPs, one should not expect that it contains a procedure that dominates every procedure in the class.

2.3 Step-down procedures

A multiple testing procedure $\mathcal{M}$ is a *step-down* procedure if, given $\mathbf{t} \in \text{Simp}^m$, whether or not $i \in \mathcal{M}(\mathbf{t})$ (i.e., $s_{\mathcal{M}}(\mathbf{t}) \geq i$), depends only on $t_1, \dots, t_i$. In other words, it is required that if $\mathbf{t}, \mathbf{t}' \in \text{Simp}^m$, $t_j = t'_j$ for all $j = 1, 2, \dots, i$, and $s_{\mathcal{M}}(\mathbf{t}) \geq i$, then also $s_{\mathcal{M}}(\mathbf{t}') \geq i$.

The *threshold step-down (TSD) procedure generated by* $\mathbf{u} \in \text{Simp}^m$ (notation: $TSD_{\mathbf{u}}$), given a nondecreasing sequence of p -values $\mathbf{t} = (t_1, t_2, \dots, t_m) \in \text{Simp}^m$, declares t_i significant if and only if $t_j \leq u_j$ for all $j = 1, \dots, i$.

Remark 2.1—A TSD procedure is a step-down procedure.

Example 2.2—The Bonferroni procedure $Bon f^{\alpha}$ ($0 < \alpha < 1$) is $TSD_{\mathbf{u}}$ with $u_i = \alpha/m$, $i = 1, \dots, m$.

Example 2.3—The Holm (1979) procedure $Holm^{\alpha}$ ($0 < \alpha < 1$) is $TSD_{\mathbf{u}}$ with $u_i = \alpha/(m - i + 1)$, $i = 1, \dots, m$.

Example 2.4—The procedure introduced by Benjamini and Liu (1999) is $TSD_{\mathbf{u}}$ with

$$u_i = \min \left(1, \frac{mq}{(m - i + 1)^2} \right), \quad 1 \leq i \leq m \quad (0 < q < 1).$$

Given two m -dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, we write $\mathbf{x} \geq \mathbf{y}$ if $x_i \geq y_i$, $i = 1, \dots, m$.

Remark 2.2— $TSD_{\mathbf{u}'} \geq TSD_{\mathbf{u}}$ if and only if $\mathbf{u}' \geq \mathbf{u}$.

Note that the class of step-down procedures is broader than that of TSD procedures.

Example 2.5—Let $m = 2$ and $a, \varepsilon \in (0, 1)$. Let procedure $\mathcal{M}$ act as follows. Given $0 \leq t_1 \leq t_2 \leq 1$, t_1 is $\mathcal{M}$ -significant if and only if $t_1 \leq \varepsilon$; t_2 is $\mathcal{M}$ -significant if and only if $t_1 \leq \varepsilon$ and $at_1 + (1 - a)t_2 \leq \varepsilon$. Then $\mathcal{M}$ is a well-defined MTP (see the end of Subsection 2.1); moreover, $\mathcal{M}$ is a step-down procedure, but not a TSD one.

This example can be modified in many ways and generalized to arbitrary $m > 1$.

2.4 Monotone procedures

We call a multiple testing procedure, $\mathcal{M}$, *monotone*, if for any $\mathbf{p}, \mathbf{p}' \in I^m$ relation $\mathbf{p}' \leq \mathbf{p}$ implies: $s_{\mathcal{M}}(\mathbf{p}') \geq s_{\mathcal{M}}(\mathbf{p})$.

To show that a given procedure is monotone, it suffices to establish the above implication only for pairs of p -vectors that belong to Simp^m , which is usually simpler.

Proposition 1—Procedure $\mathcal{M}$ is monotone, if and only if relation $\mathbf{t}' \leq \mathbf{t}$ ($\mathbf{t}, \mathbf{t}' \in \text{Simp}^m$) implies: $s_{\mathcal{M}}(\mathbf{t}') \geq s_{\mathcal{M}}(\mathbf{t})$.

This statement, in a slightly different formulation, was proven in (Gordon 2007b). It shows that the definition of monotonicity given in that work is equivalent to the definition given above.

Corollary 1—Any TSD procedure is monotone.

Remark 2.3—Procedure $\mathcal{M}$ in Example 2.5 is monotone.

2.5 Family-wise error rate

The family-wise error rate (FWER) introduced by Tukey (1953) measures the probability with which *some* type I errors occur.

Let $\mathcal{M}$ be a multiple testing procedure and P a probability distribution on the unit cube I^m . Let $\mathbf{P} = (P_1, P_2, \dots, P_m)$ be a random vector with distribution P . Denote by T the set (depending on P) of all $i \in \mathbf{K}_m = \{1, 2, \dots, m\}$ for which the i th hypothesis

$$H_i: \text{pr}\{P_i \leq x\} \leq x \text{ for all } x \in [0, 1] \quad (3)$$

is true. Let

$$\text{FWER}(\mathcal{M}, P) := \text{pr}\{\mathcal{M}(\mathbf{P}) \cap T \neq \emptyset\}. \quad (4)$$

In other words, $\text{FWER}(\mathcal{M}, P)$ is the probability for $\mathcal{M}$, given a set of m random values P_i with joint distribution P , to reject at least one true hypothesis H_i .

Procedure $\mathcal{M}$ is said to *control the FWER at level α* , if for all probability distributions P on I^m we have $\text{FWER}(\mathcal{M}, P) \leq \alpha$. This is true, if and only if the quantity

$$\text{FWER}(\mathcal{M}) := \sup_P \text{FWER}(\mathcal{M}, P), \quad (5)$$

where the supremum is taken over all probability distributions P on I^m , does not exceed α .

The quantity (5) may be viewed as the exact level at which $\mathcal{M}$ controls the FWER.

3 Optimality of the Holm procedure

The Holm procedure Holm^α (see Example 2.3), given $\mathbf{t} \in \text{Simp}^m$, finds t_i significant if and only if

$$t_j \leq \frac{\alpha}{m - j + 1} \text{ for all } j, \quad 1 \leq j \leq i.$$

Proposition 2

(Holm 1979). $\text{FWER}(\text{Holm}^\alpha) \leq \alpha$.

Therefore, Holm^α is a monotone step-down procedure (see Remark 2.1 and Corollary 1) controlling the FWER at level α . The following statement – the main result of this work – shows that Holm^α *dominates* any procedure with similar properties.

Theorem 1

Let $\mathcal{M}$ be a monotone step-down multiple testing procedure with

$$\text{FWER}(\mathcal{M}) \leq \alpha < 1. \quad (6)$$

Then $\mathcal{M} \leq \text{Holm}^\alpha$.

Proof

Let $\mathcal{M}$ be a monotone step-down procedure, such that (6) is true. Suppose that, given a nondecreasing sequence of p -values

$$\mathbf{t}=(t_1, \dots, t_m), \quad 0 \leq t_1 \leq t_2 \leq \dots \leq t_m \leq 1, \quad (7)$$

procedure $\mathcal{M}$ finds the k th of them, t_k , significant. How large can the number t_k be?

Set $l := m - k + 1$, $\tau := \min\{t_k, 1/l\}$, and

$$\beta := l\tau, \quad (8)$$

so that $\beta \leq 1$.

Define l random variables

$$P_k, P_{k+1}, \dots, P_m \quad (9)$$

as follows. First, choose at random an integer z equal to one of the numbers $k, k+1, \dots, m, 0$ with probabilities $\tau, \tau, \dots, \tau, 1-\beta$, respectively. Then generate, independently of each other (given z), the random numbers (9), so that if $z \neq 0$, P_z is uniformly distributed on the interval $[0, \tau]$, while the others are uniformly distributed on $[\tau, 1]$; if $z = 0$, all $P_i, i = k, k+1, \dots, m$, are uniformly distributed on $[\tau, 1]$.

It follows from the construction that the distribution of each random variable $P_i, i = k, k+1, \dots, m$, is supported on $[0, 1]$ and has a constant density on each of the two intervals $[0, \tau]$ and $[\tau, 1]$. On the first of them the density equals 1, hence the same is true for the other (because the integral of the density is 1). Therefore,

$$\text{each } P_i, i=k, k+1, \dots, m, \text{ is uniformly distributed on } [0, 1]. \quad (10)$$

Also, we have

$$\min_{k \leq i \leq m} P_i \leq \tau, \quad \text{if } z \neq 0. \quad (11)$$

Furthermore, let $P_1, \dots, P_{k-1}$ be such random variables that $0 \leq P_i \leq t_i, i = 1, \dots, k-1$. These inequalities and (11), where $\tau \leq t_k$, imply that the ordered random variables $P_{(i)}$ satisfy inequalities

$$P_{(i)} \leq t_i, \quad i=1, 2, \dots, k, \quad \text{if } z \neq 0. \quad (12)$$

(We use the following property of nondecreasing rearrangements: given $a_1, \dots, a_m \in \mathbf{R}$ and $c \in \mathbf{R}$, the inequality $a_{(i)} \leq c$ is equivalent to the existence of at least i values of j , such that $a_j \leq c$.)

The random vector $\mathbf{P} = (P_1, \dots, P_m)$, whose distribution on the unit cube $[0, 1]^m$ we denote by P , has the following properties:

- i. According to (10), hypotheses $H_k, H_{k+1}, \dots, H_m$ (see (3)) are true.
- ii. By (12), with probability at least β we have

$$P_{(i)} \leq t_i, \quad i=1, 2, \dots, k. \quad (13)$$

Since the procedure $\mathcal{M}$ found the first k terms t_i in (7) significant, being a *step-down* procedure, $\mathcal{M}$ will still find them significant if the remaining terms $t_{k+1}, t_{k+2}, \dots, t_m$ are replaced by 1's. Because $\mathcal{M}$ is *monotone*, the first k terms of the sequence $(P_{(1)}, \dots, P_{(m)})$ will also be found significant, if (13) holds. According to (ii), this occurs with probability at least β . By (i), no more than $k-1$ hypotheses H_i are false, so by rejecting k or more hypotheses procedure $\mathcal{M}$ commits at least one type I error. We, therefore, have $\text{FWER}(\mathcal{M}, P) \geq \beta$, and consequently $\text{FWER}(\mathcal{M}) \geq \beta$. Comparing this with (6), we see that $\beta \leq \alpha$. In particular, $\beta < 1$, so that recalling (8) we obtain $\tau = t_k$ and

$$t_k \leq \frac{\alpha}{m-k+1}.$$

Because, given p -vector (7), procedure $\mathcal{M}$ finds all $t_j, j=1, \dots, k$, significant, it follows that inequalities

$$t_j \leq \frac{\alpha}{m-j+1}, \quad j=1, \dots, k,$$

are all true; this means that the k th p -value t_k is Holm^α -significant.

We have shown that if, for a given $\mathbf{t} \in \text{Simp}^m$, its k th component t_k is $\mathcal{M}$ -significant, then it is Holm^α -significant as well. Therefore, $\mathcal{M} \leq \text{Holm}^\alpha$.

Corollary 2

(Lehmann and Romano 2005b, Chapter 9) Let $\mathcal{M} = \text{TSD}_{\mathbf{u}}$, where $\mathbf{u} \in \text{Simp}^m$. If $\text{FWER}(\mathcal{M}) \leq \alpha < 1$, then $u_i \leq \alpha/(m-i+1)$, $i=1, \dots, m$.

Proof—A TSD procedure is a step-down procedure; according to Corollary 1, it is monotone. Hence, Theorem 1 applies, and $\mathcal{M} \leq \text{Holm}^\alpha$. In view of Remark 2.2, this implies inequalities between respective thresholds of the two procedures.

The following example shows that the condition that $\mathcal{M}$ is a step-down procedure cannot be removed from Theorem 1.

Example 3.1

Let $\mathcal{M}$ be a procedure that, given a p -vector $\mathbf{t} \in \text{Simp}^m$, rejects all hypotheses, if $t_i \leq \alpha$ for all i , and accepts all hypotheses otherwise. This procedure, as is easy to see, is monotone and controls FWER at level α . Nevertheless, the relation $\mathcal{M} \leq \text{Holm}^\alpha$ is not true: if $t_i = c$ ($\alpha/m < c < \alpha$), $i=1, 2, \dots, m$, then $\mathcal{M}$ rejects all hypotheses, while Holm^α rejects none.

4 Generalized FWER of a step-down procedure

The concept of generalized FWER – the probability of rejecting k or more true hypotheses – was introduced by Victor (1982) and studied in the context of TSD procedures by Hommel and Hoffman (1988), Lehmann and Romano (2005a), and Gordon (2007a). We use notation from the latter work similar to (4) and (5): for a multiple testing procedure $\mathcal{M}$, a probability distribution P on the unit cube I^m , and an integer k ($1 \leq k \leq m$), we set

$$\text{FWER}_k(\mathcal{M}, P) := \text{pr} \{ \text{Card}(\mathcal{M}(\mathbf{P}) \cap T) \geq k \},$$

where $\mathbf{P}$ is a random vector with distribution P and T is the set (depending on P) of all $i \in \{1, 2, \dots, m\}$ for which the hypothesis (3) is true. Furthermore, we set

$$\text{FWER}_k(\mathcal{M}) := \sup_P \text{FWER}_k(\mathcal{M}, P).$$

This quantity may be interpreted as the exact level at which $\mathcal{M}$ controls FWER_k .

Obviously, $\text{FWER}_k(\mathcal{M}) \leq \text{FWER}_1(\mathcal{M}) \equiv \text{FWER}(\mathcal{M})$ for all k ($1 \leq k \leq m$). It turns out that for monotone step-down MTPs this inequality can be significantly strengthened.

Theorem 2

Let $\mathcal{M}$ be a monotone step-down procedure, such that $\text{FWER}(\mathcal{M}) \leq \alpha$ ($\alpha < 1$). Then for all k ($1 \leq k \leq m$)

$$\text{FWER}_k(\mathcal{M}) \leq C_k \alpha, \tag{14}$$

where

$$C_k = \begin{cases} 4k/(k+1)^2, & \text{if } k \text{ is odd;} \\ 4/(k+2), & \text{if } k \text{ is even.} \end{cases} \tag{15}$$

The upper bound $C_k \alpha$ in (14) is sharp.

Proof—By Theorem 1, $\mathcal{M} \leq \text{Holm}^\alpha$, so that $\text{FWER}_k(\mathcal{M}) \leq \text{FWER}_k(\text{Holm}^\alpha)$. It is shown in Gordon (2007a, Example 5.1) that $\text{FWER}_k(\text{Holm}^\alpha) = C_k \alpha$, where C_k is given by (15). This proves (14). The upper bound $C_k \alpha$ cannot be replaced by a smaller one, because it is attained when $\mathcal{M} = \text{Holm}^\alpha$.

Remark 4.1

Note that the sharp upper bound in (14) does not depend on m (the number of hypotheses, $m \geq k$). Also note that $1 = C_1 = C_2 > C_3 > C_4 > \dots$ and that $C_k < 4/k$, the two sides being asymptotically equivalent for large k .

Acknowledgements

We are grateful to Dr. Andrei Yakovlev for valuable discussions. We are also thankful to the anonymous referee whose comments and suggestions helped to improve the presentation.

References

- Benjamini, Y.; Bretz, F.; Sarkar, S., editors. Recent Developments in Multiple Comparison Procedures, Volume 47 of Lecture Notes – Monograph Series. Beathwood, Ohio: Institute of Mathematical Statistics; 2004.
- Benjamini, Y.; Liu, W. A distribution-free multiple test procedure that controls the false discovery rate. Technical report. Department of Statistics and Operations Research, Tel Aviv University; Technical report RP-SOR-99-3
- Dudoit S, Shaffer JP, Boldrich JC. Multiple hypothesis testing in microarray experiments. *Statist. Science* 2003;18:71–103.
- Gordon AY. Explicit formulas for generalized family-wise error rates and unimprovable step-down multiple testing procedures. *J. Statist. Plann* 2007a;Inference 137 (S.N.Roy Centennial Volume): 3497–3512. Available online at <http://www.sciencedirect.com>
- Gordon AY. Unimprovability of the Bonferroni procedure in the class of general step-up multiple testing procedures. *Statist. Prob. Lett* 2007b;77:117–122. available online at <http://www.sciencedirect.com>
- Hochberg, Y.; Tamhane, AC. Multiple Comparison Procedures. New York: Wiley; 1987.
- Holm S. A simple sequentially rejective multiple test procedure. *Scand. J. Statist* 1979;6:65–70.
- Hommel, G.; Hoffman, T. Controlled uncertainty. In: Bauer, P.; Hommel, G.; Sonnemann, E., editors. Multiple Hypothesis Testing. Heidelberg: Springer; 1988. p. 154-161.
- Hsu, JC. Multiple Comparisons: Theory and Methods. New York: Chapman & Hall; 1996.
- Lehmann EL, Romano JP. Generalizations of the familywise error rate. *Ann. Statist* 2005a;33:1138–1154.
- Lehmann, EL.; Romano, JP. Testing statistical hypotheses. Vol. 3rd ed.. New York: Springer; 2005b.
- Liu W. Multiple tests of a non-hierarchical family of hypotheses. *J. Roy. Statist. Soc. Ser. B* 1996;58:455–461.
- Shaffer J. Multiple hypothesis testing: A review. *Annual Review of Psychology* 1995;46:561–584.
- Tukey, JW. The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948–1983. New York: Chapman and Hall; 1953. The problem of multiple comparison. Unpublished manuscript; p. 1-300.
- Victor N. Exploratory data analysis and clinical research. *Methods of Information in Medicine* 1982;21:53–54. [PubMed: 7098979]
- Westfall, PH.; Young, SS. Resampling-based multiple testing: examples and methods for p-value adjustment. New York: Wiley; 1993.