© Board of the Foundation of the Scandinavian Journal of Statistics 1997. Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA. Vol 24: 407–418, 1997

Multiple Hypotheses Testing with Weights

YOAV BENJAMINI and YOSEF HOCHBERG

Tel Aviv University

ABSTRACT. In this paper we offer a multiplicity of approaches and procedures for multiple testing problems with weights. Some rationale for incorporating weights in multiple hypotheses testing are discussed. Various type-I error-rates and different possible formulations are considered, for both the intersection hypothesis testing and the multiple hypotheses testing problems. An optimal per family weighted error-rate controlling procedure *a lá* Spjøtvoll (1972) is obtained. This model serves as a vehicle for demonstrating the different implications of the approaches to weighting. Alternative approaches to that of Holm (1979) for family-wise error-rate control with weights are discussed, one involving an alternative procedure for family-wise error-rate control, and the other involving the control of a weighted family-wise error-rate. Extensions and modifications of the procedures based on Simes (1986) are given. These include a test of the overall intersection hypothesis with general weights, and weighted sequentially rejective procedures for testing the individual hypotheses. The false discovery rate controlling approach and procedure of Benjamini & Hochberg (1995) are extended to allow for different weights.

Key words: control weights, false discovery rate, family-wise error-rate, per-family error-rate, procedural weights, p-values

1. Introduction

Consider testing m (null) hypotheses $H_1, \ldots, H_m$ with corresponding p-values $P_1, \ldots, P_m$. The intersection hypothesis test (IHT) problem is to provide a powerful α level test of the single intersection hypothesis $H_0 = \bigcap_{i=1}^{m} H_i$ based on the separate tests of the m H_i (as e.g. in meta-analysis, testing for carcinogenicity with multiple tumor sites and in testing for a treatment effect with multiple end-points or in multiple subgroups). The aim of the multiple hypotheses test (MHT) is to provide m individual inferences on the H_i with suitable type-I error-rate control. If this control is desired in the strong sense, that is under all combinations of true and false (null) hypotheses, the IHT and the MHT problems become well separated.

To discuss the error rates to control, define $R_i = 1$ if H_i is rejected and 0 otherwise, and define $V_i = 1$ if a true H_i is (erroneously) rejected, and 0 otherwise. The traditional error-rates are: (i) per-family error-rate (PFE) $= E(\Sigma V_i)$ which is the expected number of erroneous rejections, and (ii) family-wise error-rate (FWE) $= P(\Sigma V_i > 0)$, which is the probability of making at least one erroneous rejection.

Recently Benjamini & Hochberg (1995) introduced a third type of error-rate, which is the expected proportion of erroneously rejected hypotheses among the rejected ones: (iii) the false discovery rate (FDR) $= E(\Sigma V_i / \Sigma R_i)$, where $\Sigma V_i / \Sigma R_i$ is defined to be 0 when $\Sigma R_i = 0$. It is easy to see that FDR $\leqslant$ FWE $\leqslant$ PFE. Note also that when all hypotheses are true FDR $=$ FWE.

In both IHT and MHT problems there is often a need for incorporating weights into the respective procedures. For example, in meta-analysis there is often a discussion of the differential quality of the different studies, see e.g. Rosenthal (1984, ch. 3). Holm (1979), who first introduced weights into his sequentially rejective multiple hypotheses testing procedure, described the weights as "positive contants indicating the *importance* of the hypotheses ..." He also implied that the chance of rejecting the corresponding hypothesis increases when a larger weight is assigned to that hypothesis. A similar purpose motivates the practice of allocating unequally the total allowable FWE, in an allocated Bonferroni procedure, among the tested hypotheses according to their importance. Thus each hypothesis with $P_i \leqslant \alpha_i$ is rejected, where $\sum_{i=1}^{m} \alpha_i = \alpha$. Another important example is in clinical trials with multiple end-points—multiple measures of success. A common practice is to divide these end-points into primary and secondary ones. A primary end-point is usually the more convincing measure, say mortality, while the secondary end-points deal with (secondary) aspects of the entire range of possible advantages (or disadvantages) of the suggested treatment.

In these problems weights reflecting differential attitude towards the tested hypotheses can be incorporated in several ways.

(1) Direct modification of the procedure. Such "procedural weights" may have an obvious effect on the power of the procedure, as in an "allocated Bonferroni" procedure, or a not so obvious effect as in the weighted Holm's procedure (see section 3).
(2) Modification of the error-rate under control, by introducing weights which reflect the importance of the different type I errors.
(3) Modifications of the power function to be maximized, which reflect the importance of different rejections. This function may be overall power (i.e. probability of rejecting H_0), sum of individual powers, expected number of rejections, etc.
(4) Modification of both the error-rate controlled and the power function to be maximized. These weights will usually be related.

The weights introduced according to (2), (3), or (4) raise multiple testing problems whose solution may differ from the *ad hoc* ones used in (1).

Spjøtvoll (1972) introduced a model for testing the intersection null hypothesis, where the PFE is controlled and the expected number of rejections is maximized. (In many specific problems the derived procedure is also a multiple hypotheses testing procedure). This model will serve as a vehicle for demonstrating the different implications of taking any of the above approaches to weighting. This will be done in the next section.

No theory seems to exist for optimal FWE controlling procedures (even) in the equal weights case. Therefore we limit our discussion in this case to procedural weights, introducing them into the Bonferroni type procedures which make use of the individual p-values only. In particular, in section 3, we discuss alternative step-down procedures to that of Holm (1979) for FWE control. Simes' (1986) procedure is more powerful than Holm's, but requires that test statistics be independent. Extensions of Simes' procedure to IHT and MHT problems with weights are discussed in section 4.

In section 5 we introduce weights into the FDR criterion, in accordance with approach (2), and we give a procedure that controls the weighted FDR. All of these new procedures still carry over the simplicity associated with their unweighted counterparts.

In this paper we offer a multiplicity of approaches and procedures for multiple testing problems with weights. This variety is the result of recognizing that there are different reasons for, and different implications of, incorporating weights into testing procedures of multiple hypotheses.

2. Controlling the per-family error-rate: introducing weights into Spjøtvoll's model

In this section we denote by H_{0i} and H_{1i} the ith null and alternative hypotheses, respectively. To simplify matters, suppose that under H_{0i} the test statistic X_i is distributed with density f_{0i}, with respect to the measure μ, while under the alternative H_{1i} the density is f_{1i}, $i = 1, 2, \ldots, m$. This is the simplest setting for which we can state Spjøtvoll's (1972) result:

Among all tests of $\bigcap_i H_{0i}$ constructed from m individual tests of H_{0i}, and satisfying $E_{\cap_i H_{0i}}(\sum_{i=1}^m V_i) = \alpha$, the individual likelihood ratio tests with rejection regions of the form $\{x | f_{1i}(x) > cf_{0i}(x)\}$ maximize $E_{\cap H_{1i}}(\sum_{i=1}^m R_i)$.

This model will serve as a vehicle for presenting the four possible ways of introducing weights mentioned before.

1. First note that Spjøtvoll's result does not imply that each individual test is conducted at the same level $\alpha_i = \alpha^*$. If all test statistics X_i are similarly distributed under the H_{0i}, and also similarly distributed under the H_{1i}, then the optimal procedure is to conduct a likelihood ratio test for each hypothesis at the same level α^*. Otherwise, if the distributions are different, since c is the same constant for all m regions, the further f_{1i} is away from f_{0i} the smaller $\alpha_i = P_{H_{0i}}\{x | f_{1i}(x) > cf_{0i}(x)\}$ is. So even though no weighting had been introduced in the formulation, the optimal procedure will involve weights: each H_{0i} should be tested at a separate p-value $\alpha_i = \omega_i \alpha^*$.

No procedural weights with the above form can be expected to remain optimal when other considerations, such as different costs of erroneous decisions or different prior beliefs, are important. These considerations should therefore be introduced into the problem through the weighting of the PFE criterion and the power function maximized. We suggest the following general formulations.

Let a_i and b_i, $i = 1, \ldots, m$ be two sets of positive weights, satisfying $\sum_{i=1}^m a_i = \sum_{i=1}^m b_i = m$.

Maximize

$$E_{\cap H_{1i}}\left(\sum_{i=1}^m b_i R_i\right)$$

so that

$$E_{\cap H_{0i}}\left(\sum_{i=1}^m a_i V_i\right) \leqslant \alpha.$$

Theorem 1

Among all tests satisfying $E_{\cap H_{0i}}(\sum_{i=1}^m a_i V_i) = \alpha$, the test defined by

$$\begin{aligned} \varphi_i^{ab}(x) &= 1 && f_{1i}(x) > c\frac{a_i}{b_i} f_{0i}(x) \\ &= c\frac{c_i}{a_i} && f_{1i}(x) = c\frac{a_i}{b_i} f_{0i}(x) \\ &= 0 && f_{1i}(x) < c\frac{a_i}{b_i} f_{0i}(x) \end{aligned}$$

maximizes $E_{\cap H_{1i}}(\sum_{i=1}^m b_i R_i)$.

Proof. Let $\psi_1 \cdots \psi_m$ be any other such test, satisfying $\sum_{i=1}^m \int a_i f_{0i}(x)\, d\mu(x) = \alpha$. Then

$$\sum_{i=1}^m b_i \int \varphi_i^{ac}(x) f_{1i}(x)\, d\mu(x) - \sum_{i=1}^m b_i \int \psi_i(x) f_{1i}(x)\, d\mu(x)$$

$$= \sum_{i=1}^m \int (\varphi_i^{ac}(x) - \psi_i(x))(b_i f_{1i}(x) - a_i c f_{0i}(x))\, d\mu(x) \geqslant 0$$

2. Using theorem 1 we can view the effect of introducing weights only through the error-rate to be controlled (i.e. setting $b_i \equiv 1$). Here, $\alpha_i = P_{H_{0i}}(\{x|f_{1i}(x) > ca_i f_{0i}(x)\})$ is decreasing in a_i. Note, however, that only when all nulls are similar and all alternatives are similar, then the different α_is given above reflect only the differences in the weights a_i. For example, when testing H_{0i}: $\mu = 0$ vs H_{1i}: $\mu = \mu_1$ with normally distributed test statistics of equal variance,

$$\alpha_i = 1 - \Phi\left(A + \frac{1}{\mu_1}\ln(a_i)\right),$$

where A is determined from $\sum_{i=1}^m \alpha_i = \alpha$. Otherwise the procedural weights should reflect both the weights and the distances between the nulls and the alternatives.

3. In the case where weights are introduced into the maximized power function, but not into the error-rate, the situation is the opposite. The α_is are monotonically related to the b_i, and again are "proportional" only when all testing problems are similar.

Thus using a Bonferroni procedure with reallocation of the error rate according to "importance" fits into this case. Contrary to intuition, in this setting the procedure treats equally the probabilities of type I error.

4. Theorem 1 prescribes what should affect α_i: α_i increases with the importance of rejection and with the distance of the alternative from the null, and it decreases when the importance of controlling its error increases. One important consequence, though, is evident: if we have some statement about the importance of a single hypothesis and therefore would like to treat similarly its false rejection and the power to reject it, i.e. $a_i = b_i$, then we are back at case 1 with weights reflecting the relative distances of the alternatives from the nulls *only.*

Remark. It has been sometimes suggested that the control of the per comparison error-rate (PCE) $= E(\sum V_i)/m$ should be considered in IHT and MHT problems. While the above theorem and discussion apply also to PCE controlling procedures, the mere control of PCE is inadequate for most practical purposes. In the current framework, suppose that one of the hypotheses is extremely important, as reflected by a large weight (a_i). If m is large enough, a weighted PCE controlling procedure is to always reject that hypothesis and retain all other ones, no matter what the observed values are.

In the next two sections we mostly discuss various weighted procedures designed to control the unweighted FWE. These procedures are not known to be optimal but some comparisons of their appropriateness and operating characteristics are given.

3. Controlling the family-wise error-rate: on Holm's (1979) and alternative procedures

When FWE control in the strong sense is desired, Holm's (1979) weighted procedure (WHP) can be used. Let $P_i^* = P_i/w_i$ and order $P_{(1)}^* \leqslant \cdots \leqslant P_{(m)}^*$. Let $H_{(i)}^*$, $w_{(i)}^*$, correspond to $P_{(i)}^*$. Reject $H_{(i)}^*$ when

$$P_{(j)}^* \leqslant \frac{\alpha}{\sum_{k=j}^{m} w_{(k)}^*}, \quad j = 1, \ldots, i. \tag{3.1}$$

Holm's unweighted procedure is the above procedure with all weights being equal (to 1). If the weighted procedure called for rejecting H_i when $P_i^* = P_i/w_i$ is smaller than a constant, then a larger w_i implies greater power for rejecting that hypothesis. However, because of the ordering and the use of (3.1) it is not obvious that larger w_i always gives a higher power to reject the ith hypothesis in the WHP. Consider $m = 2$ and independent $P_i \sim U[0, \delta_i]$, $\delta_i < 1$ $i = 1, 2$. The unweighted Holm procedure gives higher power for rejecting H_2 than a weighted Holm procedure with weights w_1 and w_2, where $w_1 + w_2 = 2$, if

$$2 - 2w_1 + (w_1^2 - 1)\frac{\alpha}{\delta_1} < 0. \tag{3.2}$$

When $w_1 < 1$ and for sufficiently small values of δ_1 condition (3.2) holds.

An alternative procedure is based on the ordered P_i, $P_{(1)} \leqslant \cdots \leqslant P_{(m)}$. Let $H_{(i)}$, $w_{(i)}$, correspond to $P_{(i)}$. Reject $H_{(i)}$ as long as

$$P_{(j)} \leqslant \frac{w_{(j)}}{\sum_{k=j}^{m} w_{(k)}} \cdot \alpha, \quad j = 1, \ldots, i. \tag{3.3}$$

Theorem 2
The procedure based on (3.3) *controls the FWE, in the strong sense, at* α.

Proof. Let $I \subset \{1, \ldots, m\}$ with cardinality $c(I)$ be the index set of the true null hypotheses. By letting $P_j = 0 \forall j \notin I$ with probability 1, the FWE is increased. Let $i' = m - c(I) + 1$, so $P_{(i')}$ is the smallest of the p-values corresponding to the true null hypotheses. In order for some true $H_{(i)}$ to be rejected, $H_{(i')}$ has to be rejected first. Therefore, the FWE is given by

$$P\left[P_{(i')} \leqslant \frac{w_{(i')}}{\sum_{k=i'}^{m} w_{(k)}} \alpha\right] \leqslant \sum_{i \in I} P\left[P_i \leqslant \frac{w_i}{\sum_{k \in I} w_k}\right] = \alpha$$

When $m = 2$ the WHP is preferable because its rejection regions contain those of the weighted alternative procedure (WAP). Specifically, the regions of the two procedures are identical except for the region: $w_2\alpha/2 \leqslant p_1 \leqslant w_1\alpha/2$; $w_2\alpha/2 \leqslant p_2 \leqslant p_1$ (where $w_1 > w_2$ and $w_1 + w_2 = 2$) where WHP rejects both hypotheses and WAP rejects none. This superiority however does not extend to higher dimensions as the following simple example shows: take $m = 3$, with $w_i = i$ for $\{i = 1, 2, 3\}$, and suppose that $p_1 = 0.05$, $p_2 = 0.06$, $p_3 = 0.30$. The WHP compares sequentially p_2, p_1 and p_3 with $\alpha/3$, $\alpha/4$ and α while WAP compares sequentially p_1, p_2 and p_3 with $\alpha/6$, $2\alpha/5$ and α, respectively. If $\alpha/3 < p_2 < 2\alpha/5$ then WHP does not reject any hypothesis but WAP rejects H_1 and H_2 if $p_1 < \alpha/6$.

Note that whenever the ordering of the P_is is reversed to that of the w_is then WHP will reject all hypotheses rejected by WAP and may reject some more. This favourable circumstance to the WHP is associated with situations when we give larger weights to larger deviations from null hypothesis and the ordered p-values later obtained reflect good judgement on our part. In other situations, where we are less certain of our prior knowledge one *cannot expect* better performance by the WHP (as the simple example above shows). However, the WHP is monotonous (in the sense that if one or more p-values become smaller at least the same or even more null hypotheses would be rejected) while the WAP does not satisfy this property. This can be seen from the last example mentioned above (for $m = 2$): on letting $P_1 = w_1\alpha/2 - 2\varepsilon$, $P_2 = w_1\alpha/2 - \varepsilon$, with ε sufficiently small. Then $P_{(1)} = P_1 \leqslant w_1/2\alpha$, $P_{(2)} = P_2 \leqslant \alpha$, and the WAP rejects both hypotheses. On the other hand, if one has p-values $Q_1 = w_2\alpha/2 + 2\varepsilon$, $Q_2 = w_2\alpha/2 + \varepsilon$, then $Q_1 < P_1$, $Q_2 < P_2$ but (because $Q_{(1)} = Q_2 > w_2/2\alpha$) no hypothesis is rejected. Therefore, we do not recommend the WAP.

The above procedures control the unweighted FWE. This is a natural extension even when we start with a "weighted version" of $P[\Sigma V_i > 0]$ i.e. $P[\Sigma a_i V_i > 0]$ since the latter is equal to the former for any set of non-zero weights. However, since the other natural definition of the unweighted FWE is $P[\Sigma V_i \geqslant 1]$, we obtain the alternative weighted version of the FWE as $\text{WFWE} = P[\Sigma a_i V_i \geqslant 1]$. Now WFWE $\neq$ FWE and the effect of weights can be very dramatic on the suitable procedure. Consider single-step procedures of the form:

Reject H_i when $P_i \leqslant w_i \alpha$.

Example. Take $m = 2$, $a_1 = \varepsilon$, $a_2 = 2 - \varepsilon$ then $\Sigma a_i V_i \geqslant 1 \Leftrightarrow V_2 = 1$. It then follows that the best procedure is to always reject H_1 (by choosing $w_1 \geqslant 1/\alpha$) and reject H_2 if $P_2 \leqslant \alpha$ (by choosing $w_2 = 1$).

Another problem associated with the WFWE is that even for very different choices of the a_i the criterion may be identical, and it may differ substantially for very similar values. Therefore, we do not recommend the WFWE.

4. Controlling the FWE for independent test statistics: weighted Simes type procedures

When the P_i are independent, Simes' (1986) test of the intersection null hypothesis H_0, which rejects H_0 if for at least one j

$$P_{(j)} \leqslant \frac{j}{m}\alpha, \quad j = 1, \ldots, m, \tag{4.1}$$

has an exact level α.

This test is more powerful than the Bonferroni test from which Holm's test is derived. Hence, for independent test statistics, the sequentially rejective procedures derived from Simes' test and the closure method of Marcus *et al.* (1976), such as Hochberg (1988), are more powerful than Holm's procedure.

4.1. The intersection hypothesis test problem

When allotment of weights is desired to enhance the testing of H_0 based on the H_i, then an extension of Simes' result is required. Hochberg & Liberman (1993) provided such an extension. Their procedure rejects H_0 if for some j

$$P^*_{(j)} \leqslant \frac{j}{m}\alpha, \quad j = 1, \ldots, m, \tag{4.2}$$

where the $P^*_{(j)}$ are the ordered values of P_j/w_j, $\sum_{j=1}^m w_j = m$ and $\max(w_j) \leqslant 1/\alpha$.

Following the alternative considered in section 3, we may suggest here the following alternative procedure to that of Hochberg & Liberman.

Reject H_0 when for some j

$$P_{(j)} \leqslant \frac{w_{(j)}}{m} j\alpha. \tag{4.3}$$

The problem with this suggestion is that the constraints on the w_is are too strong, and moreover, the weights depend on α. Consequently, this approach is dropped from further consideration.

A different alternative, much simpler and with some good operating characteristics, follows. Let $\sum_{j=1}^m w_j = m$ and consider the procedure:

Reject H_0 when for some j

$$P_{(j)} \leqslant \frac{\sum_{k=1}^{j} w_{(k)}}{m} \alpha. \tag{4.4}$$

Theorem 3

The extended Simes type test based on (4.4) *controls the type I error probability at α, for independent test statistics.*

Proof. By induction. For $m = 1$ it is obviously true. Assume it is true for $m - 1$ and any weights which sum up to $m - 1$ and show it implies that it is true for m. We want to show that

$$P\left[P_{(j)} > \frac{\sum_{1}^{j} w_{(k)}}{m} \alpha \;\; j = 1, \ldots, m\right] = 1 - \alpha. \tag{4.5}$$

Conditions on $P_{(m)}$, integrating with respect to its density, and summing over the m possibilities we get

$$\frac{1}{m}\sum_{i=1}^{m}\int_{p=\alpha}^{1} P\left[\frac{P_{(j)}}{p} > \frac{\alpha}{pm}\frac{\sum_{1}^{j} w_{(k)}}{m - w_i}\,\frac{(m - w_i)(m-1)}{m-1}, \;\; j = 1, \ldots, m-1\right] p^{m-1}\, dp$$

$$= \frac{1}{m}\sum_{i=1}^{m}\int_{p=\alpha}^{1} P\left[\frac{P_{(j)}}{p} > \frac{\sum_{k=1}^{j} w'_{(k)}}{m-1}\alpha'_i, \;\; j \neq m\right] p^{m-1}\, dp \tag{4.6}$$

$$w'_{(k)} = \frac{w_{(k)}(m-1)}{m - w_i}, \qquad \alpha'_i = \frac{\alpha}{p}\frac{m - w_i}{m}.$$

the $P_{(j)}/p$, $j = 1, \ldots, m-1$ are distributed like order statistics of $m - 1$ independent $U[0, 1]$ variables and by the induction hypothesis the probability within the integral in (4.6) is equal to $1 - \alpha_i$ and we get

$$= \frac{1}{m}\sum_{i=1}^{m}\int_{\alpha}^{1}\left[1 - \frac{\alpha}{p}\frac{m - w_i}{m}\right] p^{m-1}\, dp = 1 - \alpha.$$

A thorough investigation of the powers of this procedure and the one in Hochberg & Liberman is not undertaken here. Note that if the weighting turned out to be successful (in the manner discussed in section 3) i.e. the p-values observed are ordered in the same way as the weights, then the weighted procedure introduced here gives better results than the procedure in Hochberg & Liberman (1993). In the first procedure we compare $P_{(i)}$ to

$$\frac{\sum_{k=1}^{i} w_k}{i}\,\frac{i}{m}\alpha$$

while in the latter procedure we compare $P_{(i)}$ to

$$w_i \frac{i}{m}\alpha$$

and in this case the former values are larger.

4.2. Multiple hypotheses testing problems

Any weighted Simes procedure can be extended in principle to a multiple hypotheses testing procedure, by the use of the closure method. However, Hochberg and Liberman's extension and the use of the closure does not imply that Hommel's (1988) and Hochberg's (1988) procedures can be used with P_i^*s instead of P_is. Their extension as well as the extension of Rom's (1990) procedure, however, will require special constants for each new set of w_is. The other extended Simes procedures are also not easily amenable to sequential procedures. If $m - i + 1$ in Hochberg's original procedure is replaced by $\sum_{h=i}^{m} w_{(h)}$ then the resulting procedure does not control the FWE. To see this let the first $m - 1$ P_is be zero with probability 1 and $P_m \sim U[0, 1]$. FWE $= P[P_m \leqslant \alpha/w_m] > \alpha$ if $w_m < 1$.

5. Weighted false discovery rate control for independent test statistics

For ease of notation assume that the first $0 \leqslant m_0 \leqslant m$ hypotheses tested are in fact true and $m_1 = m - m_0$ are false. The false discovery rate (FDR) is

$$E(Q) = E\left(\frac{\sum_{i=1}^{m_0} V_i}{\sum_{i=1}^{m} R_i}\right) \tag{5.1}$$

which is the expected proportion of the falsely rejected hypotheses among the rejected ones. When weighting is desired, the FDR can be generalized as follows.

Definition
Let $Q(w)$ be

$$\begin{aligned} Q(w) &= \frac{\sum_{i=1}^{m_0} w_i V_i}{\sum_{i=1}^{m} w_i R_i} \qquad \sum_{i=1}^{m} w_i R_i > 0 \\ &= 0 \qquad\qquad \text{otherwise} \end{aligned} \tag{5.2}$$

then the *weighted false discovery rate* (*WFDR*) is defined to be $E(Q(w))$.

Note that if some of the weights are 0, and the others are all equal, then the WFDR is identical to the FDR for the limited problem of testing the positively weighted hypotheses. Under the intersection null hypothesis that all tested hypotheses are true, WFDR = WFWE. As in Benjamini & Hochberg (1995), it is again easy to show that WFDR $\leqslant$ WFWE, so a WFDR controlling procedure is potentially more powerful than a WFWE controlling procedure.

Consider now the following procedure:

Let k be the largest j satisfying

$$P_{(j)} \leqslant \frac{\sum_{i=1}^{j} w_{(i)}}{m} q^*, \tag{5.3}$$

then reject $H_{(1)} \cdots H_{(k)}$.

Theorem 4
For independent test statistics the procedure based on (5.3) *controls the WFDR at level* q^*.

The proof of theorem 4 is immediate from the following lemma proved in the appendix.

Lemma

For any $1 \leqslant m_0 \leqslant m$ *independent p-values corresponding to the true null hypotheses, any set of values that the* $m_1 = m - m_0$ *p-values corresponding to the false null hypotheses take, any set of weights* $w_i \geqslant 0$, $\sum_{i=1}^{m} w_i = m$, *and any constant* q^*, *the multiple testing procedure defined by* (5.3) *satisfies the inequality*

$$E(Q(w) | P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leqslant \frac{\sum_{i=1}^{m_0} w_i}{m} q^*. \tag{5.4}$$

In this procedure, the weights incorporated into the error rate are suitably accumulated to form the procedural weights. It is important to reject an hypothesis with high weight, as it considerably increases the "weight" of the total discoveries. Yet it also increases the weight of the errors. Essentially we are using approach (4) of incorporating the same weights into the loss from errors ($\Sigma w_i V_i$) and the gain from rejections $\Sigma w_i R_i$ as in Spjøtvoll's model. The two are combined differently in the WFDR approach, and the procedure is also very different. One important difference is that no change was needed in Spjøtvoll's procedure when using the same set of weights, while a change in the procedure is needed when controlling the WFDR instead of the FDR.

6. Conclusion

In several areas of applications there are identified needs for weighted multiple comparison analysis. Two prominent examples are meta-analysis and multiple end-points analysis in clinical trials. In meta-analysis there is need to weight the different studies according to their quality or sample size etc. In clinical trials with multiple end-points there is often a need to treat various end-points such as primary vs secondary differently.

In this paper we attempted new and extended approaches to the general problem. First we examined a variety of formulations for different error-rates. These formulations allow assignments of weights in response to the specific requirements of the problem at hand: "Although the assignment of weights ... is subjective, the weights allow the experimenter to include economic and ethical considerations in the data analysis" (Westfall & Young, 1992, ch. 6).

Second, alternative procedures were derived for some of the weighted multiple comparison problems and formulations which have been raised. Some comparisons were made and as a result, some procedures were eliminated.

Additional work is required on the power of the different procedures for different systems of weights. Also, comparisons with other procedures which were published in the literature are necessary. In particular one should consider Rüger (1978) weighted IHT along with its extensions to weighted MHTs. These include Hommel (1986) and a procedure for logically related improvements of Shaffer's (1986) hypotheses using weights, which were proposed by Bergmann & Hommel (1988).

Acknowledgements

We wish to thank G. Hommel and the referees for their helpful comments.

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a new and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 1289–1300.

Bergmann, B. & Hommel, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses. In: *Multiple hypothesenprüfungen—multiple hypotheses testing* (ed. P. Bauer, G. Hommel & E. Sonnemann) 100–115. Springer, Berlin, Heidelberg, New York.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–803.

Hochberg, Y. & Liberman, U. (1994). An extended Simes test. *Statist. Probab. Lett.* **21**, 101–105.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

Hommel, G. (1986). Multiple test procedures for arbitrary dependences structures. *Metrika* **33**, 321–336.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386.

Marcus, R., Peritz, E. & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663–665.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Sage, Menlo Park, CA.

Rüger, B. (1978). Das maximale Signifikanzniveau des Tests “Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablenung führen”. *Metrika* **25**, 171–178.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81**, 826–831.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Statist.* **43**, 398–411.

Westfall, P. H. & Young, S. S. (1992). *Resampling-based multiple testing*, Wiley, New York.

Received February 1995, in final form May 1996

Yoav Benjamini, Department of Statistics, School of Mathematical Sciences, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

Appendix

Proof of the lemma. The case $m = 1$ being immediate, we proceed by induction assuming that the lemma holds for any $m' \leqslant m$. If $m_0 = 0$: All null hypotheses are false, Q is identically 0, and

$$E(Q(w)|P_1 = p_1, \ldots, P_{m+1} = p_{m+1}) = 0 \leqslant \frac{\sum_{i=1}^{m_0} w_i}{m+1} q^*. \tag{1}$$

If $m_0 > 0$: Denote by $P'_{(m_0)}$ the p-value corresponding to the largest p-value among $P_1 \cdots P_{m_0}$. Since these correspond to the true null hypotheses, $P'_{(m_0)}$ is distributed as the largest of m_0 independent $U(0, 1)$ random variables, and $f_{P_{(m_0)}}(p) = m_0 p^{m_0 - 1}$ for $0 \leqslant p \leqslant 1$. Let us also define j_0 to be the largest j, $m_0 + 1 \leqslant j \leqslant m + 1$, satisfying

$$p_j \leqslant \frac{\sum_{i=1}^{j} w_i}{m+1} q^*. \tag{2}$$

Denote by p'' the value of the right side of (2) at j_0. Conditioning on $P'_{(m_0)} = p$,

$$E(Q(w)|P_{m_0+1} = p_1, \ldots, P_{m+1} = p_{m_1})$$

$$= \int_0^{p''} E(Q(w)|P'_{(m_0)} = p,\ P_{m_0+1} = p_1, \ldots, P_{m+1} = p_{m_1}) f_{P'_{(m_0)}}(p)\,dp \tag{3}$$

$$+ \int_{p''}^1 E(Q(w)|P'_{(m_0)} = p,\ P_{m_0+1} = p_1, \ldots, P_{m+1} = p_{m_1}) f_{P'_{(m_0)}}(p)\,dp. \tag{4}$$

For $p \leqslant p''$ all j_0 hypotheses are rejected, and

$$Q(w) = \frac{\sum_{i=1}^{m_0} w_i}{\sum_{i=1}^{j_0} w_i}. \tag{5}$$

Evaluating (3) we get

$$\frac{\sum_{i=1}^{m_0} w_i}{\sum_{i=1}^{j_0} w_i}(p'')^{m_0} = \frac{\sum_{i=1}^{m_0} w_i}{\sum_{i=1}^{j_0} w_i}\frac{\sum_{i=1}^{j_0} w_i}{(m+1)} q^*(p'')^{m_0-1} = \frac{\sum_{i=1}^{m_0} w_i}{m+1} q^*(p'')^{m_0-1}. \tag{6}$$

In order to evaluate (4) we further condition on the P-value at which $P'_{(m_0)}$ is achieved indexed by i', $1 \leqslant i' \leqslant m_0$.

$$\int_{p''}^1 E(Q(w)|P'_{(m_0)} = p,\ P'_{(m_0)} = P_i,\ P_{m_0+1} = p_1, \ldots, P_{m+1} = p_{m_1}) f_{P'_{(m_0)}}(p)\,dp$$

$$= \frac{1}{m_0}\sum_{i'=1}^{m_0}\int_{p''}^1 E(Q(w)|P'_{(m_0)} = p,\ P'_{(m_0)} = P_{i'},\ P_{m_0+1} = p_1, \ldots, P_{m+1} = p_{m_1}) f_{P'_{(m_0)}}(p)\,dp \tag{7}$$

Consider each case: $P_j \leqslant P'_{(m_0)} = P_{i'} = p < p_{j+1}$ for $j > j_0$, or $p'' < P'_{(m_0)} = P_{i'} = p < p_{j_0+1}$. From the definition of j_0 and p'', no hypothesis can be rejected because of the value of p, $p_{j+1}, \ldots, p_{m+1}$. Therefore, when all hypotheses—true and false—are considered together and their p-values are ordered, a hypothesis $H_{(i)}$ can be rejected only if there exists a k, $i \leqslant k \leqslant j-1$, for which

$$\frac{P_{(k)}}{p} \leqslant \frac{\sum_{i=1}^{k} w_{(i)}}{(m+1)p} q^*. \tag{8}$$

Equivalently, $H_{(i)}$ will be rejected if this k satisfies

$$\frac{P_{(k)}}{p} \leqslant \frac{(j-1)\sum_{i=1}^{k} w_{(i)}}{(j-1)\sum_{i=1}^{j-1} w_{(i)}}\left(\frac{\sum_{i=1}^{j-1} w_i}{(m+1)p} q^*\right) = \frac{\sum_{i=1}^{k} w^*_{(i)}}{(j-1)}\left(\frac{\sum_{i=1}^{j-1} w_i}{(m+1)p} q^*\right) \tag{9}$$

with

$$w^*_{(i)} = \frac{(j-1)}{\sum_{i=1}^{j-1} w_{(i)}} w_i. \tag{10}$$

These weights satisfy $w_i^* \geqslant 0$, and $\sum_{i=1}^{j-1} w_i^* = j-1$; since we conditioned on $P_{i'} = P'_{(m_0)} = p$, the $m_0 - 1$ other P_i/p are distributed as independent $U(0, 1)$ random variables; p_i/p for $i = m_0+1, \ldots, j$ are numbers between (0, 1) corresponding to false null hypotheses. Hence using (5.3) to test the $j-1 = m' \leqslant m$ hypotheses, of which $m_0 - 1$ are true, is equivalent to using the procedure with the constant

$$\frac{\sum_{i=1}^{j-1} w_i}{(m+1)p} q^* \tag{11}$$

taking the role of q^*.

Now applying the induction hypothesis we have that

$$E(Q(w)|P'_{(m_0)} = p,\, P'_{(m_0)} = P_{i'},\, P_{m_0+1} = p_1, \ldots, P_{m+1} = p_{m_1})$$

$$\leqslant \frac{\sum_{\substack{i=1\\ i\neq i'}}^{m_0} w_i^*}{(j-1)} \left(\frac{\sum_{i=1}^{j-1} w_i^*}{(m+1)p} q^* \right) \tag{12}$$

$$= \frac{\sum_{i=1}^{m_0} w_i - w_{i'}}{(m+1)p} q^* \tag{13}$$

where (13) is derived after replacing w_i^* with their definition in (10).

The bound in (13) depends on p, but not on the segment j $p_j < p < p_{j+1}$ for which it was evaluated (recall that for j_0 the range for p is $p'' < p \leqslant p_{j_0+1}$). It does depend on i'. Therefore, integrating (13) over $(p'', 1]$, while still conditioning on i' we get

$$\int_{p''}^{1} \frac{\sum_{i=1}^{m_0} w_i - w_{i'}}{(m+1)p} q^* m_0 p^{m_0-1}\, dp = \frac{m_0}{m_0-1} \frac{\left(\sum_{i=1}^{m_0} w_i - w_{i'} \right)}{(m+1)} q^*(1 - (p'')^{m_0-1}). \tag{14}$$

Averaging now over i' we get from (7) and (14)

$$\int_{p''}^{1} E(Q(w)|P_{m_0+1} = p_1, \ldots, P_{m+1} = p_{m_1}) f_{P'_{(m_0)}}(p)\, dp$$

$$= \frac{1}{m_0} \sum_{i'=1}^{m_0} \frac{m_0}{m_0-1} \frac{\left(\sum_{i=1}^{m_0} w_i - w_{i'} \right)}{(m+1)} q^*(1-(p'')^{m_0-1}) = \frac{\sum_{i=1}^{m_0} w_i}{(m+1)} q^*(1-(p'')^{m_0-1}). \tag{15}$$

Finally adding (15) and (6) we get the desired inequality (1) for $m+1$.

Remark. Theorem 3 follows from this lemma as well, by letting $m = m_0$ and recalling that in this case FWE = FDR. Since the proof of the special case is considerably less complicated, it was given separately.