The Annals of Statistics
2001, Vol. 29, No. 4, 1165–1188

THE CONTROL OF THE FALSE DISCOVERY RATE IN MULTIPLE TESTING UNDER DEPENDENCY

BY YOAV BENJAMINI[1] AND DANIEL YEKUTIELI[2]

Tel Aviv University

Benjamini and Hochberg suggest that the false discovery rate may be the appropriate error rate to control in many applied multiple testing problems. A simple procedure was given there as an FDR controlling procedure for independent test statistics and was shown to be much more powerful than comparable procedures which control the traditional familywise error rate. We prove that this same procedure also controls the false discovery rate when the test statistics have positive regression dependency on each of the test statistics corresponding to the true null hypotheses. This condition for positive dependency is general enough to cover many problems of practical interest, including the comparisons of many treatments with a single control, multivariate normal test statistics with positive correlation matrix and multivariate t. Furthermore, the test statistics may be discrete, and the tested hypotheses composite without posing special difficulties. For all other forms of dependency, a simple conservative modification of the procedure controls the false discovery rate. Thus the range of problems for which a procedure with proven FDR control can be offered is greatly increased.

1. Introduction.

1.1. *Simultaneous hypotheses testing.* The control of the increased type I error when testing simultaneously a family of hypotheses is a central issue in the area of multiple comparisons. Rarely are we interested only in whether all hypotheses are jointly true or not, which is the test of the intersection null hypothesis. In most applications, we infer about the individual hypotheses, realizing that some of the tested hypotheses are usually true—we hope not all—and some are not. We wish to decide which ones are not true, indicating (statistical) discoveries. An important such problem is that of multiple endpoints in a clinical trial: a new treatment is compared with an existing one in terms of a large number of potential benefits (endpoints).

EXAMPLE 1.1 (Multiple endpoints in clinical trials). As a typical example, consider the double-blind controlled trial of oral clodronate in patients with bone metastases from breast cancer, reported in Paterson, Powles, Kanis, McCloskey, Hanson and Ashley (1993). Eighteen endpoints were compared

Received February 1998; revised April 2001.

[1]Supported by FIRST foundation of the Israeli Academy of Sciences and Humanities.

[2]This article is a part of the author's Ph.D. dissertation at Tel Aviv University, under the guidance of Yoav Benjamini.

AMS 2000 *subject classifications*. 62J15, 62G30, 47N30.

Key words and phrases. Multiple comparisons procedures, FDR, Simes' equality, Hochberg's procedure, MTP_2 densities, positive regression dependency, unidimensional latent variables, discrete test statistics, multiple endpoints many-to-one comparisons, comparisons with control.

between the treatment and the control groups. These endpoints included, among others, the number of patients developing hypercalcemia, the number of episodes, the time the episodes first appeared, number of fractures and morbidity. As is clear from the condensed information in the abstract, the researchers were interested in all 18 particular potential benefits of the treatment.

The traditional concern in such multiple hypotheses testing problems has been about controlling the probability of erroneously rejecting even one of the true null hypotheses, the familywise error-rate (FWE). Books by Hochberg and Tamhane (1987), Westfall and Young (1993), Hsu (1996) and the review by Tamhane (1996) all reflect this tradition. The control of the FWE at some level α requires each of the individual m tests to be conducted at lower levels, as in the Bonferroni procedure where α is divided by the number of tests performed.

The Bonferroni procedure is just an example, as more powerful FWE controlling procedures are currently available for many multiple testing problems. Many of the newer procedures are as flexible as the Bonferroni, making use of the p-values only, and a common thread is their stepwise nature (see recent reviews by Tamhane (1996), Shaffer (1995) and Hsu (1996)). Still, the power to detect a specific hypothesis while controlling the FWE is greatly reduced when the number of hypotheses in the family increases, the newer procedures notwithstanding. The incurred loss of power even in medium size problems has led many practitioners to neglect multiplicity control altogether.

EXAMPLE 1.1 (Continued). Paterson et al. (1993) summarize their results in the abstract as follows:

> In patients who received clodronate, there was a significant reduction compared with placebo in the total number of hypercalcemic episodes (28 v 52; $p \leq .01$), in the number of terminal hypercalcemic episodes (7 v 17; $p \leq .05$), in the incidence of vertebral fractures (84 v 124 per 100 patient-years; $p \leq .025$), and in the rate of vertebral deformity (168 v 252 per 100 patient-years; $p \leq .001$)...

All six p-values less than 0.05 are reported as significant findings. No adjustment for multiplicity was tried nor even a concern voiced.

While almost mandatory in psychological research, most medical journals do not require the analysis of the multiplicity effect on the statistical conclusions, a notable exception being the leading *New England Journal of Medicine*. In genetics research, the need for multiplicity control has been recognized as one of the fundamental questions, especially since entire genome scans are now common [see Lander and Botstein (1989), Barinaga (1994), Lander and Kruglyak (1995), Weller, Song, Heyen, Lewin and Ron (1998)]. The appropriate balance between lack of type I error control and low power ["the choice

between Scylla and Charybdis" in Lander and Kruglyak (1995)] has been heavily debated.

1.2. *The false discovery rate.* The false discovery rate (FDR), suggested by Benjamini and Hochberg (1995) is a new and different point of view for how the errors in multiple testing could be considered. The FDR is the expected proportion of erroneous rejections among all rejections. If all tested hypotheses are true, controlling the FDR controls the traditional FWE. But when many of the tested hypotheses are rejected, indicating that many hypotheses are not true, the error from a single erroneous rejection is not always as crucial for drawing conclusions from the family tested, and the proportion of errors is controlled instead. Thus we are ready to bear with more errors when many hypotheses are rejected, but with less when fewer are rejected. (This frequentist goal has a Bayesian flavor.) In many applied problems it has been argued that the control of the FDR at some specified level is the more appropriate response to the multiplicity concern: examples are given in Section 2.1 and discussed in Section 4.

The practical difference between the two approaches is neither trivial nor small and the larger the problem the more dramatic the difference is. Let us demonstrate this point by comparing two specific procedures, as applied to Example 1.1. To fix notation, let us assume that of the m hypotheses tested $\{H_1^0, H_2^0, \ldots, H_m^0\}$, m_0 are true null hypotheses, the number and identity of which are unknown. The other $m - m_0$ hypotheses are false. Denote the corresponding random vector of test statistics $\{X_1, X_2, \ldots, X_m\}$, and the corresponding p-values (observed significance levels) by $\{P_1, P_2, \ldots, P_m\}$ where $P_i = 1 - F_{H_i^0}(X_i)$.

Benjamini and Hochberg (1995) showed that when the test statistics are independent the following procedure controls the FDR at level $q \cdot m_0/m \leq q$.

THE BENJAMINI HOCHBERG PROCEDURE. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered observed p-values. Define

$$k = \max\left\{i\colon p_{(i)} \leq \frac{i}{m}q\right\}, \tag{1}$$

and reject $H_{(1)}^0 \cdots H_{(k)}^0$. If no such i exists, reject no hypothesis.

In the case that all tested hypotheses are true, that is, when $m_0 = m$, this theorem reduces to Simes' global test of the intersection hypothesis proved first by Seeger (1968) and then independently by Simes (1986). However, when $m_0 < m$ the procedure does not control the FWE. To achieve FWE control, Hochberg (1988) constructed a procedure from the global test, which has the same stepwise structure but each $P_{(i)}$ is compared to $\frac{q}{m-i+1}$ instead of $\frac{iq}{m}$. The constants for the two procedures are the same at $i = 1$ and $i = m$ but elsewhere the FDR controlling constants are larger.

EXAMPLE 1.1 (Continued). Compare the two procedures conducted at the 0.05 level in the multiple endpoint example. Hochberg's FWE controlling pro-

cedure rejects the two hypotheses with p-values less than 0.001, just as the Bonferroni procedure does. The FDR controlling procedure rejects the four hypotheses with p-values less than 0.01. In this study the ninth p-value is compared with 0.005 if FWE control is required, with 0.025 if FDR control is desired.

More details about the concept and procedures, other connections and historical references are discussed in Section 2.2.

1.3. *The problem.* When trying to use the FDR approach in practice, dependent test statistics are encountered more often than independent ones, the multiple endpoints example of the above being a case in point. A simulation study by Benjamini, Hochberg and Kling (1997) showed that the same procedure controls the FDR for equally positively correlated normally distributed (possibly Studentized) test statistics. The study also showed, as demonstrated above, that the gain in power is large. In the current paper we prove that the procedure controls the FDR in families with positively dependent test statistics (including the case investigated in the mentioned simulation study). In other cases of dependency, we prove that the procedure can still be easily modified to control the FDR, although the resulting procedure is more conservative.

Since we prove the theorem for the case when not all tested hypotheses are true, the structure of the dependency assumed may be different for the set of the true hypotheses and for the false. We shall obviously assume that at least one of the hypotheses is true, otherwise the FDR is trivially 0. The following property, which we call *positive regression dependency on each one from a subset* I_0, or PRDS on I_0, captures the positive dependency structure for which our main result holds. Recall that a set D is called increasing if $x \in D$ and $y \geq x$, implying that $y \in D$ as well.

PROPERTY PRDS. For any increasing set D, and for each $i \in I_0$, $P(\mathbf{X} \in D \mid X_i = x)$ is nondecreasing in x.

The PRDS property is a relaxed form of the positive regression dependency property. The latter means that for any increasing set D, $P(\mathbf{X} \in D \mid X_1 = x_1, \ldots, X_i = x_i)$ is nondecreasing in $(x_1, \ldots, x_i)$ [Sarkar (1969)]. In PRDS the conditioning is on one variable only, each time, and required to hold only for a subset of the variables. If $\mathbf{X}$ is MTP_2, $\mathbf{X}$ is positive regression dependent, and therefore also PRDS over any subset (details in Section 2.3), a property we shall simply refer to as PRDS.

1.4. *The results.* We are now able to state our main theorems.

THEOREM 1.2. *If the joint distribution of the test statistics is PRDS on the subset of test statistics corresponding to true null hypotheses, the Benjamini Hochberg procedure controls the FDR at level less than or equal to* $\frac{m_0}{m} q$.

In Section 2 we discuss in more detail the FDR criterion, the historical background of the procedure and available results and review the relevant notions of positive dependency. This section can be consulted as needed. In Section 3 we outline some important problems where it is natural to assume that the conditions of Theorem 1.2 hold. In Section 4 we prove the theorem. In the course of the proof we provide an explicit expression for the FDR, from which many more new properties can be derived, both for the independent and the dependent cases. Thus issues such as discrete test statistics, composite null hypotheses, general step-up procedures and general dependency can be addressed. This is done in Section 5. In particular we prove there the following theorem.

THEOREM 1.3. *When the Benjamini Hochberg procedure is conducted with* $q/(\sum_{i=1}^{m} \frac{1}{i})$ *taking the place of* q *in* (1), *it always controls the FDR at level less than or equal to* $\frac{m_0}{m}q$.

As can be seen from the above summary, the results of this article greatly increase the range of problems for which a powerful procedure with proven FDR control can be offered.

2. Background.

2.1. *The FDR criterion.* Formally, as in Benjamini and Hochberg (1995), let $\mathbf{V}$ denote the number of true null hypotheses rejected and $\mathbf{R}$ the total number of hypotheses rejected, and let $\mathbf{Q}$ be the unobservable random quotient,

$$\mathbf{Q} = \begin{cases} \mathbf{V}/\mathbf{R}, & \text{if } \mathbf{R} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then the FDR is simply $E(\mathbf{Q})$. Their approach calls for controlling the FDR at a desired level q, while maximizing $E(\mathbf{R})$.

If all null hypotheses are true (the intersection null hypothesis holds) the FDR is the same as the probability of making even one error. Thus controlling the FDR controls the latter, and q is maybe chosen at the conventional levels for α. Otherwise, when some of the hypotheses are true and some are false, the FDR is smaller [Benjamini and Hochberg (1995)]. The control of FDR assumes that when many of the tested hypotheses are rejected it may be preferable to control the proportion of errors rather than the probability of making even one error.

The FDR criterion, and the step-up procedure that controls it, have been used successfully in some very large problems: thresholding of wavelets coefficients [Abramovich and Benjamini (1996)], studying weather maps [Yekutieli and Benjamini (1999)] and multiple trait location in genetics [Weller et al. (1998)], among others. Another attractive feature of the FDR criterion is that if it is controlled separately in several families at some level, then it is also controlled at the same level at large (as long as the families are large enough, and do not consist only of true null hypotheses).

Although the FDR controlling procedure has been implemented in standard computer packages (MULTPROC in SAS), one of its merits is the simplicity with which it can be performed by succinct examination of the ordered list of p-values from the largest to the smallest, and comparing each $p_{(i)}$ to i times q/m stopping at the first time the former is smaller than the latter and rejecting all hypotheses with smaller p-values. Rough arithmetic is usually enough.

2.2. *Positive dependency.* Lehmann (1996) first suggested a concept for bivariate positive dependency, which is very close to the above one and amounts to being PRDS on every subset. Generalizing his concept from bivariate distributions to the multivariate ones was done by Sarkar (1969). A multivariate distribution is said to have positive regression dependency if for any increasing set D, $P(\mathbf{X} \in D \mid X_1 = x_1, \ldots, X_i = x_i)$ is nondecreasing in $(x_1, \ldots, x_i)$.

A stricter condition, implying positive regression dependency, is multivariate total positivity of order 2, denoted MTP_2: $\mathbf{X}$ is MTP_2 if for all $\mathbf{x}$ and $\mathbf{y}$,

$$f(\mathbf{x}) \cdot f(\mathbf{y}) \leq f(\min(\mathbf{x}, \mathbf{y})) \cdot f(\max(\mathbf{x}, \mathbf{y})), \tag{2}$$

where f is either the joint density or the joint probability function, and the minimum and maximum are evaluated componentwise. While being a strong notion of dependency, MTP_2 is widely used, as this property is easier to show. Positive regression dependence implies in turn that X is positive associated, in the sense that for any two functions f and g, which are both increasing (or both decreasing) in each of the coordinates, $\mathrm{cov}(f(\mathbf{X})g(\mathbf{X})) \geq 0$.

PRDS has two properties in which it is different from the above concept. First, monotonicity is required after conditioning only on one variable at a time. Second, the conditioning is done only on any one from a subset of the variables. Thus if $\mathbf{X}$ is MTP_2, or if it is positive regression dependent, then it is obviously positive regression dependent on each one from any subset. Nevertheless, PRDS and positive association do not imply one another, and the difference is of some importance. For example, a multivariate normal distribution is positively associated iff all correlations are nonnegative. Not all correlations need be nonnegative for the PRDS property to hold (see Section 3.1, Case 1 below). On the other hand, a bivariate distribution may be positively associated, yet not positive regression dependent [Lehmann (1966)], and therefore also not PRDS on any subset. A stricter notion of positive association, Rosenbaum's (1984) *conditional (positive) association*, is enough to imply PRDS: $\mathbf{X}$ is conditionally associated, if for any partition $(\mathbf{X_1}, \mathbf{X_2})$ of $\mathbf{X}$, and any function $h(\mathbf{X_1}), \mathbf{X_2}$ given $h(\mathbf{X_1})$ is positively associated.

It is important to note that all of the above properties, including PRDS, remain invariant to taking comonotone transformations in each of the coordinates [Eaton (1986)]. Note also that D is increasing iff $\overline{D}$ is decreasing, so the PRDS property can equivalently be expressed by requiring that for any decreasing set C, and for each $i \in I_0$, $P(\mathbf{X} \in C \mid X_i = x)$ is nonincreasing in x. Therefore, whenever the joint distribution of the test statistics is PRDS

on some I_0 so is the joint distribution of the corresponding p-values, be they right-tailed or left-tailed. Background on these concepts is clearly presented in Eaton (1986), supplemented by Holland and Rosenbaum (1986).

2.3. *Historical background and related results.* The FDR controlling multiple testing procedure [Benjamini and Hochberg (1995)], given by (1), is a step-up procedure that involves a linear set of constants on the p-value scale (step-up in terms of test statistics, not p-values). The FDR controlling procedure is related to the global test for the intersection hypothesis, which is defined in terms of the same set of constants: reject the single intersection hypothesis if there exist an i s.t. $p_{(i)} \leq \frac{i}{m}\alpha$. Simes (1986) showed that when the test statistics are continuous and independent, and all hypotheses are true, the level of the test is α. The equality is referred to as Simes' equality, and the test has been known in recent years as Simes' global test. However the result had already been proved by Seeger (1968) [Shaffer (1995) brought this forgotten reference to the current literature.] See Sen (1999a, b) for an even earlier, though indirect, reference.

Simes (1986) also suggested the procedure given by (1) as an informal multiple testing procedure, and so did Elkund, some 20 years earlier [Seeger (1968)]. The distinction between a global test and a multiple testing procedure is important. If the single intersection hypothesis is rejected by a global test, one cannot further point at the individual hypotheses which are false. When some hypotheses are true while other are false (i.e., when $m_0 < m$), Seeger (1968) showed, referring to Elkund, and Hommel (1988) showed, referring to Simes, that the multiple testing procedure does not necessarily control the FWE at the desired level. Therefore, from the perspective of FWE control, it should not be used as a multiple testing procedure. Other multiple testing procedures that control the FWE have been derived from the Seeger–Simes equality, for example, by Hochberg (1988) and Hommel (1988).

Interest in the performance of the global test when the test statistics are dependent started with Simes (1986), who investigated whether the procedure is conservative under some dependency structures, using simulations. On the negative side, it has been established by Hommel (1988) that the FWE can get as high as $\alpha \cdot (1 + 1/2 + \cdots + 1/m)$. The joint distribution for which this upper bound is achieved is quite bizarre, and rarely encountered in practice. But even with tamed distributions, the global test does not always control the FWE at level α. For example, when two test statistics are normally distributed with negative correlation the FWE is greater than α, even though the difference is very small for conventional levels [Hochberg and Rom (1995)]. On the other hand, extensive simulation studies had shown that for positive dependent test statistics, the test is generally conservative. These results were followed by efforts to extend theoretically the scope of conservativeness, starting with Hochberg and Rom (1995). These efforts have been reviewed in the most recent addition to this line of research by Sarkar (1998). An extensive discussion with many references can be found in Hochberg and Hommel (1998).

Directly relevant to our work are the two strongest results for positive dependent test statistics: Chang, Rom and Sarkar (1996) proved the conservativeness for multivariate distributions with MTP_2 densities. The condition for positive dependency is weaker in the first but the proof applies to bivariate distributions only. Theorem 1.2, when applied to the limited situation where all null hypotheses are true, generalizes the result of Chang, Rom and Sarkar (1996) to multivariate distributions. Although the final result is somewhat stronger than that of Sarkar (1998), the generalization is hardly of importance for the limited case in which all tested hypotheses are true. The full strength of Theorem 1.2 is in the situation when some hypotheses may be true and some may be false, where the full strength of a multiple testing procedure is needed. For this situation the results of Section 2.1 for independent test statistics are the only ones available.

3. Applications. In the first part of this section we establish the PRDS property for some commonly encountered distributions. Recall the sets of variables we have: test statistics for which the tested hypotheses are true and test statistics for which they are false. We are inclined to assume less about the joint distribution of the latter, as will be reflected in some of the following results. In the second part we review some multiple hypotheses testing problems where controlling the FDR is desirable, and where applying Theorem 1.2 shows that using the procedure is a valid way to control it. We emphasize the normal distribution and its related distributions in the first part. For many of the examples in the second part, using normal distribution assumptions for the test statistics is only a partial answer, as methods which are based on other distributions for the test statistics are sometimes needed (such as nonparametric). These issues are beyond the scope of this study.

3.1. *Distributions.*

CASE 1 (Multivariate normal test statistics). Consider $\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$ a vector of test statistics each testing the hypothesis $\mu_i = 0$ against the alternative $\mu_i > 0$, for $i = 1, \ldots, m$. For $i \in I_0$, the set of true null hypotheses, $\mu_i = 0$. Otherwise $\mu_i > 0$.

Assume that for each $i \in I_0$, and for each $j \neq i$, $\mathbf{\Sigma}_{ij} \geq 0$, then the distribution of $\mathbf{X}$ is PRDS over I_0.

PROOF. For any $i \in I_0$, denote by $\mathbf{X}^{(i)}$ the remaining $m-1$ test statistics, $\mu_{(i)}$ is its mean vector, $\mathbf{\Sigma}_{(i),i}$ is the column of covariances of X_i with $\mathbf{X}^{(i)}$, and $\mathbf{\Sigma}_{(i,i)}$ is $\mathbf{\Sigma}$ after dropping the ith row and column.

The distribution of $\mathbf{X}^{(i)}$ given $X_i = x_i$ is $N(\mu^{(i)}, \mathbf{\Sigma}^{(i)})$, where

$$\mathbf{\Sigma}^{(i)} = \mathbf{\Sigma}_{(i,i)} - \mathbf{\Sigma}_{(i),i}\mathbf{\Sigma}_{i,i}^{-1}\mathbf{\Sigma}'_{(i),i} \quad \text{and} \quad \mu^{(i)} = \mu_{(i)} + \mathbf{\Sigma}_{(i),i}\mathbf{\Sigma}_{i,i}^{-1}(x_i - \mu_i).$$

Thus if $\mathbf{\Sigma}_{(i),i}$ is positive, the conditional means increase in x_i. Since the covariance remains unchanged, the conditional distribution increases stochastically

as x_i increases; that is, for any increasing functions f, if $x_i \le x_i'$ then

$$E(f(\mathbf{X}^{(i)}) \mid X_i = x_i) \le E(f(\mathbf{X}^{(i)}) \mid X_i = x_i'). \tag{3}$$

Hence the PRDS over I_0 holds.

Note that the intercorrelations among the test statistics corresponding to the false null hypotheses need not be nonnegative. The fact that less structure is imposed under the alternative hypotheses may be important in some applications; see, for example, the multiple endpoints problem in the following section.

CASE 2 (Latent variable models). In monotone latent variable models, the distribution of $\mathbf{X}$ is assumed to be the marginal distribution of some $(\mathbf{X}, \mathbf{U})$, where the components of $\mathbf{X}$ given $\mathbf{U} = u$ are (a) independent, and (b) stochastically comonotone in $\mathbf{u}$.

If, furthermore, $\mathbf{U}$ is univariate, $\mathbf{X}$ is said to have a unidimensional latent variable distribution [Holland and Rosenbaum (1986)]. Holland and Rosenbaum (1986) show that a unidimensional latent distribution is conditionally positively associated. Therefore it is also PRDS on any subset.

It is interesting to note that the distributions for which Sarkar and Chang (1997) prove their result are all unidimensional latent variable distributions.

For the multivariate latent variable model, if $\mathbf{U}$ is MTP_2, and each $X_i \mid \mathbf{U} = u$ is MTP_2 in x_i and $\mathbf{u}$, then the distribution of $\mathbf{X}$ is MTP_2 (called latent MTP_2.) See again Holland and Rosenbaum (1986), based on a lemma of Karlin and Rinott (1980). While MTP_2 is not enough to imply conditional positive association, it is enough to assure PRDS over any subset.

We shall now generalize the unidimensional latent variable models, to distributions in which the conditional distribution of $\mathbf{X}$ given U is not independent but PRDS on a subset I_0. In this class of distributions the random vector $\mathbf{X}$ is expressed as a monotone transformation of a PRDS random vector $\mathbf{Y}$ and an independent latent variable U, the components of $\mathbf{X}$ are $X_j = g_j(Y_j, U)$.

LEMMA 3.1. *If* (a) $\mathbf{Y}$ *is a continuous random vector, PRDS on a subset* I_0; (b) U *an independently distributed continuous random variable*; (c) *for* $j = 1 \cdots m$ *the components of* $\mathbf{X}$, $X_j = g_j(Y_j, U)$ *are strictly increasing continuous functions of the coordinates* Y_j *and of* U; (d) *for* $i \in I_0, U$ *and* Y_i *are PRDS on* X_i; *then* $\mathbf{X}$ *is PRDS on* I_0.

The proof of this lemma is somewhat delicate and lengthy and is given in the Appendix. Condition (d) of the lemma depends on both the transformation g_i and the distribution of Y_i and U. In the following example condition (d) is asserted via the stronger TP_2 condition.

EXAMPLE 3.2. U_0 and U_1 are independent chi-square or inverse chi-square random variables, $W = U_0 \cdot U_1$. We show that U_i is PRDS on W by showing

the TP_2 property for each pair (U_i, W), $i = 0, 1$. Since for $i = 0, 1$,

$$f_{U_i, W}(x_1, x_2) = 1/x_1 \cdot f_{U_i}(x_1) \cdot f_{U_{1-i}}(x_2/x_1),$$

it is sufficient to assert that $f_{U_{1-i}}(x_2/x_1)$ is TP_2 in x_1 and x_2. It is easy to check that this property holds for both the chi-square and inverse chi-square distributions.

COROLLARY 3.3. *If* $\mathbf{Y}$ *is multivariate normal,* $|Y|$ *PRDS on the subset* I_0 *for which* $\mu_i = 0$ *and* S^2 *is an independently distributed* χ^2_ν*, then* $|\mathbf{X}| = |\mathbf{Y}|/S$ *is PRDS on* I_0.

PROOF. Using Example 3.2, setting $U_0 = |Y_i|^2$ and $U_1 = 1/S^2$, condition (d) holds so we can apply Lemma 3.1.

CASE 3 (Absolute values of multivariate normal and t). $\mathbf{Y} \sim N(\mu, \Sigma)$ and consider two-sided tests: $\mu_i = 0$ against the alternative $\mu_i \neq 0$. Test statistics are multivariate t, obtained by dividing $|\mathbf{Y}|$ by an independent (pooled) chi-square distributed estimator $S > 0$. According to Corollary 3.3 if $|\mathbf{Y}|$ is PRDS over the set of true null hypotheses then $|\mathbf{Y}|/S$ is also PRDS over the set of true null hypotheses.

If $\Sigma = I$, the components of $|\mathbf{Y}|$ are independent and thus PRDS over any subset. For $\Sigma \neq I$, $|\mathbf{Y}|$ is known to be MTP_2 under some conditions [see Karlin and Rinott (1981)], but only when all $\mu_i = 0$. This case was already covered by Sarkar (1998) and is an uncommon example in which all null hypotheses are true, hence the FDR equals the FWE.

$\mathbf{Y}$ can also contain a subset of dependent $\mu = 0$ components of the above form and a subset of $\mu \neq 0$ components, each component corresponding to $\mu = 0$ independent of all $\mu \neq 0$ components; $|\mathbf{Y}|$ is then PRDS over the subset for which $\mu = 0$.

CASE 4 (Studentized multivariate normal). Consider now $\mathbf{Y}$ multivariate normal as in Case 1, Studentized as in Case 3 by S. Because the direction of monotonicity of Y_i/S in S changes as the sign of Y_i changes, $\mathbf{Y}/S$ is not PRDS. Yet we will now show that if q, the level of the test, is less than $1/2$, the Benjamini Hochberg procedure applied to $\mathbf{Y}/S$ offers FDR control.

We will show this by introducing a new random vector $S^+(\mathbf{Y}, S)$ defined as follows: if $Y_j > 0$ then $S^+(Y_j, S) = Y_j/S$, otherwise $S^+(Y_j, S) = Y_j$. The transformation $S^+(\mathbf{Y}, S)$ is increasing in both Y_j and in $1/S$, which satisfies condition (c) in Lemma 3.1. Condition (d) of Lemma 3.1 is also kept, but only for positive values of Y_i, for which we can express $S^+(Y_i, S) = |Y_i|/S$. According to Remark A.4 in the Appendix, $S^+(\mathbf{Y}, S)$ is PRDS, but only when the conditioning is on positive values of $S^+(Y_i, S)$.

According to Remark 4.2, the PRDS condition must only hold for $P_i \in [0, q]$. For $q < 1/2$ this means positive value of $S^+(Y_i, S)$. Hence when applied to $S^+(\mathbf{Y}, S)$ procedure (1) controls the FDR.

Finally notice that since $q < 1/2$ all the critical values of procedure (1) are positive, and for $\mathbf{Y} > 0$, $S^+(\mathbf{Y}, S) \equiv \mathbf{Y}/S$. Hence the outcome of applying

procedure (a) on $\mathbf{Y}/S$ is identical to the outcome of applying procedure (1) on $S^+(\mathbf{Y}, S)$, therefore procedure (1) will also control the FDR when applied to $\mathbf{Y}/S$.

3.2. *Applied problems.*

PROBLEM 1 [Subgroup (subset) analysis in the comparison of two treatments]. When comparing a new treatment to a common one, it is usually of interest to find subgroups for which the new treatment may prove to be better. If there is no "pooling" across subgroups involved, then the test statistics are independent. More typically, averages are compared within the subgroups, yet a pooled estimator of the standard deviation S_{pooled} is used. Hence we have test statistics which are independent and approximately normal, conditionally on S_{pooled}. These (usually) one-sided correlated t-tests fall under Case 4, and thus Theorem 1.2 applies.

PROBLEM 2 (Screening orthogonal contrasts in a balanced design). Consider a balanced factorial experiment with m factorial combinations and n repetitions per cell, which is performed for the purpose of screening many potential factors for their possible effect on a quantity of interest. Such experiments are common, for example, in industrial statistics when screening for possible factors affecting quality characteristics, and in the pharmaceutical industry when screening for potentially beneficial compounds. In the above two, economic considerations make it clear that in identifying a set of hypotheses for further research, allowing a controlled proportion of errors in the identified pool is desirable. In fact the chosen level for q may be higher than the levels usually used for α. The distributional model is that of (usually) two-sided correlated t-tests, which thus fall under Case 3.

PROBLEM 3 (Many-to-one comparisons in clinical trials). Differently phrased this is the problem of comparing a few treatments with a single control, using one-sided tests. See the recent review by Tamhane and Dunnett (1999) for the many approaches and procedures that control the FWE. If the interest lies in recommending one of the tested treatments based solely on the current experiment, FWE should be controlled. But if the conclusion is closer in nature to the conclusion of Problem 2, the control of FDR is appropriate [see detailed discussion in Benjamini, Hochberg and Kling (1993)].

In the normal model, $X_i = (Y_i - Y_0)/c_i S, Y_i, i = 0, 1, \ldots, m$ independent normal random variables, with variances $c_i\sigma^2$ which are known up to σ, S^2, an independent estimator such that $S^2/\sigma^2 \sim \chi^2_\nu/\nu$. $(Y_i - Y_0)/c_i$ is multivariate normal with $\rho_{ij} > 0$, hence PRDS, thus according to Case 4, $\mathbf{X}$ is PRDS on the set of true null hypotheses.

EXAMPLE 3.4. The study of uterine weights of mice reported by Steel and Torrie (1980) and discussed in Westfall and Young (1993) comprised a comparison of six groups receiving different solutions to one control group. The

lower-tailed p-values of the pooled variance t-statistics are 0.183, 0.101, 0.028, 0.012, 0.003, 0.002. Westfall and Young (1993) show that, using p-value resampling and step-down testing, three hypotheses are rejected at FWE 0.05. Four hypotheses are rejected when applying procedure (1) using FDR level of 0.05.

PROBLEM 4 (Multiple endpoints in clinical trials). Multiple endpoints, that is, the multiple outcomes according to which the therapeutic properties of one treatment are compared with those of an established treatment, raises one of the most serious multiplicity control problems in the design and analysis of clinical trials. For a recent review, see Wassmer, Reitmer, Kieser and Lehmacher (1998). Eighteen outcomes were studied in Example 1.1, but the number may reach hundreds, so addressing this problem by controlling the FWE is overwhelmingly conservative. A common remedy is to specify very few *primary endpoints* on which the conclusion will be based and give a lesser standing to the conclusions from the other *secondary endpoints*, for which FWE is not controlled. However, it is not uncommon to find the advocated features of a new treatment to come mostly from the secondary endpoints.

The FDR approach is very natural for this problem, and the emphasise on primary endpoints is no longer essential [but feasible as in Benjamini and Hochberg (1997)].

The test statistics of the different endpoints are usually dependent. Their dependency is in most cases neither constant nor known, and stems both from correlated treatment effect (for nonnull treatment effects) and a latent individual component affecting the value of all endpoints of the same person. The individual component introduces a latent positive dependence between all test statistics. Thus test statistics of null hypotheses are positively correlated with all other test statistics. Treatment effect may introduce negative correlation between the affected endpoints, which may dominate the latent positive dependency. Thus we want to allow those endpoints which are affected by the treatment to have whatever dependence structure occurs among themselves.

Then, using the results of Cases 1, 2 and 4 above, Theorem 1.2 applies for the one-sided tests, be they normal tests or t-tests. The situation with two-sided tests is more complicated, as Case 3 requires a stronger assumption.

EXAMPLE 3.5 (Low lead levels and IQ). Needleman, Gunnoe, Leviton, Reed, Presie, Maher and Barret (1979) studied the neuropsychologic effects of unidentified childhood exposure to lead by comparing various psychological and classroom performances between two groups of children differing in the lead level observed in their shed teeth. While there is no doubt that high levels of lead are harmful, Needleman's findings regarding exposure to low lead levels, especially because of their contribution to the Environmental Protection Agency's review of lead exposure standards, are controversial. Needleman's study was attacked on the ground of methodological flaws; for details see Westfall and Young (1993). One of the methodological flaws pointed out is control of multiplicity. Needleman et al. (1979) present three families of

TABLE 1

Family	*p*-values (omitting sum score *p*-values)	FWE Rej. thrshld.	FWE # of rej.	FDR Rej. thrshld.	FDR # of rej.
Teacher's behavioral ratings	0.003 0.05 0.05 0.14 0.08 0.01 0.04 0.01 0.05 0.003 0.003	0.005	3	0.02	5
Score of Wechsler Intelligence Scale for Children (revised)	0.04 0.05 0.02 0.49 0.08 0.36 0.03 0.38 0.15 0.90 0.37 0.54	0.004	0	0.004	0
Verbal processing and reaction times	0.002 0.03 0.07 0.37 0.90 0.42 0.05 0.04 0.32 0.001 0.0010.01	0.004	3	0.016	4
The three families jointly		0.001	2	0.012	9

endpoints, and comment on the results of separate multiplicity adjustments within each family as summarized in Table 1 (under the FWE heading).

The critics argue that multiplicity should be controlled for all families jointly. Using Hochberg's method at 0.05 level, correcting within each family, six hypotheses are rejected. Correcting for all 35 responses, lead is found to have an adverse effect in only two out of 35 endpoints.

Applying procedure (1) at 0.05 FDR level, the attack on Needleman findings on grounds of inadequate multiplicity control is unjustified; whether analyzed jointly or each family separately, lead was found to have an adverse effect in more than a quarter of the endpoints.

4. Proof of theorem. For ease of exposition let us denote the set of constants in (1), which define the procedure, by

$$q_i = \frac{i}{m} q, \qquad i = 1, 2, \ldots, m. \tag{4}$$

Let $A_{v,s}$ denote the event that the Benjamini Hochberg procedure rejects exactly v true and s false null hypotheses. The FDR is then

$$E(\mathbf{Q}) = \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \Pr(A_{v,s}). \tag{5}$$

In the following lemma, $\Pr(A_{v,s})$ is expressed as an average.

LEMMA 4.1.

$$\Pr(A_{v,s}) = \frac{1}{v} \sum_{i=1}^{m_0} \Pr\big(\{P_i \le q_{v+s}\} \cap A_{v,s}\big). \tag{6}$$

PROOF. For a fixed v and s, let ω denote a subset of $\{1\cdots m_0\}$ of size v, and $A_{v,s}^{\omega}$ the event in $A_{v,s}$ that the v true null hypotheses rejected are ω. Note that $\Pr\{P_i \le q_{v+s} \cap A_{v,s}^{\omega}\}$ equals $\Pr\{A_{v,s}^{\omega}\}$ if $i \in \omega$, and is otherwise 0.

$$\begin{aligned}\sum_{i=1}^{m_0} \Pr(\{P_i \le q_{v+s}\} \cap A_{v,s}) &= \sum_{i=1}^{m_0} \sum_{\omega} \Pr\left(\{P_i \le q_{v+s}\} \cap A_{v,s}^{\omega}\right) \\ &= \sum_{\omega} \sum_{i=1}^{m_0} \Pr\left(\{P_i \le q_{v+s}\} \cap A_{v,s}^{\omega}\right) \\ &= \sum_{\omega} \sum_{i=1}^{m_0} I(i \in \omega) \Pr\{A_{v,s}^{\omega}\} \\ &= \sum_{\omega} v \cdot \Pr\left(A_{v,s}^{\omega}\right) = v \cdot \Pr\{A_{v,s}\}.\end{aligned} \tag{7}$$

Combining equation (5) with Lemma 4.1, the FDR is

$$\begin{aligned}E(\mathbf{Q}) &= \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \left\{ \sum_{i=0}^{m_0} \frac{1}{v} \Pr(\{P_i \le q_{v+s}\} \cap\ A_{v,s}) \right\} \\ &= \sum_{i=0}^{m_0} \left\{ \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{1}{v+s} \Pr(\{P_i \le q_{v+s}\} \cap A_{v,s}) \right\}\end{aligned} \tag{8}$$

Now that the dependency of the expectation on v is only through $A_{v,s}$; we reconstruct $A_{v,s}$ from events that depend on i and $k = v + s$ only, so the FDR may be expressed similarly.

For $i = 1\cdots m_0$, let $\mathbf{P}^{(i)}$ be the remaining $m - 1$ p-values after dropping P_i. Let $C_{v,s}^{(i)}$ denote the event in which if P_i is rejected then $v - 1$ true null hypotheses and s false null hypotheses are rejected alongside with it. That is, $C_{v,s}^{(i)}$ is the projection of $\{P_i \le q_{v+s}\} \cap A_{v,s}$ onto the range of $\mathbf{P}^{(i)}$, and expanded again by cross multiplying with the range of P_i. Thus we have

$$\{P_i \le q_{v+s}\} \cap A_{v,s} = \{P_i \le q_{v+s}\} \cap C_{v,s}^{(i)}. \tag{9}$$

Denote by $C_k^{(i)} = \bigcup\{C_{v,s}^{(i)}\colon v + s = k\}$. For each i the $C_k^{(i)}$ are disjoint, so the FDR can be expressed as

$$E(\mathbf{Q}) = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \Pr\left(P_i \le q_k \cap C_k^{(i)}\right), \tag{10}$$

where the expression no longer depends on v and s, as desired.

In the last part of the proof we construct an expanding series of increasing sets, on which we use the PRDS property to bound the inner sum in (8) by q/m. For this purpose, define $D_k^{(i)} = \bigcup\{C_j^{(i)}\colon j \le k\}$ for $k = 1\cdots m$. $D_k^{(i)}$

can also be described using the ordered set of the p-values in the range of $\mathbf{P}^{(i)}$, $\{p^{(i)}_{(1)} \le \cdots \le p^{(i)}_{(m-1)}\}$, in the following way:

$$D_k = \left\{\mathbf{p}\colon q_{k+1} < p^{(i)}_{(k)}, q_{k+2} < p^{(i)}_{(k+1)} \cdots q_m < p^{(i)}_{(m-1)}\right\} \tag{11}$$

for $k = 1 \ldots m-1$, and $D^{(i)}_m$ is simply the entire space. Expressing $D^{(i)}_k$ as above, it becomes clear that for each k, $D^{(i)}_k$ is a nondecreasing set.

We now shall make use of the PRDS property, which states that for $p \le p'$,

$$\Pr(D \mid P_i = p) \le \Pr(D \mid P_i = p'). \tag{12}$$

Following Lehmann (1996), it is easy to see that for $j \le l$ since $q_j \le q_l$,

$$\Pr(D \mid P_i \le q_j) \le \Pr(D \mid P_i \le q_l), \tag{13}$$

for any nondecreasing set D, or equivalently,

$$\frac{\Pr\left(\{P_i \le q_k\} \cap D^{(i)}_k\right)}{\Pr(P_i \le q_k)} \le \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap D^{(i)}_k\right)}{\Pr(P_i \le q_{k+1})}. \tag{14}$$

Invoking (14) together with the fact that $D^{(i)}_{j+1} = D^{(i)}_j \cup C^{(i)}_{j+1}$ yields for all $k \le m-1$,

$$\begin{aligned} &\frac{\Pr\left(\{P_i \le q_k\} \cap D^{(i)}_k\right)}{\Pr(P_i \le q_k)} + \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap C^{(i)}_{k+1}\right)}{\Pr(P_i \le q_{k+1})} \\ &\quad \le \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap D^{(i)}_k\right)}{\Pr(P_i \le q_{k+1})} + \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap C^{(i)}_{k+1}\right)}{\Pr(P_i \le q_{k+1})} \\ &\quad = \frac{\Pr\left(\{P_i \le q_{k+1}\} \cap D^{(i)}_{k+1}\right)}{\Pr\left(P_i \le q_{k+1}\right)}. \end{aligned} \tag{15}$$

Now, start by noting that $C_1 = D_1$, and repeatedly use the above inequality for $i = 1, \ldots, m-1$, to fold the sum on the left into a single expression,

$$\sum_{k=1}^{m} \frac{\Pr\left(\{P_i \le q_k\} \cap C^{(i)}_k\right)}{\Pr(P_i \le q_k)} \le \frac{\Pr\left(\{P_i \le q_m\} \cap D^{(i)}_m\right)}{\Pr(P_i \le q_m)} = 1, \tag{16}$$

where the last equality follows because $D^{(i)}_m$ is the entire space.

Going back to expression (10) for the FDR,

$$\begin{aligned} E(\mathbf{Q}) &= \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \Pr\left(\{P_i \le q_k\} \cap C^{(i)}_k\right) \\ &\le \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{q}{m} \cdot \frac{\Pr\left(\{P_i \le q_k\} \cap C^{(i)}_k\right)}{\Pr(P_i \le q_k)}, \end{aligned} \tag{17}$$

because $\Pr(P_i \le q_k) \le q_k = \frac{k}{m}q$ under the null hypothesis (with equality for continuous test statistics where each P_i is uniform), so finally, invoking (16),

$$\frac{q}{m}\sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{\Pr\left(\{P_i \le q_k\} \cap C_k^{(i)}\right)}{\Pr(P_i \le q_k)} \le \frac{m_0}{m}q. \tag{18}$$

REMARK 4.2. Note that PRDS is a sufficient but not a necessary condition. In particular the PRDS property need not hold for all monotone sets D and all values of p_i. According to inequality (12), it is enough that they hold for monotone sets of the form of (11) and $P_i \in [0,\ q]$.

This remark is used to establish that Theorem 1.2 holds for one-sided multivariate t and $q < 1/2$, even though the distribution is not PRDS.

5. Generalizations and further results. If the test statistics are jointly independent, the FDR as expressed in (10) is

$$\begin{aligned} E(\mathbf{Q}) &= \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\Pr\left(\left\{P_i \le \frac{k}{m}q\right\} \cap C_k^{(i)}\right) \\ &= \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\Pr\left(P_i \le \frac{k}{m}q\right)\cdot\Pr\left(C_k^{(i)}\right) \qquad (19) \\ &= \sum_{i=1}^{m_0}\frac{\alpha}{m}\cdot\sum_{k=1}^{m}\Pr\left(C_k^{(i)}\right) = \frac{m_0}{m}q, \qquad (20) \end{aligned}$$

which yields an alternative (and possibly simpler) proof of the result in Benjamini and Hochberg (1995). Moreover, the proof there depends critically on the assumption that the P-values are uniformly distributed under the null hypotheses, and therefore do not apply to discrete test statistics. However, for discrete test statistics, we have that

$$\Pr\left(P_i \le \frac{k}{m}q\right) \le \frac{k}{m}q, \qquad i = 1, 2, \ldots, m_0. \tag{21}$$

Therefore, when passing from (19) to (20), we need only change the equality to inequality in order to complete the proof of the following theorem.

THEOREM 5.1. *For independent test statistics, the Benjamini Hochberg procedure controls the FDR at level less or equal to* $\frac{m_0}{m}q$. *If the test statistics are also continuous, the FDR is exactly* $\frac{m_0}{m}q$.

The argument leading to the above theorem used only the fact that for discrete test statistics the tail probabilities are smaller. Thus, in a similar way, it follows that the FDR is controlled when the procedure is used for testing composite null hypotheses, as in one-sided tests.

THEOREM 5.2. *For independent one-sided test statistics, if the distributions in each of the composite null hypothesis are stochastically smaller than the null distribution under which each p-value is computed, the Benjamini Hochberg procedure controls the FDR at level less or equal to* $\frac{m_0}{m}q$.

The surprising part of Theorem 5.1 is that equality holds no matter what the distributions of the test statistics corresponding to the false null hypotheses are. The following theorem shows that this is a unique property of the step-up procedure which uses the constants $\{\frac{k}{m}q\}$. More generally, we can define step-up procedures SU($\boldsymbol{\alpha}$), using any other monotone series of constants $\alpha_1 \le \alpha_1 \le \cdots \le \alpha_m$: let $k = \max\{i\colon p_{(i)} \le \alpha_i\}$, and if such k exists reject $H_{(1)} \cdots H_{(k)}$.

THEOREM 5.3. *Testing m hypotheses with SU*($\boldsymbol{\alpha}$), *assume that the distribution of the P-values,* $\mathbf{P} = (\mathbf{P_0}, \mathbf{P_1})$ *is jointly independent.*

(i) *If the ratio* α_k/k *is increasing in* k, *as the distribution of* $\mathbf{P_1}$ *increases stochastically the FDR decreases.*

(ii) *If the ratio* α_k/k *is decreasing in* k, *as the distribution of* $\mathbf{P_1}$ *increases stochastically the FDR increases.*

PROOF. Given the set of critical values $\boldsymbol{\alpha}$ for $k = 1, \ldots, m$ we define the following sets:

$$C_k(\boldsymbol{\alpha}) = \left\{\mathbf{P}^{(i)}\colon P^{(i)}_{(k-1)} \le \alpha_k, \ldots, P^{(i)}_{(k)} > \alpha_{k+1}, \ldots, P^{(i)}_{(m-1)} > \alpha_m\right\}. \tag{22}$$

Thus if $\mathbf{P}^{(i)} \in C_k(\boldsymbol{\alpha})$ and $P_i \le \alpha_k$ then H_i^0 is rejected along with $k-1$ other hypotheses, but if $P_i > \alpha_k$, H_i^0 is not rejected. Notice that sets $C_k(\boldsymbol{\alpha})$ are ordered. If $\mathbf{P}^{(i)} \in C_k(\boldsymbol{\alpha})$ and $\mathbf{P}^{(i)} \le \mathbf{P}'^{(i)}$, then all ordered coordinates of $\mathbf{P}'^{(i)}$ are greater or equal to corresponding coordinates of $\mathbf{P}^{(i)}$. Therefore for $j = 1 \cdots m-1$, $P'^{(i)}_{(j)} \ge \alpha_j$, thus $\mathbf{P}'^{(i)} \in C_l(\boldsymbol{\alpha})$ for some $l \le k$.

Next we define the function $f_{\boldsymbol{\alpha}}$, $f_{\boldsymbol{\alpha}}\colon [0,1]^{m-1} \to \mathscr{R}$,

$$f_{\boldsymbol{\alpha}}(\mathbf{P}^{(i)}) = \alpha_k/k \quad \text{for } \mathbf{P}^{(i)} \in C_k(\boldsymbol{\alpha}). \tag{23}$$

The FDR of all step-up procedures can be expressed similarly to expression (10). Start deriving Lemma 4.1 by substituting α_k in place of $\alpha k/m$ throughout the proof. Then, denoting the FDR of $SU(\boldsymbol{\alpha})$ by $E(\mathbf{Q}(\boldsymbol{\alpha}))$, we use the independence of the test statistics to get

$$E(\mathbf{Q}(\boldsymbol{\alpha})) = \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\Pr\left(\{P_i \le \alpha_k\} \cap \mathbf{P}^{(i)} \in C_k(\boldsymbol{\alpha})\right) \tag{24}$$

$$= \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\Pr(P_i \le \alpha_k)\Pr\left(\mathbf{P}^{(i)} \in C_k(\boldsymbol{\alpha})\right) \tag{25}$$

$$= \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{\alpha_k}{k}\Pr\left(\mathbf{P}^{(i)} \in C_k(\boldsymbol{\alpha})\right) = \sum_{i=1}^{m_0} E_{\mathbf{P}^{(i)}} f_{\boldsymbol{\alpha}}. \tag{26}$$

Note that the distribution of the test statistics corresponding to the m_0 true null hypotheses is fully specified as $U[0,1]$. If α_k/k increases in k, the function $f_{\boldsymbol{\alpha}}$ is a decreasing function. Stochastic increase in the distribution of $\mathbf{P}^{(i)}$ is characterized by the decrease of the expectation of all decreasing functions, in particular a decrease in all the summands of the right side of (26). Thus if $\mathbf{P_1}$ increases stochastically, the FDR decreases. If α_k/k decreases in k, the function $f_{\boldsymbol{\alpha}}$ is an increasing function. Thus if $\mathbf{P_1}$ increases stochastically the FDR increases. (The case where α_k/k is constant has been covered by Theorem 5.1) $\square$

These more general step-up procedures are especially important in particular settings, where the structure of dependency can be precisely specified. In such a case a specific set of constants can be used for designing a step-up procedure which exactly achieves the desired FDR at the specified distribution. Troendle (1996) took this route, calculating a monotone series of constants, which upon being used in the above fashion, control the FDR for normally distributed test statistics which are equally and positively correlated. His calculations were done under the unproven assertion that when the nonzero means are set at infinity the FDR is maximized. In order to use Theorem 5.3 for that purpose it should be generalized first to hold under some joint distribution other than independent, say PRDS. We do not have yet such a result.

An important question that remains to be answered is the scope of problems for which the two-sided tests retain the same level of control. Another important open question is whether the same procedure controls the FDR when testing pairwise comparisons of normal means, either Studentized or not. Simulation studies, by Williams, Jones and Tukey (1999) and by Benjamini, Hochberg and Kling (1993), and some limited calculations in the latter, show that this is the case. It is known that the distribution of the test statistics is not MTP_2. The PRDS condition does not hold as well.

When facing such problems, it is always comforting to have a fallback procedure. The available FWE controlling procedure can be modified by working at level $\alpha/\sum_{j=1}^{m}\frac{1}{j}$, and it will then control the FWE at level α for any joint distribution of the test statistics—as long as the hypotheses are all true [Hommel (1988)]. Similarly, Theorem 1.3 establishes that the same modification of the procedure controls the FDR at the desired level, for any joint distribution of the test statistics.

PROOF OF THEOREM 1.3. For simplicity of the exposition we shall use q in (1), and show that the FDR is increased by no more than $\sum_{j=1}^{m}\frac{1}{j}$.

Denote $p_{ikj} = \Pr(\{P_i \in [\frac{(j-1)}{m}q, \frac{j}{m}q]\} \cap C_k^{(i)})$. Note that,

$$\sum_{k=1}^{m} p_{ijk} = \Pr\left(\left\{P_i \in \left[\frac{(j-1)}{m}q, \frac{j}{m}q\right]\right\} \cap \left(\bigcup_{k=1}^{m} C_k^{(i)}\right)\right) = \frac{q}{m}. \tag{27}$$

Returning to expression (10), the FDR can be expressed as

$$E(\mathbf{Q}) = \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\sum_{j=1}^{k} p_{ijk} = \sum_{i=1}^{m_0}\sum_{j=1}^{m}\sum_{k=j}^{m}\frac{1}{k}p_{ijk} \tag{28}$$

$$\leq \sum_{i=1}^{m_0}\sum_{j=1}^{m}\sum_{k=j}^{m}\frac{1}{j}p_{ijk} \leq \sum_{i=1}^{m_0}\sum_{j=1}^{m}\frac{1}{j}\sum_{k=1}^{m} p_{ijk} = m_0\sum_{j=1}^{m}\frac{1}{j}\frac{q}{m}. \tag{29}$$

□

Obviously, as the main thrust of this paper shows, the adjustment by $\sum_{i=1}^{m}\frac{1}{i} \approx \log(m) + \frac{1}{2}$ is very often unneeded, and yields too conservative a procedure. Still, even if only a small proportion of the tested hypotheses are detected as not true [approximately $\log(m)/m$], the procedure is more powerful than the comparable FWE controlling procedure of Holm (1979). The ratio of the defining constants can get as high as $(m+1)/4\log(m)$ in favor of the FDR controlling procedure, so its advantage can get very large.

It should be noted that throughout all results of this work, the procedure controls the FDR at a level too low by a factor of m_0/m. Loosely speaking, the procedure actually controls the false discovery likelihood ratio,

$$E\left(\frac{\frac{\mathbf{V}}{m_0}}{\frac{\mathbf{R}}{m}}\right) \leq q. \tag{30}$$

Other procedures, which get closer to controlling the FDR at the desired level, have been offered for independent test statistics in Benjamini and Hochberg (2000), and in Benjamini and Wei (1999). Only little is known about the performance of the first for dependent test statistics [Benjamini, Hochberg and Kling (1997)], and nothing about the second.

Finally, recall the resampling based procedure of Yekutieli and Benjamini (1999), which tries to cope with the above problem and at the same time utilize the information about the dependency structure derived from the sample. The resampling based procedure is more powerful, at the expense of greater complexity and only approximate FDR control.

APPENDIX

PROOF OF LEMMA 3.1. For each $i \in I_0$ and increasing set D, we have to show that

$$\Pr(\mathbf{X} \in D \mid X_i = x)$$

is increasing in x. We will achieve this by expressing

$$\Pr(\mathbf{X} \in D \mid X_i = x) = E_{U|X_i=x}\Pr(\mathbf{X} \in D \mid X_i = x, U) \tag{31}$$

and showing that for $x \leq x'$,

$$E_{U|X_i=x}\Pr(\mathbf{X} \in D \mid X_i = x, U) \leq E_{U|X_i=x'}\Pr(\mathbf{X} \in D \mid X_i = x', U). \tag{32}$$

We prove the lemma in two steps.

1. For each $x \leq x'$ we construct a new random variable U' whose marginal distribution is stochastically smaller than the marginal distribution of U, but its conditional distribution given $X_i = x'$ is identical to the conditional distribution of U given $X_i = x$.
2. We show that the newly defined random variable U' satisfies

$$\Pr(\mathbf{X} \in D \mid X_i = x, U = u) \leq \Pr(\mathbf{X} \in D \mid X_i = x', U' = u). \tag{33}$$

By re-expressing the second term in inequality (32) in terms of U' and then using inequality (33), the proof is complete:

$$\begin{aligned} E_{U|X_i=x'} \Pr(\mathbf{X} \in D \mid X_i = x', U) &= E_{U'|X_i=x'} \Pr(\mathbf{X} \in D \mid X_i = x', U') \\ &\geq E_{U|X_i=x} \Pr(\mathbf{X} \in D \mid X_i = x, U). \end{aligned}$$

STEP 1. The construction of U': according to condition (d) of this lemma, U is PRDS on X_i; this means that the cdf of $U \mid X_i = x'$ is less or equal to the cdf of $U \mid X_i = x$,

$$F_{U|X_i=x'} \leq F_{U|X_i=x}. \tag{34}$$

In order to avoid technicalities let us assume that $U \mid X_i = x$ has the same support as U for any x. Now the following increasing transformation is well defined, and satisfies

$$h_{x,x'}(u) = F^{-1}_{U|X_i=x}(F_{U|X_i=x'}(u)) \leq F^{-1}_{U|X_i=x}(F_{U|X_i=x}(u)) = u, \tag{35}$$

because of (34). The new random variable U' is defined as

$$U' = h_{x,x'}(U)$$

and is, from (35), stochastically smaller than U. Because g, $\mathbf{Y}$ and U are continuous, the conditional distribution of U given X_i is continuous, hence $h_{x,x'}$ and its inverse $h_{x',x}$ can be defined. Using the notation

$$u' = h_{x',x}(u), \tag{36}$$

we can state the following properties:

(i) $u \leq u'$, again because of (35), and $h_{x',x}$ being its inverse.
(ii) $F_{U|X_i=x}(u) = F_{U|X_i=x'}(u')$, which follows directly from the definition of $h_{x,x'}$.
(iii) The events $U \leq u'$ and $U' \leq u$ are identical, as U' is a monotone function of U.

Combining (i), (ii) and (iii), we get

$$\begin{aligned} \Pr(U \leq u \mid X_i = x) &= \Pr(U \leq u' \mid X_i = x') \\ &= \Pr(U' \leq u \mid X_i = x'). \end{aligned}$$

Hence $U \mid X_i = x$ and $U' \mid X_i = x'$ are identically distributed.

STEP 2. A proof of inequality (33): the function g_i is one-to-one, so the values of U and X_i uniquely determine the value of Y_i. Thus for each u, and the corresponding u' defined in expression (36), denote y and y' those values of Y_i which satisfy

$$g_i(y, u) = x \quad \text{and} \quad g_i(y', u') = x'.$$

We now establish that for the pair $x' \geq x$, and the pair $u' \geq u$ as above, we also have that $y' \geq y$. As g_i is strictly increasing in both components, fixing X_i then $Y_i \leq y$ iff $U \geq u$, thus

$$\Pr(Y_i \leq y \mid X_i = x) = \Pr(U \geq u \mid X_i = x) = 1 - F_{U|X_i=x}(u).$$

Similarly, $Y_i \leq y'$ iff $U \geq u'$,

$$\Pr(Y_i \leq y' \mid X_i = x') = \Pr(U \geq u' \mid X_i = x') = 1 - F_{U|X_i=x'}(u').$$

As $F_{U|X_i=x'}(u') = F_{U|X_i=x}(u)$, y and y' are quantiles corresponding to the same probability. Returning to condition (d) of the lemma, Y_i is PRDS on X_i, therefore $Y_i \mid X_i = x'$ is stochastically greater than $Y_i \mid X_i = x$, thus $y \leq y'$.

We now define

$$Y(D, u) := \{\mathbf{Y}\colon g(\mathbf{Y}, u) \in D\}.$$

Note that if D is an increasing set then $Y(D, u)$ is an increasing set. We can now proceed to complete the proof of Step 2:

$$\begin{aligned}
\Pr(\mathbf{X} \in D \mid X_i = x, U = u) &= \Pr(\mathbf{Y} \in Y(D, u) \mid Y_i = y, U = u) \\
&\leq \Pr(\mathbf{Y} \in Y(D, u) \mid Y_i = y', U = u) \qquad (37) \\
&\leq \Pr(\mathbf{Y} \in Y(D, u') \mid Y_i = y', U = u') \qquad (38) \\
&= \Pr(\mathbf{X} \in D \mid X_i = x', U = u') \\
&= \Pr(\mathbf{X} \in D \mid X_i = x', U' = u) \qquad (39)
\end{aligned}$$

Inequality (37) holds because $\mathbf{Y}$ is PRDS and independent of U. Using again the independence, and the fact that if $u \leq u'$ then $Y(D, u) \subseteq Y(D, u')$, we get inequality (38). Finally as $U' = u$ iff $U = u'$ we get the equality in expression (39). This completes the proof of Step 2, and thereby the proof of Lemma 3.2. □

REMARK A.1. Note that the seemingly simple route of proving Lemma 3.1 via showing

$$\Pr(\mathbf{X} \in D \mid X_i = x, U = u) \leq \Pr(\mathbf{X} \in D \mid X_i = x', U = u) \qquad (40)$$

does not yield the desired result, because the distribution of $U \mid X_i = x$ is different than the the distribution of $U \mid X_i = x'$.

REMARK A.2. In the course of the proof we established the monotonicity of

$$\Pr(\mathbf{X} \in D \mid Y_i = y, U = u)$$

in y and in u. However, because g_i is increasing, fixing X_i and increasing U will decrease Y_i, because $\mathbf{Y}$ is PRDS, and

$$\Pr(\mathbf{X} \in D \mid X_i = x, U = u) \tag{41}$$

does not necessarily increase in u. If expression 41 increases in u, for example when the components of $\mathbf{Y}$ are independent, proof of Lemma 3.2 is immediate because the distribution of $U \mid X_i = x'$ is stochastically greater than the distribution of $U \mid X_i = x$.

REMARK A.3. The assumption that $U \mid X_i = x$ has the same support as U is not critical. With appropriate definition of the inverse of the conditional cdf of U, $F^{-1}_{U|X_i}$, $h_{x,x'}$ can be well defined over the entire range of U. Also $h_{x',x}$ can be defined similarly. It will be the inverse of $h_{x,x'}$ only on the respective ranges. Properties (i)–(iii) still hold under this more complicated construction.

REMARK A.4. If conditions (a)–(c) of the lemma are met, while condition (d), U and Y_i, are PRDS on X_i is only true for X_i such that $X_i \geq x_i$ then altering the proof accordingly, $\mathbf{X}$ is PRDS on $X_i \geq x_i$.

Acknowledgments. We are grateful to Ester Samuel-Cahn, Yosef Rinott and David Gilat for their helpful comments and to a referee for keeping us honest.

REFERENCES

ABRAMOVICH, F. and BENJAMINI, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.* **22** 351–361.

BARINAGA, M. (1994). From fruit flies, rats, mice: evidence of genetic influence. *Science* **264** 1690–1693.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

BENJAMINI, Y. and HOCHBERG, Y. (1997). Multiple hypotheses testing with weights. *Scand. J. Statist.* **24** 407–418.

BENJAMINI, Y. and HOCHBERG, Y. (2000). The adaptive control of the false discovery rate in multiple hypotheses testing. *J. Behav. Educ. Statist.* **25** 60–83.

BENJAMINI, Y., HOCHBERG, Y. and KLING, Y. (1993). False discovery rate control in pairwise comparisons. Working Paper 93-2, Dept. Statistics and O.R., Tel Aviv Univ.

BENJAMINI, Y., HOCHBERG, Y. and KLING, Y. (1997). False discovery rate control in multiple hypotheses testing using dependent test statistics. Research Paper 97-1, Dept. Statistics and O.R., Tel Aviv Univ.

BENJAMINI, Y. and WEI, L. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* **82** 163–170.

CHANG, C. K., ROM, D. M. and SARKAR, S. K. (1996). A modified Bonferroni procedure for repeated significance testing. Technical Report 96-01, Temple Univ.

EATON, M. L. (1986). *Lectures on topics in probability inequalities*. *CWI Tract* **35**.

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

HOCHBERG, Y. and HOMMEL, G. (1998). Step-up multiple testing procedures. *Encyclopedia Statist. Sci. (Supp.)* **2**.

HOCHBERG, Y. and ROM, D. (1995). Extensions of multiple testing procedures based on Simes' test. *J. Statist. Plann. Inference* **48** 141–152.

HOCHBERG, Y. and TAMHANE, A. (1987). *Multiple Comparison Procedures*. Wiley, New York.

HOLLAND, P. W. and ROSENBAUM, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Ann. Statist.* **14** 1523–1543.

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

HOMMEL, G. (1988). A stage-wise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75** 383–386.

HSU, J. (1996). *Multiple Comparisons Procedures*. Chapman and Hall, London.

KARLIN, S. and RINOTT, Y. (1980). Classes of orderings of measures and related correlation inequalities I. Multivariate totally positive distributions. *J. Multivariate Statist.* **10** 467–498.

KARLIN, S. and RINOTT, Y. (1981). Total positivity properties of absolute value multinormal variable with applications to confidence interval estimates and related probabilistic inequalities. *Ann. Statist.* **9** 1035–1049.

LANDER E. S. and BOTSTEIN D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121** 185–190.

LANDER, E. S. and KRUGLYAK L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11** 241–247.

LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137–1153.

NEEDLEMAN, H., GUNNOE, C., LEVITON, A., REED, R., PRESIE, H., MAHER, C. and BARRET, P. (1979). Deficits in psychologic and classroom performance of children with elevated dentine lead levels. *New England J. Medicine* **300** 689–695.

PATERSON, A. H. G., POWLES, T. J., KANIS, J. A., MCCLOSKEY, E., HANSON, J. and ASHLEY, S. (1993). Double-blind controlled trial of oral clodronate in patients with bone metastases from breast cancer. *J. Clinical Oncology* **1** 59–65.

ROSENBAUM, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika* **49** 425–436.

SARKAR, T. K. (1969). Some lower bounds of reliability. Technical Report, 124, Dept. Operation Research and Statistics, Stanford Univ.

SARKAR, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of Simes' conjecture. *Ann. Statist*. **26** 494–504.

SARKAR, S. K. and CHANG, C. K. (1997). The Simes method for multiple hypotheses testing with positively dependent test statistics. *J. Amer. Statist. Assoc.* **92** 1601–1608.

SEEGER, (1968). A note on a method for the analysis of significances en mass. *Technometrics* **10** 586–593.

SEN, P. K. (1999a). Some remarks on Simes-type multiple tests of significance. *J. Statist. Plann. Inference*, **82** 139–145.

SEN, P. K. (1999b). Multiple comparisons in interim analysis. *J. Statist. Plann. Inference* **82** 5–23.

SHAFFER, J. P. (1995). Multiple hypotheses-testing. *Ann. Rev. Psychol.* **46** 561–584.

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

STEEL, R. G. D. and TORRIE, J. H. (1980). *Principles and Procedures of Statistics: A Biometrical Approach*, 2nd ed. McGraw-Hill, New York.

TAMHANE, A. C. (1996). Multiple comparisons. In *Handbook of Statistics* (S. Ghosh and C. R. Rao, eds.) **13** 587–629. North-Holland, Amsterdam.

TAMHANE, A. C. and DUNNETT, C. W. (1999). Stepwise multiple test procedures with biometric applications. *J. Statist. Plann. Inference* **82** 55–68.

TROENDLE, J. (2000). Stepwise normal theory tests procedures controlling the false discovery rate. *J. Statist. Plann. Inference* **84** 139–158.

WASSMER, G., REITMER, P., KIESER, M. and LEHMACHER, W. (1999). Procedures for testing multiple endpoints in clinical trials: an overview. *J. Statist. Plann. Inference* **82** 69–81.

WELLER, J. I., SONG, J. Z., HEYEN, D. W., LEWIN, H. A. and RON, M. (1998). A new approach to the problem of multiple comparison in the genetic dissection of complex traits. *Genetics* **150** 1699–1706.

WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling Based Multiple Testing*, Wiley, New York.

WILLIAMS, V. S. L., JONES, L. V. and TUKEY, J. W. (1999). Controlling error in multiple comparisons, with special attention to the National Assessment of Educational Progress. *J. Behav. Educ. Statist.* **24** 42–69.

YEKUTIELI, D. and BENJAMINI, Y. (1999). A resampling based false discovery rate controlling multiple test procedure. *J. Statist. Plann. Inference* **82** 171–196.

SCHOOL OF MATHEMATICAL SCIENCES
DEPARTMENT OF STATISTICS
AND OPERATIONS RESEARCH
TEL AVIV UNIVERSITY
RAMAT AVIV, 69978 TEL AVIV
ISRAEL
E-MAIL: benja@math.tau.ac.il
yekutiel@post.tau.ac.il