Journal of Educational and Behavioral Statistics
Spring 2000, Vol. 25, No. 1, pp. 60–83

On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics

Yoav Benjamini
Tel Aviv University

Yosef Hochberg
Tel Aviv University

Keywords: *bonferroni type procedures, meta-analysis, multi-center, multiple comparison procedures,* p-*values, stepwise procedures, subset analysis*

A new approach to problems of multiple significance testing was presented in Benjamini and Hochberg (1995), which calls for controlling the expected ratio of the number of erroneous rejections to the number of rejections—the False Discovery Rate *(FDR). The procedure given there was shown to control the FDR for independent test statistics. When some of the hypotheses are in fact false, that procedure is too conservative. We present here an adaptive procedure, where the number of true null hypotheses is estimated first as in Hochberg and Benjamini (1990), and this estimate is used in the procedure of Benjamini and Hochberg (1995). The result is still a simple stepwise procedure, to which we also give a graphical companion. The new procedure is used in several examples drawn from educational and behavioral studies, addressing problems in multi-center studies, subset analysis and meta-analysis. The examples vary in the number of hypotheses tested, and the implication of the new procedure on the conclusions. In a large simulation study of independent test statistics the adaptive procedure is shown to control the FDR and have substantially better power than the previously suggested FDR controlling method, which by itself is more powerful than the traditional familywise error-rate controlling methods. In cases where most of the tested hypotheses are far from being true there is hardly any penalty due to the simultaneous testing of many hypotheses.*

Multiple independent tests of related hypotheses are often encountered in practice, for example when the same theoretical effect is studied under changing conditions in substudies, and interest lies both at the overall conclusion and at the level of the particular substudies. A frequently encountered case of such a problem is that of subgroup analysis in a single study. When autonomous test statistics are computed for predetermined subgroups, one actually gets multiple independent tests of related hypotheses. Clearly, inference about the separate subgroups is of interest. At the same time, conducting the analysis for many subgroups and highlighting or reaching decisions about the selected few that come out to be statistically significant raises a danger that the conclusions from the study will not be a result of a real phenomenon but merely reflect the selection of the extremes among the extensively tested noise.

In order to reduce this danger, the tests should not be conducted on a per comparison basis. Classical Multiple Comparison Procedures (MCPs) aim, instead, at controlling the probability of committing even a single type-I error within the tested family of hypotheses, hereafter refered to as FWE (Family-Wise Error-rate.) By controlling the familywise error-rate, these procedures also assure the simultaneous correctness of all rejection decisions—all individual statistical discoveries. Thus FWE controlling procedures answer both concerns that, according to Cox (1965), should lead a researcher to control for the multiplicity effect: selection bias, and simultaneous correctness. For example, FWE control has been recommended by Tukey (1977) for the specific problem of subgroup analysis in clinical-trials.

The main problem with such classical MCPs, which hinder their application in applied research, is that since they control the FWE, they tend to have substantially less power than the per comparison procedures of the same levels. Yet often, this 'double feature' offered by FWE controlling methods is not quite needed. This is the case when the overall conclusion from the various individual inferences is not necessarily erroneous as soon as one of them is, yet erroneous selection effect is still of concern. We present later three examples of such cases in some detail. In these instances, lack of multiplicity control is too permissive; the full protection resulting from controlling the FWE is too restrictive.

In Benjamini and Hochberg (1995) we introduced the False Discovery Rate (FDR)—the expected ratio of erroneous rejections to the number of rejected hypotheses—as an appropriate error rate to control in many problems where the control of the familywise error rate is not necessarily implied. The FDR is equal to the familywise error rate when the number of true null hypotheses m_0 equals the number of all hypotheses under test m, so in such a situation controlling the FDR controls the FWE as well. But the FDR criterion is adaptive, in the sense that when some of the tested hypotheses are not true (i.e., $m_0 < m$), the FDR is smaller, and more so when more of the hypotheses are not true. Hence FDR controlling procedures can be more powerful than FWE controlling procedures at the same level. In Benjamini and Hochberg (1995), hereafter BH, such a procedure was given for independent test statistics and the gain in power was shown to be large.

Returning to the classical procedures, the maximal familywise error rate is usually achieved when $m_0 = m$. When $m_0 < m$ the familywise error rate of a single step procedure is usually lower than the designated level, suggesting a potential gain in power with adaptive familywise error rate controlling procedures. In Hochberg and Benjamini (1990) an adaptive familywise error rate controlling procedure was constructed to utilize this potential: based on a modification of the graphical method of Schweder and Spjøtvoll (1982), the number of true hypotheses m_0 was estimated from the observed p-values, and was used to modify in a simple way the Bonferroni and the Bonferroni type procedures of Holm (1979) and Hochberg (1988).

A similar phenomenon occurs with the FDR controlling procedure of BH. It was also shown to control the FDR at a level smaller than the desired q, in fact, smaller than $(m_0/m)q$. Therefore, not only is the FDR criterion adaptive, but here too there is room for gaining in power when $m_0 < m$, and such an adaptive FDR controlling procedure is developed and studied in this paper. It is based on the fact that if m_0 was known, the FDR controlling procedure in Benjamini and Hochberg (1995) could be simply adjusted to control the FDR at exactly level q. Instead, the unknown m_0 is first estimated from the data, and then utilized by the adaptive procedure. The end result is still a simple stepwise procedure, which is formulated in this paper both as an algorithm and as a graphical procedure. Three examples, varying in the kind of problem addressed, the number of hypotheses tested, and the implication of the new procedure on the conclusions, are presented in detail, and are used to demonstrate the computations and graphics involved in carrying out the procedure. The operating characteristics of the procedure are studied in the following sections of this paper. We start, in the next section, with a definition of the FDR, a brief review of the non-adaptive procedure of BH and its properties, and some historical background.

The False Discovery Rate and the Non-Adaptive Procedure

Definition

Consider the problem of simultaneously testing m (null) hypotheses, of which m_0 hypotheses are true. Define R to be the number of hypotheses rejected by a given testing procedure, and V is the number of the true null hypotheses erroneously rejected. These are random variables so small letters are used for the realized values. Actually, V is unobservable (unknown) even after the experiment, and the conclusions. In terms of these random variables, the effective per comparison error rate is $E(V/m_0)$, and the familywise error-rate is $P(V \geq 1)$. Testing individually each hypothesis at level α guarantees $E(V/ m_0) \leq \alpha$. The simple Bonferroni procedure, in which each hypothesis is tested at level α/m, guarantees $P(V \geq 1) \leq \alpha$, although better methods that achieve this goal are available (see Shaffer, 1997 for a recent review.)

In BH, the error committed by falsely rejecting null hypotheses is captured by the random variable $Q = V/R$ – the proportion of the rejected null hypotheses which are erroneously rejected (also known as the proportion of false-positive). Naturally, we define $Q = 0$ when $R = 0$, as no error of false rejection can be committed. The *False Discovery Rate* (FDR) is,

$$\text{FDR} = E(Q) = E\left(\frac{V}{R}\right).$$

In testing null values of comparisons of interest, the decision about the direction of the deviation from the null hypothesis, which practically follows a rejection, is often the major goal. Viewing a true null hypothesis as an almost impossible exact equality, a rejection is considered erroneous if it leads to a

wrong conclusion about the direction of the deviation from the null. This approach was emphasized by Tukey (1991), discussed and reviewed in Benjamini et al. (1998), and its extension to the FDR criterion can also be found in Williams et al. (1999): They define a proportion counting the relative number of directional errors among the rejections. A second version, considered here, acknowledges that exact null values may occur, so adds their erroneous rejections to the count of directional errors. The resulting criterion can still be refered to as the (directional) FDR, this time FDR standing for False Directional Decision Rate.

Properties

Two properties of the FDR, which are easy to show (see BH), are very important:

(a) If all null hypotheses are true, the FDR is equivalent to the familywise error.

(b) When $m_0 < m$, the FDR is smaller than or equal to the familywise error, and the two error rates can be quite different. As a result, any procedure which controls the familywise error rate also controls the FDR. However, if FDR control rather than familywise error rate control is desired, then a gain in power can be expected.

Note that as the number of type I errors increases, so does Q (if R is fixed), but with the same number of type I errors committed, the more hypotheses are rejected the less severe the associated errors are conceived: making 5 errors in 10 rejections might not be acceptable whereas making 5 errors out of 100 might well be.

Procedure

Denote by $P_{(1)} \leq \ldots \leq P_{(i)} \leq \ldots \leq P_{(m)}$ the ordered set of p-values corresponding to the tested hypotheses, and by $H_{(1)} \leq \ldots \leq H_{(i)} \leq \ldots \leq H_{(m)}$ the corresponding hypotheses. The following multiple testing procedure has been shown in BH to control the FDR at level q for independent test statistics:

$$\text{Let } k \text{ be the largest } i \text{ for which } P_{(i)} \leq \frac{i}{m} q, \quad \text{then reject all } H_{(i)}\ i = 1,2,\ldots,k. \tag{1}$$

It follows from Seeger (1968) and Hommel (1988) that this procedure does not control the FWE in the strong sense (under all configurations of true and false hypotheses). The power study in the latter part of this paper compares the power of this FDR controlling procedure with that of several familywise error rate controlling procedures, and shows it to enjoy substantially more power.

Historical Perspective

Our proposal for FDR control in BH was motivated by the paper of Soriç (1989) which was a strong and emotional call for the necessity to controlled inference because of the increased error resulting from multiple inferences.

Otherwise, he warned, the expected number of 'false discoveries' becomes large relative to the number of discoveries. Since BH has been published we have learned of independent previous efforts in the direction in which we went: looking for suitable error control in the face of multiplicity, when the full protective power of the FWE is not necessary.

Shaffer (1995) has noted that an informal effort in this direction had already been attempted by Elkund in an unpublished work in Swedish. This work has been reported by Seeger (1968) who also attributes the procedure in the previous section to Elkund. Seeger proved that when all tested null hypotheses are true the procedure controls the FWE at level q, but when some hypotheses are true while other are false (i.e. when $m_0 < m$), this is not the case. Apparently Seeger's second result, that the procedure does not always control the FWE at the desired level, had diminished the interest in the procedure at the time it was proposed, to the point it became completely unknown (e.g., no mentioning in Hochberg & Tamhane, 1987).

Independently, Simes (1986) proposed a global test of the single intersection hypothesis that all hypotheses are true: reject the intersection hypothesis if there exists an i for which $P_{(i)} \leq \frac{i}{m} q$. He gave a nice proof (by induction) of the error controlling property of the test, which is essentially Seeger's first result. Simes suggested Equation (1) as an informal multiple testing procedure, but then Hommel (1988) showed—as Seeger had done before—that it does not control the FWE in the strong sense. Therefore, in the realm of FWE control, the procedure cannot be used for making the multiple inferences about the individual hypotheses. It can be, and was, used to derive several other testing procedures, for example by Hochberg (1988) and Hommel (1988), but these procedures are less powerful. Sen (1999a) points out that this equality is actually the classical Ballot Theorem related to uniform order statistics. Revived interest in the procedure as a multiple testing procedure came in view of the FDR criterion it controls (BH): see, for example, its implementation in the new SAS MULTPROC software. For a review of the global testing procedure and its extensions see Hochberg and Hommel (1997).

Additional efforts were brought to our attention by Professor Victor. In an editorial note (Victor, 1982), he had discussed the need to relax the control of the FWE, in order to be able to use exploratory data analysis in clinical research, an activity that requires the testing of many hypotheses. He proposed informally a few possible directions: to control $P(V \geq k) \leq \alpha$, for some prechosen $k < 1$, to control $P(V \geq \gamma m) \leq \alpha$ for some prechosen proportion γ, or to supplement a set of individually rejected hypotheses by reporting an observed FDR-like quantity (which is actually $Q_e(\alpha)$ in Section 3 below). Harvånek and Chytil (1983) gave a procedure that controls $P(V \geq \gamma m)$ in the independent case. Hommel and Hoffmann (1987) gave single step and stepwise procedures that allows to control $P(V \geq k) \leq \alpha$, or the multiple k-level α, as they called this error-rate (see also Helperin et al., 1988). However, they also mentioned that they could not find any suitable procedure satisfying the FDR-like criterion.

An Adaptive FDR Controlling Approach

The proof that the procedure in (1) controls the FDR is based on the lemma that for any m_0 independent test statistics corresponding to true null hypotheses, and for any values that the $m_1 = m - m_0$ values $P_{m_0+1}, \ldots, P_m$ corresponding to the false null hypotheses may take, the multiple test procedure satisfies the inequality

$$\mathrm{E}(Q) \mid \mathrm{P}_{m_0+1} = p_{1,,}, \ldots, P_m = p_{m_1}) = \frac{m_0}{m} q. \tag{2}$$

Therefore, when not all hypotheses are true, that is when $m_0 < m$, the procedure controls the FDR at a level too low. So even though the procedure is sensitive to the configuration of the hypotheses because of its step-wise nature, further gain can be made by seeking an approach which is more adaptive to the actual configuration. Such an adaptive approach is outlined below.

When the individual tests are conducted at level α, we can use the results of the experiment to estimate the False Discovery Rate. If m_0 was known then $\mathrm{E}(V) = m_0\alpha$, $r(\alpha) = v + s$ is an observation on the random variable in the denominator, so an estimate of $Q_e(\alpha)$ is $m_0\alpha / r(\alpha)$. If m_0 is not known but can be estimated, a prediction for V can be based on an estimate of m_0 through $\hat{v}(\alpha) = \hat{m}_0\alpha$, so finally, an estimate of the FDR Q_e (α) is

$$\hat{Q}_e(\alpha) = \hat{v}(\alpha) / r(\alpha) = \hat{m}_0\alpha / r(\alpha).$$

The level α can now be adapted to the (estimated) configuration of the tested hypotheses, by maximizing the observed number of rejections $r(\alpha)$ subject to the constraint on the estimated FDR. The following three components summarize this adaptive approach for controlling the FDR in multiple comparisons.

(1) Specify an allowable false discovery rate q (prior to experimentation).

(2) Use the results of the experiment to estimate m_0 $(\hat{m}_0)$.

(3) Choose α that maximizes $r(\alpha)$ under the constraint

$$\hat{Q}_e(\alpha) = \alpha\hat{m}_0 / r(\alpha) \leq q.$$

This maximization problem can be shown (as in BH Theorem 2) to have the following solution:

$$\begin{gathered}\text{Let } k \text{ be the largest } i \text{ for which } P_{(i)} \leq \frac{i}{\hat{m}_0} q, \\ \text{then reject all } H_{(i)}\ i = 1,2, \ldots, k.\end{gathered} \tag{3}$$

It is obvious that if $\hat{m}_0 \geq m_0$ (with probability 1) and if the test statistics remain independent given $\hat{m}_0$, it follows from the lemma that

$$E(Q \mid P_{m_0+1} = p_{1,,}, \ldots, P_m = p_{m_1}) \leq \frac{m_0}{\hat{m}_0} q \leq q.$$

The inequality can be integrated over the distribution of the test statistics corresponding to the non null hypotheses, and so the procedure in Equation (3) controls the FDR at q.

Thus, if an estimator of m_0 fulfils the above conditions, then controlling $\hat{Q}_e$ in that manner (i.e., examining $\hat{Q}_e(\alpha)$ post factum, then choosing α, etc.) is justified in the independent case. In fact using $\hat{m}_0 \equiv m$ is equivalent to the procedure in (1). The actual performance of procedures which are based on estimators of m_0 that are not almost surely conservative but only upward biased, and which depends somewhat on the p-values corresponding to true null hypotheses, has to be checked and demonstrated for finite samples under various model configurations and parameters. In the next section we present a simple procedure for estimating m_0.

Remark 1

It is important to emphasize that estimating, or even knowing, the number of true null hypotheses, and thus the number of false, is not equivalent to concluding which ones are false. Accordingly, there is no need to postulate any requirement of equality between the estimated $m - \hat{m}_0$ and r, because the latter is the number of hypotheses for which we can conclude that they are false. To demonstrate this point, consider $m_0 = m - 1$, and p'_1 being the p-value corresponding to the single non null hypothesis. Suppose it is known that there is exactly one false hypothesis, and therefore it is required that exactly one hypothesis has to be rejected. Obviously the hypothesis having the smallest p-value is rejected. Also for any fixed p'_1, and for large enough m, $p_1 \leq p'_1$ with high probability. Therefore, if we insist on consistency no error rate can be controlled. This remark, however, does not imply that it might not turn out useful to add some requirement of consistency between r and $m - \hat{m}_0$ in order to improve the performance of a specific procedure.

Remark 2

In the current approach an hypothesis may be rejected even though its p-value is not below q. Consider, for example, twenty tested hypotheses, 19 false with p-values close to 0, and one true hypothesis whose p-value is p_1. When $p_1 \leq .5$, the later defined estimator yields $\hat{m}_0 = 2$, H_1 is rejected because $p_1 = p_{(20)} \leq .05*20/2$. This need not be of concern, since we view the various comparisons in a simultaneous framework, and against the background of the many hypotheses rejected in such a study controlling the FDR allows to reject such an hypothesis as well. Nevertheless, one may wish to require that a necessary condition for rejecting an hypothesis is that its p-value be less than some α^* (possibly $\alpha^* = q$ or even $\alpha^* = .5$). One justification for such a point of view might be the concern that a rejected hypothesis will be further highlighted, and/or separated from the rest—either by reporting or by quoting. A second justification is the concern against contaminating a large set of rejected hypotheses of which we are quite sure, with other hypotheses which are quite surely null because of their very high p-values. If these issues are of concern, our suggested approach can be easily amended by adding a second constraint in the maximization problem (3):

(3)′ Choose α that maximizes $r(\alpha)$ under the constraints

$$(3.1)\quad \hat{Q}_e(\alpha) = \alpha \hat{m}_0 / r(\alpha) \leq q$$

$$(3.2)\quad \alpha \leq \alpha^*.$$

The procedures suggested below can all be easily modified to allow for the additional constraint (3.2), because $r(\alpha)$ is monotone in α. Practical experience with the method will tell whether this additional constraint is important enough in order to use (3)′ rather than (3).

Estimating the Number of True Null Hypotheses m_0

The method for estimating m_0 is motivated by the graphical approach proposed by Schweder and Spjøtvoll (1982), as developed and presented in Hochberg and Benjamini (1990).

If all null hypotheses are true, that is, $m = m_0$, and the test statistics independent, the set of observed p-values $p_{(i)}$'s can be considered as a realization of an ordered sample from the uniform distribution over [0,1]. The expected value of the i-th p-value is thus $E(P_{(i)}) = i/(m + 1)$. The plot of $p_{(i)}$ versus i (the quantile plot of the p-values) should exhibit linear relationship, along a line of slope $S = 1/(m + 1)$ passing through the origin and the point $(m + 1,1)$.

When $m_0 < m$, the p-values corresponding to the false null hypotheses tend to be smaller than the p-values corresponding to the true null hypotheses, so they concentrate on the left side of the plot. The relationship over the right side of the plot remains approximately linear, with slope $\beta = 1/(m_0 + 1)$. Using a suitable set of the largest p-values, fit a straight line through the point $(m + 1, 1)$ with slope $\hat{\beta}$, and use it to estimate m_0 by $\hat{m}_0 = 1/\hat{\beta}$. This estimator of m_0 is biased, when $\hat{\beta}$ is an unbiased estimator of β, but the bias is upwards, thereby leaning towards the conservative side. Figure 1 displays the quantile plots of the $p_{(i)}$'s for three data sets discussed in a later section. The largest 6 among the 9 on the top panel, the largest 3 or 4 among the 34 in the middle panel, and the largest 40 to 50 $p_{(i)}$'s among the 91 displayed in the bottom one, do seem to lie on some line through the upper-right corner of the plot.

How many of the largest $p_{(i)}$'s should be used to fit the line? When the points are along the line, the estimated slopes corresponding to subsets tend to vary asystematically. On the other side, the p-values which correspond to the false hypotheses are small, and the slopes will tend to be smaller (see an example in Figure 1). A possible approach would be to use some change point detection method to find the end of the linear part. Our concern for a conservative estimator, though, suggests starting from the smallest $p_{(i)}$'s, and working towards the linear part. Therefore, as in Hochberg and Benjamini (1990), we fit lines progressively to smaller number of the largest $p_{(i)}$'s and the first decrease in the estimated slope is an indication of where to stop excluding additional $p_{(i)}$'s. The successive lines are demonstrated in Figure 2, for the first panel of Figure 1. For more rationale see the points below.

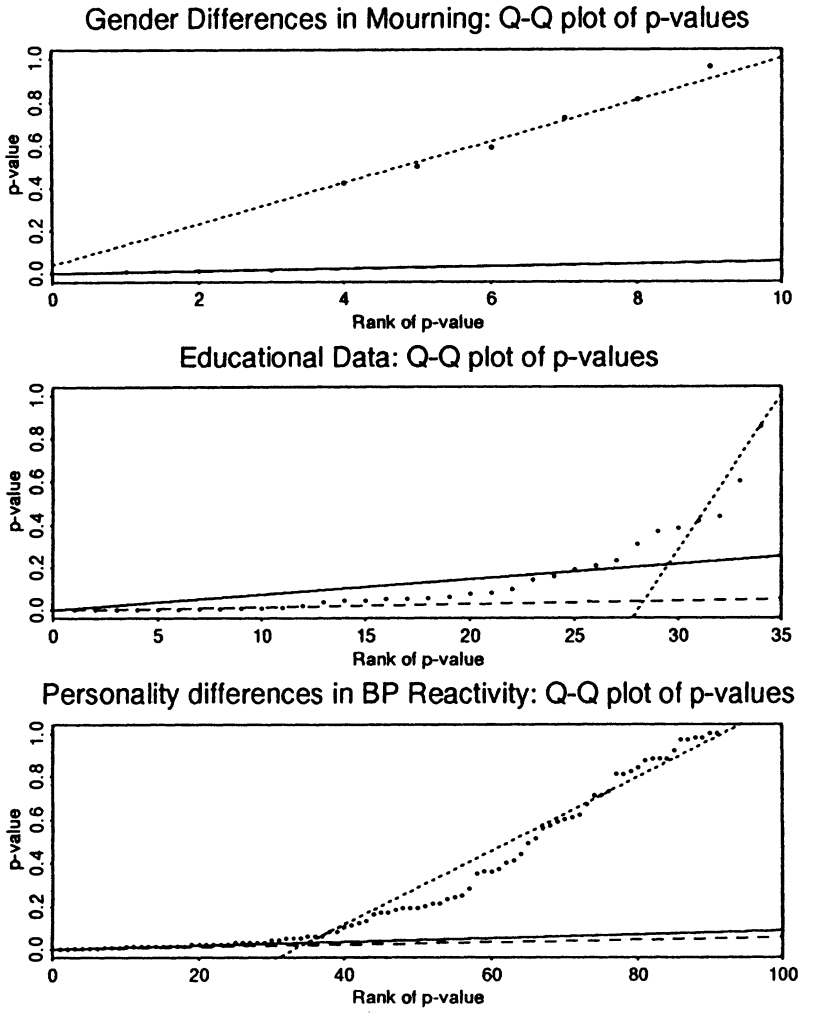

FIGURE 1. *The quantile plot of the* $p_{(i)}$*'s for three data sets. The largest 6 among the 9 on the top panel, the largest 3 or 4 among the 34 in the middle panel, and the larget 40 to 50* $p_{(i)}$*'s among the 91 displayed in the upper-right corner of the plots. This slope is used to estimate the number of true null hypotheses* m_0 *for each study. Below the solid line are points which satisfy the constraint* $p_{(i)} \leq 0.05i/\hat{m}_0$*. Below the dashed line are points which satisfy the constraint* $p_{(i)} \leq 0.05i/m$.

FIGURE 2. *The graphical method for estimating* m_0. *On the quantile plot of the* $p_{(i)}$*'s for the top data set of Figure 1, we fit lines progressively to 9, 8, and 7 largest* $p_{(i)}$*'s, and the slope increases. Fitting the line next to the 6 largest* $p_{(i)}$*'s, and the slope first decreases. Based on this line the estimated number is 9.*

We have tried different methods for estimating m_0 from $p_{(i)}, p_{(i+1)}, \ldots, p_{(m)}$. These included, among others, ordinary (unweighted) least squares estimator of the slope of the line restricted to pass through the point $(m + 1,1)$, and the Lowest SLope (LSL) estimator $(1 - p_{(i)})/(m + 1 - i)$ which is the slope of the line passing through the points $(m + 1,1)$ and $(i, p_{(i)})$ (see also property [a] below). The simple LSL method was finally chosen, and in conjunction with the stopping rule we get the following simple stepwise algorithm:

Calculate $S_i = (1 - p_{(i)})/(m + 1 - i)$, the i-th slope estimate.
Starting with $i = 1$, proceed towards larger i as long as $S_i \geq S_{i-1}$,
stop when the first time $S_j < S_{j-1}$, and use

$$\hat{m}_0 = min[(1/S_j + 1), m],$$

that is, the smaller between m and the integer larger than the inverse of the slope. (The latter step introduces further bias in the conservative direction).

This method, in addition to being simple to perform (either graphically or by a calculator), has other good properties as well.

(a) Assuming that the largest $m + 1 - j$ p-values $p_{(j)}, \ldots, p_{(m)}$ correspond to some of the m_0 p-values of the truly null hypotheses, the gaps $g_{(i)} = p_{(i+1)} - p_{(i)}$ are independent and exponentially distributed, each with the expectation of $1/(m_0 + 1)$ (using the notation $g_{(m)} = 1 - p_{(m)}$). The average of the $m + 1 - j$ gaps between the largest p-values is exactly $(1 - p_{(j)})/(m + 1 - j) = S_j$. Thus the LSL is also the unbiased maximum likelihood estimator of the slope. It is also approximately a weighted median of slopes, as can be found in Benjamini and Hochberg (1993).

(b) The condition $S_i \geq S_{i-1}$ is equivalent to $(p_{(j)} - p_{(j-1)})/(1 - p_{(j-1)}) \leq 1/[m + 1 - (j - 1)]$. The value $1/[m + 1 - (j - 1)]$ is the expected value of the normalized gap on the left under the assumption that all p-values greater or equal to $p_{(j-1)}$ correspond to true null hypotheses. Thus the stopping rule is equivalent to dropping a small p-value from the estimation of m_0 if its gap to its larger neighbor is smaller than its expected value, indicating a suspected false hypothesis.

(c) If $p_{(j)} \leq q/m$ for $j = 1, 2, \ldots, h$, so that these h hypotheses are rejected even according to the conservative Bonferroni procedure, then the LSL estimate of the slope (according to the stopping rule) will not depend on the corresponding leftmost h $p_{(i)}$'s. This follows easily from property (b).

(d) If the slope is estimated from $p_{(i)}$, $i = j, j + 1, \ldots, m$, then $\hat{m}_0 \geq 1/S_j = (m + 1 - j)/(1 - p_{(j)}) \geq m + 1 - j$, so the estimate is always larger than the number of p-values used for the estimation.

It is evident that any estimating procedure might run into difficulties when $m_0 = m$, as bounding $\hat{m}_0$ by m makes the expected value of $\hat{m}_0$ less than m. We protect against this situation by rejecting anything by the adaptive method only if some hypothesis is rejected by the procedure of BH (i.e., using first $\hat{m}_0 = m$).

It should be emphasized that in addition to the attractive features discussed above, the most important property of this estimation procedure of m_0 is its good performance as the first step in the procedure outlined below. It was chosen among the many more procedures tried according to how well the combined procedure increased the power while controlling the FDR.

Remark 3

Although this work deals with independent test statistics, it is worth mentioning that this graphically motivated estimation procedure can turn out to be useful in cases of dependency as well. Without the independence assumption we do not have the distributional properties of the p_i's and their gaps, which were used above, but the global relationship remains similarly linear. Since $i = \#\{j \mid P_{(j)} \leq P_{(i)}\} = n(P_{(i)})$, the linear relationship is between $E(P_{(i)})$ and the fixed $n(P_{(i)})$. If p is fixed instead, $n(p)$ becomes random and the linear relationship holds between p and $E(n(p))$, $p = E(n(p))/m$, where the latter holds even in case of dependency. See also Schweder and Spjøtvoll (1982). While the adaptive

procedure is more powerful by construction, both its FDR control properties and the extent of the power benefit in the dependent case are not studied in this work.

Remark 4

The above remark applies to the case of discrete test statistics as well. The gaps are not exponentially distributed but the above relationship holds. Practically, discrete test statistics may cause the stoping rule to be exercised too soon, leading to valid but possibly too conservative estimation of m_0.

The Adaptive Procedure

The Procedure

Bundling together the stepwise solution to the maximization problem, with the LSL version of estimating m_0, we get the following adaptive procedure:

1. Order the p-values.
2. Compare each $p_{(i)}$ to iq/m. If none is found to be smaller do not reject any hypothesis and stop.
3. Calculate the slopes $S_i = (1 - p_{(i)})/(m + 1 - i)$.
4. Starting with $i = 1$, proceed as long as $S_i \geq S_{i-1}$. When for the first time $S_j < S_{j-1}$ stop. Set $\hat{m}_0 = min([1/S_j + 1], m)$.
5. Starting now with the largest p-value $p_{(m)}$, compare each $p_{(i)}$ to $iq/\hat{m}_0$ until reaching the first p-value that satisfies $p_{(k)} \leq kq/\hat{m}_0$. Reject all k hypotheses having p-values smaller than $p_{(k)}$.

For a detailed computational example see Table 1 below.

A Graphical Implementation of the Procedure

1. Draw the quantile plot of $p_{(i)}$ versus i.
2. Draw lines from the point $(m + 1, 1)$ to each $(i, q_{(i)})$, starting with $i = 1$. Proceed as long as the newly considered point $(i, q_{(i)})$ is below the previous line. When for the first time the new $(j, q_{(j)})$ is above the line—stop.
3. Continue the line to $(j, q_{(j)})$ till it intersects the horizontal axis. The number of integers to the right of the intersection point until and including $m + 1$ is the estimate $\hat{m}_0$.
4. Draw two lines starting from the origin, one passing through the point (m, Q) the other passing through the point $(\hat{m}_0, Q)$.
5. As long as there is at least one point below the lower line, choose among the points below the upper line the rightmost point. The corresponding p-value is the desired α. Reject all hypotheses having smaller p-values.

It is immediate to confirm that this graphical version is equivalent to the stepwise one discussed earlier, because the comparisons of ratios become comparisons of angles. See Figure 2 for a detailed example of the graphical procedure. A short *S* program that performs the above procedure on a given set of p-values (including the graphics) can be found at the homepage http:\\www.math.tau.ac.il/~benja.

TABLE 1

Subgroup	$p_{(i)}$	i	S_i	ABH/BH
Opponents G11-12	0.9600	9	0.0400	0.0500
Supporters G7-8	0.8094	8	0.0953	0.0444
No Stand G11-12	0.7240	7	0.0920	0.0389
No Stand G7-8	0.5870	6	0.1033	0.0333
Opponents G7-8	0.4989	5	0.1002	0.0278
Supporters G11-12	0.4241	4	0.0960	0.0222
Opponents G9-10	0.0133	3	0.1410	0.0167*
No Stand G9-10	0.0098	2	0.1238	0.0111*
Supporters G9-10	0.0074	1	0.1103	0.0056*

Examples

Three examples where the new adaptive procedure is used are given below. The first involves testing hypotheses about gender difference in mourning in 9 subgroups; in the second example an educational change is tested in 34 states; the last one is a meta-analysis of 91 studies.

Youths' Mourning After Rabin's Assassination

The assassination of the Israeli Prime Minister Itzhak Rabin in November of 1995, following the peace effort he had been leading, has resulted in shock and grief among the Israelies, and the youth in particular. One of the most visible expressions of mourning was candle lighting in various public locations all over Israel, most notably at the square where he was shot.

Raviv et al. (1998) used this mass expression of grief to study many psychological and behavioral aspects of the mourning behavior, by interviewing a large sample of high school children. They were interested in different forms of expressing grief and what factors affected behavior. The original study is very comprehensive, but for the purpose of this example we restrict our attention to a single question: are there gender differences in the proportion of children choosing to express their mourning by lighting candles in public, within subsets of the sample, generated by stratifying on age group (identified by school grades 7th and 8th, 9th and 10th, 11th, and 12th) and political orientation regarding Rabin's Peace Policy (supporters, opponents, and those with no stand). This specific question was raised by the investigators after studying the effects of these factors and identifying an interaction. The proportion participating in candle lighting was compared between boys and girls in each of the 9 subsets generated, and the individual statistical significance assessed at each subgroup using a chi-square test. The results are given in Table 1, the nine subgroups are ordered from the largest observed p-value, at the top, to the smallest at the bottom.

Using the Bonferroni procedure at the .05 level yields no significant gender difference: the lowest p-value is .0074 > .0055 = .05/9. Using Hochberg's proce-

dure leaves us again with no significant gender difference in any subset, so controlling the FWE does not allow the detection of any significant difference. Turning to the FDR controlling procedures, each $p_{(i)}$ is compared to $0.05i/9$. The third p-value, 0.013, is the largest p-value smaller than that, that is, smaller than $3*.05/9 = 0.0166$. The three subgroups for which there is a gender difference are those three corresponding to the middle age group, where girls show higher participation than boys. Trying to use the current adaptive procedure we get the first decrease in slope at S_4, and an estimate of 9 for m_0 (even though only 6 points do seem to lie on a line). Therefore the set of constants to which the p-values are compared is the same. Even though the FDR approach has shown to be useful in this example, the adaptive procedure offers no further improvement (and using 6 as an estimate would not have changed our conclusions either.)

NAEP Trial State Assessment

Williams et al. (1999) discuss the problems of error control in large studies giving specific attention to problems arising in the National Assessment of Educational Progress (NAEP). The change in the average eighth-grade mathematics achievement scores for the 34 states that participated in both the 1990 and the 1992 NAEP Trial State Assessment is adapted from their study, and displayed in Table 2. Both the overall change at the national and the changes in specific States are of interest, since the methods used to enhance mathematics achievements in the individual States are not the same. Williams et al. (1999) discuss the FDR as an appropriate and attractive error to control in this and similar situations, and recommend its use. The 34 states are ordered from the largest observed p-value at the top, to the smallest at the bottom. Both Bonferroni and Hochberg's FWE controlling procedures identify only 4 significant results, those with p-value less than $p_{(4)} = .0002$ (bold type in column 2 of Table 2). Using the FDR controlling procedure of BH (under the heading of which the constants $.05i/34$ appear) yields significant results for 11 states, those with p-value less than .01618. Note that the S_i in column 4 of Table 2 is increasing in i all the way up to S_{33} and decreases to S_{34}. Thus m_0 is estimated from S_{34} to be $min\ ([1/0.14372 + 1],\ 34) = 7$. The column under the heading ABH displays $.05i/7$ to which the $p_{(i)}$ are compared stepping from $p_{(34)}$ downwards. The first p-value to be smaller than its constant is $p_{(24)}$. Hence, using the new adaptive FDR controlling procedure finds 24 significant results, all those with p-value less or equal to 0.1572, and more than would have been found if the individual .05 level was used. This study is a case where individual inferences may get separated from the combined study and used separately, by the authority of an individual state, so there seems to be a good reason to further limit declaring a significant result, to the 16 states with p-value $\leq .05$. Thus, using the new procedure to control for possible multiplicity effect in this example incurred no loss in power.

TABLE 2

State	*Change*	*i-th p-value*	*i*	S_i	*ABH*	*BH*	*BON*
GA	−0.323	0.85628	34	0.14372	0.24286	0.05000	
AR	−0.777	0.60282	33	0.19859	0.23571	0.04853	
AL	−1.568	0.44008	32	0.18664	0.22857	0.04706	
NJ	1.565	0.41998	31	0.14501	0.22143	0.04559	
NE	1.334	0.38640	30	0.12272	0.21429	0.04412	
ND	1.526	0.36890	29	0.10518	0.20714	0.04265	
DE	1.374	0.31162	28	0.09834	0.20000	0.04118	
MI	2.215	0.23522	27	0.09560	0.19286	0.03971	
LA	2.637	0.20964	26	0.08782	0.18571	0.03824	
IN	2.149	0.19388	25	0.08061	0.17857	0.03676	
WI	2.801	0.15872	24	0.07648	0.17143*	0.03529	
VA	2.859	0.14374	23	0.07136	0.16429*	0.03382	
WV	2.331	0.10026	22	0.06921	0.15714*	0.03235	
MD	3.399	0.08226	21	0.06555	0.15000*	0.03088	
CA	3.777	0.07912	20	0.06139	0.14286*	0.02941	
OH	3.466	0.06590	19	0.05838	0.13571*	0.02794	
NY	4.893	0.05802	18	0.05541	0.12857*	0.02647	
PA	4.303	0.05572	17	0.05246	0.12143*	0.02500	
FL	3.784	0.05490	16	0.04974	0.11429*	0.02353	
WY	2.226	0.04678	15	0.04766	0.10714*	0.02206	
NM	2.334	0.04650	14	0.04540	0.10000*	0.02059	
CT	3.204	0.04104	13	0.04359	0.09286*	0.01912	
OK	4.181	0.02036	12	0.04259	0.08571*	0.01765	
KY	4.327	0.00964	11	0.04127	0.07857*	0.01618*	
AZ	4.994	0.00904	10	0.03964	0.07143*	0.01471*	
ID	2.956	0.00748	9	0.03817	0.06429*	0.01324*	
TX	5.645	0.00404	8	0.03689	0.05714*	0.01176*	
CO	4.326	0.00282	7	0.03561	0.05000*	0.01029*	
IA	4.811	0.00200	6	0.03441	0.04286*	0.00882*	
NH	4.422	0.00180	5	0.03327	0.03571*	0.00735*	
NC	7.265	0.00002	4	0.03226	0.02857*	0.00588*	*
HI	5.550	0.00002	3	0.03125	0.02143*	0.00441*	*
MN	6.421	0.00002	2	0.03030	0.01429*	0.00294*	*
RI	5.097	0.00000	1	0.02941	0.00714*	0.00147*	*

A Meta-Analysis of Differences in Reactivity Between Type A and B Personalities

Individuals with the type A personality are characterized as being competitive, aggressive, and hostile, and as having a sense of time urgency. Prospective studies have linked type A personality with increased risk of coronary heart disease. Individuals with type B personality are comparatively more relaxed and easygoing, and are less at risk for developing coronary heart disease. The dangers associated with type A personality are frequently attributed to the

hypothesized excessive physiological reactivity to stressful situations of people of this type. More specifically, the view is that in a stressful sitation, type As show greater change in systolic blood pressure, diastolic blood pressure, and heart rate reactivity than type Bs. In a large meta-analysis, Lyness (1993) reviews the problem, the research, and the experimental evidence for such differences between the types. Following a search for relevant literature, 99 studies were identified, with 78 articles studying reactivity differences. Ninety-one studies regarding systolic blood pressure differences in reactivity to stress between the two types were available. The effect sizes estimated d-values, their confidence intervals, and the respective p-values are given in Table 4 of Lyness (1993). The main use of meta-analysis is to see whether the combined evidence from published studies provides a conclusive evidence for an overall effect, and such a significant overall difference between the types was identified in this study. A secondary goal in some meta-analyses, is to identify possible reasons for observing different effects in different studies, possibly due to their design, measurement techniques, populations under study, or even investigators involved. In the Lyness study no such inference was attempted regarding individual studies, may be because the results of only two individual studies out of the 91 came out to be statistically significant at the conventional .05 level, when controlling for the FWE using Bonferroni (the cutoff for significance being .00055). Using Hochberg's procedure leaves us with the same two statistically significant studies.

Using the FDR criterion rather than FWE, many more studies can be identified individually as having enough evidence to yield a significant result. The bottom panel of Figure 1 displays the quantile plot of the p-values, and the line from which m_0 is estimated. The two lines starting at the origin correspond to the constants in BH and to the adaptive ones. Using the new adaptive procedure the cutoff is at .02, and 24 of the reviewed studies gave significant results. While protecting against an overall error at level .05, the new procedure allows the investigator to get into individual studies and try to learn about the sources of the positive statistically significant result.

The control of FDR

The adaptive approach suggested in this paper involves maximizing the number of rejected hypotheses subject to a constraint on the estimated False Discovery Rate. The first question is whether the adaptive method does in fact control the true FDR. Again we discuss only the case of independent test statistics (under the null hypotheses).

For the case of two hypotheses $\hat{m}_0$ is always m, the procedure does not differ from that of BH, and so the actual FDR is less than q. For other values of m, the procedure is designed to control the FDR for $m_0 = m$. For other configurations of tested hypotheses an answer is given by a large simulation study.

Simulation Study Design

The equality of the expectations of independent normally distributed random variables to zero is tested by individual z-tests. The configurations of the hypotheses involve $m = 4$, 8, 16, 32, and 64 hypotheses, and $m_0 = m$, $m - 1$, $m - 2$, $m - 3$, $3m/4$, $m/2$, $m/4$, and 0. The non-zero expectations are placed at four locations over (0,L), at L/4, 2L/4, 3L/4 and L. Three configurations are further considered for the number of non-zero expectations placed at each of these locations:

(E) equal number placed at each location.

(D) decreasing number placed away from 0, and

(I) increasing number placed away from 0.

The variance of the noise was set to 1, and L was chosen at two levels—5 and 10—varying thereby the signal to noise ratio.

The simulation study involved 20000 repetitions. The estimated standard errors are about .0008–.0016. As the same noise was used in a single repetition across all configurations with the same number of hypotheses, and the alternatives in different configurations were monotonically related, a positive correlation was induced. This correlation reduces the variance of a comparison between two methods or two configurations to below twice the variance of a single one. The positive correlation was also used by constructing a difference estimator of the FDR for $m_0 \leq m$ by subtracting the difference between the average simulation based FDR for the nonadaptive procedure and its known expected upper bound.

Identifying Extreme Configurations

In Figure 3 we display the results of the simulations for testing $m = 8$ hypotheses, for configurations (I) and (D). The behavior of the FDR evident in Figure 3 was found across all configurations:

(1) For a fixed number of hypotheses m, and for every configuration, the FDR is monotonically increasing in m_0, and achieves its maximum under the intersection null hypotheses that all hypotheses are true, where it takes the desired level by design.

(2) Comparing across configurations we noticed that for $m_0 < m$ the lowest FDR was recorded for configuration (D), and for small L. This may seem odd at first, as this configuration has its non-null hypotheses clustering close to the null ones. A second thought reveals that because of this closeness more p-values of the non null interfere with those of the truly null, introducing further upward bias in the estimator of m_0 and so the procedure becomes more conservative. The FDR for configuration (E) (not shown in the figure) is in between, as expected.

FDR Control at Extreme Configurations

The monotonicity of the FDR in m_0 allows us to limit further investigation of the control of this error-rate only to m_0 which are close to m. Figure 4 displays

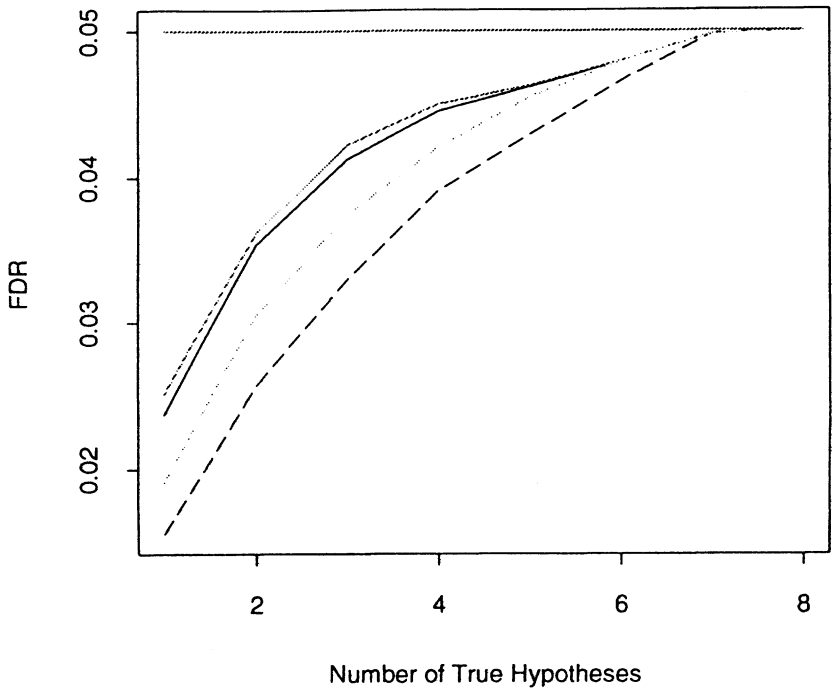

FIGURE 3. *The simulation based estimates of the FDR control by the adaptive method when testing* m = *8 hypotheses,* m_0 *of which are true. Two configurations of non-null hypotheses (I) and (D) and two values for L are shown: L = 5 at (D)—dashed; L = 10 at (D)—solid line; L = 5 at (I)—dashed dotted; L = 10 at (I)—dotted. The FDR is monotonically increasing in* m_0, *is maximal when all hypotheses are true, where it takes the desired level by design.*

the result of the simulations for $m \geq 2$, when $m_0 = m$, $m - 1$, $m - 2$, and $m - 3$ for the least conservative configuration (I) with L = 10. Our simulation study thus establishes that the bounds mentioned before on the actual FDR achieved by the adaptive method, using the LSL version for estimating m_0, hold for the entire range of the configurations investigated.

The Control of the False Directional Rate

Although we have not studied this issue in great detail, the following two observations are important. As noted above the most critical configurations for FDR control turned out to be the cases where m_0 is close to m, and the few false hypotheses are far from being true. It is exactly there that the probability of a directional error is close to 0, so that the directional FDR is also controlled. In the other extreme, consider the case where the directional error is close to its

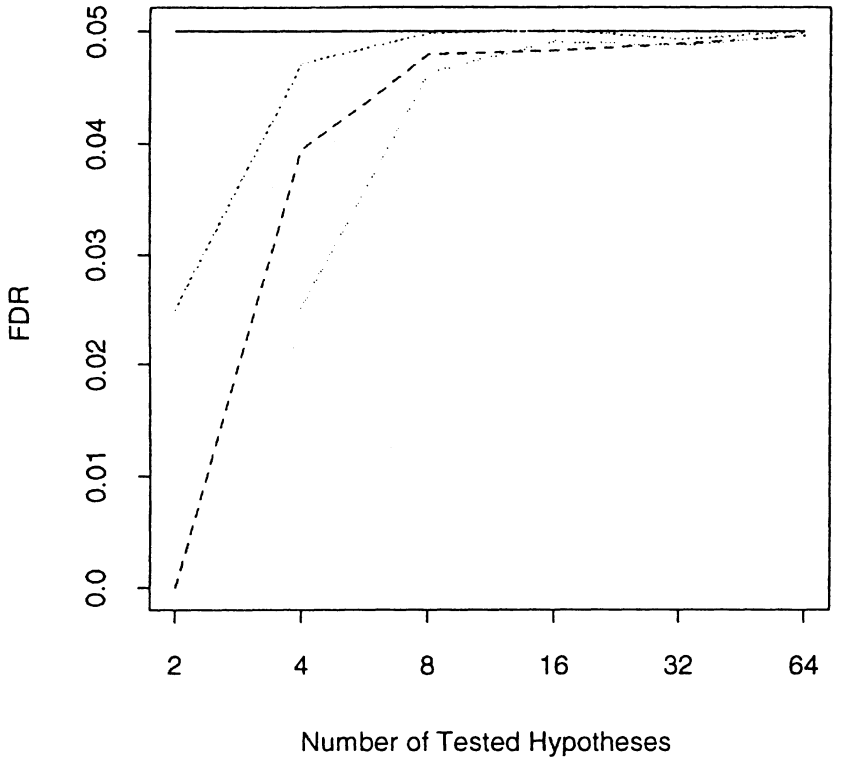

FIGURE 4. *The simulation-based estimates of the FDR control by the adaptive method testing* $m \geq 2$ *hypotheses, when* $m_0 = m$, $m - 1$, $m - 2$, *and* $m - 3$ *(solid, dotted, dashed, and dashed dotted line, respectively), for the least conservative configuration* $L = 10$ *at (I). The FDR is appropriately controlled for the entire range of the configurations investigated.*

maximal level .5, when the individual false hypotheses are almost indistinguishable from the true ones. In this case they will also be indistinguishable in the estimation step of m_0, and $\hat{m}_0$ will tend to overestimate m_0 by their number, say m'_1. Hence, in such a configuration, we have roughly

$$\text{FDR} = E\left(\frac{V}{R}\right) \leq E\left(\frac{V + \frac{1}{2}m'}{R}\right) \leq \frac{m}{m_0 + m'} \frac{m_0 + \frac{1}{2}m'}{m} q \leq q.$$

Further research is required to establish that the above two configurations are actually the extreme (being least favorable).

Some Power Comparisons

A different issue is how the adaptive procedure that controls the FDR compares in terms of power to the non-adaptive procedure that controls the FDR, and to those controlling the familywise error rate. It is clear from the discussion in sections 2 and 6 that the methods that control the FDR are generally more powerful than their counterpart which controls the familywise error rate (in the strong sense), and the gain has been shown in BH to be large. It is also clear that the adaptive FDR control is more powerful than the non-adaptive one. The question remains whether the magnitude of the difference justifies the introduction of adaptiveness. We shall study this question using the setup of the previous section, the number of truly null hypotheses being $3m/4$, $m/2$, $m/4$, and 0. As mentioned before, under the overall null hypotheses the FDR controlling methods control the familywise error rate at level q so we calibrate all methods to control the familywise error rate weakly at the same level by using $q = \alpha$. We do not use the restriction on the individual p-values to be less than α in the ABH (remark 2 of Section 3).

Figure 5 presents the simulation based estimates of the average power (the proportion of the non null hypotheses which are correctly rejected) for four methods: the adaptive FDR controlling method suggested here (AFDR), the non-adaptive FDR controlling method of BH (FDR), the regular Bonferroni (BON), and the adaptive Bonferroni (ABON) of Hochberg and Benjamini (1990), where m_0 is estimated by the LSL.

The power of the methods is ordered as expected uniformly across all configurations: the adaptive methods are more powerful than the non-adaptive ones, the FDR controlling more powerful than their familywise error rate controlling counterparts. Furthermore the non-adaptive FDR is more powerful than the adaptive Bonferroni. When the proportion of true null hypotheses among the tested ones is large, the adaptive versions have very little advantage over the non-adaptive methods. The gain in power due to the adaptation increases as the proportion of truly null hypotheses decreases (down the rows of the display). It also increases when the alternatives are more separated from the null (from left to right across the columns of the display).

The power of the FDR controlling methods is much larger than the familywise error rate controlling procedures, even when the proportion of the true null hypotheses is large. Here, too, the advantage increases as the number of true hypotheses decreases, but it does not change much across configurations.

Studying the power as a function of the number of tested hypotheses we arrive at the most interesting results. It can be clearly observed from each scatterplot that the power of the familywise error rate controlling procedures decreases sharply the more hypotheses are tested—the cost of the control of multiplicity is clearly evident. For the FDR nonadaptive method this behavior is much less pronounced. The loss of power is attenuated the more hypotheses are false, and the further away from the null the false ones are (towards the right-lower display).

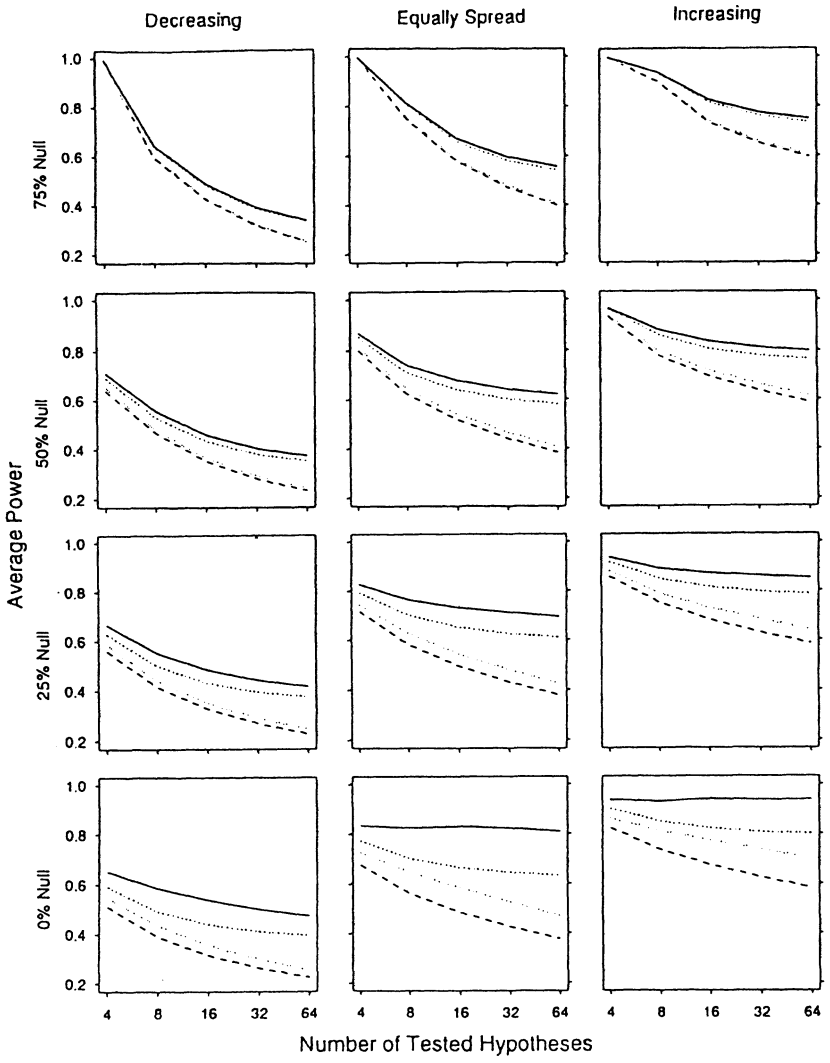

FIGURE 5. *The simulation based estimates of the average power, as a function of the number of hypotheses tested* m, *their proportion which are non-null, and their configuration, for 4 procedures: the adaptive FDR controlling (AFDR)—solid lines; the non-adaptive FDR controlling (FDR)—dotted lines; the adaptive Bonferroni (ABON)—dotted dashed lines; and Bonferroni (BON)—dashed lines. The gain in power due to the adaptation increases down the rows and increases from left to right across the columns. The power of the familywise error rate controlling methods decreases sharply as* m *increases. For (FDR) the loss of power is attenuated towards the lower-right. For (AFDR) the loss from increasing* m *is even smaller, and there is no loss of power at the lower-right scatterplot.*

Turning to the adaptive FDR method we find that the effect of increase of the number of tested hypotheses on the power is even smaller. In fact when all hypotheses are false, and are reasonably well separated from the null, there is no loss of power at all as a result of increasing the number of hypotheses tested! Even before that, say when half of the hypotheses are true, the loss of power from testing 64 hypotheses rather than 16 is small.

Note that the advantage in using the adaptive control of FDR is in some situations extremely large. For example: testing 32 hypotheses, none of which are true, the power of the Bonferroni is .4; the adaptive Bonferroni increases it to .5; the power of the FDR controlling method is .65; and the power of the suggested adaptive procedure is .82, more than double that of the Bonferroni; testing as few as 4 hypotheses a quarter of which are true, the 4 values are about equally spread between .72 and .84 respectively.

Casting these results into a different form, it may be seen that between a quarter to three quarters of the non null hypotheses which were not rejected by the Bonferroni procedure, are now rejected by the adaptive FDR controlling method. This holds for all configurations and number of tested hypotheses when at least half of the hypotheses are non null. And even when only 25% of the hypotheses are non null, the gain in power is such, that about a third of the equally spaced hypotheses which were not rejected before are now rejected.

In Closing

The adaptive approach to multiple significance testing presented in this paper is philosophically different from the classical approach that requires a conservative type-I error rate control, in the sense that it controls the FDR instead, and thus also the familywise error rate, but only in the weak sense.

We have already seen in BH that this point of view on the multiplicity problem is an appropriate one for some practical problems, and at the same time allows much more powerful procedures than those that control the familywise error rate. In this paper we continued this line of research by introducing an adaptive FDR controlling procedure. It was also seen that the advantage of the adaptive FDR controlling method is large, even when compared with the already powerful non-adaptive method, and the cost paid for the control of multiplicity is minimal if many hypotheses are non null.

In this paper we have still limited our attention to methods that control the FDR for independent test statistics. Our initial investigations into this question show that the adaptive procedure suggested here works well in the case of pairwise comparisons and other structured problems. This, however, need not be the only direction: other, more appropriate methods, have been constructed for specific problems (Yekutieli & Benjamini, 1999; Troendle, 2000). Still, work is needed to construct more procedures that control the FDR, whether adaptively or not, for dependent test statistics.

Acknowledgments

We would like to express our special thanks to J. W. Tukey. His comments and questions about many earlier drafts, dating back to 1989, helped us crystallize the approach and the procedure presented here. Joining him are Val Williams and Lyle Jones at NISS, whom we thank for making their results available to us at an early stage for re-analysis. Our thanks are extended to Alona and Amiram Raviv for offering their data and analysis. Finally we thank Yetty Varon for her programming during the initial phase of this study.

References

Benjamini, Y., & Hochberg, Y. (1993). The adaptive control of the false discovery rate in multiple independent testing problems. *Series in Statistics 93.1, Technical Report of the Department of Statistics and OR*, Tel Aviv University, Tel Aviv, Israel.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate—a new and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57,* 289–300.

Benjamini, Y., Hochberg, Y., & Stark, P. B. (1998). Confidence Intervals with more power to determine the sign: Two ends constrain the means. *Journal of the American Statistical Association, 93,* 309–317.

Cox, D. R. (1965). A remark on multiple comparison methods. *Technometrics 2,* 149–156.

Harvånek & Chytil (1983). Mechanizing hypotheses formation—a way for computerized exploratory data analysis? *Bulletin of the International Statistical Institution 50,* 1, 104–121.

Helperin, M., Lan G.K.K., & Hamdy, M. I. (1988). Some implications of an alternative definition of the multiple comparison problem. *Biometrika, 75,* 773–778.

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

Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine, 9,* 811–818.

Hochberg, Y., & Hommel, G. (1997). Step-up multiple testing procedures: Encyclopedia for Statistical Sciences, Supplementary Volume 2 (Ed. Kotz, S).

Hochberg, Y., & Tamhane, A. (1987). *Multiple Comparison Procedures.* NY: Wiley & Sons.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

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

Hommel, G., & Hoffmann, T. (1987). Controlled uncertainty. In P. Bauer, G. Hommel, and E. Sonnemann, (Eds.), *Multiple hypotheses testing* (pp. 154–161). Heidelberg: Springer.

Lyness, S. A. (1993). Predictors of differences between type A and B individuals in heart rate and blood pressure reactivity. *Psychological Bulletin, 114,* 2, 266–295.

Raviv, A., Sadeh, A., Raviv, A., & Silberstein, O. (1998). The reaction of the youth in Israel to the assassination of prime minister Yizhak Rabin. *Political Psychology, 19,* 255–278.

Schweder, T., & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika, 69*, 493–502.

Seeger, P. (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. *Journal of Statistical Planning and Inference, 82* (1–2), 139–145.

Sen, P. K. (1999b). Multiple comparisons in interim analysis. *Journal of Statistical Planning and Inference, 82* (1–2), 5–23.

Shaffer, J. P. (1995). Multiple hypothesis-testing. *Annual Review of Psychology, 46,* 561–584.

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

Soriç B. (1989). Statistical discoveries and effect size estimation. *Journal of the American Statistical Association, 84*, 608–610.

Troendle, J. F. (2000). Stepwise normal theory multiple test procedures controlling the false discovery rate. *Journal of Statistical Planning and Inference, 84* (1–2), 139–158.

Tukey, J. W. (1991). The Philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.

Tukey, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science 198,* 697–684.

Victor, N., (1982). Exploratory data analysis and clinical research. *Methods of Information in Medicine, 21*, 53–54.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1994). Controlling error in multiple comparisons, with special attention to the National Assessment of Educational Progress. *Technical Report #33. Research Triangle Park, NC: National Institute of Statistical Sciences.*

Williams, V.S.L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24,* 42–69.

Yekutieli, D. & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference, 82* (1–2), 171–196.

Authors

YOAV BENJAMINI is Associate Professor of Statistics and Operations Research at Tel Aviv University. His research interests include model selection, signal and image processing, wavelets analysis, genetics, and data mining.

YOSEF HOCHBERG is Associate Professor of Statistics and Operations Research at Tel Aviv University. His research interests include Multiple Comparisons and Biostatistics.

References

Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing

Yoav Benjamini; Yosef Hochberg

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1. (1995), pp. 289-300.

Stable URL:

http://links.jstor.org/sici?sici=0035-9246%281995%2957%3A1%3C289%3ACTFDRA%3E2.0.CO%3B2-E

Confidence Intervals with More Power to Determine the Sign: Two Ends Constrain the Means

Y. Benjamini; Y. Hochberg; P. B. Stark

Journal of the American Statistical Association, Vol. 93, No. 441. (Mar., 1998), pp. 309-317.

Stable URL:

http://links.jstor.org/sici?sici=0162-1459%28199803%2993%3A441%3C309%3ACIWMPT%3E2.0.CO%3B2-M

A Remark on Multiple Comparison Methods

D. R. Cox

Technometrics, Vol. 7, No. 2. (May, 1965), pp. 223-224.

Stable URL:

http://links.jstor.org/sici?sici=0040-1706%28196505%297%3A2%3C223%3AAROMCM%3E2.0.CO%3B2-U

LINKED CITATIONS

Some Implications of an Alternative Definition of the Multiple Comparison Problem
Max Halperin; K. K. Gordon Lan; Mohamed I. Hamdy
Biometrika, Vol. 75, No. 4. (Dec., 1988), pp. 773-778.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198812%2975%3A4%3C773%3ASIOAAD%3E2.0.CO%3B2-2

A Sharper Bonferroni Procedure for Multiple Tests of Significance
Yosef Hochberg
Biometrika, Vol. 75, No. 4. (Dec., 1988), pp. 800-802.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198812%2975%3A4%3C800%3AASBPFM%3E2.0.CO%3B2-N

A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test
G. Hommel
Biometrika, Vol. 75, No. 2. (Jun., 1988), pp. 383-386.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198806%2975%3A2%3C383%3AASRMTP%3E2.0.CO%3B2-%23

Plots of P-Values to Evaluate Many Tests Simultaneously
T. Schweder; E. Spjotvoll
Biometrika, Vol. 69, No. 3. (Dec., 1982), pp. 493-502.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198212%2969%3A3%3C493%3APOTEMT%3E2.0.CO%3B2-%23

A Note on a Method for the Analysis of Significances en masse
Paul Seeger
Technometrics, Vol. 10, No. 3. (Aug., 1968), pp. 586-593.
Stable URL:
http://links.jstor.org/sici?sici=0040-1706%28196808%2910%3A3%3C586%3AANOAMF%3E2.0.CO%3B2-J

An Improved Bonferroni Procedure for Multiple Tests of Significance
R. J. Simes
Biometrika, Vol. 73, No. 3. (Dec., 1986), pp. 751-754.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198612%2973%3A3%3C751%3AAIBPFM%3E2.0.CO%3B2-A