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

Author(s): Yoav Benjamini and Yosef Hochberg

Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1 (1995), pp. 289-300
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2346101
Accessed: 23-10-2018 20:54 UTC

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

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

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

SUMMARY

The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypotheses—the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is desired, there is potential for a gain in power. A simple sequential Bonferroni-type procedure is proved to control the false discovery rate for independent test statistics, and a simulation study shows that the gain in power is substantial. The use of the new procedure and the appropriateness of the criterion are illustrated with examples.

Keywords: BONFERRONI-TYPE PROCEDURES; FAMILYWISE ERROR RATE; MULTIPLE-COMPARISON PROCEDURES; *p*-VALUES

1. INTRODUCTION

When pursuing multiple inferences, researchers tend to select the (statistically) significant ones for emphasis, discussion and support of conclusions. An unguarded use of single-inference procedures results in a greatly increased false positive (significance) rate. To control this multiplicity (selection) effect, classical multiple-comparison procedures (MCPs) aim to control the probability of committing any type I error in families of comparisons under simultaneous consideration. The control of this familywise error rate (FWER) is usually required in a strong sense, i.e. under all configurations of the true and false hypotheses tested (see for example Hochberg and Tamhane (1987)).

Even though MCPs have been in use since the early 1950s, and in spite of the advocacy for their use (e.g. being mandatory for some journals, as well as in some institutions such as the Food and Drug Administration of the USA), researchers have not yet widely adopted these procedures. In medical research, for example, Godfrey (1985), Pocock *et al.* (1987) and Smith *et al.* (1987) examined samples of reports of comparative studies from major medical journals. They found that researchers overlook various kinds of multiplicity, and as a result reporting tends to exaggerate treatment differences (Pocock *et al.*, 1987).

Some of the difficulties with classical MCPs which cause their underutilization in applied research are as follows.

†*Address for correspondence*: Department of Statistics, School of Mathematical Sciences, Sackler Faculty for Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel.
E-mail: benja@math.tav.ac.il

 0035–9246/95/57289

(a) Much of the methodology of FWER controlling MCPs concerns comparisons of multiple treatments and families whose test statistics are multivariate normal (or t). In practice, many of the problems encountered are not of the multiple-treatments type, and test statistics are not multivariate normal. In fact, families are often combined with statistics of different types.

(b) Classical procedures that control the FWER in the strong sense, at levels conventional in single-comparison problems, tend to have substantially less power than the per comparison procedure of the same levels.

(c) Often the control of the FWER is not quite needed. The control of the FWER is important when a conclusion from the various individual inferences is likely to be erroneous when at least one of them is. This may be the case, for example, when several new treatments are competing against a standard, and a single treatment is chosen from the set of treatments which are declared significantly better than the standard. However, a treatment group and a control group are often compared by testing various aspects of the effect (different end points in clinical trials terminology). The overall conclusion that the treatment is superior need not be erroneous even if some of the null hypotheses are falsely rejected.

The first difficulty has been partially addressed by the recent line of research advancing Bonferroni-type procedures, which use the observed individual p-values, while remaining faithful to FWER control: Simes (1986), Hommel (1988), Hochberg (1988) and Rom (1990). The other two difficulties still present a serious problem. This is probably why a per comparison error rate (PCER) approach, which amounts to ignoring the multiplicity problem altogether, is still recommended by some (e.g. Saville (1990)).

In this work we suggest a new point of view on the problem of multiplicity. In many multiplicity problems the number of erroneous rejections should be taken into account and not only the question whether any error was made. Yet, at the same time, the seriousness of the loss incurred by erroneous rejections is inversely related to the number of hypotheses rejected. From this point of view, a desirable error rate to control may be the expected proportion of errors among the rejected hypotheses, which we term the false discovery rate (FDR). This criterion integrates Spjøtvoll's (1972) concern about the number of errors committed in multiple-comparison problems, with Soriç's (1989) concern about the probability of a false rejection given a rejection. We use the term FDR after Soriç (1989), who identified a rejected hypothesis with a 'statistical discovery'.

After some preliminaries, we present in Section 2.1 a formal definition of the FDR. Two immediate but important consequences of controlling this error rate are given: it implies weak control of FWER and it admits more powerful procedures. In Section 2.2 we present some examples where the control of the FDR is desirable. In Section 3 we present a simple Bonferroni-type FDR controlling procedure and the rest of the section is devoted to a discussion and demonstration of its properties. Section 4 presents a simulation study of the power of the procedure.

2. FALSE DISCOVERY RATE

Consider the problem of testing simultaneously m (null) hypotheses, of which m_0 are true. $\mathbf{R}$ is the number of hypotheses rejected. Table 1 summarizes the

TABLE 1
Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

situation in a traditional form. The specific m hypotheses are assumed to be known in advance. **R** is an observable random variable; **U**, **V**, **S** and **T** are unobservable random variables. If each individual null hypothesis is tested separately at level α, then $\mathbf{R} = \mathbf{R}(\alpha)$ is increasing in α. We use the equivalent lower case letters for their realized values.

In terms of these random variables, the PCER is $E(\mathbf{V}/m)$ and the FWER is $\mathbf{P}(\mathbf{V} \geqslant 1)$. Testing individually each hypothesis at level α guarantees that $E(\mathbf{V}/m) \leqslant \alpha$. Testing individually each hypothesis at level α/m guarantees that $P(\mathbf{V} \geqslant 1) \leqslant \alpha$.

2.1. *Definition of False Discovery Rate*

The proportion of errors committed by falsely rejecting null hypotheses can be viewed through the random variable $\mathbf{Q} = \mathbf{V}/(\mathbf{V} + \mathbf{S})$—the proportion of the rejected null hypotheses which are erroneously rejected. Naturally, we define $\mathbf{Q} = 0$ when $\mathbf{V} + \mathbf{S} = 0$, as no error of false rejection can be committed. **Q** is an unobserved (unknown) random variable, as we do not know v or s, and thus $q = v/(v + s)$, even after experimentation and data analysis. We define the FDR Q_e to be the expectation of **Q**,

$$Q_e = E(\mathbf{Q}) = E\{\mathbf{V}/(\mathbf{V} + \mathbf{S})\} = E(\mathbf{V}/\mathbf{R}).$$

Two properties of this error rate are easily shown, yet are very important.

(a) If all null hypotheses are true, the FDR is equivalent to the FWER: in this case $s = 0$ and $v = r$, so if $v = 0$ then $\mathbf{Q} = 0$, and if $v > 0$ then $\mathbf{Q} = 1$, leading to $P(\mathbf{V} \geqslant 1) = E(\mathbf{Q}) = Q_e$. Therefore control of the FDR implies control of the FWER in the weak sense.

(b) When $m_0 < m$, the FDR is smaller than or equal to the FWER: in this case, if $v > 0$ then $v/r \leqslant 1$, leading to $\chi_{(\mathbf{V} \geqslant 1)} \geqslant \mathbf{Q}$. Taking expectations on both sides we obtain $P(\mathbf{V} \geqslant 1) \geqslant Q_e$, and the two can be quite different. As a result, any procedure that controls the FWER also controls the FDR. However, if a procedure controls the FDR only, it can be less stringent, and a gain in power may be expected. In particular, the larger the number of the non-true null hypotheses is, the larger **S** tends to be, and so is the difference between the error rates. As a result, the potential for increase in power is larger when more of the hypotheses are non-true.

2.2. *Examples*

The following examples show the relevance of FDR control in some typical situations. In addition they indicate the desirability of a large number of rejections (discoveries). Both aspects are addressed by the procedure given later in Section 3.

One type of multiple-comparison problem involves an overall decision (conclusion, recommendation, etc.) which is based on multiple inferences. An example of this type of problems is the 'multiple end points problem', which was used earlier to show that FWER control is not always needed. In this example the overall decision problem is whether to recommend a new treatment over a standard treatment. Discoveries here are rejections of null hypotheses claiming that treatment is no better than standard on specified end points. These conclusions about different aspects of the benefit of the new treatment are of interest *per se*, but the set of discoveries will be used to reach an overall decision regarding the new treatment. We wish therefore to make as many discoveries as possible (which will enhance a decision in favour of the new treatment), subject to control of the FDR. Control of the probability of any error is unnecessarily stringent, as a small proportion of errors will not change the overall validity of the conclusion.

Another type of problems involves multiple separate decisions without an overall decision being required. An example of this type is the multiple-subgroups problem, where two treatments are compared in multiple subgroups, and separate recommendations on the preferred treatments must be made for all subgroups. As usual we wish to discover as many as possible significant differences, thereby reaching operational decisions, but would be willing to admit a prespecified proportion of misses, i.e. willing to use an FDR controlling procedure.

The third type involves screening problems, where multiple potential effects are screened to weed out the null effects. One example is screening of various chemicals for potential drug development. Another example is testing multiple factors in an experimental (2^k say) design. In such examples we want to obtain as many as possible discoveries (candidates for drug developments, factors that affect the quality of a product) but again wish to control the FDR, because too large a fraction of false leads would burden the second phase of the confirmatory analysis.

2.3. *Alternative Formulations*

We have suggested to capture the error rate vaguely described as 'the proportion of false discoveries' using the FDR. At this point it might be illuminating to discuss alternative formulations of this concept, and thus to motivate our choice of FDR further.

Undoubtedly, controlling the random variable $\mathbf{Q}$ at each realization is most desirable. This is impossible, for if $m_0 = m$ and even if a single hypothesis is rejected $v/r = 1$ and $\mathbf{Q}$ cannot then be controlled. Controlling $(\mathbf{V}/\mathbf{R}|\mathbf{R} > 0)$ has the same problem—it is identically 1 in the above configuration. Therefore $E(\mathbf{V}/\mathbf{R}|\mathbf{R} > 0)$ cannot be controlled. The FDR, instead, is $P(\mathbf{R} > 0)E(\mathbf{V}/\mathbf{R}|\mathbf{R} > 0)$, and, as will be shown later, this is possible to control.

Second, consider the formulation that Soriç (1989) gave to 'the proportion of false discoveries among the discoveries' as $Q' = E(\mathbf{V})/r$. This quotient is neither the random variable $\mathbf{Q}$ nor its expectation but is a mixture of expectations and realizations. It is not even the conditional expectation of $\mathbf{Q}$, namely $E(\mathbf{Q}|\mathbf{R} = r) = E(\mathbf{V}|\mathbf{R} = r)/r$, which has again the problem of control for $m_0 = m$.

Third, consider $E(\mathbf{V})/E(\mathbf{R})$. When all hypotheses are true it is identically 1, and again impossible to control. A remedy may be given by either adding 1 to the denominator, a somewhat artificial solution, or by changing the denominator to $E(\mathbf{R}|\mathbf{R} > 0)$. Modifying both numerator and denominator in the same way will again run into problems of control when $m_0 = m$.

3. FALSE DISCOVERY RATE CONTROLLING PROCEDURE

3.1. *The Procedure*

Consider testing $H_1, H_2, \ldots, H_m$ based on the corresponding p-values $P_1, P_2, \ldots, P_m$. Let $P_{(1)} \leqslant P_{(2)} \leqslant \ldots \leqslant P_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. Define the following Bonferroni-type multiple-testing procedure:

$$\text{let } k \text{ be the largest } i \text{ for which } P_{(i)} \leqslant \frac{i}{m} q^*;$$

$$\text{then reject all } H_{(i)}\ i = 1, 2, \ldots, k. \tag{1}$$

Theorem 1. For independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at q^*.

Proof. The theorem follows from the following lemma, whose proof is given in Appendix A.

Lemma. For any $0 \leqslant m_0 \leqslant m$ independent p-values corresponding to true null hypotheses, and for any values that the $m_1 = m - m_0$ p-values corresponding to the false null hypotheses can take, the multiple-testing procedure defined by procedure (1) above satisfies the inequality

$$E(\mathbf{Q}|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leqslant \frac{m_0}{m} q^*. \tag{2}$$

Now, suppose that $m_1 = m - m_0$ of the hypotheses are false. Whatever the joint distribution of $P_1'', \ldots, P_{m_1}''$ which corresponds to these false hypotheses is, integrating inequality (2) above we obtain

$$E(\mathbf{Q}) \leqslant \frac{m_0}{m} q^* \leqslant q^*,$$

and the FDR is controlled.

Remark. Note that the independence of the test statistics corresponding to the false null hypotheses is not needed for the proof of the theorem.

This procedure was mentioned by Simes (1986) as an exploratory extension to his procedure for rejection of the intersection hypothesis that all null hypotheses are true if, for some i, $P_{(i)} \leqslant i\alpha/m$. Whereas Simes (1986) showed that his procedure controls the FWER under the intersection null hypothesis, Hommel (1988) showed that the extended procedure for inference on individual hypotheses does not control the FWER in the strong sense: for some configuration of the false null hypotheses, the probability of an erroneous rejection is greater than α.

Hochberg (1988) has suggested a different way to utilize Simes's procedure so that it does control the FWER in the strong sense, by offering the following procedure:

$$\text{let } k \text{ be the largest } i \text{ for which } P_{(i)} \leqslant \frac{i}{m+1-i}\alpha;$$

$$\text{then reject all } H_{(i)}\ i = 1, 2, \ldots, k.$$

Note the relationship between Hochberg's procedure and the FDR controlling procedure when q^* is chosen to equal α. Both Hochberg's procedure and the FDR controlling procedure are step-down procedures, which start by comparing $p_{(m)}$ with α, and if smaller all hypotheses are rejected—as if a PCER approach had been taken. If $p_{(m)} > \alpha$, proceed to smaller p-values until one satisfies the condition. The procedures end, if not terminated earlier, by comparing $p_{(1)}$ with α/m, as in a pure Bonferroni comparison. At the two ends the procedures are similar, but, in between, the sequence of $p_{(i)}$s is compared with $\{1 - (i - 1)/m\}\alpha$ in the current procedure, rather than with $\{1/(m + 1 - i)\}\alpha$ in Hochberg's procedure. The series of linearly decreasing constants of the FDR controlling method is always larger than the hyperbolically decreasing constants of Hochberg, and the extreme ratio is as large as $4m/(m + 1)^2$ at $i = (m + 1)/2$. This shows that the suggested procedure rejects samplewise at least as many hypotheses as Hochberg's method and therefore has also greater power than other FWER controlling methods such as Holm's (1979).

3.2. *Example of False Discovery Rate Controlling Procedure*

Thrombolysis with recombinant tissue-type plasminogen activator (rt-PA) and anisoylated plasminogen streptokinase activator (APSAC) in myocardial infarction has been proved to reduce mortality. Neuhaus *et al.* (1992) investigated the effects of a new front-loaded administration of rt-PA *versus* those obtained with a standard regimen of APSAC, in a randomized multicentre trial in 421 patients with acute myocardial infarction. Four families of hypotheses can be identified in the study:

(a) base-line comparisons (11 hypotheses), where the problem is of showing equivalence;
(b) patency of infarct-related artery (eight hypotheses);
(c) reocclusion rates of patent infarct-related artery (six hypotheses);
(d) cardiac and other events after the start of thrombolitic treatment (15 hypotheses).

In this last family FDR control may be desired: we do not wish to conclude that the front-loaded treatment is better if it is merely equivalent to the previous treatment in all respects.

In the paper, however, there is no attention to the problem of multiplicity (the only exception being the division of the end points into primary and secondary). The individual p-values are reported as they are, with no word of warning regarding their interpretation. The authors conclude that

> 'Compared to APSAC treatment, despite more early reocclusions, the clinical course with rt-PA treatment is more favorable with fewer bleeding complications and a substantially lower in-hospital mortality rate, presumably due to improved early patency of the infarct-related artery'.

The statement about the mortality is based on a p-value of 0.0095.

Consider now the fourth family, which contains the comparison of mortality and 14 other comparisons. The ordered $p_{(i)}$s for the 15 comparisons made are

$$0.0001,\ 0.0004,\ 0.0019,\ 0.0095,\ 0.0201,\ 0.0278,\ 0.0298,\ 0.0344,$$
$$0.0459,\ 0.3240,\ 0.4262,\ 0.5719,\ 0.6528,\ 0.7590,\ 1.000.$$

Controlling the FWER at 0.05, the Bonferroni approach, using $0.05/15 = 0.0033$, rejects the three hypotheses corresponding to the smallest p-values. These hypotheses correspond to reduced allergic reaction, and to two different aspects of bleeding; they do not include the comparison of mortality. Using Hochberg's procedure leaves us with the same three hypotheses rejected. Thus the statement about a significant reduction in mortality is unjustified from the classical point of view.

Using the FDR controlling procedure with $q^* = 0.05$, we now compare sequentially each $p_{(i)}$ with $0.05i/15$, starting with $p_{(15)}$. The first p-value to satisfy the constraint is $p_{(4)}$ as

$$p_{(4)} = 0.0095 \leqslant \frac{4}{15}0.05 = 0.013.$$

Thus we reject the four hypotheses having p-values which are less than or equal to 0.013. We may support now with appropriate confidence the statements about mortality decrease, of which we did not have sufficiently strong evidence before.

3.3. *Another Look at False Discovery Rate Controlling Procedure*

The above FDR controlling procedure can be viewed as a *post hoc* maximizing procedure, as the following theorem suggests.

Theorem 2. The FDR controlling procedure given by expression (1) is the solution of the following constrained maximization problem:

$$\text{choose } \alpha \text{ that maximizes the number of rejections at this level, } r(\alpha),$$
$$\text{subject to the constraint } \alpha m/r(\alpha) \leqslant q^*. \tag{3}$$

Proof. Observe that, for each α, if $p_{(i)} \leqslant \alpha < p_{(i+1)}$, then $r(\alpha) = i$. Furthermore, as the ratio on the left-hand side of constraint (3) increases in α over the range on which $r(\alpha)$ is constant, it is enough to investigate αs which are equal to one of the $p_{(i)}$s. This $\alpha = p_{(k)}$ satisfies the constraint because $\alpha/r(\alpha) = p_{(k)}/k \leqslant q^*/m$. By considering the largest potential αs (largest $p_{(i)}$s) first, the procedure yields the α with the largest $r(\alpha)$ satisfying the constraint.

Thus the procedure has also the appearance of a simultaneous maximization of **R** and FDR control being attempted after experimentation. When each hypothesis is tested individually at level α, the expected number of wrong rejections satisfies $E(V) \leqslant \alpha m$. So, after observing the outcome of the experiment, an upper bound estimate of Q_e is $\alpha m/r(\alpha)$. In view of the observed p-values, the level α can now be chosen, by maximizing the observed number of rejections $r(\alpha)$ subject to the constraint on the implied FDR-like bound. As noted in the examples of Section 2.2 this aspect of the procedure is desirable.

4. SOME POWER COMPARISONS

We compare the power of our FDR controlling procedure with some other Bonferroni-type procedures which control the FWER. Under the overall null hypothesis the proposed method controls the FWER at level q^*. We take the FWER controlling methods and the FDR controlling method to control the FWER weakly at the same level, using $q^* = \alpha$, and compare the power of the methods from the two different approaches under different configurations. It is clear from the comment in Section 2.1, property (b), that a method which controls the FDR is generally more powerful than its counterpart which controls the FWER (in the strong sense). The magnitude of the difference remains to be investigated.

4.1. *The Setting*

We studied this question by using a large simulation study, where the family of hypotheses is the expectations of m independent normally distributed random variables being equal to 0. Each individual hypothesis is tested by a z-test, and the test statistics are independent. We use $q^* = \alpha = 0.05$. The configurations of the hypotheses involve $m = 4$, 8, 16, 32 and 64 hypotheses, and the number of truly null hypotheses being $3m/4$, $m/2$, $m/4$ and 0. The non-zero expectations were divided into four groups and placed at $L/4$, $L/2$, $3L/4$ and L in the following three ways:

(a) linearly decreasing (D) number of hypotheses away from 0 in each group;
(b) equal (E) number of hypotheses in each group;
(c) linearly increasing (I) number of hypotheses away from 0 in each group.

These expectations were fixed (per configuration) throughout the experiment.

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

Each simulation involved 20000 repetitions. The estimated standard errors of the power are about 0.0008–0.0016. As the same normal deviates were 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 method.

4.2. *Results*

Fig. 1 presents the estimates of the average power (the proportion of the false hypotheses which are correctly rejected) for three methods. The two FWER controlling methods, the Bonferroni (dotted curves) and Hochberg's (1988) method (broken curves), are compared with the new FDR controlling procedure (full curves). The following observations can be made from the results displayed.

(a) The power of all the methods decreases when the number of hypotheses tested increases—this is the cost of multiplicity control.
(b) The power is smallest for the D-configuration, where the non-null hypotheses are closer to the null, and is largest for I (which is obvious but mentioned for completeness).

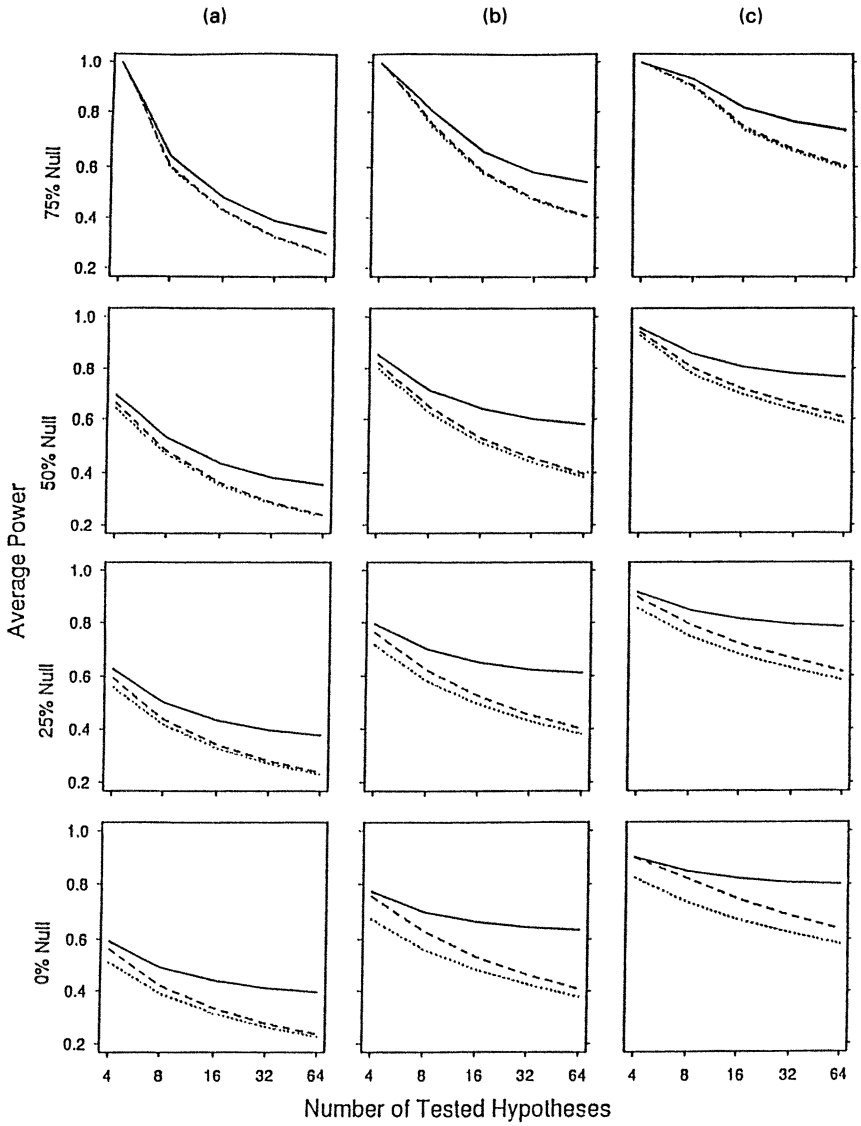

Fig. 1. Simulation-based estimates of the average power (the proportion of the false null hypotheses which are correctly rejected) for two FWER controlling methods, the Bonferroni (·····) and Hochberg's (1988) (-----) methods, and the FDR controlling procedure (———): (a) decreasing; (b) equally spread; (c) increasing

(c) The power of the FDR controlling method is uniformly larger than that of the other methods.
(d) The advantage increases with the number of non-null hypotheses.
(e) The advantage increases in m. Therefore, the loss of power as m increases is relatively small for the FDR controlling method in the E- and I-configurations.

(f) The advantage in some situations is extremely large. For example, testing 32 hypotheses, equally spread in four clusters from 1.25 to 5 so that none is true, the power of the Bonferroni method is 0.42. The procedure suggested increases the power to 0.65; testing as few as four hypotheses, half of which are true, the values are 0.62 and 0.70 respectively.

(g) It is known that Hochberg's method offers a more powerful alternative to the traditional Bonferroni method. Nevertheless, it is important to note that the gain in power due to the control of the FDR rather than the FWER is much larger than the gain of the FWER controlling method over the Bonferroni method.

Casting these results into a different form, it may be seen that in some configurations up to half the non-null hypotheses which were not rejected by the Bonferroni procedure are now rejected by the FDR controlling method, when at least half of the tested hypotheses are non-null. Even when only 25% of the hypotheses are non-null, the gain in power is such that about a quarter of the equally spaced hypotheses which were not rejected before are now rejected.

Fig. 1 allows us also to answer a question raised by a referee, about how $E(V/m_0)$ is controlled by the FDR controlling method. This error rate is 0 when $m_0 = 0$ and otherwise can be approximated by the average level α_{ave} at which the individual hypotheses are tested. Obviously α_{ave} is always less than q^*, but for m_0 away from m more can be said: let R_{ave} be the average number of rejections and f_{ave} be the average power (displayed in Fig. 1). It follows that $m\alpha_{ave} \leqslant R_{ave}q^* \cong (m_0\alpha_{ave} + m_1 f_{ave})q^*$, so

$$\alpha_{ave} \leqslant f_{ave}q^* \frac{m_1}{m_1 + m_0(1 - q^*)}.$$

Therefore $E(V/m_0)$ looks as in Fig. 1 but is smaller by a factor of q^* or even less. For $m = m_0$ the error is much closer to q^*/m_0 than to q^*: 0.0132, 0.0063, 0.0033, 0.0017 and 0.0009 for the four, eight, 16, 32 and 64 hypotheses tested respectively.

5. CONCLUSION

The approach to multiple significance testing in this paper is philosophically different from the classical approaches. The classical approach requires the control of the FWER in the strong sense, a conservative type I error rate control against any configuration of the hypotheses tested. The new approach calls for the control of the FDR instead, and thereby also the control of the FWER in the weak sense. In many applications this is the desirable control against errors originating from multiplicity.

Within the framework suggested, other procedures may be developed, including procedures which utilize the structure of specific problems such as pairwise comparisons in analysis of variance. A different direction, which we have already pursued, is to develop an adaptive method which incorporates the ideas of Hochberg and Benjamini (1990). In this paper, however, we have only focused on presenting and motivating the new approach that calls for controlling the FDR, and we have demonstrated that it can be developed into a simple and powerful procedure. Thus the cost paid for the control of multiplicity need not be large. This might contribute

considerably to the proliferation of a greater awareness of multiple-comparison problems, and of cases where something is done about it.

ACKNOWLEDGEMENTS

We would like to express our thanks to J. P. Shaffer and J. W. Tukey, whose comments and questions about an earlier draft helped us to crystallize the approach presented here. We thank Yetty Varon for her participation in the programming of the simulation study and Yechezkel Kling for pointing out a need to tighten the original proof of theorem 1.

APPENDIX A: PROOF OF LEMMA

The proof of the lemma is by induction on m. As the case $m = 1$ is immediate, we proceed by assuming that the lemma is true for any $m' \leqslant m$, and showing it to hold for $m + 1$.

If $m_0 = 0$, all null hypotheses are false, $\mathbf{Q}$ is identically 0 and

$$E(\mathbf{Q} \mid P_1 = p_1, \ldots, P_m = p_m) = 0 \leqslant \frac{m_0}{m+1} q^*.$$

If $m_0 > 0$, denote by P_i', $i = 1, 2, \ldots, m_0$, the p-values corresponding to the true null hypotheses, and the largest of these by $P'_{(m_0)}$. These are $U(0, 1)$ independent random variables. For ease of notation assume that the m_1 p-values that the false null hypotheses take are ordered $p_1 \leqslant p_2 \leqslant \ldots \leqslant p_{m_1}$. Finally, define j_0 to be the largest $0 \leqslant j \leqslant m_1$ satisfying

$$p_j \leqslant \frac{m_0 + j}{m+1} q^*, \tag{4}$$

and denote the right-hand side of inequality (4) at j_0 by p''.

Conditioning on $P'_{(m_0)} = p$,

$$\begin{aligned} E(\mathbf{Q} \mid P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) = \int_0^{p''} & E(\mathbf{Q} \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, \\ & P_m = p_{m_1}) f'_{P_{(m_0)}}(p)\, dp \\ + \int_{p''}^1 & E(\mathbf{Q} \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, \\ & P_m = p_{m_1}) f'_{P_{(m_0)}}(p)\, dp \end{aligned} \tag{5}$$

with $f_{P'_{(m_0)}}(p) = m_0 p^{(m_0 - 1)}$.

In the first part $p \leqslant p''$. Thus all $m_0 + j_0$ hypotheses are rejected, and $\mathbf{Q} \equiv m_0/(m_0 + j_0)$. Evaluating the integral first, and then using inequality (4), we obtain

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

In the second part of equation (5), consider separately each $p_{j_0} < p_j \leqslant P'_{(m_0)} = p < p_{j+1}$, along with $p_{j_0} \leqslant p'' < P'_{(m_0)} = p < p_{j_0+1}$. It is important to note that, because of the way by which j_0 and p'' are defined, no hypothesis can be rejected as a result of the values of p, p_{j+1}, p_{j+2}, $\ldots$, p_{m_1}. Therefore, when all hypotheses—true and false—are considered together, and their p-values thus ordered, a hypothesis $H_{(i)}$ can be rejected only if there exists k, $i \leqslant k \leqslant m_0 + j - 1$, for which $p_{(k)} \leqslant \{k/(m+1)\} q^*$, or equivalently

$$\frac{p_{(k)}}{p} \leqslant \frac{k}{m_0 + j - 1} \frac{m_0 + j - 1}{(m + 1)p} q^*. \tag{7}$$

When conditioning on $P'_{(m_0)} = p$, the P'_i/p for $i = 1, 2, \ldots, m_0 - 1$ are distributed as $m_0 - 1$ independent $U(0, 1)$ random variables, and the p_i/p for $i = 1, 2, \ldots, j$ are numbers corresponding to false null hypotheses between 0 and 1. Using inequality (7) to test the $m_0 + j - 1 = m' \leqslant m$ hypotheses is equivalent to using procedure (1), with the constant $\{(m_0 + j - 1)/(m + 1)p\}q^*$ taking the role of q^*. Applying now the induction hypothesis, we have

$$E(\mathbf{Q} | P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leqslant \frac{m_0 - 1}{m_0 + j - 1} \frac{m_0 + j - 1}{(m + 1)p} q^* = \frac{m_0 - 1}{(m + 1)p} q^* \tag{8}.$$

The bound in inequality (8) depends on p, but not on the segment $p_j < p < p_{j+1}$ for which it was evaluated, so

$$\int_{p''}^{1} E(\mathbf{Q} | P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f'_{P_{(m_0)}}(p)\, dp \leqslant \int_{p''}^{1} \frac{m_0 - 1}{(m + 1)p} q^* m_0 p^{(m_0 - 1)}\, dp$$

$$= \frac{m_0}{m + 1} q^* \int_{p''}^{1} (m_0 - 1) p^{(m_0 - 2)}\, dp = \frac{m_0}{m + 1} q^* \{1 - p''^{(m_0 - 1)}\}. \tag{9}$$

Adding inequalities (6) and (9) completes the proof of the lemma.

REFERENCES

Godfrey, K. (1985) Comparing the means of several groups. *New Engl. J. Med.*, **311**, 1450–1456.

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

Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Statist. Med.*, **9**, 811–818.

Hochberg, Y. and Tamhane, A. (1987) *Multiple Comparison Procedures.* New York: Wiley.

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

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

Neuhaus, K. L., Von Essen, R., Tebbe, U., Vogt, A., Roth, M., Riess, M., Niederer, W., Forycki, F., Wirtzfeld, A., Maeurer, W., Limbourg, P., Merx, W. and Haerten, K. (1992) Improved thrombolysis in acute myocardial infarction front-loaded administration of Alteplase: results of the rt-PA–APSAC patency study (TAPS). *J. Am. Coll. Card.*, **19**, 885–891.

Pocock, S. J., Hughes, M. D. and Lee, R. J. (1987) Statistical problems in reporting of clinical trials. *J. Am. Statist. Ass.*, **84**, 381–392.

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

Saville, D. J. (1990) Multiple comparison procedures: the practical solution. *Am. Statistn*, **44**, 174–180.

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

Smith, D. E., Clemens, J., Crede, W., Harvey, M. and Gracely, E. J. (1987) Impact of multiple comparisons in randomized clinical trials. *Am. J. Med.*, **83**, 545–550.

Soriç, B. (1989) Statistical "discoveries" and effect size estimation. *J. Am. Statist. Ass.*, **84**, 608–610.

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