British Journal of Mathematical and Statistical Psychology (2002), 55, 27–39

www.bps.org.uk

Controlling the rate of Type I error over a large set of statistical tests

H.J. Keselman[1]*, Robert Cribbie[1] and Burt Holland[2]
[1]University of Manitoba, Canada
[2]Temple University, USA

When many tests of significance are examined in a research investigation with procedures that limit the probability of making at least one Type I error—the so-called familywise techniques of control—the likelihood of detecting effects can be very low. That is, when familywise error controlling methods are adopted to assess statistical significance, the size of the critical value that must be exceeded in order to obtain statistical significance can be extremely large when the number of tests to be examined is also very large. In our investigation we examined three methods for increasing the sensitivity to detect effects when family size is large: the false discovery rate of error control presented by Benjamini and Hochberg (1995), a modified false discovery rate presented by Benjamini and Hochberg (2000) which estimates the number of true null hypotheses prior to adopting false discovery rate control, and a familywise method modified to control the probability of committing two or more Type I errors in the family of tests examined—not one, as is the case with the usual familywise techniques. Our results indicated that the level of significance for the two or more familywise method of Type I error control varied with the testing scenario and needed to be set on occasion at values in excess of 0.15 in order to control the two or more rate at a reasonable value of 0.01. In addition, the false discovery rate methods typically resulted in substantially greater power to detect non-null effects even though their levels of significance were set at the standard 0.05 value. Accordingly, we recommend the Benjamini and Hochberg (1995, 2000) methods of Type I error control when the number of tests in the family is large.

*Requests for reprints should be addressed to H.J. Keselman, Department of Psychology, University of Manitoba, 190 Dysart Road, Winnipeg, Manitoba, Canada R3T 2N2.

1. Introduction

It is common to compute many tests of significance in a typical research investigation (see, for example, Barton & Huberty, 1989; Knoop, 1986; Schippman & Prien, 1986). Indeed, not only do researchers examine all possible correlations in, say, 16×16 and 21×21 correlation matrices, but we have found a paper in which 444 400 tests were examined (Drigalenko & Elston, 1997; Mallet, Mazoyer, & Martinot, 1998)! It is well known that the probability of committing one or more Type I errors increases as the number of tests examined in the family of tests increases. The prevailing sentiment is that when many tests of significance are to be computed the error rate should be controlled familywise, that is, over the entire set (family) of tests. This opinion is diametrically opposite to the view that the error rate should be set on the individual test (the per-test approach) and not on the entire set of tests. Those who favour the per-test approach do so based on power considerations. That is, as the number of tests in the family increases there is a concomitant increase in the size of the critical value that must be exceeded to obtain statistical significance (see Miller, 1981). Thus, though the prevailing view is for familywise error (FWE) control, there is still a minority of opinion that argues fervently for per-test control (see Rothman, 1990; Saville, 1990; Wilson, 1962). However, other alternatives also exist for researchers.

FWE methods control the probability of committing one or more Type I errors, but when many tests of significance are computed is it reasonable to set such a stringent criterion? Indeed, previous authors have suggested that such a criterion could reasonably be relaxed when the number of tests in the family is substantial. Specifically, researchers can choose to control the probability of committing *two or more*, or perhaps *three or more* (or *four or more*, etc.) Type I errors when the number of tests is large (Halperin, Lan & Hamdy, 1988).

Another approach to controlling errors in the multiple-testing situation which affords researchers greater power to detect true effects than conventional FWE methods is the false discovery rate (FDR) presented by Benjamini and Hochberg (1995, 2000) (for other FDR-type procedures, see Benjamini & Liu, 1999; Yekutieli & Benjamini, 1999). The FDR is defined by these authors as the expected proportion of the number of erroneous rejections to the total number of rejections. The motivation for such control, as Shaffer (1995) suggests, stems from a common misconception regarding the overall error rate. That is, some believe that the overall rate applied to a family of hypotheses indicates that on average 'only a proportion α of the rejected hypotheses are true ones, i.e., are falsely rejected' (Shaffer, 1995, p. 567). This is clearly a misconception, for as Shaffer notes, if all hypotheses are true, 'then 100% of rejected hypotheses are true, i.e., are rejected in error, in those situations in which any rejections occur' (p. 567). Such a misconception, however, suggests setting a rate of error for the proportion of rejections which are erroneous, hence the FDR.

Suppose we have J ($j = 1, \ldots, J$) means, $\mu_1, \mu_2, \ldots, \mu_J$, and our interest is in testing m hypotheses of which m_0 are true. Let S equal the number of correctly rejected hypotheses from the set of R rejections; the number of falsely rejected hypotheses will be V. In terms of the random variable V, the per-comparison error rate is $E(V/M)$, while the familywise rate is given by $P(V \geq 1)$. Thus, testing each and every comparison at α guarantees that $E(V/M) \leq \alpha$, while testing each and every comparison at α/M (Bonferroni) guarantees $P(V \geq 1) \leq \alpha$.

According to Benjamini and Hochberg (1995) the proportion of errors committed by falsely rejecting null hypotheses can be expressed by the random variable

$Q = V/(V + S)$. It is important to note that Q is defined to be zero when $R = 0$; that is, the error rate is zero when there are no rejections. The FDR was defined by Benjamini and Hochberg as the mean of Q, that is

$$E(Q) = E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right)$$

$$= E\left(\frac{\text{Number of false rejections}}{\text{Number of rejections}}\right);$$

thus, FDR is the mean of the proportion of the falsely declared tests among all tests declared significant.

As Benjamini and Hochberg (1995) indicate, this error rate has a number of important properties. For example, when $m_0 < m$, the FDR is smaller than or equal to the FWE because in this case the latter is given by $P(R \geq 1) \geq E(V/R) = E(Q)$. This indicates that if the FWE is controlled for a procedure, then FDR is as well. Moreover, and most importantly for the purposes of this paper, if one adopts a procedure which provides strong (i.e., over all possible mean configurations) FDR control, rather than strong FWE control, then based on the preceding relationship, a gain in power can be expected.

In addition to these characteristics, Benjamini, Hochberg, and Kling (1994) provide a number of illustrations where FDR control seems more reasonable than FWE or per-test control. Exploratory research, for example, would be one area of application for FDR control. That is, in new areas of inquiry where we are merely trying to see what parameters might be important for the phenomenon under investigation, a few errors of inference should be tolerable; thus, one can reasonably adopt the less stringent FDR method of control which does not completely ignore the multiple-testing problem, as does per-test control, and yet provides greater sensitivity than FWE control. Only at later stages in the development of our conceptual formulations does one need more stringent FWE control. Another area where FDR control might be preferred over FWE control, suggested by Benjamini and Hochberg (1995), would be when two treatments (say, treatments for dyslexia) are being compared in multiple subgroups (say, children of different ages). In studies of this sort, where an overall decision regarding the efficacy of the treatment is not of interest, but rather where separate recommendations would be made within each subgroup, researchers may well be willing to tolerate a few errors of inference and accordingly would profit from adopting FDR rather than FWE control.

Simulation studies comparing the power of the Benjamini–Hochberg (BH) procedure to several FWE controlling procedures (for detecting non-null pairwise mean differences) have shown that as the number of treatment groups increases (beyond $J = 4$), the power advantage of the BH procedure over the FWE controlling procedures becomes increasingly large (Benjamini *et al.*, 1994; Keselman, Cribbie, & Holland, 1999; Williams, Jones, & Tukey, 1999). The power of FWE controlling procedures is highly dependent on the family size (i.e., number of comparisons), decreasing rapidly with larger families (Holland & Cheung, 2002; Miller, 1981). Therefore, control of the FDR results in more power than FWE controlling procedures in experiments with many treatment groups, yet provides more control over Type I errors than per-test controlling procedures.

As Hochberg and Benjamini (1990) and Benjamini and Hochberg (2000) note, the FDR can result in conservative rates of Type I error when some of the tested hypotheses are indeed false. Accordingly, these authors developed an adaptive FDR (AFDR)

controlling procedure which estimates the number of true null hypotheses and then subsequently applies the estimate with the FDR method of control. Benjamini and Hochberg (2000) demonstrate that the AFDR can result in greater power to detect effects than the usual FDR method of control.

Based on the preceding, the purpose of our investigation was to identify the two or more Type I error properties of an FWE method. Specifically, as a first step in exploring this approach to Type I error control, we wanted to determine what the FWEs would need to be in various multiple-testing scenarios such that the probability of making two or more Type I errors would be controlled at some reasonable value. Then, based on these results, and some ancillary comparisons between it and the FDR methods of control, we could make recommendations regarding the preferred method for multiple-testing scenarios. In the remainder of this paper we let γ stand for the probability of making two or more Type I errors, that is, $\gamma = P(V \geq 2)$.

2. Procedures

A Monte Carlo study was conducted to determine the FWE rates necessary for controlling γ at 0.01 with Hochberg's (1988) Bonferroni-type (HB) procedure. We chose the HB method over other Bonferroni-type procedures because Olejnik, Li, Supattathum, and Huberty (1997) found relatively small power differences between them and, most importantly, because the HB procedure is very simple to apply. However, researchers who prefer to use the other more complicated, and slightly more powerful, Bonferroni-type FWE methods (Hommel, 1988; Rom, 1990) may refer to Westfall, Tobias, Rom, Wolfinger, and Hochberg (1999).

2.1. The HB Procedure

In this procedure, the $p_{(i)}$-values corresponding to the m statistics for testing the hypotheses $H_1, \ldots, H_m$ are ordered from smallest to largest. Then, for any $i = m, m-1, \ldots, 1$, if $p_{(i)} \leq \alpha/(m - i + 1)$, the HB procedure rejects all $H_{i'}$ $(i' \leq i)$. According to this procedure, therefore, one begins by assessing the largest $p_{(i)}$-value, $p_{(m)}$. If $p_{(m)} \leq \alpha$, all hypotheses are rejected. If $p_{(m)} > \alpha$, then H_m is accepted and one proceeds to compare $p_{(m-1)}$ to $\alpha/2$. If $P_{(m-1)} \leq \alpha/2$, then all H_i $(i = m-1, \ldots, 1)$ are rejected; if not, then $H_{(m-1)}$ is accepted and one proceeds to compare $p_{(m-2)}$ with $\alpha/3$, and so on. HB has been shown to control the FWE for several situations of dependent tests, that is, for a wide variety of multivariate distributions that make HB applicable to most testing situations psychologists might encounter. In particular, Sarkar and Chang (1997) proved that HB controls the FWE when the distribution of the test statistics has exchangeable positive dependence, and Sarkar (1998) proved that HB controls the FWE when the distribution satisfies the ordered multivariate totally positive of order 2 (MTP_2) condition (see also Benjamini & Yekutieli, 2001). These distributional conditions include those we study in our examples (to be discussed shortly).

2.2. The BH Procedure

Benjamini and Hochberg (1995) proposed controlling the FDR, instead of the often conservative FWE or the often liberal per-test error rate. In this procedure, the $p_{(i)}$-values corresponding to the m statistics for testing the hypotheses $H_1, \ldots, H_m$ are also ordered from smallest to largest, $p_{(1)} \leq p_{(2)} \leq, \ldots, \leq p_{(m)}$. Let k be the largest value of i for which

$p_{(i)} \leq (i/m)\alpha$, then reject all H_i, $i = 1, 2, \ldots, k$. According to this procedure one begins by assessing the largest $p_{(i)}$-value, $p_{(m)}$, proceeding to smaller $p_{(i)}$-values as long as $p_{(i)} > (i/m)\alpha$. Testing stops when $p_{(k)} \leq (k/m)\alpha$.

The BH procedure has been shown to control the FDR for several situations of dependent tests, that is, for a wide variety of multivariate distributions that make FDR applicable to most testing situations social scientists might encounter (see Sarkar, 1998; Sarkar & Chang, 1997). In addition, simulation studies comparing the power of the BH procedure to several FWE controlling procedures have shown that as the number of treatment groups increases (beyond $J = 4$), the power advantage of the BH procedure over the FWE controlling procedures becomes increasingly large (Benjamini *et al.*, 1994; Keselman *et al.*, 1999). The power of FWE controlling procedures is highly dependent on the family size (i.e., number of tests), decreasing rapidly with larger families (Holland & Cheung, 2002; Miller, 1981). Therefore, control of the FDR results in more power than FWE controlling procedures in experiments with many treatment groups, but yet provides more control over Type I errors than per-test controlling procedures.

2.3. The AFDR Procedure

Benjamini and Hochberg (2000) also presented a modified (adaptive) version of their original procedure that uses the data to estimate the number of true hypotheses ($\hat{m}_0$). (The adaptive BH procedure has only been demonstrated, not proven, to control FDR, and only in the independent case.) With the original procedure, when the number of true null hypotheses, is less than the total number of hypotheses, the FDR rate is controlled at a level less than that specified (α).

To compute the AFDR procedure according to Benjamini and Hochberg (2000) one would perform the following steps:

1. Order the $p_{(i)}$-values.
2. Compare each $p_{(i)}$ to $\alpha i/m$ (as in the BH procedure). If all H_i are retained, testing stops.
3. If any H_i is rejected with the criterion of the BH procedure, then testing continues by estimating the slopes $S_i = (1 - p_{(i)})/(m - i + 1)$.
4. Beginning 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 $\hat{m}_0 = \min([1/S_j + 1], m)$, where $[x]$ is the largest integer less than or equal to x.
5. Starting with the largest $p_{(i)}$-value, $p_{(m)}$, compare each $p_{(i)}$ to $\alpha(i/\hat{m}_0)$. Testing stops when $p_{(k)} \leq (k/\hat{m}_0)\alpha$.

One disadvantage of the AFDR procedure, noted by both Benjamini and Hochberg (2000) and Holland and Cheung (2002), is that it is possible for an H_i to be rejected with $p_{(i)} > \alpha$. Therefore, it is suggested, by both authors, that H_i only be rejected if the hypothesis satisfies the rejection criterion of the AFDR, and $p_{(i)} \leq \alpha$. To illustrate this procedure, assume a researcher has conducted a study with $J = 4$ and $\alpha = 0.05$. The ordered $p_{(i)}$-values associated with the $m = 6$ pairwise comparisons are: $p_{(1)} = 0.0014$, $p_{(2)} = 0.0044$, $p_{(3)} = 0.0097$, $p_{(4)} = 0.0145$, $p_{(5)} = 0.0490$ and $p_{(6)} = 0.1239$. The first stage of the AFDR procedure would involve comparing $p_{(6)} = 0.1239$ to $\alpha(i/m) = 0.05(6/6) = 0.05$. Since $0.1239 > .05$, the procedure would continue by comparing $p_{(5)} = 0.0490$ to $\alpha(i/m) = 0.05(5/6) = 0.0417$. Again, since $0.0490 > 0.0417$, the procedure would continue by comparing $p_{(4)} = 0.0145$ to $\alpha(i/m) = 0.05(4/6) = 0.0333$. Since $0.0145 < 0.0333$, H_4 would be rejected. Because at least one H_i was rejected during the

first stage, testing continues by estimating each of the slopes, $S_i = (1 - p_{(i)})/(m - i + 1)$, for $i = 1, \ldots, m$. The calculated slopes for this example are: $S_1 = 0.1664$, $S_2 = 0.1991$, $S_3 = 0.2475$, $S_4 = 0.3285$, $S_5 = 0.4755$ and $S_6 = 0.8761$. Given that all $S_i > S_{i-1}$, S_j is set at $m = 6$. The estimated number of true nulls is then determined by $\hat{m}_0 = \min([1/S_j + 1], m) = \min([1/6 + 1], 6) = \min([1.1667], 6) = 1$. Therefore, the AFDR procedure would compare $p_{(6)} = 0.1239$ to $\alpha(i/\hat{m}_0) = 0.05(6/1) = 0.30$. Since $0.1239 < 0.30$, but $0.1239 > \alpha$, H_6 would not be rejected and the procedure would continue by comparing $p_{(5)} = 0.0490$ to $0.05(5/1) = 0.25$. Since $0.0490 < 0.25$ and $0.0490 < \alpha$, H_5 would be rejected; in addition, all H_k would also be rejected (i.e., H_1, H_2, H_3 and H_4).

2.4. The two or more FWE procedure

In our study the FWE properties of γ were investigated for three different testing scenarios. First, this rate was determined for testing the hypotheses that each of 50, 100 and 150 population proportions were equal to 0.5. In our second scenario, we investigated the family of hypotheses that each population correlation coefficient equalled zero in a 16×16 (family size 120) correlation matrix. In our last scenario, we examined the family of all possible pairwise comparisons, testing the hypotheses H_0: $\mu_j = \mu_{j'}$ $(j \neq j')$ in a completely randomized design containing ten groups of subjects (i.e., family size of 45). In all cases, we initially were only interested in testing complete null hypotheses.

Our family of all possible proportions was generated in the following manner. Let Y_1, Y_2, Y_3 and Y_4 be multivariate Bernoulli with respective category probabilities π_i, $\pi_1 - \pi_i$, $\pi_2 - \pi_i$ and $1 - \pi_1 - \pi_2 + \pi_i$. Here, π_1 and π_2 are probabilities of 'yes' for each of two correlated questionnaire responses. The values of $\pi_1 = \pi_2 = 0.5$ were selected, and π_i was obtained from the relationship

$$\pi_i = \rho\sqrt{[(\pi_1 - \pi_1^2)(\pi_2 - \pi_2^2)]} + \pi_1\pi_2.$$

Accordingly, the value of ρ determines π_i. We considered a low and high value for ρ, namely $\rho = 0.3$ and $\rho = 0.8$. In particular, to obtain, say, $\rho = 0.8$ we selected a two-digit random number from 00 to 99 and if that number was between 00 and 44, then $Y_1 = 1$ and the other three Ys were set to 0; if between 45 and 49, then $Y_2 = 1$ and the other three Ys were 0; if between 50 and 54, then $Y_3 = 1$ and the other three Ys were 0; if between 55 and 99, then $Y_4 = 1$ and the other three Ys were 0.

To illustrate, consider a sample of N items from the above distribution. Define X_1 as the sum of all the Y_1s plus the sum of all the Y_2s and X_2 as the sum of all the Y_1s plus the sum of all the Y_3s. Then X_1 is $Binomial(N, \pi_1)$, X_2 is $Binomial(N, \pi_2)$ and $\text{Corr}(X_1, X_2) = \rho$. By generating data in this manner we were able to obtain correlated pairs of questions (Q_1, Q_2), (Q_3, Q_4), etc. where, say, Q_1 is uncorrelated with any other questions apart from Q_2. A SAS/IML (SAS Institute, 1989) program was written to generate data from this distribution.

To generate our null and non-null correlation (16×16) matrices we followed the procedures discussed in Olejnik *et al.* (1997). Specifically, in the null case ($\rho_{ij} = 0$ for all $i \neq j$), data were generated for 16 mutually independent variables, each having a standard normal distribution. For the non-zero (non-null) cases, like Olejnik *et al.*, we adopted the procedure described by Kaiser and Dickman (1962).

For the pairwise multiple-comparison problem, to generate pseudo-random normal variates, we used the SAS generator RANNOR (SAS Institute, 1999). If Z_{ij} is a standard

normal variate, then $X_{ij} = \mu_j + \sigma_j \times Z_{ij}$ is a normal variate with mean equal to μ_j and variance equal to σ_j^2 ($j = j'$).

In addition to exploring what FWE rates would need to be in order to obtain 0.01 control for γ, we compared the power rates of the two or more HB procedures to the BH methods. (We will have more to say about this in Section 3.) For the family of proportion tests we created a non-null case by setting half of the true proportions to 0.59. In the 16×16 all possible pairwise correlations family we created a non-null case by setting the correlations between the first 11 variables to 0.15. For the pairwise comparisons problem we investigated three non-null cases for $\mu_1, \mu_2, \ldots, \mu_{10}$: (a) 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, (b) 0, 0, 0, 0, 0, 1.3, 1.3, 1.3, 1.3, 1.3, and (c) 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.3.

Two power rates were collected: (a) the probability of detecting all possible non-null tests (all-tests/pairs rate) and (b) the average of the per-test/pair power rates. Non-null parameters were chosen such that comparisons between the procedures would, if possible, avoid floor and ceiling effects.

The sample size for the first two scenarios investigated was 500, while for the third scenario we examined the cases $n = 10$ and $n = 20$ per group. Ten thousand simulations were performed for each investigated condition.[1]

3. Results

3.1. Type I error rates

Scenario 1

H_O: P = 0.5. Table 1 contains estimates of the FWE that were necessary to obtain $\gamma \simeq 0.01$ control for the HB procedure when testing 50, 100 and 150 proportions. The FWE values ranged from 0.069 to 0.152, with larger values occurring when the correlation among the items (questions) was smaller. That is, when $\rho = 0.3$ the FWE value necessary to obtain 0.01 γ control was approximately equal to 0.15, twice the value when $\rho = 0.8$.

Table 1. Hochberg (1988) FWE rates of Type I error such that $\gamma \leq 0.01$ (null hypotheses $P = 0.5$)

	Family size		
ρ	50	100	150
0.3	0.152	0.150	0.140
0.8	0.070	0.069	0.072

Note: $\gamma = P(V \geqslant 2) \leq 0.01$.

[1] *As previously indicated, the FDR error rate is zero when there are no rejections. Thus, for accurate estimation of empirical rates of error, the number of replications should be large when simulating FDR rates. To check the stability of our FDR values we conducted a number of additional simulations based on 100 000 replications and found that these FDR rates were virtually identical to those we report based on 10 000 replications.*

Scenario 2

H_0: $\rho = 0$. For the family of 120 tests of correlation coefficients from the 16×16 correlation matrix, similar results were obtained. That is, a FWE value of approximately 0.15 was required to achieve approximately 0.01 control for the FWE procedure.

Scenario 3

H_0: $\mu_j = \mu_{j'}$ $(j \neq j')$. In our last scenario we examined all possible pairwise comparisons (45) in a ten-group completely randomized design. As we found with the other two testing scenarios, the FWE Type I error rates that were necessary to achieve 0.01 γ control varied with the conditions investigated. For example, when $n = 10$, the HB FWE value was approximately equal to 0.08, while when $n = 20$ the value was approximately equal to 0.05.

In summary, our Type I error results indicated that with at most a 0.15 FWE value the γ rate could be controlled at the 0.01 level. This result applied to testing the hypotheses that each of 50, 100 and 150 population proportions were equal to 0.5 and the family of hypotheses that each population correlation coefficient equalled zero in a 16×16 (family size 120) correlation matrix, though in our tests of proportions scenario smaller FWE values could be adopted when the correlation between the questions was large.[2] With the all possible pairwise comparison problem, the FWE values needed to achieve 0.01 γ control was approximately 0.08 and 0.05 for the two cases of sample size investigated, $n = 10$ and $n = 20$, respectively.

Based on comments from reviewers on an earlier draft of our paper, we decided that before probing further into the two or more errors method of Type I error FWE control we could first compare the sensitivity of the method to the BH approaches. That is, if the FDR and AFDR methods of Type I error control afford greater sensitivity for detecting non-null effects with a more traditional level of significance (e.g., 0.05), then the two or more FWE Type I error method of control would not be attractive to researchers and accordingly it would not be worth pursuing this investigation further. Interestingly, though, it is also worth noting that we had also collected two or more Type I error data for the BH and AFDR methods and found that in all cases, the rates necessary to achieve 0.01 control were *much* smaller than the values reported in Table 1 (i.e., ranging from 0.022 to 0.072).

3.2. Power rates

As just indicated, we had compared the two or more Type I error HB controlling methods to the two or more Type I error BH methods. Thus, not only did we gather empirical rates of Type I error but also power rates. These power rates always favoured, frequently by a substantial amount, the FDR approaches. However, because we would not be recommending that the two or more approach to Type I error control be applied

[2] *The effect that the correlation between items had on the FWE rates seems explainable. That is, when $\rho = 0.8$ and one makes a Type I error, there is likely to be a similar test statistic that also leads to a Type I error, and hence there are likely to be at least two Type I errors. However, if ρ is small (i.e., $\rho = 0.3$), the existence of one Type I error does not strongly imply the existence of another Type I error. Since it is easier to make two Type I errors with larger ρ than smaller ρ, less control needs to be placed on FWE with larger ρ than with smaller ρ to guarantee equivalent tight control on γ. We examined this hypothesis by re-examining our tests of proportions scenario; however, we set $\rho = 0.5$. The FWE values necessary to achieve 0.01 γ control were between the values reported for $\rho = 0.3$ and $\rho = 0.8$, though closer to the $\rho = 0.3$ values because $\rho = 0.5$ is closer to $\rho = 0.3$ than it is to $\rho = 0.8$.*

to the FDR methods, we accordingly went on to compare the power of the procedures using $\alpha = 0.05$ with the FDR methods. The 0.05 criterion was selected because it is a familiar and accepted standard; however, 0.05 is also a representative Type I error value for the values that were found when the two or more criterion was applied to BH and AFDR. Thus, from these perspectives, the power of the approaches is being compared under 'comparable' conditions of Type I error control.

Scenario 1

H_0: $P = 0.5$. Table 2 contains the all-tests and average per-test power values for the family of proportions tests. Most evident from Table 2 is that the power values for the FWE procedure were always less than the BH and AFDR values. In particular, the AFDR method was always more powerful than the BH method which in turn was always more powerful than the HB FWE method. The differences between the all-tests rates could be described as substantial (i.e., greater than 0.20) (see Einot & Gabriel, 1975). On the other hand, the per-test rates, though still favouring the AFDR method over the BH method and the BH method over the HB method, were not always as dramatically different (i.e., less than 0.20).

Table 2. Power rates for testing null hypotheses that $P = 0.5$

	Family size					
	50		100		150	
Procedure	All-tests	Per-test	All-tests	Per-test	All-tests	Per-test
$\rho = 0.3$						
HB	0.09	0.90	0.00	0.85	0.00	0.81
BH	0.46	0.97	0.20	0.96	0.09	0.96
AFDR	0.65	0.98	0.41	0.98	0.28	0.98
$\rho = 0.8$						
HB	0.05	0.85	0.00	0.78	0.00	0.75
BH	0.52	0.96	0.27	0.96	0.15	0.96
AFDR	0.68	0.98	0.48	0.98	0.34	0.98

Scenario 2

H_0: $\rho = 0$. The power values for the family of 120 tests of correlations from the 16×16 matrix were not dramatically different as they were for the tests of all possible proportions, nor did they always favour FDR control over FWE control. When power was defined as the probability of detecting all non-null hypotheses, the HB procedure had a slight power advantage over the BH and AFDR methods (HB = 0.09 vs. BH = 0.01 and AFDR = 0.04, respectively). (Remember $\alpha \simeq 0.15$ for HB, while $\alpha = 0.05$ for the FDR methods.) However, the per-test rates favoured the FDR approaches (AFDR = 0.90, BH = 0.85 and HB = 0.80).

Scenario 3

H_0: $\mu_j = \mu_{j'}$ $(j \neq j')$. Table 3 contains all–pairs and per-pair power rates for the all possible pairwise comparison problem for just one case of sample size, namely $n = 10$ ($n = 20$ results were very similar in pattern) for the three non-null cases investigated. Once again, the AFDR rates were largest followed by the BH method and then by the HB method. Furthermore, as with the tests of all possible proportions data, many of the differences could be described as substantial (e.g., per-pair power rates were HB = 0.38, BH = 0.75 and AFDR = 0.86 for the non-null case 0, 0, 0, 0, 0, 1.3, 1.3, 1.3, 1.3, 1.3).

Table 3. Power rates for all possible pairwise tests ($J = 10$)

Procedure	All-pairs	Per-pair
$\mu_1, \ldots, \mu_{10} = 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5$		
HB	0.00	0.61
BH	0.00	0.77
AFDR	0.00	0.79
$\mu_1, \ldots, \mu_{10} = 0, 0, 0, 0, 0, 1.3, 1.3, 1.3, 1.3, 1.3$		
HB	0.01	0.38
BH	0.13	0.75
AFDR	0.26	0.86
$\mu_1, \ldots, \mu_{10} = 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.3$		
HB	0.05	0.35
BH	0.21	0.53
AFDR	0.33	0.62

4. Discussion

When many tests of significance are to be examined in a research investigation, controlling the probability of a Type I error with traditional FWE procedures can result in substantial reductions in power to detect effects (see Miller, 1981). While multiple comparisonists are certainly aware of this phenomenon, applied researchers may not be and therefore may simply routinely apply currently popular methods of control. Though this issue has been discussed over the years in the applied and statistical literatures, it has recently been readdressed by Benjamini and Hochberg (1995) by way of their FDR approach to Type I error control. Their approach is intended to provide greater power to detect effects than the currently popular FWE procedures and be more stringent with respect to Type I error control than would be the case if the Type I error rate is set on each individual test. Furthermore, they also presented a modified FDR, the adaptive FDR, which also provides Type I error protection and even greater sensitivity to detecting non-null effects than even their FDR approach to multiple testing.

Because of this renewed interest in increasing sensitivity to detecting effects when a large family of tests is to be examined, we decided to examine an approach to Type I error control that had been suggested earlier in the literature, namely controlling the rate of error at, say, t or more errors, where in our investigation, we set t at 2 (see

Halperin *et al.*, 1988). That is, traditional FWE methods protect against any error while the two or more error rate does not consider the presence of only one error to be serious when the family of tests is large in number.

We examined three scenarios involving many tests of significance: (a) testing that all proportions equal 0.5 when the family size of proportions tests was 50, 100 and 150; (b) testing that all of the 120 correlation coefficients equal zero from a 16×16 correlation matrix; and (c) testing all pairwise comparisons in a ten-group completely randomized design. The first of these scenarios would typify many clinical/medical-type research studies where responses from a large questionnaire are examined for some population of subjects, the second scenario has been reported (see Olejnik *et al.*, 1997), and the third would not be that uncommon in behavioural science investigations (Keselman *et al.*, 1998).

Specifically, we were interested in determining what the FWE value would need to be set at in order to provide 0.01 γ protection for Hochberg's (1988) step-up Bonferroni-type controlling procedure when all tested hypotheses were true. Finally, in the second phase of our investigation we compared the all-tests/pairs and average per-test/pair power rates of the HB and FDR procedures.

Our results indicated that with at most a 0.15 FWE value the γ rate could be controlled at the 0.01 level. This result applied to testing the hypotheses that each of 50, 100 and 150 population proportions were equal to 0.5 and the family of hypotheses that each population correlation coefficient equalled zero in a 16×16 (family size 120) correlation matrix, though in our tests of proportions scenario smaller FWE values could be adopted when the correlation between the questions was large. With the all possible pairwise comparison problem, the FWE values needed to achieve 0.01 γ control was approximately 0.08 and 0.05 for the two cases of sample size investigated, $n = 10$ and $n = 20$, respectively.

Our second major finding was that the power to detect effects, either per-test/pair or for all tests/pairs, was typically larger when adopting FDR control than for the FWE HB procedure. Indeed, though the FWE HB procedure occasionally used levels of significance that were considerably larger (e.g., $\alpha = 0.15$) than 0.05 (in order to give $\gamma = 0.01$), the BH and AFDR power values based on a 0.05 level of significance nonetheless were typically substantially larger (i.e., more than 0.20 percentage points) than the FWE power values. Accordingly, at this time, we do not see the need to explore further the two or more FWE method of Type I error control and therefore strongly recommend the BH methods of Type I error control, particularly the adaptive FDR.[3]

References

Barton, R. M., & Huberty, C. J. (1989, April). *Multiple testing and correlation matrices: An application of three procedures*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289–300.

[3] *Naturally, the power to detect effects can be increased with an FWE procedure by allowing the value to achieve 0.01 control for two or more errors to be set at a larger FWE value. However, based on our results, these values would be larger than could be taken seriously with regard to reasonable Type I error control.*

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*, 60–83.

Benjamini, Y., Hochberg, Y., & Kling, Y. (1994). *False discovery rate controlling procedures for pairwise comparisons*. Unpublished manuscript.

Benjamini, Y., & Liu, W. (1999). A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, *82*, 163–170.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1152–1175.

Drigalenko, E. I., & Elston, R. C. (1997). False discoveries in genome scanning. *Genetic Epidemiology*, *14*, 779–784.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, *70*, 574–583.

Halperin, M., Lan, K. K. G., & 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*, 800–802.

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

Holland, B., & Cheung, S. H. (2002). Family size robustness criteria for multiple comparison procedures. *Journal of the Royal Statistical Society, B*, *54*, 63–77.

Hommel, G. (1988). A comparison of two modified Bonferroni procedures. *Biometrika*, *75*, 383–386.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, *27*, 179–182.

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise/comparisonwise Type I error control. *Psychological Methods*, *4*, 58–69.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*, *68*, 350–386.

Knoop, R. (1986). Job involvement: An elusive concept. *Psychological Reports*, *59*, 451–456.

Mallet, L., Mayozer, B., & Martinot, J. L. (1998). Functional connectivity in depressive, obsessive-compulsive, and schizophrenic disorders: An explorative correlational analysis of regional cerebral metabolism. *Psychiatry Research: Neuroimaging Section*, *82*, 83–93.

Miller, R. G., Jr (1981). *Simultaneous statistical inference*, 2nd ed. New York: Springer-Verlag.

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistics*, *22*, 389–406.

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

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*, 43–46.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: A proof of the Simes conjecture. *Annals of Statistics*, *26*, 494–504.

Sarkar, S. K., & Chang, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, *92*, 1601–1608.

SAS Institute (1989). *SAS/IML software: Usage and reference, Version 6*. Cary, NC: SAS Institute, Inc.

SAS Institute (1999). *SAS/STAT user's guide, Version 7-1*. Cary, NC: SAS Institute, Inc.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *American Statistician*, *44*, 174–180.

Schippman, J. S., & Prien, E. P. (1986). Psychometric evaluation of an integrated assessment procedure. *Psychological Reports*, *59*, 111–122.

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

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests*. Cary, NC: SAS Institute, Inc.

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.

Wilson, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons. *Psychological Bulletin*, *59*, 296–300.

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*, 171–196.

Received 16 March 2000; revised version received 12 December 2000