ELSEVIER

Journal of Statistical Planning and Inference 82 (1999) 163–170

journal of statistical planning and inference

www.elsevier.com/locate/jspi

A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence

Yoav Benjamini[a,*], Wei Liu[b]

[a] *Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, 69978 Tel Aviv, Israel*

[b] *Southampton University, UK*

Received 1 October 1997; accepted 1 February 1998

Abstract

For the problems of multiple hypotheses testing, Benjamini and Hochberg (1995, J. Roy. Statist. Soc. Ser. B 57, 289–300), proposed the control of the expected ratio of the number of erroneous rejections to the number of total rejections, the false discovery rate (FDR). The step-up procedure given in that paper controls the FDR when the test statistics are independent. In this paper, a new step-down procedure is presented, and it also controls the FDR when the test statistics are independent. The step-down procedure neither dominates nor is dominated by the step-up procedure. In a large simulation study of the power of the two procedures, the step-down procedure turns out to be more powerful when the number of tested hypotheses is small and many of the hypotheses are far from being true. An example is given to illustrate the step-down procedure.

MSC: 62J15

Keywords: FDR; Multiple comparison procedures; Stepwise procedures

1. Introduction

Multiple tests of related hypotheses using independent test statistics are frequently encountered in practice. One example is the screening of factors for their possible effect on an outcome: in the pharmaceutical industry we may wish to screen many compounds for their possible therapeutical effect, and in quality research we may want to assess different design elements that may have a negative effect on some quality characteristic of a product. A second example is subset analysis in experiments. Another similar problem is multi-center analysis, in which both the overall and center-specific

* Corresponding author. Tel.: +972-3-640-8756; fax: +972-3-640-9357.
E-mail address: benja@math.tau.ac.il (Y. Benjamini)

0378-3758/99/$ - see front matter
PII: S0378-3758(99)00040-3

assessments are desirable. In these examples, the hypotheses to be tested are the same between-treatment comparison, but under changing conditions from study to study, from sub-group to sub-group, from center to center; and the test statistics are independent. Even in meta-analysis, where the main interest is the overall conclusion from many separate studies, there is still interest in identifying individual studies which remain statistically significant when considered simultaneously.

In order to control the multiplicity effect when testing such a family of hypotheses simultaneously, multiple comparison procedures are classically designed to control the type-I familywise error (FWE) rate. This error rate is the probability of one or more false rejections of true hypotheses, irrespective of how many hypotheses are true and what values the parameters of the false hypotheses take (see e.g. Hochberg and Tamhane, 1987). The control of the FWE rate is, however, at the expense of substantially lower power in detecting false hypotheses. Furthermore, the control of the FWE rate may not always be necessary, as in the examples above, since the final overall conclusion derived from the individual inferences may still be correct even if some of the individual inferences are wrong. This led to the proposal (Benjamini and Hochberg, 1995) of controlling the false discovery rate (FDR), which is reviewed next.

Let $H_1,\ldots,H_m$ be the m null hypotheses under consideration, and $P_1,\ldots,P_m$ the corresponding p-values which are assumed to be independent. Let $P_{(1)}\leqslant\cdots\leqslant P_{(m)}$ be the ordered p-values, and $H_{(1)},\ldots,H_{(m)}$ the corresponding hypotheses. The type-I error committed by a multiple testing procedure is viewed through the random variable $\mathbf{Q}=\mathbf{V}/\mathbf{R}$, where $\mathbf{R}$ denotes the number of hypotheses rejected, and $\mathbf{V}$ the number of the true hypotheses erroneously rejected, by the testing procedure. Define $\mathbf{Q}$ to be 0 when $\mathbf{R}=0$, since no error of false rejection is committed in this case. The FDR, Q_e, is then defined to be

$$Q_e=E(\mathbf{Q})=E(\mathbf{V}/\mathbf{R}). \tag{1.1}$$

The value of FDR is less than the FWE rate, i.e.

$$E(\mathbf{V}/\mathbf{R})\leqslant P(\mathbf{V}\geqslant 1). \tag{1.2}$$

Let m_0 denote the number of true hypotheses throughout this paper. Then the second inequality in (1.2) becomes an equality when $m_0=m$. When $m_0<m$, an FDR controlling procedure can be considerably more powerful than an FWE rate controlling procedure at the same level.

The following step-up procedure was shown, in Benjamini and Hochberg (1995), to control the FDR at level q when the P_i are independent.

Let k be the largest i for which $P_{(i)}\leqslant(i/m)q$. Reject $H_{(1)},\ldots,H_{(k)}$.

Note that if $P_{(i)}>(i/m)q$ for $i=1,\ldots,m$ then no hypothesis is rejected. This procedure does not control the FWE rate at q, as can be seen from Hommel (1988). It is a step-up procedure in the sense that it starts from the largest p-value $P_{(m)}$ and proceeds to

smaller p-values by comparing each p-value with the corresponding critical constant, until it finds the first $P_{(i)}$ satisfying $P_{(i)} \leqslant (i/m)q$.

In this paper we propose a new step-down multiple testing procedure that also controls the FDR when the test statistics are independent. By comparing the critical constants of the step-up and step-down procedures, it is pointed out that neither one dominates the other in terms of power. A power comparison using simulation indicates that the step-down test is more powerful than the step-up test when m is small and a large proportion of the hypotheses are false.

2. A step-down FDR controlling procedure

Define the m critical values by

$$\delta_i \equiv 1 - \left[1 - \min\left(1, \frac{m}{m-i+1}q\right)\right]^{1/(m-i+1)}, \quad 1 \leqslant i \leqslant m. \tag{2.1}$$

It is clear that $0 < \delta_1 \leqslant \cdots \leqslant \delta_m \leqslant 1$. The step-down procedure then operates in the following way.

Let k be the smallest i for which $P_{(i)} > \delta_i$. Reject $\mathrm{H}_{(1)}, \ldots, \mathrm{H}_{(k-1)}$.

Note that if $P_{(i)} \leqslant \delta_i$ for $i = 1, \ldots, m$ then all the m hypotheses are rejected. This procedure can again be rephrased as a stepwise algorithm. Start with the smallest p-value $P_{(1)}$, by comparing it with δ_1. If $P_{(1)} > \delta_1$ then stop and reject no hypotheses. Otherwise, step up towards larger p-values, by rejecting $\mathrm{H}_{(i)}$ so long as $P_{(i)} \leqslant \delta_i$, and stop rejecting any more hypotheses when for the first time $P_{(i)} > \delta_i$.

Theorem. *If $P_1, \ldots, P_m$ are independent, then the step-down procedure controls the FDR at level q.*

Proof. Firstly, if $m_0 = 0$ then $\mathbf{V} = 0$, $\mathbf{Q} = 0$ and so $Q_e = 0$. Secondly, if $m_0 = m$ then $\mathbf{V} = \mathbf{R}$ and so

$$Q_e = E(I_{\mathbf{V}>0}) = P(\mathbf{V} > 0) = P(P_{(1)} \leqslant \delta_1) = 1 - [P(P_1 > \delta_1)]^m = 1 - (1 - \delta_1)^m = q.$$

We shall, therefore assume in the rest of the proof that $1 \leqslant m_0 \leqslant m - 1$. Let $m_1 = m - m_0 > 0$, $P'_1, \ldots, P'_{m_1}$ denote the p-values corresponding to the m_1 false hypotheses, and $P^*_1, \ldots, P^*_{m_0}$ denote the p-values corresponding to the m_0 true hypotheses. Define the expectation of $\mathbf{Q}$ conditioning on the values of $P'_1, \ldots, P'_{m_1}$ by

$$Q_e(P'_1, \ldots, P'_{m_1}) \equiv E(\mathbf{Q} \mid P'_1, \ldots, P'_{m_1}).$$

Next, we show that $Q_e(P'_1, \ldots, P'_{m_1}) \leqslant q$, from which the theorem clearly follows.

For this, let $P'_{(1)} \leqslant \cdots \leqslant P'_{(m_1)}$ denote the ordered values of $P'_1, \ldots, P'_{m_1}$. Define S, $0 \leqslant S \leqslant m_1$, to be the largest integer j satisfying $P'_{(1)} \leqslant \delta_1, \ldots, P'_{(j)} \leqslant \delta_j$; $S=0$ if $P'_{(1)} > \delta_1$.

Now we have

$$
\begin{aligned}
Q_e(P'_1,\ldots,P'_{m_1}) &= E\left(\frac{\mathbf{V}}{\mathbf{R}} I_{\mathbf{V}>0} \mid P'_1,\ldots,P'_{m_1}\right) \\
&\leqslant E\left(\frac{\mathbf{V}}{S+\mathbf{V}} I_{\mathbf{V}>0} \mid P'_1,\ldots,P'_{m_1}\right) \qquad (2.4)\\
&\leqslant \frac{m_0}{S+m_0} E(I_{\mathbf{V}>0} \mid P'_1,\ldots,P'_{m_1}) \\
&\leqslant \frac{m_0}{S+m_0} P(\min(P^*_1,\ldots,P^*_{m_0}) \leqslant \delta_{S+1}) \\
&= \frac{m_0}{S+m_0}[1-(1-\delta_{S+1})^{m_0}] \\
&\leqslant \frac{m_0}{S+m_0}[1-(1-\delta_{S+1})^{m-S}] \qquad (2.5)\\
&= \frac{m_0}{S+m_0}\min\left(1, \frac{m}{m-S} q\right) \\
&\leqslant \frac{m_0 m}{(S+m_0)(m-S)} q \\
&\leqslant q, \qquad (2.6)
\end{aligned}
$$

where inequality (2.4) follows from the relationship $\mathbf{R} \geqslant S + \mathbf{V}$, and inequalities (2.5 and 2.6) follow from the fact that $m_0 + S \leqslant m$. The proof is thus completed. □

Remark 1. It is clear from the definition of δ_i in (2.1) that $\delta_i = 1$ for $m(1-q)+1 \leqslant i \leqslant m$. Consequently, the hypotheses $H_{(i)}$, $m(1-q)+1 \leqslant i \leqslant m$ will be rejected, no matter how large the corresponding p-values are. This is not surprising, since all the hypotheses are tested simultaneously and the control of FDR allows a few erroneous rejections if many correct rejections have already been made.

There are situations where one does not want to reject those hypotheses with large p-values. One reason is that a rejected hypothesis might be highlighted for further investigation. Another reason is to avoid contaminating a set of rejected hypotheses which are most likely false by those hypotheses which are almost certainly true. If this is the case, one may simply add the constraint that a hypothesis cannot be rejected if the corresponding p-value is larger than some prespecified value. This, of course, will not inflate the FDR.

Remark 2. Note that δ_1 cannot be improved upon, as can be seen from the proof when $m_0 = m$. At the other end, $\delta_m = \min(1, mq)$ is also stringent; this can be seen from the proof by setting $m_1 = m - 1$ and letting the p-values $P'_1,\ldots,P'_{m_1}$ approach zero almost surely. No similar statement, however, can be made on $\delta_2,\ldots,\delta_{m-1}$, except for those $\delta_i = 1$, $m(1-q)+1 \leqslant i \leqslant m$.

3. An example

The national Surgical Adjuvant Breast and Bowel Project published data from a trial of L-phenylalanine mustard, 5-fluorouracil, and tamoxifen (PFT) versus L-phenylalanine mustard and 5-fluorouracil (PF) in 1891 patients with primary operable breast cancer and positive nodes (Fisher et al., 1983). These investigators found "evidence for heterogeneity in response to PFT therapy that is age and progesterone receptor dependent". We use these data (see Table 1 below), as described and analyzed by Gail and Simon (1985), to demonstrate the step-down FDR controlling procedure, and compare the results with that of the step-up procedure.

The observed p-values are for two-sided z-tests, and the four test statistics are independent. Gail and Simon have developed in their paper a test of the null hypothesis of no cross-over effect, also termed as no qualitative interaction, which means that the direction of difference is the same in all subgroups. Using their test, they have found that this single hypothesis can be rejected at the 0.1 level, but not at the 0.05 level.

However, a medical practitioner might be more interested in a decision about each subgroup, in order to prescribe an appropriate treatment. Thus, the testing of the single overall hypothesis is, while important for research purposes, not enough. It is desirable to come up with statements about the individual subgroups, and a multiple testing procedure can be employed for this purpose.

To perform the step-down procedure at the 0.1 level, the ordered p-values 0.0058, 0.0362, 0.0972 and 0.4440 are compared with the critical values 0.0260, 0.0466, 0.1056 and 0.4 stepwisely. Start with $0.0058 \leqslant 0.0260$, until $0.4440 > 0.4$ for the first time. So the three hypotheses corresponding to the three smaller p-values are rejected, i.e. the hypotheses of no difference in all subgroups except for the subgroup of "Age less than 50 and PR greater than 10" are rejected. It is noted that, in this example, the step-down multiple testing procedure rejects the same three hypotheses as the procedure that tests each hypothesis separately at the level 0.1 (so no loss of power is incurred by the simultaneous consideration). At the 0.05 level the step-down test compares the p-values with 0.0127, 0.0227, 0.0513 and 0.02, and only the hypothesis corresponding to the smallest p-value is rejected.

Table 1
Analysis of proportions free of breast cancer at 3 years in 4 subgroups

Subgroup	Progesterone	Age < 50 PR < 10	Age > 50 PR < 10	Age < 50 PR > 10	Age > 50 PR > 10
Proportion	PF	0.599	0.526	0.651	0.639
	PFT	0.436	0.639	0.698	0.790
Standard error	PF	0.0542	0.0510	0.0431	0.0386
	PFT	0.0572	0.0463	0.0438	0.0387
Difference		0.163	−0.114	−0.047	−0.151
SE of difference		0.0778	0.0689	0.0614	0.0547
Observed z-statistic		2.095	−1.655	−0.7655	−2.7605
p-values (two-sided)		0.0362	0.0972	0.4440	0.0058

The step-up procedure of 0.1 level compares the ordered p-values with 0.025, 0.05, 0.075 and 0.1. Since $0.4440 > 0.1$ and then $0.0972 > 0.075$, and the second smallest p-value is the first one that is smaller than the corresponding critical value, $0.0362 \leqslant 0.05$. Hence, the two hypotheses corresponding to the two smaller p-values are rejected. So in this case the step-up procedure is less powerful than the step-down. At the 0.05 level, the step-up procedure rejects the same single hypothesis that the step-down procedure does.

4. Power comparison

A simple comparison of power can be made by comparing the critical constants of the procedures. The step-down procedure that controls the FWE rate at q, for independent test statistics, uses the critical constants

$$\delta_i' \equiv 1 - (1-q)^{1/(m-i+1)}, \quad 1 \leqslant i \leqslant m. \tag{3.1}$$

This procedure is sometimes referred to as a Holm-type procedure (Holm, 1979). It is clear that $\delta_i' < \delta_i$ for $i = 2, 3, \ldots, m$ and $\delta_1' = \delta_1$. So the FDR controlling step-down procedure is uniformly more powerful than the corresponding FWE controlling one, as one would expect.

The power comparison between the step-down and step-up FDR controlling procedures is more complicated. Even if all the critical constants of the step-down procedure are larger than the corresponding critical constants of the step-up procedure, the "more powerful direction of stepping" of the step-up procedure does not guarantee that the step-down procedure is more powerful. A comparison between the δ_i in (2.1) and the $c_i = (i/m)q$ of the step-up procedure leads to the following observations. Obviously, $\delta_1 > c_1$ for all m, though by a very small amount. For $m \geqslant 4$, $\delta_i < c_i$ for all i upto some $i_0(m)$, and then δ_i becomes larger. Comparing the first term in the series expansion of $(1-c_i)^{m+1-i}$ with $(1-\delta_i)^{m+1-i}$, the two are equal when $i_0(m)$ satisfies $i_0(m)(m+1-i_0(m))^2 \approx m^2$. For small m that is the case when $i_0(m) \approx (m+1) - \sqrt{m+1}$; for large m this happens when $i_0(m) = m(1-q)$. It may, therefore, be concluded that, beyond small m, as long as the number of false null hypothesis is not larger than the $i_0(m)$ given above, the step-up procedure will be more powerful than the step-down procedure. On the other hand, if most null hypotheses are far from being true then the step-down procedure should be more powerful.

To make a more detailed comparison of power, we have carried out a large simulation study. The four procedures compared are the step-up procedure (SUFDR), the step-down procedure (SDFDR), the step-down FWE rate-controlling procedure using the critical constants in (3.1) (SDFWE) and the adaptive stepwise FDR controlling procedure (AFDR) of Benjamini and Hochberg (1999). Note that the AFDR has been shown to control the FDR only by a simulation study similar to the one given below (i.e. normally distributed random variables), and that the four procedures control the

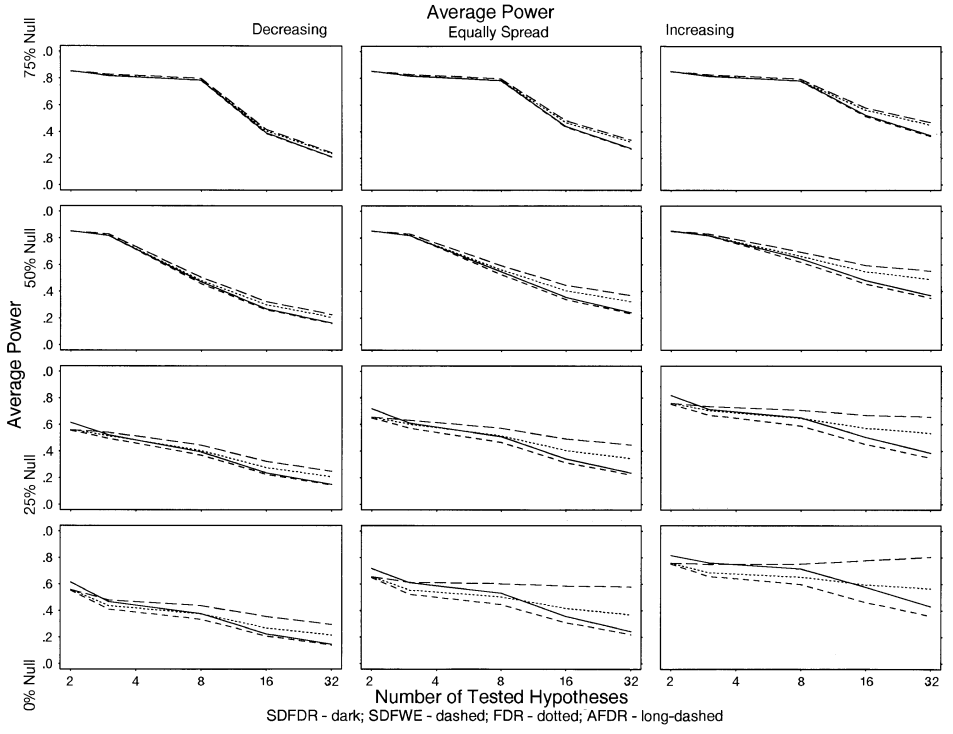

Fig. 1. Simulation results on the average power of the step-down FDR controlling (SDFDR) – solid lines, the step-up FDR controlling (SUFDR) – dotted lines, the step-down FWE controlling (SDFWE) – short dashed lines, and the adaptive FDR controlling (AFDR) – long dashed lines.

FWE rate at the same level when all the hypotheses are true. The criterion for comparison is the average power, the proportion of the false hypotheses that are rejected.

The hypotheses were $H_i: \theta_i = 0$, tested (two-sided) at $q = 0.05$ by using independent statistics T_i with normal distributions $T_i \sim N(\theta_i, 1)$, $1 \leqslant i \leqslant m$. The number of hypotheses was $m = 4, 8, 16, 32$ and 64, with $m_0 = 3m/4, m/2, m/4$ and 0. Three configurations of the nonzero θ_i were considered: (E) equally spaced over $(0, L]$; (D) placed with decreasing density away from zero over $(0, L]$; (I) placed with increasing density away from zero over $(0, L]$. The value of L was set at levels, 2, 3, 5 and 10, varying therefore the signal-to-noise ratio. The simulation study used 20 000 repetitions, with the estimated standard errors about 0.0008 to 0.0016. 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 reduced the variance of a comparison between two procedures or two configurations to less than twice the variance of a single one.

Fig. 1 presents the simulation results on the average power for $L = 3$. It can be seen that the SDFDR performs best for $m = 2$ under all configurations, and is generally more powerful for $m = 3$ and 4. The AFDR performs better overall than both the

SDFDR and SUFDR for $m > 4$, but the control of FDR of the AFDR has not been proved analytically thus far. SDFDR is uniformly more powerful than SDFWE, which agrees with the theoretical result. Between SUFDR and SDFDR, SUFDR is generally more powerful for larger values of m and smaller proportion of which are false (the upper-left of Fig. 1). However, the power of SDFDR can be substantially larger than that of SUFDR, when the value of m is small and most of the hypotheses are false. For example, when four hypotheses are tested, one of which is true, the power of the SDFDR is the same as that of the AFDR, and higher than that of the SUFDR. In the extreme configuration, when all hypotheses are far from being true (the lower-right of Fig. 1) the SDFDR is more powerful than SUFDR even when 32 hypotheses are tested. For 16 hypotheses the difference in power is as much as 0.1. More generally though, for 4–16 tested hypotheses, the SDFDR is superior only when the proportion of the true null hypotheses is below half.

The conclusion is that, between the two mathematically proved FDR controlling procedures, the SDFDR is recommendable if $m \leqslant 4$, or if $m \leqslant 16$ and most of the hypotheses are false. Otherwise, the SUFDR is recommended.

Acknowledgements

The authors would like to thank Tony Hayter and Yosi Hochberg, discussions with whom initiated this work. The second author would like to thank the ASG and RS for some financial support on his work.

References

Benjamini, Y., Hochberg, Y., 1999. On the adaptive control of the false discovery rate in multiple testing with independent statistics, J. Behav. Educ. Statist., in press.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate – a new and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57, 289–300.

Fisher, B., Redmond, C., Brown, A., Wickerham, D.L., Wolmark, N., Allegra, J., Escher, G., Lippman, M., Savlov, E., Wittliff, J., Fisher, E.R., 1983. Influence of tumor estrogen and progesterone receptor levels on the response to Tamoxifen and chemotherapy in primary breast cancer. J. Clini. Oncol. 1, 227–241.

Gail, M., Simon, R., 1985. Testing for qualitative interactions between treatment effects and patient subset. Biometrics 41, 361–372.

Hochberg, Y., Tamhane, A., 1987. Multiple Comparison Procedures. Wiley, New York.

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.