A distribution-free multiple-test procedure that controls the false discovery rate

Yoav Benjamini
Tel Aviv University, Israel
Wei Liu
Southampton University, UK

Abstract

For the problems of testing a family of hypotheses $H_1, \cdots, H_m$ simultaneously, Benjamini and Hochberg (1995) proposed the control of the False Discovery Rate (FDR), instead of the classical familywise error rate (FWER). Most FDR-control multiple-test procedures require certain assumptions on the joint distribution of the test statistics. The only distribution-free FDR-control multiple-test procedure is given by Benjamini and Yekutieli (BY) (1999). In this paper, another distribution-free FDR-control multiple-test procedure is provided. This new procedure is often more powerful then BY's procedure when the number of hypotheses m is small and most of the false hypotheses are far from being true. An example is given to illustrate the new procedure.

Key Words and Phrases: FDR, Multiple-test Procedures, Stepwise procedures.

AMS 1991 Subject Classification: 62J15.

Running title: Distribution-free FDR-control Procedure.

Corresponding Author: Yoav Benjamini, Dept. of Statistics and OR, Tel Aviv University, Tel Aviv, Israel. Email: benja@math.tau.ac.il. Fax:972-3-6409357.

1. Introduction

When testing m related null hypotheses $H_1, \cdots, H_m$ simultaneously, it is often required to control the multiple testing effect to avoid excessive number of false rejections. Multiple test procedures are classically designed to control the type I familywise error rate (FWER). It is the probability of false rejections of one or more true hypotheses, irrespective of how many and which hypotheses are true and what values the parameters of the false hypotheses take. See Hochberg and Tamhane (1987) for a general introduction. The control of FWER may however be unnecessary stringent and consequently reduce the power in detecting false hypotheses. This leads to Benjamini and Hochberg's (BH) (1995) approach of controlling the false discovery rate (FDR).

In this approach, the type I error committed by a multiple test procedure is quantified 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. 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 $Q_e = \mathrm{E}(\mathbf{Q}) = \mathrm{E}(\mathbf{V}/\mathbf{R})$. See Benjamini and Hochberg (1995) for a rational behind the control of FDR. The value of FDR is in between the per comparison error rate $\mathrm{E}(\mathbf{V}/m)$ and the FWER $P(\mathbf{V} \geq 1)$: $\mathrm{E}(\mathbf{V}/m) \leq \mathrm{E}(\mathbf{V}/\mathbf{R}) \leq \mathrm{P}(\mathbf{V} \geq 1)$. Let m_0 denote the number of true hypotheses throughout this paper. When $m_0 = m$, $\mathrm{E}(\mathbf{V}/\mathbf{R}) = \mathrm{P}(\mathbf{V} \geq 1)$. When $m_0 < m$, a FDR-control test can be considerably more powerful than a FWER-control test at the same level.

Several FDR-control test procedures have appeared in the literature under various assumptions on the joint distribution of the test statistics. When the test statistics are independent, there are BH's (1995) step-up test and Benjamini and Liu's (1997) step-down test. For comparisons between several treatments and a control, Troendle (1998) proposed a simulation-based step-up test. When the test statistics have a certain positive regression dependence, Benjamini and Yekutieli (BY) (1999) show that BH's step-up test still controls the FDR. The only procedure available that controls FDR without any assumption on the joint distribution of the test statistics is also given by BY (1999) and stated in the next section.

In this paper we provide another distribution-free multiple-test procedure that controls the FDR. This new procedure is often more powerful than BY's procedure when the number of hypotheses m is small and most of the false hypotheses are far from being true.

2. A distribution-free FDR-control test

Let P_i be the p-value for testing H_i, $i = 1, \cdots, m$. Suppose $P_{(1)} \leq \cdots \leq P_{(m)}$ are the ordered p-values, and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses. Define the m critical values by

$$d_i \equiv \min\left(1, \frac{m}{(m-i+1)^2} q\right), \quad 1 \leq i \leq m \tag{2.1}$$

which satisfy $0 < d_1 \leq \cdots \leq d_m \leq 1$. Then the new test $\mathcal{P}_1$

rejects $H_{(1)}, \cdots, H_{(k-1)}$ where k is the smallest i for which $P_{(i)} > d_i$.

Note that if $P_{(i)} \leq d_i$ for $i = 1, \cdots, m$ then all the m hypotheses are rejected. This is a step-down test. It starts with the smallest p-value $P_{(1)}$, by comparing it with d_1. If $P_{(1)} > d_1$ then stop and reject no hypotheses. Otherwise, step up towards larger p-values , by rejecting $H_{(i)}$ so long as $P_{(i)} \leq d_i$, and stop rejecting any more hypotheses when for the first time $P_{(i)} > d_i$.

Theorem. Test $\mathcal{P}_1$ 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

$$Q_e = \mathrm{E}(I_{\mathbf{V}>0}) = \mathrm{P}\{\mathbf{V} > 0\} = \mathrm{P}\{P_{(1)} \leq d_1\} \leq \sum_{i=1}^{m} \mathrm{P}\{P_i \leq d_1\} = md_1 = m\frac{m}{m^2}q = q.$$

We shall therefore assume in the rest of the proof that $1 \leq m_0 \leq m-1$. Let $m_1 = m - m_0 > 0$, $P_1', \cdots, P_{m_1}'$ denote the p-values corresponding to the m_1 false hypotheses, and $P_1^*, \cdots, P_{m_0}^*$ denote the p-values corresponding to the m_0 true hypotheses. Denote the expectation of $\mathbf{Q}$ conditioning on the values of $P_1', \cdots, P_{m_1}'$ by

$$Q_e(P_1', \cdots, P_{m_1}') \equiv \mathrm{E}(\mathbf{Q} \mid P_1', \cdots, P_{m_1}').$$

Next we show that $Q_e(P_1', \cdots, P_{m_1}') \leq q$, from which the theorem clearly follows.

For this, let $P_{(1)}' \leq \cdots \leq P_{(m_1)}'$ denote the ordered values of $P_1', \cdots, P_{m_1}'$. Define S $(0 \leq S \leq m_1)$ to be the largest integer j satisfying $P_{(1)}' \leq d_1, \cdots, P_{(j)}' \leq d_j$; $S = 0$ if $P_{(1)}' > d_1$. Now we have

$$\begin{aligned}
Q_e(P_1', \cdots, P_{m_1}') &= \mathrm{E}\left(\frac{\mathbf{V}}{\mathbf{R}} I_{\mathbf{V}>0} \mid P_1', \cdots, P_{m_1}'\right) \\
&\leq \mathrm{E}\left(\frac{\mathbf{V}}{S+\mathbf{V}} I_{\mathbf{V}>0} \mid P_1', \cdots, P_{m_1}'\right) \qquad (2.4)\\
&\leq \frac{m_0}{S+m_0}\mathrm{E}\left(I_{\mathbf{V}>0} \mid P_1', \cdots, P_{m_1}'\right) \\
&\leq \frac{m_0}{S+m_0}\mathrm{P}\{\min(P_1^*, \cdots, P_{m_0}^*) \leq d_{S+1}\} \\
&\leq \frac{m_0}{S+m_0}\sum_{i=1}^{m_0}\mathrm{P}\{P_i^* \leq d_{S+1}\} \\
&\leq \frac{m_0}{S+m_0}(m-S)d_{S+1} \qquad (2.5)\\
&= \frac{m_0(m-S)}{S+m_0}\min\left(1, \frac{m}{(m-S)^2}q\right) \\
&\leq \frac{m_0 m}{(S+m_0)(m-S)}q \\
&\leq q, \qquad (2.6)
\end{aligned}$$

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

The distribution-free test of BY (1999), denoted as $\mathcal{P}_2$, is to

reject $H_{(1)}, \cdots, H_{(k)}$ where k is the largest i for which $P_{(i)} \leq u_i$,

where $u_i = \frac{i}{m(1+1/2+\cdots+1/m)}q$. Note that if $P_{(i)} > u_i$ for $i = 1, \cdots, m$ then no H_i is rejected. This is a step-up test. 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 $P_{(i)} \leq u_i$.

By comparing the critical values of $\mathcal{P}_1$ and $\mathcal{P}_2$, it is clear that neither procedure dominates the other. However, the critical values d_i of $\mathcal{P}_1$ are considerably larger than the corresponding critical values u_i of $\mathcal{P}_2$. This indicates that procedure $\mathcal{P}_1$ will often be more powerful than procedure $\mathcal{P}_2$ at least when the number of hypotheses m is small and most of the false hypotheses are far from being true.

3. An example

To compare the tensile strength (lb/in.^{2}) of portland under four different mixing techniques, a completely randomized experiment with the four mixing techniques and four replicates has been run (Montgomery, 1991, page 89). The four sample means are given by $\bar{Y}_1 = 2666.3$, $\bar{Y}_2 = 2933.8$, $\bar{Y}_3 = 2971.0$, $\bar{Y}_4 = 3156.3$, and the sample std is 113.25 with 12 df. The p-values for the six pairwise comparisons, based on 2-sided t-tests with 12 df, are given by: $p_{12} = 0.006$, $p_{13} = 0.003$, $p_{14} = 0.000, p_{23} = 0.651$, $p_{24} = 0.017$, $p_{34} = 0.039$. (Here p_{ij} is the p-value for testing the hypothesis H_{ij} that treatments i and j have the same mean values.)

To carry out a multiple-test of H_{ij} $(1 \leq i < j \leq 4)$, the six p-values are ordered: 0.000, 0.003, 0.006, 0.017, 0.039, 0.651. For $m = 6$ and $\alpha = 0.05$, the critical values d_i of $\mathcal{P}_1$ are given by 0.008, 0.012, 0.019, 0.033, 0.075, 0.300, and the critical values u_i of $\mathcal{P}_2$ are given by 0.003, 0.007, 0.010, 0.014, 0.017, 0.020. So $\mathcal{P}_1$ does not reject only H_{23} which has $p_{23} = 0.651$, while $\mathcal{P}_2$ does not reject H_{23}, H_{34} and H_{24} which have the corresponding p-values $p_{23} = 0.651, p_{34} = 0.039$ and $p_{24} = 0.017$. The new procedure $\mathcal{P}_1$ rejects two more hypotheses than BY's procedure $\mathcal{P}_2$. Finally, we note that the $\alpha = 0.05$ step-up FWER-control multiple test of Liu (1997) does not reject H_{23} and H_{34}.

References

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate - a new and powerful approach to multiple testing, *JRSS B*, **57**,289-300.

Benjamini, Y. and Liu, W. (1997). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J Statistical Planning and Inference*, to appear.

Benjamini, Y. and Yekutieli, D. (1999). The control of false discovery rate in multiple testing under dependency. Manuscript, Department of Mathematical Science, University of Tel Aviv.

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

Liu, W. (1997). On step-up tests for comparing several treatments. *Statistica Sinica*, **7(4)**, 957-972.

Montgomery, D.C. (1991). *Design and Analysis of Experiments*, Wiley & Sons.

Troendle, J.F. (1998). Stepwise normal theory multiple test procedures controlling the false discovery rate. Manuscript, Biometry and Mathematical Statistics Branch, Division of Epidemiology, Statistics, and Prevention Research, National Institute of Child Health and Human Development.