The International Journal of Biostatistics

Volume 3, Issue 1

2007

Article 11

The Benjamini-Hochberg Method in the Case of Discrete Test Statistics

José A. Ferreira, *Vrije University Medical Centre,
Amsterdam*

Recommended Citation:

Ferreira, José A. (2007) "The Benjamini-Hochberg Method in the Case of Discrete Test Statistics," *The International Journal of Biostatistics*: Vol. 3: Iss. 1, Article 11.

DOI: 10.2202/1557-4679.1065

The Benjamini-Hochberg Method in the Case of Discrete Test Statistics

José A. Ferreira

Abstract

We present a reformulation of the Benjamini-Hochberg method that is useful in 'large-scale' multiple testing problems based on discrete test statistics and derive its basic asymptotic (as the number of hypotheses tends to infinity) properties, subsuming earlier results. A set of gene expression data is used to illustrate the workings of the method in a multiple testing problem based on Kolmogorov-Smirnov and Mann-Whitney statistics.

KEYWORDS: multiple testing, false discovery rate, average power, non-parametric tests

Author Notes: I thank Prof. A. Richardson for information about the data used in Section 4, and two reviewers for useful comments on the paper.

1 Introduction

It has been suggested by several authors that the Benjamini-Hochberg method is not completely adequate to treat multiple testing problems involving discrete statistics; see, for example, Kulinskaya and Lewin (2006) and Pounds and Cheng (2006). This is understandable since the method, as originally developed by Benjamini and Hochberg (1995), presupposes that the p-values computed under the set of null hypotheses are uniformly distributed on $[0, 1]$. Certain ‘pathologies’ arise for example in multiple non-parametric tests based on small samples, and they have to do with the discreteness and non-uniformity of the distribution of p-values. The discrete nature of the p-values sometimes causes one to reject either no hypotheses or too many hypotheses (e.g. p. 5120 of Tusher, Tibshirani, and Chu (2001)), which in practical terms means either making absolutely no use of the data or incurring an intolerable false discovery rate. However, this inconvenience is inherent to the use of small samples and essentially reflects the relative weakness of the effects one is trying to detect, so it is hard to see how it could be obviated. On the other hand, discrepancies between the null distribution and the uniform distribution could affect the estimates or bounds on the false discovery rate and average power (see the definition of these concepts below) and therefore could very well lead to misleading or sub-optimal results.

Figure 1 shows histograms of the p-values of 10,000 simulated Kolmogorov-Smirnov (KS) and Mann-Whitney (MW) tests comparing two independent random samples of seven observations each under the null hypothesis that the samples have the same continuous distribution. The histograms illustrate the extent of non-uniformity one is likely to encounter in ‘density estimates’ constructed from p-values of non-parametric tests associated to null hypotheses. They show that applying the Benjamini-Hochberg method to the p-values of multiple KS or MW tests will cause one to overestimate the proportion of ‘small’ p-values coming from null hypotheses and thus reject fewer p-values than one is entitled to at a given false discovery rate.

Multiple testing problems with small samples happen to be rather common in the statistical analysis of microarray experiments, and there are good reasons why the analysis of microarray data should sometimes be based on non-parametric tools such as the KS and MW statistics; for instance, the use of a t-test may not be appropriate with certain data, or one may simply want to double-check the validity of ‘model-based’ results by means of a distribution-free method. It seems worthwhile, therefore, to investigate the behaviour of the Benjamini-Hochberg in the case of discrete test statistics. With this in mind, in this article we present a reformulation of the Benjamini-Hochberg method and prove its basic asymptotic

Figure 1: Histograms of 10,000 p-values corresponding to KS (left panel) and MW (right panel) tests comparing two independent samples of size seven each under the null hypothesis that the samples have the same continuous distribution.

properties—namely lower and upper bounds for the false discovery rate and average power as the number of hypotheses tends to infinity—in full generality, so as to cover both the continuous and the discrete case. It will be seen that the Benjamini-Hochberg method *is* adequate to deal with discrete test statistics, in so far as it controls the false discovery rate at a given level and it does so as sharply as possible under each set of circumstances; see Remark (iv) at the end of Section 2.

In Section 2 we present the reformulation of the Benjamini-Hochberg method and then state our main result in the form of a proposition. A complete and simple proof of the proposition is given in Section 3. Finally, in Section 4 we use a set of microarray data to illustrate the application of the method in the discrete case.

2 The reformulation, and the main result

We assume that we are given two sequences $\{T'_j\}$ and $\{T''_j\}$ of random variables and two distribution functions (d.f.'s) F and G on $\mathbb{R}$ such that

$$F_n(t) := \frac{1}{n_0} \sum_{j=1}^{n_0} 1_{\{T'_j \leq t\}} \xrightarrow{P} F(t) \quad (2.1)$$

and

$$G_n(t) := \frac{1}{n_1} \sum_{j=1}^{n_1} 1_{\{T''_j \leq t\}} \xrightarrow{P} G(t) \quad (2.2)$$

uniformly in t as $n_0, n_1 \rightarrow \infty$ and $\gamma_n := n_0/n \rightarrow \gamma$, where $n = n_0 + n_1$ and γ is some number in $(0, 1)$.

For simplicity, we write $T_j = T'_j$ for $j = 1, \dots, n_0$ and $T_{n_0+j} = T''_j$ for $j = 1, \dots, n_1$. Then, for each n , $T_1, \dots, T_n$ are interpreted as the *test statistics*; the first n_0 statistics are regarded as being computed under *the null hypothesis*, and accordingly their empirical d.f. F_n is assumed to converge to the ‘null d.f.’ F , and the remaining n_1 are regarded as being computed under *alternative hypotheses*. The null d.f. F is regarded as known; γ , the *limiting proportion of null hypotheses*, and G are regarded as unknowns.

For convenience, and in order to conform with the original framework of Benjamini and Hochberg (1995), we *assume in all that follows that small values of the test statistics indicate a departure from the null*, or equivalently that each test rejects the null hypothesis for small values of the test statistic. In this way, the p-values are defined as $F(T_j)$ and can be seen as special cases of test statistics.

Assumptions (2.1) and (2.2) imply the convergence of the empirical d.f. of the test statistics $T_1, \dots, T_n$, which we denote by H_n , to a mixture of F and G :

$$H_n \equiv \gamma_n F_n + (1 - \gamma_n) G_n \xrightarrow{P} H := \gamma F + (1 - \gamma) G. \quad (2.3)$$

Example. In ‘genomics’ one is often interested in detecting systematic differences between gene expression measurements from two ‘groups’ (e.g. tumour and control patients) across a large number n of genes (see McLachlan, Do, and Ambroise (2004), for example). Suppose a proportion of $\gamma_n \approx \gamma \in (0, 1)$ genes are equally expressive in the two groups. Then $n_0 = \gamma_n n$ statistics are to be computed under the null hypothesis of equality of distributions, and $n_1 = (1 - \gamma_n)n$ statistics are to be computed under (generally different) alternative hypotheses.

To be specific, suppose that, for each gene, two samples of m observations each are to be compared by means of the KS statistic. Then one may define the test statistics by $T = m(1 - D_m)$, where D_m is the supremum distance between the empirical d.f.’s of the two groups, in which case T takes values in $\{0, 1, \dots, m - 1\}$ and small values of T indicate a departure from the null hypothesis. In the setting originally proposed by Benjamini and Hochberg (1995), the statistics would really be the p-values, namely $1 - \varphi_{D_m}(D_m -)$, where φ_{D_m} is the d.f. of D_m under the null hypothesis. In order to derive meaningful approximate statements about the Benjamini-Hochberg method in this context, we need to assume that (2.1) and (2.3) hold with $F = \varphi_{D_m}$. $\square$

Theorem 2 of Benjamini and Hochberg (1995), Lemma 1 of Storey, Taylor, and Siegmund (2004), and Theorem 3.1 of Ferreira and Zwinderman (2006b) suggest the following reformulation of the Benjamini-Hochberg method: Given $\delta \in (0, \bar{\gamma})$ and $\bar{\gamma} \in [\gamma, 1]$, reject all the hypotheses whose test statistics are *strictly smaller* than the *threshold*

$$t_n(\bar{\gamma}, \delta) := \sup \left\{ t : F(t) < \frac{\delta}{\bar{\gamma}} H_n(t) \right\}, \quad (2.4)$$

where the supremum is defined to be $-\infty$ if the set is empty. Because asymptotically (as stated in the proposition below) this procedure keeps the false discovery rate below δ , and possibly strictly below δ , the number δ will be called the **nominal false discovery rate**.

Before explaining why the procedure is essentially equivalent to Benjamini and Hochberg's, let us note that the number $\bar{\gamma}$ is meant to be an upper bound on γ . If no such upper bound can be set in a given application, then one simply takes $\bar{\gamma} = 1$. It will be seen in Section 4 that obtaining a reliable upper bound from the data is particularly easy in the discrete case.

If F is standard uniform, H_n is the empirical d.f. of the p-values and $\bar{\gamma} = 1$, the procedure is the Benjamini-Hochberg method as formulated by Storey et al. (2004).

If F is strictly increasing then the empirical d.f. of the p-values is given by $W_n(w) = H_n(F^{-1}(w))$, $w \in [0, 1]$, so

$$\begin{aligned} t_n(\bar{\gamma}, \delta) &= \sup \left\{ t : F(t) < \frac{\delta}{\bar{\gamma}} H_n(t) \right\} \\ &= F^{-1} \left(\sup \left\{ w : w < \frac{\delta}{\bar{\gamma}} H_n(F^{-1}(w)) \right\} \right) = F^{-1}(w_n(\bar{\gamma}, \delta)), \end{aligned}$$

where

$$w_n(\bar{\gamma}, \delta) := \sup \left\{ w : w < \frac{\delta}{\bar{\gamma}} W_n(w) \right\} \quad (2.5)$$

is the threshold corresponding to the Benjamini-Hochberg method. This shows that if F is strictly increasing and $\bar{\gamma} = 1$ then the procedure based on (2.4) is equivalent to the original one. (However, we note that the use of (2.4) in place of (2.5) allows greater flexibility and may simplify actual numerical computations.)

Finally, we show that even when F is discrete (2.4) and (2.5) will typically (though not always) yield the same number of rejections. Suppose F has a finite number of jumps. Then F/H_n is a right-continuous step function which increases where F does (and in spite of an eventual increase of H_n), so $t_n(\bar{\gamma}, \delta)$ is necessarily a point of increase of F . In contrast, since $w \rightarrow w/W_n(w)$ cannot increase when W_n does, $w_n(\bar{\gamma}, \delta)$ is not a point of increase of W_n . If we assume that W_n jumps wherever F does (which happens with probability tending to one as $n \rightarrow \infty$), it follows from these two observations and from the definitions of $t_n(\bar{\gamma}, \delta)$ and $w_n(\bar{\gamma}, \delta)$ that $F(t_n^*(\bar{\gamma}, \delta)) < w_n(\bar{\gamma}, \delta) < F(t_n(\bar{\gamma}, \delta))$, where $t_n^*(\bar{\gamma}, \delta)$ is the point of increase of F immediately preceding $t_n(\bar{\gamma}, \delta)$. If we further assume that W_n has no points of increase other than those of F (which will be the case if all statistics have the

same range), so that no test statistics lie in $(t_n^*(\bar{\gamma}, \delta), t_n(\bar{\gamma}, \delta))$, we may conclude that rejecting all the p-values smaller than or equal to $w_n(\bar{\gamma}, \delta)$ is equivalent to rejecting all p-values smaller than or equal to $F(t_n^*(\bar{\gamma}, \delta))$, which is equivalent to rejecting all the statistics smaller than or equal to $t_n^*(\bar{\gamma}, \delta)$ and equivalent to rejecting all the statistics *strictly smaller* than $t_n(\bar{\gamma}, \delta)$.

The usefulness of the Benjamini-Hochberg method in multiple testing problems involving large numbers of hypotheses derives mainly from its ability to specify approximately a value for the **false discovery rate**, the r.v.

$$\Pi_{1,n}(\bar{\gamma}, \delta) := \frac{n_0 F_n(t_n(\bar{\gamma}, \delta) -)}{\max \{n H_n(t_n(\bar{\gamma}, \delta) -), 1\}},$$

and to estimate, or obtain approximate bounds for, **average power**, namely

$$\Pi_{2,n}(\bar{\gamma}, \delta) := \frac{n H_n(t_n(\bar{\gamma}, \delta) -) - n_0 F_n(t_n(\bar{\gamma}, \delta) -)}{n_1}.$$

$\Pi_{1,n}(\bar{\gamma}, \delta)$ and $\Pi_{2,n}(\bar{\gamma}, \delta)$ are respectively the ratio of the number of incorrectly rejected hypotheses to the total number of rejected hypotheses and the ratio of the number of correctly rejected hypotheses to the total number of alternative hypotheses (those one would like to detect) that result from the application of the Benjamini-Hochberg method. They are usually defined with $F_n(t_n(\bar{\gamma}, \delta))$ and $H_n(t_n(\bar{\gamma}, \delta))$ in place of $F_n(t_n(\bar{\gamma}, \delta) -)$ and $H_n(t_n(\bar{\gamma}, \delta) -)$; the modification is irrelevant in the continuous case, but in the discrete case it is needed in order to guarantee the control of the false discovery rate at a specified value.

To state our main result—which consists of a couple of asymptotic results about $t_n(\bar{\gamma}, \delta)$, $\Pi_{1,n}(\bar{\gamma}, \delta)$ and $\Pi_{2,n}(\bar{\gamma}, \delta)$ —we need to define

$$t_0(\bar{\gamma}, \delta) = \sup \left\{ t : F(t) < \frac{\delta}{\bar{\gamma}} H(t) \right\} = \sup \left\{ t : G(t) < \frac{(\bar{\gamma} - \delta\gamma)}{\delta(1 - \gamma)} F(t) \right\} \quad (2.6)$$

for $\delta \in (0, \bar{\gamma})$, $\bar{\gamma} \in [\gamma, 1]$. We also put $T_\Psi = \inf\{t : \Psi(t) > 0\}$ for any d.f. Ψ , and define $S_{F,G}$ as the set of points $(\delta, \bar{\gamma})$ satisfying $\delta \in (0, \bar{\gamma})$, $\bar{\gamma} \in [\gamma, 1]$, and

$$F(t) = \frac{\delta}{\bar{\gamma}} H(t) \quad \Rightarrow \quad t \leq t_0(\bar{\gamma}, \delta). \quad (2.7)$$

In words, the pairs of δ 's and $\bar{\gamma}$'s in $S_{F,G}$ are those for which the equation in (2.7) has no solutions greater than $t_0(\bar{\gamma}, \delta)$. If F and G are continuous, $(\delta, \bar{\gamma}) \in S_{F,G}$ implies that $t_0(\bar{\gamma}, \delta)$ is the largest solution to that equation; if they are step functions, $(\delta, \bar{\gamma}) \in S_{F,G}$ implies that $t_0(\bar{\gamma}, \delta)$ is *not* a solution (unless $t_0(\bar{\gamma}, \delta) = T_H$).

Proposition. Under assumptions (2.1) and (2.2), we have

$$t_n(\bar{\gamma}, \delta) \rightarrow_P t_0(\bar{\gamma}, \delta) \quad (2.8)$$

as $n \rightarrow \infty$ for all $(\delta, \bar{\gamma}) \in S_{F,G}$ such that $t_0(\bar{\gamma}, \delta) > T_H$; moreover,

$$\delta \frac{\gamma}{\bar{\gamma}} \frac{F(t_0(\bar{\gamma}, \delta) -)}{F(t_0(\bar{\gamma}, \delta))} \leq \liminf_{n \rightarrow \infty} \Pi_{1,n}(\bar{\gamma}, \delta) \leq \limsup_{n \rightarrow \infty} \Pi_{1,n}(\bar{\gamma}, \delta) \leq \delta \frac{\gamma}{\bar{\gamma}} \leq \delta$$

and

$$\begin{aligned} \frac{(\bar{\gamma} - \gamma\delta)}{\delta(1 - \gamma)} F(t_0(\bar{\gamma}, \delta) -) &\leq \liminf_{n \rightarrow \infty} \Pi_{2,n}(\bar{\gamma}, \delta) \\ &\leq \limsup_{n \rightarrow \infty} \Pi_{2,n}(\bar{\gamma}, \delta) \leq \frac{[\bar{\gamma}F(t_0(\bar{\gamma}, \delta)) - \gamma\delta F(t_0(\bar{\gamma}, \delta) -)]}{\delta(1 - \gamma)} \end{aligned}$$

in probability for all $(\delta, \bar{\gamma}) \in S_{F,G}$ such that $t_0(\bar{\gamma}, \delta) > T_F$.

Remarks. (i) Clearly, (2.8) can fail only for a countable number of choices of $(\bar{\gamma}, \delta)$ satisfying $\delta \in (0, \bar{\gamma})$ and $\bar{\gamma} \in [\gamma, 1]$. If F and G have finite supports (the case in most applications based on discrete test statistics) then (2.8) fails only for a finite number of such values of $(\bar{\gamma}, \delta)$, namely those for which $\delta/\bar{\gamma}$ is in the range of the right-continuous step function F/H . This statement on the convergence of $t_n(\bar{\gamma}, \delta)$ to $t_0(\bar{\gamma}, \delta)$ generalizes Theorem 5 of Storey et al. (2004) and an earlier result of Genovese and Wasserman (2002).

(ii) If F is continuous, the asymptotic bounds amount to

$$\Pi_{1,n}(\bar{\gamma}, \delta) \rightarrow_P \delta \frac{\gamma}{\bar{\gamma}} \quad \text{and} \quad \Pi_{2,n}(\bar{\gamma}, \delta) \rightarrow_P \frac{(\bar{\gamma} - \gamma\delta)}{\delta(1 - \gamma)} F(t_0(\bar{\gamma}, \delta)),$$

which generalize the analogous statements in Corollary 3.2 and Corollary 3.3 of Ferreira and Zwiderman (2006b).

(iii) The first part of the proposition can be seen to hold also when $t_0(\bar{\gamma}, \delta) = T_H = -\infty$, since $\limsup_n t_n(\bar{\gamma}, \delta) \leq T_H$ in probability in that case; but it may fail when $t_0(\bar{\gamma}, \delta) = T_H > -\infty$. The upper bound on $\Pi_{1,n}(\bar{\gamma}, \delta)$ and both bounds on $\Pi_{2,n}(\bar{\gamma}, \delta)$ hold even if $t_0(\bar{\gamma}, \delta) \leq T_F$, for in that case $\Pi_{1,n}(\bar{\gamma}, \delta), \Pi_{2,n}(\bar{\gamma}, \delta) \rightarrow_P 0$.

(iv) Any *threshold-based* multiple testing procedure will, by definition, yield the bounds given in the proposition, but the relationship between the triple $(\delta, \bar{\gamma}, \gamma)$ and the particular threshold used may not be apparent; as far as the control of the false discovery rate is concerned, the Benjamini-Hochberg method is best seen as the canonical threshold-based multiple testing procedure, since it expresses that relationship explicitly and since the bounds cannot be improved (using one $\bar{\gamma}$ rather than another does not count as improvement).

To show that the bounds on $\Pi_{1,n}(\bar{\gamma}, \delta)$ cannot be improved, we note in the first place that the ‘more liberal’ procedure of rejecting all the statistics $\leq t_n(\bar{\gamma}, \delta)$ (rather than just those $< t_n(\bar{\gamma}, \delta)$) yields an asymptotic false discovery rate of *at least* $\delta_{\bar{\gamma}}^{\gamma}$. Consequently, if the proportion of statistics equal to $t_n(\bar{\gamma}, \delta)$ is asymptotically zero, then the more liberal procedure and the Benjamini-Hochberg method are asymptotically equivalent and hence their false discovery rate must tend to $\delta_{\bar{\gamma}}^{\gamma}$. Secondly, if the statistics are discrete—if they take integer values, say—and one wants to apply the more liberal procedure, then the ideal way to ensure that $\delta_{\bar{\gamma}}^{\gamma}$ is not exceeded, and in fact of minimizing the false discovery rate, would be to delete from the more liberal list of rejections those statistics computed under the null which happen to equal $t_n(\bar{\gamma}, \delta)$. This ‘rectified’ procedure is not always realizable because one does not know for sure which observations come from the null and which do not. However, if the statistics coming from alternative hypotheses are asymptotically all $< t_n(\bar{\gamma}, \delta)$, then the rectified procedure is asymptotically equivalent to the Benjamini-Hochberg method. Since in the more liberal list of rejections there is roughly a proportion $p_0 := [F(t_0(\bar{\gamma}, \delta)) - F(t_0(\bar{\gamma}, \delta) -)] / F(t_0(\bar{\gamma}, \delta))$ of statistics from the null equal to $t_0(\bar{\gamma}, \delta)$, both the rectified and Benjamini-Hochberg procedures will, under the required assumption, yield a false discovery rate of $\delta_{\bar{\gamma}}^{\gamma}(1 - p_0)$, which is the lower bound on $\Pi_{1,n}(\bar{\gamma}, \delta)$. $\square$

3 Proof of the proposition

For simplicity, we prove (2.8) by assuming that (2.3) holds for each realization.

Suppose $t_0(\bar{\gamma}, \delta) > T_H$ (so that in particular $t_0(\bar{\gamma}, \delta) > -\infty$). By definition of $t_0(\bar{\gamma}, \delta)$, we have $F(t) \geq \delta/\bar{\gamma}H(t)$ for all $t \geq t_0(\bar{\gamma}, \delta)$, and $F(t_0(\bar{\gamma}, \delta) - \varepsilon) < (\delta/\bar{\gamma})H(t_0(\bar{\gamma}, \delta) - \varepsilon)$ for all sufficiently small $\varepsilon > 0$. In fact, by (2.7) we even have

$$\frac{F(t_0(\bar{\gamma}, \delta) - \varepsilon)}{H(t_0(\bar{\gamma}, \delta) - \varepsilon)} < \frac{\delta}{\bar{\gamma}} < \frac{F(t_0(\bar{\gamma}, \delta) + \eta)}{H(t_0(\bar{\gamma}, \delta) + \eta)}$$

for all $\eta > 0$ and sufficiently small $\varepsilon > 0$. By (2.3), this implies that, given $\eta > 0$,

$$\frac{F(t_0(\bar{\gamma}, \delta) - \varepsilon)}{H_n(t_0(\bar{\gamma}, \delta) - \varepsilon)} < \frac{\delta}{\bar{\gamma}} < \frac{F(t_0(\bar{\gamma}, \delta) + \eta')}{H_n(t_0(\bar{\gamma}, \delta) + \eta')}$$

for all sufficiently large n and all $\eta' \geq \eta$, and hence, by definition of $t_n(\bar{\gamma}, \delta)$, that

$$t_n(\bar{\gamma}, \delta) \in [t_0(\bar{\gamma}, \delta) - \varepsilon, t_0(\bar{\gamma}, \delta) + \eta]$$

for all sufficiently large n and given $\varepsilon, \eta > 0$. Since ε and η can be made arbitrarily small, this shows that $t_n(\bar{\gamma}, \delta) \rightarrow t_0(\bar{\gamma}, \delta)$.

In order to obtain the asymptotic upper bound on $\Pi_{1,n}(\bar{\gamma}, \delta)$ suppose (2.8) holds with $t_0(\bar{\gamma}, \delta) > T_F(\geq T_H)$, so that $\liminf_n F(t_n(\bar{\gamma}, \delta)-) > 0$ in probability. By the definitions of $\Pi_{1,n}(\bar{\gamma}, \delta)$ and $t_n(\bar{\gamma}, \delta)$ we then have

$$\begin{aligned}\Pi_{1,n}(\bar{\gamma}, \delta) &\sim \frac{n_0 F_n(t_n(\bar{\gamma}, \delta)-)}{n H_n(t_n(\bar{\gamma}, \delta)-)} < \gamma_n \frac{F_n(t_n(\bar{\gamma}, \delta)-)}{\frac{\bar{\gamma}}{\delta} F(t_n(\bar{\gamma}, \delta)-)} \\ &\leq \delta \frac{\gamma_n \sup_t |F_n(t) - F(t)| + F(t_n(\bar{\gamma}, \delta)-)}{F(t_n(\bar{\gamma}, \delta)-)} \rightarrow \delta \frac{\gamma}{\bar{\gamma}} \leq \delta\end{aligned}$$

in probability.

To get the lower bound, we note that by definition of $t_n(\bar{\gamma}, \delta)$ we have $F(t_n(\bar{\gamma}, \delta) + \varepsilon) \geq H_n(t_n(\bar{\gamma}, \delta) + \varepsilon)\delta/\bar{\gamma}$ for $\varepsilon \geq 0$, so that

$$\begin{aligned}\Pi_{1,n}(\bar{\gamma}, \delta) &\sim \frac{n_0 F_n(t_n(\bar{\gamma}, \delta)-)}{n H_n(t_n(\bar{\gamma}, \delta)-)} \geq \frac{n_0 F_n(t_n(\bar{\gamma}, \delta) - \varepsilon)}{n H_n(t_n(\bar{\gamma}, \delta) + \varepsilon)} \\ &\geq \delta \frac{\gamma_n}{\bar{\gamma}} \frac{F_n(t_n(\bar{\gamma}, \delta) - \varepsilon)}{F(t_n(\bar{\gamma}, \delta) + \varepsilon)} \rightarrow \delta \frac{\gamma}{\bar{\gamma}} \frac{F(t_0(\bar{\gamma}, \delta) - \varepsilon)}{F(t_0(\bar{\gamma}, \delta) + \varepsilon)}\end{aligned}$$

in probability, for all $\varepsilon > 0$ for which F is continuous at $t_0(\bar{\gamma}, \delta) \pm \varepsilon$, whenever (2.8) holds with $t_0(\bar{\gamma}, \delta) > T_F$. Since the $\varepsilon > 0$ here can be made arbitrarily small, this yields the lower bound on $\Pi_{1,n}(\bar{\gamma}, \delta)$.

The asymptotic bounds on $\Pi_{2,n}$ follow almost directly from those on $\Pi_{1,n}$. Indeed, under the required assumptions

$$\begin{aligned}\Pi_{2,n}(\bar{\gamma}, \delta) &= \frac{n H_n(t_n(\bar{\gamma}, \delta)-) - n_0 F_n(t_n(\bar{\gamma}, \delta)-)}{n_1} \\ &\sim \frac{n}{n_1} H_n(t_n(\bar{\gamma}, \delta)-) \left[1 - \frac{n_0}{n} \frac{F_n(t_n(\bar{\gamma}, \delta)-)}{H_n(t_n(\bar{\gamma}, \delta)-)} \right] \\ &= \frac{1}{1 - \gamma_n} H_n(t_n(\bar{\gamma}, \delta)-) [1 - \Pi_{1,n}(\bar{\gamma}, \delta)] \\ &> \frac{1}{1 - \gamma_n} \frac{\bar{\gamma}}{\delta} F(t_n(\bar{\gamma}, \delta)-) [1 - \Pi_{1,n}(\bar{\gamma}, \delta)] \\ &\gtrsim \frac{1}{1 - \gamma_n} \frac{\bar{\gamma}}{\delta} F(t_0(\bar{\gamma}, \delta) - \varepsilon) [1 - \Pi_{1,n}(\bar{\gamma}, \delta)]\end{aligned}$$

in probability for every $\varepsilon > 0$, which yields

$$\liminf_{n \rightarrow \infty} \Pi_{2,n}(\bar{\gamma}, \delta) \geq \frac{(\bar{\gamma} - \gamma\delta)}{\delta(1 - \gamma)} F(t_0(\bar{\gamma}, \delta)-).$$

Similarly,

$$\begin{aligned}
 \Pi_{2,n}(\bar{\gamma}, \delta) &\sim \frac{n}{n_1} H_n(t_n(\bar{\gamma}, \delta)-) \left[1 - \frac{n_0}{n} \frac{F_n(t_n(\bar{\gamma}, \delta)-)}{H_n(t_n(\bar{\gamma}, \delta)-)} \right] \\
 &\leq \frac{1}{1-\gamma_n} H_n(t_n(\bar{\gamma}, \delta) + \varepsilon) [1 - \Pi_{1,n}(\bar{\gamma}, \delta)] \\
 &\leq \frac{1}{1-\gamma_n} \frac{\bar{\gamma}}{\delta} F(t_n(\bar{\gamma}, \delta) + \varepsilon) [1 - \Pi_{1,n}(\bar{\gamma}, \delta)] \\
 &\sim \frac{1}{1-\gamma} \frac{\bar{\gamma}}{\delta} F(t_0(\bar{\gamma}, \delta) + \varepsilon) [1 - \Pi_{1,n}(\bar{\gamma}, \delta)]
 \end{aligned}$$

in probability for every $\varepsilon > 0$ for which F is continuous at $t_0(\bar{\gamma}, \delta) + \varepsilon$, and from this follows that

$$\begin{aligned}
 \limsup_{n \rightarrow \infty} \Pi_{2,n}(\bar{\gamma}, \delta) &\leq \frac{\left[\bar{\gamma} - \gamma \delta \frac{F(t_0(\bar{\gamma}, \delta)-)}{F(t_0(\bar{\gamma}, \delta))} \right]}{\delta(1-\gamma)} F(t_0(\bar{\gamma}, \delta)) \\
 &= \frac{\bar{\gamma} F(t_0(\bar{\gamma}, \delta)) - \gamma \delta F(t_0(\bar{\gamma}, \delta)-)}{\delta(1-\gamma)}.
 \end{aligned}$$

4 The discrete case: an illustration

We begin by showing how an upper bound on γ , needed in the calculation of the threshold $t_n(\bar{\gamma}, \delta)$ of (2.4), can be obtained in practice. As is well known (e.g. Ferreira and Zwinderman (2006b), Section 4), the use of a non-trivial bound on γ (i.e., such that $\bar{\gamma} < 1$) increases the power of the Benjamini-Hochberg method, whose original version corresponds to taking $\bar{\gamma} = 1$.

We consider a situation where the null distribution of the test statistics is concentrated on a finite set S . As a reference we may take the example of Section 2, in which the test statistics are of the form $T = m(1 - D_m)$, D_m being the supremum distance between two empirical d.f.'s: provided the samples used to construct the empirical d.f.'s have continuous distributions, T takes values in $S = \{0, 1, \dots, m-1\}$.

Then (2.1), (2.2) and (2.3) are equivalent to the analogous pointwise results about probability mass functions in which F_n , G_n , H_n , F , G and H are replaced by f_n , g_n , h_n , f , g and h , respectively, where $f_n(t) = F_n(t) - F_n(t-)$, $f(t) = F(t) - F(t-)$, etc. In particular, our assumptions imply

$$h_n = \gamma_n f_n + (1 - \gamma_n) g_n \rightarrow_P h = \gamma f + (1 - \gamma) g.$$

Here, the expression for h gives

$$\gamma = \frac{h(s) - (1 - \gamma) g(s)}{f(s)} < \frac{h(s)}{f(s)}, \quad s \in S.$$

Let $s_+ = \max S$ and suppose $g(s_+)$ is ‘small’ relative to $f(s_+)$; this will typically be the case in applications because by our convention values of test statistics associated with alternative hypotheses are unlikely to be large. Then $(1 - \gamma)g(s_+)/f(s_+)$ is also small and hence γ is not much smaller than $h(s_+)/f(s_+)$, which suggests setting

$$\bar{\gamma}_n := \frac{h_n(s_+)}{f(s_+)}$$

and taking $\bar{\gamma} := \min\{\bar{\gamma}_n, 1\}$ as an upper bound on γ .

Clearly,

$$\gamma \leq \lim_{n \rightarrow \infty} \min\{\bar{\gamma}_n, 1\}$$

in probability, and if $g(s_+) < f(s_+)$ —the case in all realistic applications—then $\bar{\gamma} < 1$ for large n , as required.

This method of (over)estimating γ is essentially due to Schweder and Spjøtvoll (1982), Benjamini and Hochberg (2000) and Storey (2002); it seems particularly appropriate in the discrete case because the definition of $\bar{\gamma}_n$ is unique, unlike that of the homologous quantity in the ‘continuous’ case, which is based on a density or histogram estimate and to a certain extent depends on the choice of smoothing parameter or bin width. (For estimators of γ in the continuous case see for instance Pounds and Morris (2003), Langaas, Lindqvist, and Ferkingstad (2005) and Ferreira and Zwinderman (2006a).)

We now illustrate the determination of $\bar{\gamma}$ and the working of the Benjamini-Hochberg method using part of a set of gene expression data from the study of Richardson et al. (2006). Since the biomedical interpretation of the data (and of the associated lists of rejections resulting from the application of the Benjamini-Hochberg method) is well beyond our competence, we will limit ourselves mostly to addressing statistical issues. The data, publicly available at the GEO website (<http://www.ncbi.nlm.nih.gov/geo>), consist of gene expression measurements on $n := 54,675$ gene probes taken on two groups of patients: a ‘T group’ (‘breast tumour specimens’) and an ‘NB group’ (‘bulk normal breast tissue’, a combination of breast epithelium and breast stroma). According to Prof. A. Richardson (personal communication), the gene expression in the two groups is expected to be quite different. The NB group contains $m := 7$ patients each, and for purposes of comparison with the T group we have randomly divided the latter into four subgroups of $m := 7$ patients. We thus have four ‘T datasets’, which may be referred to by T1, T2, T3, T4, and one ‘NB dataset’, and we may compare pairs of groups across the n genes in order to select lists of genes whose expressions supposedly depend on group classification. To compare each sample of m gene expression measurements from one group with a homologous sample of m measurements from another group we shall use both KS and MW statistics.

From the classification of patients into two groups, one would expect to detect differences in gene expression between NB and T datasets, but not between any two T datasets. We can investigate whether this is indeed the case by, for example, comparing group NB with group T1, and then T1 with T2 and T3 with T4. An important question that can be addressed by comparing pairs of T groups is whether (2.1) is a reasonable assumption with this particular type of microarray data; in other words, does the empirical d.f. of test statistics corresponding to non-differentially expressive genes remain close to the null d.f. (e.g. the d.f. of $m(1 - D_m)$ in the case of the KS statistic) as one varies the sample of patients on which measurements are taken? On the other hand, because only one NB group is available in this case, it is not really possible to check the equally important assumption (2.3), which in the presence of (2.1) is equivalent to (2.2) and which therefore is necessary to guarantee the reproducibility of findings from multiple testing procedures. [As pointed out by one reviewer, for purposes of detecting differences between the NB and T groups it would probably be better to compare the NB data with the whole T data set; but our purpose here is really to illustrate an application of the Benjamini-Hochberg method with small samples and to partially check assumption (2.1).]

Figure 2 shows the histograms of the KS and MW test statistics obtained from the comparison of groups NB and T1 together with the corresponding null probability functions. (As used here, the MW statistic is the minimum of the sum of ranks of the two groups; with 7 observations per group, its distribution is concentrated on $\{28, 29, \dots, 52\}$.) Clearly, in each case the empirical distribution is much more concentrated on the lower classes than the null distribution, suggesting the existence of big differences between the gene expressions of NB and T groups across a substantial proportion of genes. In fact, the corresponding values of $\bar{\gamma}_n$, given in Table 1, show that the percentage of differentially expressive genes in the two groups could be about 60% or more.

In contrast, the histograms that result from comparing T1 with T2 and T3 with T4, shown together with the null probability functions in Figure 3, suggest a rather good agreement between different T datasets, which is also reflected in the associated values of $\bar{\gamma}_n$: in the case of the KS statistic these are 0.96 and 1.01 when comparing T1 with T2 and T3 with T4, respectively; in the case of the MW statistic they are 1.05 and 0.95.

Of course, this is a somewhat subjective assessment, and one may very well ask whether, given the rather large value of n , some of these histograms and estimates of $\bar{\gamma}$ are, despite their apparent conformity with a ‘null situation’, in fact too different from the null probability function or unusually smaller than 1. Thus, a second thought might suggest that an estimate like $\bar{\gamma}_n = 0.95$ actually

points to a value of $\gamma < 1$ and that the histogram gotten from the comparison of T1 with T2 with the MW statistic (lower panel of Figure 3) deviates too much from the null probability function. Though natural and important, these questions cannot be addressed with standart tools (e.g. chi-square tests, confidence intervals) because, as is well known (e.g. Qiu et al. (2005)), gene expression measurements cannot be assumed to be independent nor identically distributed.

	KS statistic	MW statistic
$\bar{\gamma}_n$	0.41	0.55
$t_n(\bar{\gamma}_n, 0.05)$	2 (1)	33 (31)
no. of rejections	5136 (1641)	5024 (3074)
$F(t_n(\bar{\gamma}_n, 0.05))$	0.0530	0.0111
$F(t_n(\bar{\gamma}_n, 0.05) -)$	0.0082	0.0070

Table 1: Some of the results of applying the Benjamini-Hochberg method to compare the NB and T1 groups at a nominal false discovery rate of 0.05.

Some of the results of applying the Benjamini-Hochberg method to compare the NB and T1 datasets are given in Table 1. The nominal false discovery rate has been fixed at $\delta := 0.05$, and in each case the upper bound on γ has been set equal to $\bar{\gamma}_n$. The fact that the KS and MW statistics lead to different values of $\bar{\gamma}_n$ should not be seen as contradictory since such values represent estimates of an upper bound on γ (rather than estimates of γ) and since the null hypothesis implied by the MW statistic is somewhat vaguer than the null hypothesis of equality of distributions implied by the KS statistic.

In the table, the figures in brackets refer to the threshold and number of rejections in the original Benjamini-Hochberg method: with both statistics, taking $\bar{\gamma} = 1$ induces a substantial decrease in the number of rejections and hence a substantial decrease in power. The fact that this loss of power is considerably greater in the case of the KS statistic is in agreement with the differences in the left tails of the histograms of the test statistics (Figure 2).

In each case, the genes declared ‘differentially expressive’—equivalently: the hypotheses that are rejected—are those whose test statistics fall strictly below the threshold $t_n(\bar{\gamma}_n, 0.05)$. With the KS statistic we get about 100 more rejections than with the MW statistic, so at a nominal false discovery rate of 0.05 the Benjamini-Hochberg method could be seen as slightly more powerful when based on the former than on the latter. The theoretical ‘gap’ between the upper and lower bounds on the false discovery rate may be quantified by means of the number $1 - F(t_n(\bar{\gamma}_n, 0.05))/F(t_n(\bar{\gamma}_n, 0.05) -)$; this is 0.370 in the case of the MW statistic

Figure 2: Null ('Exact') probability functions and histograms of the KS (upper panel) and MW (lower panel) test statistics obtained from the comparison of the NB and T1 groups.

Figure 3: Null ('Exact') probability functions and histograms of the KS (upper panel) and MW (lower panel) test statistics obtained from the comparisons between the T1 and T2 groups and between the T3 and T4 groups.

and 0.846 in the case of the KS statistic, a clear sign that the actual false discovery rate is closer to the nominal $\delta = 0.05$ with the latter statistic.

To give a concrete idea of what a false discovery rate of 0.05 entails we may observe that among the 5136 genes selected on the basis of the KS statistic there could be as many as $0.05 \times 5136 \approx 257$ false positives. This last figure may be quite conservative, but at least it tells us that of the 5136 genes selected $5136 - 257 = 4879$ or more should be true positives. If the true γ were 0.41, this would lead to a power of about $4879 / (54,675 \times 0.59) = 0.151$; on the other hand, in the worst case we would have $\gamma = 0$ and a power of about $4879 / (54,675 \times 1.00) = 0.089$. In the present example this type of calculation gives somewhat sharper lower bounds on power than those provided by the proposition: indeed, replacing γ by 0.41 and 0 in $(\bar{\gamma} - \gamma\delta)F(t_n(\bar{\gamma}_n, 0.05) -) / [\delta(1 - \gamma)]$ —a decreasing function of γ —and using the value of $F(t_n(\bar{\gamma}_n, 0.05) -)$, yields lower bounds on power of 0.109 and 0.067.

References

- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*(57), 289-300.
- Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*(25), 60-83.
- Ferreira, J., and Zwinderman, A. (2006a). Approximate power and sample size calculations with the Benjamini-Hochberg method. *The International Journal of Biostatistics*, Vol. 2, Issue 1, Article 8. (Available online at <http://www.bjbs.org.uk/alex/randFDRn4.pdf>)
- Ferreira, J., and Zwinderman, A. (2006b). On the Benjamini-Hochberg method. *The Annals of Statistics*, Vol. 34(4).
- Genovese, C., and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Stat. Soc. Ser. B*(64), 499-517.
- Kulinskaya, E., and Lewin, A. (2006, May). *On randomised FWER and FDR procedures for discrete distributions*. (Available online at <http://www.bgx.org.uk/alex/randFDRn4.pdf>)

- Langaas, M., Lindqvist, B., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. Roy. Stat. Soc. Ser. B*, 67(4), 555-572.
- McLachlan, G., Do, K.-A., and Ambrose, C. (2004). *Analysing Microarray Gene Expression Data*. Wiley.
- Pounds, S., and Cheng, C. (2006). Robust estimation of the false discovery rate. *Bioinformatics*, 22(16), 1979-1987.
- Pounds, S., and Morris, S. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10), 1236-1242.
- Qiu, X., Brooks, A., Klebanov, L., and Yakovlev, A. (2005). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*(6), 120.
- Richardson, A., Wang, Z., de Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J., Livingstone, D., and Ganesan, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*(9), 121-132.
- Schweder, T., and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69, 493-502.
- Storey, J. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc. Ser. B*(64), 479-498.
- Storey, J., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Stat. Soc. Ser. B*(66), 187-205.
- Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98(9), 5116-5121.