On randomised FWER and FDR procedures for discrete distributions

Elena Kulinskaya[*], Alex Lewin [†]

May 18, 2006

Abstract

Fuzzy multiple comparisons procedures are introduced as a solution to the problem of multiple comparisons for discrete test statistics. The critical function of the randomised p-values is proposed as a measure of evidence against the null hypotheses. The classical concept of randomised tests is extended to multiple comparisons. This approach makes all theory of multiple comparisons developed for continuously distributed statistics automatically applicable to the discrete case. Examples of both FWER (Bonferroni(1935)) and FDR (Benjamini-Hochberg (1995)) procedures are discussed.

keywords: Bonferroni procedure, false discovery rate, fuzzy decision-making, multiple comparisons, randomised tests

[*] Statistical Advisory Service, Room G06, SAF Building, Imperial College, London SW7 2AZ, UK. e-mail: e.kulinskaya@ic.ac.uk

[†] Dept. of Epidemiology and Public Health, Imperial College, St Mary's Campus Norfolk Place London W2 1P, UK. e-mail: a.m.lewin@imperial.ac.uk

1 Introduction

The problem of multiple comparisons spans a long time and a multitude of publications, starting from Bonferroni (1935),(1936) [4], [5]. Bonferroni is the most conservative of the Family Wise Error Rate (FWER) procedures, i.e. procedures controlling the probability of committing any type 1 error in families of comparisons under simultaneous consideration. Less conservative FWER procedures using the observed individual p-values were introduced by Simes (1986) [15], Hochberg(1988)[11] and Rom (1990) [12].

For discrete distributions the level α is not attainable. The procedures are more conservative, and therefore less powerful. To overcome inherent difficulties in working with discrete distributions, Tarone (1990) [18] managed to reduce a number of comparisons by disregarding the hypotheses which have no chance of achieving significance after the adjustment. Roth (1999) [14] showed that Tarone's procedure is not α-consistent, i.e. a hypothesis accepted at a particular α–level can be rejected at a smaller level. He introduced an improved Tarone procedure along with modifications of Hochberg's (1988) [11] and Rom's (1990) [12] procedures for discrete distributions.

The lack of power of traditional multiple comparisons procedures motivated Benjamini and Hochberg(1995)[2] to introduce a novel class of procedures controlling the False Discovery Rate (FDR). This implies weak control of the FWER and admits more powerful procedures. The proofs use the fact that the p-values have a uniform distribution under the null. Their procedure is referred to as the BH procedure in what follows. Benjamini and Yekutieli

(2001) [3] studied the FDR procedures under dependency. An alternative approach is to estimate the FDR, and Storey (2002) [16] and Storey, Taylor and Sigmund (2004) [17] propose a family of point estimates, which has been proved to be less conservative than BH procedure.

The use of discrete test statistics and the necessity of multiple comparisons is especially widespread in novel genomics applications, such as genetics or microarray experiments. Gilbert (2005) [10] describes a typical genetics example where the goal is to identify the 'differentially polymorphic' positions, i.e. the positions at which the probability of a non-consensus amino-acid differs between the two sequence sets. In his example Fisher's exact test was used 118 times for 118 positions with no perfect consensus between two sequences. Another common application is testing gene functional categories for independence with respect to differential gene expression (eg. Al-Shahrour et al. 2004 [1]). Here again Fisher's tests are used.

Even for continuous data, test statistics can be discrete. Microarray experiments typically have few replicates (though many tests), 3-5 replicates being a rule not an exception. One of the main objectives is the identification of genes that are differentially expressed under two experimental conditions. Rank or permutation tests are very often the statistical tests of choice (eg. the SAM software [19]). With such small number of replicates even a permutation t-test has a very discrete distribution. These examples make the development of multiple comparisons procedures for discrete data paramount.

Benjamini and Yekutieli (2001) [3] considered a case of discrete test statistics and proved that the BH procedure is then conservative. Gilbert (2005) [10] developed an FDR procedure for discrete distributions combining the Tarone (1990) [18] ideas with the BH type procedure.

We use a different approach to multiple comparisons procedures, based on the traditional idea of randomised tests (Cox and Hinkley (1974) [6]). Multiple tests are randomised independently. When a test critical function may take on values between 0 and 1, the fuzzy sets terminology readily comes to mind. This connection was discussed in Dollinger, Kulinskaya, and Staudte (1996) [7] and applied recently to randomised tests and p-values by Geyer and Meeden (2004) [8], [9]. We advocate using the test critical function as a fuzzy measure of evidence against the null hypothesis.

The theory in this paper applies directly only to one-sided p-values or p-values from symmetric distributions. Treatment of p-values for two-sided tests with non-symmetric distributions is somewhat more technically involved, see Geyer and Meaden [8], [9], and is not discussed.

In Section 2 we recap the notion of randomised or fuzzy p-values. Section 3 introduces a conceptually simple level–α randomised Bonferroni procedure. Section 4 deals with the somewhat more complicated randomised BH procedure. Discussion is in Section 5.

2 Randomised p-values

Level–α statistical tests are the staple notion of statistics. When the distribution is discrete there is a difficulty in attaining an exact level. For a discrete statistic with ordered values $\{x_1, x_2, \cdots, x_n, \cdots\}$ taking on the value $X = x_i$, the traditional ('crisp') p-value for a one-sided test is $p_i \equiv P(X \geq x_i)$. Under the null this p-value has a discrete uniform distribution, i.e. $P(P \leq p_i) = p_i$, as opposed to the continuous Unif$(0, 1)$ distribution for p-values of continuously distributed statistics.

This difficulty may be solved by the introduction of randomised statistical tests. Consider a discrete null distribution of a test statisic X. Let $P(X > c) < \alpha$ and $P(X \geq c) > \alpha$. Then a critical function of a one-sided randomised level-α test is equal to 0 for $\{x < c\}$, to 1 for $\{x > c\}$, and p when $x = c$, with $P(X > c) + p \times P(X = c) = \alpha$ (Cox&Hinkley, 1974)[6].

Traditionally this was interpreted as a need for an extra Bernoulli experiment with probability of rejection p when $X = c$. An alternative interpretation is that of a p-value itself becoming a random variable, uniformly distributed between two discrete consecutive values. The *randomised p-value* is $P(c) = P(X > c) + UP(X = c)$ for $U \sim$ Unif$(0, 1)$ (Cox&Hinkley, 1974) [6].

In this work, we denote the crisp p-value for observation x_i by $p_i \equiv P(X \geq x_i)$. The crisp p-value can be thought of as a function of the observed test statistic. We will also need to know the *previously attainable* p-value, denoted by $p_{i-} \equiv P(X > x_i) = P(X \geq x_{i+1}),\ \ p_{i-} < p_i$. With this notation, the randomised p-value is $P_i|x_i \equiv p_{i-} + U(p_i - p_{i-}) \sim \text{Unif}(p_{i-}, p_i)$ conditionally on x_i. A self-evident but important corollary of this definition is that the randomised p-value has *unconditionally* a $\text{Unif}(0,1)$ distribution. Therefore all properties of multiple comparison procedures for continuous test statistics are automatically fulfilled. This is the main justification for our proposal to use randomised p-values in the multiple comparisons context.

Multiple tests are randomised independently, i.e. *conditionally* random variables $P_i|x_i,\ i = 1, \cdots, m$ are independent. Calculations of rejection probabilities for the p-values in Sections 3 and 4 use this conditional independence.

3 Fuzzy Bonferroni procedure

We shall start from a level–α Bonferroni procedure. For m tests this procedure selects the p-values less than α/m. For a discrete distribution level α is not attainable and there are various attempts to rectify this (see Tarone(1990) [18], Roth (1999)[14]). We propose to use a fuzzy Bonferroni procedure. This is a combination of m independently randomised α/m–level tests. For each of the m crisp p-values $p_i,\ i = 1, \cdots, m$ corresponding to test statistics X_i with values $x_i,\ i = 1, \cdots, m$ let p_{i-} be the previous attainable p-value for test statistic X_i. Then, by definition, the randomised p-value P_i corresponding to p_i is uniformly distributed on the interval $I_i = (p_{i-}, p_i)$.

Definition 3.1 *The fuzzy Bonferroni procedure is defined by the marginal*

n_i	k_i	p_i	p_{i-}	$p_i - p_{i-}$	rank	$\alpha * rank/7$	$\tau_B(p_i)$
8	0	0.003906	0	0.003906	1	0.007143	1
10	1	0.010742	0.000977	0.009766	2	0.014286	0.631429
6	0	0.015625	0	0.015625	3	0.021429	0.457143
8	1	0.035156	0.003906	0.03125	4	0.028571	0.103571
10	2	0.054688	0.010742	0.043945	5	0.035714	0
6	1	0.109375	0.015625	0.09375	6	0.042857	0
8	2	0.144531	0.035156	0.109375	7	0.05	0

Table 1: **Fuzzy Bonferroni procedure example.** $p_i = P(X_i \leq k_i)$ is a p-value from a 1-sided binomial test, $X_i \sim Bin(n_i; 0.5)$; p_{i-} is the previous attainable p-value, $p_i - p_{i-}$ is the length of the interval I_i, rank is the rank of the p-value p_i, $\tau_B(p_i)$ is the probability of rejection by the fuzzy Bonferroni procedure.

critical functions of the randomised tests:

$$\tau_B(p_i) = \begin{cases} 0, & \alpha/m < p_{i-}; \\ \frac{\alpha/m - p_{i-}}{p_i - p_{i-}}, & p_{i-} \leq \alpha/m \leq p_i; \\ 1, & \alpha/m > p_i; \end{cases}$$

Example 1. Fuzzy Bonferroni procedure on Binomial tests

Consider the results of 7 one-sided Binomial tests rejecting for small values of $X_i \sim Bin(n_i; 0.5), i = 1, \cdots, 7$. The 7 p-values are given in Table 1, and the support intervals are plotted in Figure 1. The standard level–0.05 Bonferroni procedure compares p-values to $0.05/7 = 0.00714$. Only the smallest p-value is rejected in this case. The fuzzy procedure has three more candidates for rejection, with probabilities provided in the last column of Table 1. The data analyst would most likely consider the 2nd test a candidate for further investigation, and possibly also the 3rd, since these have reasonably large probabilities of rejection.

So far we have not considered the question of ties which is one of the main features of dealing with discrete distributions since there is a finite (or enumerable) set of possible p-values. Consider a tie of length l: $p = p_{i+1} = \cdots = p_{i+l}$. Assume for simplicity that p_- is the same for all p-values in the tie. When $p_- \leq \alpha/m < p$ we would reject each of the p-values in the tie with probability $\tau_B(p) = P(P < \alpha/m) = (\alpha/m - p_-)/(p - p_-)$, exactly as above. The probability of zero rejections out of l is (under independence) $T_{0,l}(p) = [1 - \pi_B(p)]^l$, and the probability of rejecting at least one of the l p-values in the tie is $1 - T_{0,l}(p)$. Arguably this is a more appropriate probability of rejection of the tie than the smaller individual probabilities $\tau_B(p)$.

Example 2. Consider the same data as in Example 1. Now let there be two copies of p_3 and three copies of p_4, so that the total $m = 10$. The level for individual comparisons is 0.005. Each copy of p_2, p_3 and p_4 should be rejected with probabilities $\tau_B(p_2) = 0.412$, $\tau_B(p_3) = 0.320$ and $\tau_B(p_4) = 0.035$. For rejection of the ties as a whole we have $1 - T_{0,2}(p_3) = 0.538$ and $T_{0,3}(p_4) = 0.101$.

4 Controlling FDR for a discrete distribution

In the usual continuous case, the BH procedure consists of ordering the p-values, then examining them in turn starting from the largest one. Each p-value i is compared with $rank(i)\alpha/m$. In general the largest p-values will be accepted. *As soon as one p-value is rejected, all smaller p-values are also rejected.*

This procedure can be written as:

- Order the m p-values $p_1 \leq p_2 \leq \cdots \leq p_m$;
- find $i_0 = \max\{i : \ p_i \leq i\alpha/m\}$; $[i = rank(p_i)]$;
- reject all $H_i, i \leq i_0$.

We start from a simple case of ordered non-overlapping support intervals for randomised p-values (section 4.1). In this case the main difference from the continuous case is due to the existence of ties. In the next subsection (4.2) we proceed to the general discrete case, where the support intervals may overlap. The problem of ties for the general case is revisited in the last subsection.

4.1 Ordered non-overlapping support intervals

In this section we consider the simplest case when the same test is performed m times independently, and the sample sizes are the same. The test statistics have exactly the same discrete null distribution. There is a finite set of possible p-values. Ties are likely.

We start by ordering the support intervals. In the case of a tie, the support interval will contain more than one randomised p-value. For randomised p-value i, we will call the probability of rejection τ_i. Within each interval j, all randomised p-values have the same probability π_j of being rejected ($\tau_i = \pi_j$ for all p-values i in interval j). In the non-overlapping case we only need to calculate the π_j.

In a similar manner to the continuous BH procedure, we examine each support interval in turn, starting with the interval corresponding to the largest observed p-value. In general the intervals of the largest p-values will be accepted. Then there will be some so-called *fuzzy intervals*, which are rejected with some probability $0 < \pi_j < 1$ for interval j. As soon as one interval is fuzzily rejected, all preceding intervals are fuzzily rejected, until an interval is *crisply rejected* ($\pi_j = 1$). Then all preceding intervals are also crisply rejected.

First we must decide which intervals are fuzzy. Consider a tie of length l starting from $p = p_{i+1}$ and denoted by $\mathcal{T}_l(p)$: $p = p_{i+1} = \cdots = p_{i+l}$. Assume, without loss of generality, that all hypotheses corresponding to p-values larger than p are accepted. Note that the preceding value p_- is the same for all p-values in the tie, and all l randomised p-values are uniformly distributed on the support interval $I(p) = (p_-, p]$ of length $|I(p)| = p - p_-$. The ranks of these randomised p-values vary from $R_- = R_-(p) = i + 1$ to $R_+ = R_+(p) = i + l$. Denote the probability of rejection for each of l hypotheses in the tie by $\pi(p)$.

There are 3 possibilities:

- $p \leq \frac{R_+}{m}\alpha$, the tie is crisply rejected, i.e. the probability of rejection is $\pi(p) = 1$ for each hypothesis;
- $p_- < \frac{R_+}{m}\alpha < p$, the tie is fuzzily rejected, $0 < \pi(p) < 1$;
- $\frac{R_+}{m}\alpha \leq p_-$, the tie is accepted, i.e. $\pi(p) = 0$.

Consider the fuzzy rejection case in more detail. Order the l-tuple of randomised p-values for a particular realisation of the tie. The levels these ordered p-values are to be compared to are $\alpha_1 = \frac{R_-}{m}\alpha, \cdots, \alpha_l = \frac{R_+}{m}\alpha$. There may be $0 \leq k \leq l$ p-values rejected. Denote a probability of exactly k randomised p-values rejected out of l by $T_{k,l}(p_-, p)$, $0 \leq k \leq l$. Let also $q_k = \max(0, \frac{\alpha_k - p_-}{|I(p)|})$ for $k = 1, \cdots, l$. Then

$$\begin{aligned} T_{k,l}(p_-, p) &= P\{U_{(k)} < \alpha_k, U_{(k+1)} > \alpha_{k+1}, \cdots, U_{(l)} > \alpha_l\} \qquad (1) \\ &= \frac{l!}{k!} q_k^k \int_{q_{k+1}}^1 du_{k+1} \int_{\max(u_{k+1}, q_{k+2})}^1 du_{k+2} \cdots \int_{\max(u_{l-1}, q_l)}^1 du_l \end{aligned}$$

can be calculated by integrating over the joint distribution of the order statistics from Unif$(0, 1)$: see Appendix for details. Given k rejections the probability that a particular hypothesis is rejected is $\binom{l-1}{k-1}/\binom{l}{k} = k/l$. Therefore the unconditional probability that a particular hypothesis in a tie of length l is rejected is $\pi(p) = l^{-1} \sum_{k=1}^{l} k T_{k,l}(p_-, p)$, i.e. the expected proportion of rejections in $\mathcal{T}_l(p)$).

Next consider decisions about the p-values in previous intervals in each of the above cases.

- If the p-values in $I(p)$ are crisply rejected, all the p-values in preceding intervals are also crisply rejected.
- If $I(p)$ is a fuzzy interval, with probability $T_{0l}(p_-, p)$ no hypotheses in $I(p)$ are rejected, so the preceding interval $I(p^{prec})$ may be accepted or be crisply/fuzzily rejected on its own merit. With probability $1 - T_{0l}(p_-, p)$ at least one hypothesis in $I(p)$ is rejected, in which case the preceding interval is crisply rejected. Therefore the probability

of rejection for the preceding interval is $\pi(p^{prec}) = (1 - T_{0l}(p_-, p)) + T_{0l}(p_-, p)l^{-1}\sum_{k=1}^{l} kT_{k,l}(p_-^{prec}, p^{prec})$ and the probability of no rejections in the preceding interval is $T_{0l}(p_-, p)T_{0l}(p_-^{prec}, p^{prec})$.

- When the interval $I(p)$ is accepted, the previous interval needs to be examined as above.

More generally, we do not need the same null distribution; we need the ordered non-overlapping support intervals for randomised p-values. This BH type procedure will be generalized to overlapping support intervals (arising in a case of different discrete distributions) in Subsection 4.2.

Definition 4.1 Fuzzy BH procedure for ordered non- overlapping support intervals. *Let m ordered p-values have $J \leq m$ unique values $p_1, \cdots, p_J$, with ties of length $l_j, j = 1, \cdots, J$, $\sum l_j = m$. Let each corresponding randomised p-value be uniformly distributed on a support interval $I_j = I(p_j) = (p_{j-}, p_j]$, where the intervals $I_j, j = 1, \cdots, J$ are non-overlapping and are ordered by value of p_j. Let the ranks of the p-values in the j-th tie be from $R_{j-} = \sum_{t<j} l_t + 1$ to $R_{j+} = \sum_{t\leq j} l_t$.*

Define $s_f = max\{j : \quad p_{j-} \leq \frac{R_{j+}}{m}\alpha\}$ and $s_c = max\{j : \quad p_j \leq \frac{R_{j+}}{m}\alpha\}$, $s_c \leq s_f$. Then all p-values in the interval $J_{reject} = \cup\{J_k,\ k \leq s_c\}$ are crisply rejected and all p-values in the interval $J_{accept} = \cup\{J_k,\ k > s_f\}$ are accepted. The fuzzy interval is defined as $\mathcal{F} = \{J_k,\ s_c < k \leq s_f\}$.

Let π_j denote the unconditional probability of rejecting the p-values in interval j (see Algorithm 1 for calculation). Then τ_i for p-value i is equal to π_j where j is the label of the interval corresponding to p-value i.

Algorithm 1. Calculation of rejection probabilities in each interval.

Let interval j be $(p_{j1}, p_{j2}]$. (For the non-overlapping intervals case $p_{j1}, p_{j2} = p_{j-}, p_{j}$.) Let π_j denote the unconditional probability of rejecting the randomised p-values in interval j, and η_j be the probability of no p-values in interval j being rejected.

- For $j = J, J-1, ..., s_f + 1$,
  $\pi_j = 0$, $\eta_j = 1$

- For $j = s_f, s_f - 1, ..., s_c + 1$,
  $\pi_j = (1 - \eta_{j+1}) + \eta_{j+1} l^{-1} \sum_{k=1}^{l} k T_{k,l_j}(p_{j1}, p_{j2})$
  $\eta_j = \eta_{j+1} T_{0,l_j}(p_{j1}, p_{j2})$

- For $j = s_c, ..., 1$,
  $\pi_j = 1$

Exact calculation of the $T_{k,l_j}(p_{j1}, p_{j2})$ is given in the Appendix.

Lemma 4.1 *For independent test statistics, and for $m_0 \leq m$ true null hypotheses, the above randomised BH procedure controls FDR at exactly level $\frac{m_0}{m}\alpha$.*

Proof. This is part of theorem 5.1 from [3], applicable to any continuous test statistics. Generate an m-tuple of randomised p-values. They have the continuous uniform distribution, and theorem 5.1 holds. Since the intervals I_j are ordered, the p-values outside of the 'fuzzy subset' $\mathcal{F} = \{I_{s_c}+1, \cdots, I_{s_f}\}$ are rejected or accepted regardless of their generated values. The FDR is exactly $\frac{m_0}{m}\alpha$, conditional on any generated realisation within the fuzzy subset

$\mathcal{F}$. The proof follows by integrating over all possible realisations.

Note that any other result for BH-type or similar multiple comparisons procedures proven for the continuous case is applicable to the case of ordered support intervals in exactly the same way as was shown above.

Example 3. Fuzzy BH procedure for the same discrete distribution. Consider $m = 10$ one-sided sign tests for $n = 8$ subjects, $S_i \sim Bin(8, .5)$. Set the FDR level $\alpha = 0.05$.
The p-values are $0.004,\;\; 0.035 \times 3,\;\; 0.145 \times 2,\;\; 0.363 \times 4$.
For $p = p_2$ the interval $I_2 = (p_{2-}, p_2] = (0.004,\; 0.035]$ contains $l = 3$ p-values, $\frac{R_{2-}}{m}\alpha = 0.01$ and $\frac{R_{2+}}{m}\alpha = 0.02$. Therefore $s_c = 1$ and $s_f = 2$.
The q-values are .194, .355, .516 respectively. We obtain

$$\begin{aligned} T_{1,3}(p_2) &= 6q_1(q_3 - q_2)(1 - q_3) + 3q_1(1 - q_3)^2 = .227, \\ T_{2,3}(p_2) &= 3q_2^2(1 - q_3) = .183, \\ T_{3,3}(p_2) &= q_3^3 = .137. \end{aligned}$$

For each of the three hypotheses with p-value of 0.035 the probability of rejection is $\pi_2 = \pi(.035) = 3^{-1} \sum kT_{k,3}(p_2) = .335$ and the probability of rejecting at least one of the three hypotheses is $1 - T_{0,3}(.035) = .547$. The p-value $p_1 = 0.004$ is crisply rejected.

4.2 General case.

The case considered in Subsection 4.1 is straightforward because the support intervals are ordered. Consider now what happens with fuzzy p-values $\{P_i,\; i = 1, \cdots, m\}$ when they originate from different distributions.

Let the observed p-values p_i and corresponding intervals $I_i = (p_{i-}, p_i]$ originate from different distributions. This means that these intervals may intersect, and there is no strict ordering between them. There may be some partial ordering. Ties are still possible but they are not the main problem. If we generate an m-tuple of new randomised p-values, each from its respective Unif(I_i) distribution, the new ordering may be substantially different, and the values crisply rejected earlier can now be accepted. In this Subsection we calculate probabilities of rejection for each of the fuzzy p-values.

Ordering all $2m$ values of $\{p_{i-}, p_i,\ i = 1, \cdots, m\}$ we obtain a partition of $[0, 1]$ into $J + 1 \leq 2m + 1$ ordered subintervals $[0, 1] = \bigcup D_j,\ j = 1, \cdots, J + 1$, where $D_j = (D_{j-}, D_{j+}]$. The $(J + 1)st$ interval $(D_{J+}, 1]$ has no p-values and can be ignored. Each interval $I_i = (p_{i-}, p_i]$ is partitioned by a connected subset of intervals D_j (see Figure 1 for an example), each randomised p-value P_i belongs to each of intervals D_j with probability $\phi_{ij} = |D_j \cap I_i|/|I_i|$, and $\sum_j \phi_{ij} = 1$. Let also $\delta_{ij} = I(\phi_{ij} > 0)$. Then the fuzzy p-value P_i can belong to $\Delta_i = \sum_j \delta_{ij}$ intervals D_j. Let $\mathcal{A} = \{\mathcal{A}_d, d = 1, \cdots, \Delta\}$ be the set of all possible allocations of all m p-values to the intervals D_j, where $\Delta = \prod_i \Delta_i$. Each of these allocations $\mathcal{A}_d$ provides a partial ordering of m p-values (not a full ordering because some p-values may be allocated in the same interval). Denote by z_i^d the label j of the interval to which randomised p-value i is allocated in allocation d. The probability of an allocation $\mathcal{A}_d$ is $\mathcal{P}_d = P(\mathcal{A}_d) = \prod_i \phi_{i,z_i^d}$.

Given an allocation $\mathcal{A}_d$, we can use the BH procedure from Subsection 4.1.

All randomised p-values allocated to the same interval D_j are uniformly distributed on D_j and can be treated as a tie was in Subsection 4.1. The only change in notation required is that the probabilities of rejection are indexed from now on by the interval D_j, and not by the observed p-values as previously.

Each p-value in an interval D_j has the same (conditional on $\mathcal{A}_d$) probability π_j^d of being rejected, be it 0,1, or (in the fuzzy subset) an intermediate value. The full probability of rejecting a particular p-value P_i is

$$\tau_{BH}(P_i) = \sum_{d=1}^{\Delta} P(\mathcal{A}_d)\pi^d_{z_i^d}. \tag{2}$$

To implement this calculation we need to enumerate all allocations $\mathcal{A}_d$. It would be rather straightforward to introduce an ordering on all allocations, and to enumerate them in this order, but there are quite a few of them.

A simpler approach is to find the maximum and minimum possible ranks across all allocations for each interval D_j. These values are associated with a subset of randomised p-values which can attain these ranks. We can proceed as follows.

Definition 4.2 Fuzzy BH procedure in the general case of overlapping support intervals.

For each subinterval $D_j, j = 1, \cdots, J$ denote the maximum and the minimum possible ranks across all allocations $\mathcal{A}_d$ by $\mathcal{R}_{j+}$ and $\mathcal{R}_{j-}$. Define $s_f = max\{j : \quad D_{j-} \leq \frac{\mathcal{R}_{j+}}{m}\alpha\}$ and $s_c = max\{j : \quad D_{j+} \leq \frac{\mathcal{R}_{j-}}{m}\alpha\}$, $s_c \leq s_f$.

Then all p-values in the interval $D_{reject} = \cup\{D_j,\ j \le s_c\}$ *are crisply rejected; all p-values in the interval* $D_{accept} = \cup\{D_j,\ j > s_f\}$ *are accepted; only p-values which can be allocated to the 'fuzzy subset'* $\mathcal{F} = \{D_j,\ s_c < j \le s_f\}$ *should be investigated further.*

For each allocation $\mathcal{A}_d$*, the rejection probabilities for each interval* π_j^d *are calculated using Algorithm 1. Then* τ_i *for p-value* i *is given by Equation 2.*

The calculation of s_c and s_f makes the enumeration of the allocations simpler: we have a partition of $[0, 1]$ into $s_f - s_c + 2 \le J$ intervals, as D_{accept} and D_{reject} are each treated as one of the subintervals. For each randomised p-value P_i which may belong to $\mathcal{F}$, the resulting partition of the support interval I_i contains $\bar{\Delta}_i \le \min(\Delta_i, s_f - s_c + 2)$ subintervals into which P_i may be allocated. Overall there are $\bar{\Delta} = \prod_i \bar{\Delta}_i$ allocations to enumerate.

j	D_{j-}	D_{j+}	$\lvert D_j\rvert$	p-values	$\mathcal{R}_{j-}$	$\mathcal{R}_{j+}$	A_{j-}	A_{j+}
1	0.000	0.001	0.001	1,3	1	2	0.007	0.014
2	0.001	0.004	0.003	1,2,3	1	3	0.007	0.021
3	0.004	0.011	0.007	2,3,4	2	4	0.014	0.029
4	0.011	0.016	0.005	3,4,5	3	5	0.021	0.036
5	0.016	0.035	0.020	4,5,6	4	6	0.029	0.043
6	0.035	0.055	0.020	5,6,7	5	7	0.036	0.05
7	0.055	0.109	0.055	6,7	6	7	0.043	0.05
8	0.109	0.145	0.035	7	7	7	0.050	0.05

Table 2: **Fuzzy BH procedure example.** See data in Table 1. j is the number of an interval $D_j = (D_{j-}, D_{j+}]$, $|D_j|$ is its length; 'p-values' provides the list of p-values which can belong to D_j, $\mathcal{R}_{j-}$ and $\mathcal{R}_{j+}$ are the smallest and the largest ranks in D_j, $A_{j\pm} = \mathcal{R}_{j\pm}\alpha/7$

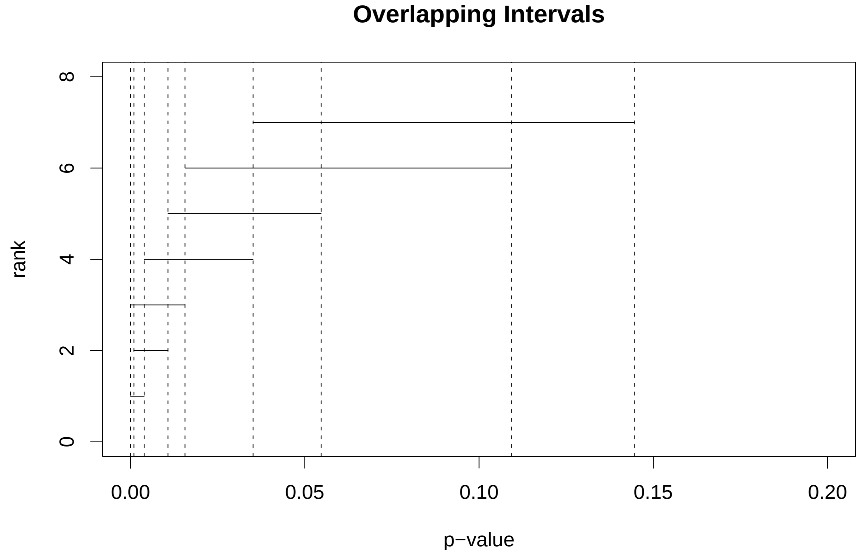

Figure 1: *Plot of p-values and related intersecting support intervals for the data from Example 4, given in Table 2. Support intervals (horizontal segments) are ordered by the ranks of respective p-values on the vertical axis. Vertical dashed lines specify 8 subintervals $D_j,\ j = 1, \cdots, 8$ on the horizontal axis.*

Example 4. Fuzzy BH procedure

Consider the 7 p-values from a mixture of Binomial distributions, given in Table 1. The support intervals partition $[0, 1]$ into the 8 subintervals $D_j, j = 1, ..., 8$ given in Table 2 and plotted in Figure 1. Here $s_c = 4,\ s_f = 6$. Denote $A_{j\pm} = \mathcal{R}_{j\pm}\alpha/7$. The first 4 intervals have $D_{j+} < A_{j-}$ and therefore constitute D_{reject}; intervals 7 and 8 constitute D_{accept}; intervals 5 and 6 are the fuzzy subset $\mathcal{F}$. Note that though $D_{5+} < A_{5+}$ this is not sufficient for crisp rejection of D_5 as we shall see below. P-values which may end up in the fuzzy subset are p-values 4 to 7. Each can belong to 3 different subintervals, therefore $3^4 = 81$ allocations are possible. Since we do not need to distinguish between different allocations in intervals before 5 and after 6, this number is reduced to $\bar{\Delta} = 36$: the p-value 4 may belong to D_5 or to D_{reject} ; p-value 7 may belong to D_6, or to D_{accept}, thus $\bar{\Delta}_4 = \bar{\Delta}_7 = 2$. Allocations of the first three p-values do not change the ranks of the last four p-values, and are therefore ignored. Given an allocation $\mathcal{A}_d$, any p-values allocated to D_6 will be fuzzily rejected with probability $\pi_6|\mathcal{A}_d$. When $R_{5+} > 4$, which happens every time two or three p-values belong to D_5, we have $D_{5+} < A_{5+}$ and $s_c = 5$. Thus every p-value in D_5 will be crisply rejected, $\pi_5 = 1$. When there is only one p-value with rank 4 in D_5, it is fuzzily rejected with probability $\pi_5 = 1 - T_{0,l_6}(D_6) + T_{0,l_6}(D_6)\sum_{k=1}^{l} kT_{k,l_5}(D_5)$. Of course this happens only when p-value 4 on its own belongs to D_5, with p-value 5 in D_6, and p-values 6 and 7 in D_6 or D_{accept}; this occurs in 4 possible allocations with l_6 varying from 1 to 3. Summing up the probabilities of rejection for each p-value we obtain $\tau_{BH}(P_1) = \tau_{BH}(P_2) = \tau_{BH}(P_3) = 1$, $\tau_{BH}(P_4) = 0.941$, $\tau_{BH}(P_5) = 0.632$, $\tau_{BH}(P_6) = 0.281$, $\tau_{BH}(P_7) = 0.080$. The standard BH procedure rejects the first three p-values. Note the very high probability of rejection for the p-value 4; p-value 7 has a low probability of rejection, it can be rejected only if it is allocated to D_6.

4.3 Ties in the BH procedure.

We have not discussed the ties specifically in Subsection 4.2 since there is no difference, theoretically speaking, between 'original' ties i.e. the same p-values, and 'virtual' ties i.e. p-values allocated to the same subinterval D_j. (The regions D_{accept} and D_{reject} constitute one interval each.) Suppose we have a tie of length l for a p-value $P = P_i$, and P can be allocated to $t = \bar{\Delta}_i < \Delta_i$ intervals D_j. This means t^l allocations of l copies of P, increasing the volume of computation. There is no reason though to treat these copies of P as distinguishable. Then the number of distinguishable allocations is the number of solutions to an equation $n_1 + n_2 + \cdots + n_t = l$, i.e. $\binom{t+l-1}{l}$, a considerable saving. The probability of each distinguishable allocation should be multiplied by its multiplicity $\binom{l}{n_1 n_2 \cdots n_t}$, where n_j is the number of copies in D_j.

Let f p-values $P_i, i = 1, \cdots, f$ with values in $\mathcal{F}$ be subdivided into f_d ties of length $l_i, \sum_i l_i = f$. Then the full order of enumeration for fuzzy BH procedure is $\Delta = \prod_{i=1}^{f_d} \binom{\bar{\Delta}_i + l_i - 1}{l_i}$.

Example 5. Fuzzy BH procedure with ties. Consider the data from Example 2. Here again $D_{reject} = \cup_1^4 D_j$, $\mathcal{F} = D_5 \cup D_6$, and $D_{accept} = D_7 \cup D_8$. Only p-values 4 to 7 may belong to $\mathcal{F}$. Three copies of P_4 can be in three intervals (D_3 to D_5), but $t = \bar{\Delta}_4 = 2$ since P_4 is either in D_{reject} or in D_5. Thus $t = 2,\ l = 3$, and the $3^3 = 27$ original possible allocations of P_4 are reduced first to $2^3 = 8$, and then to $\binom{2+3-1}{3} = 4$. There can be 3 copies of p-value 4 in D_{reject} or D_5 (2 allocations of multiplicity 1), or two in one and one in another interval (2 allocations, multiplicity 3 each). P-values 5 and 6 may belong to three intervals, and P_7 to two, making the total of $2 \times 2 \times 3 \times 3 \times 2 = 72$ configurations to enumerate. Once more, the only 'fuzziness' in rejecting P_4 results from 4 possible configurations with all three copies of P_4 in D_5 on their own, whereas P_5 is in D_6.

5 Discussion

Fuzzy multiple comparisons procedures are rather attractive from several different perspectives. Firstly, they extend the classical concept of randomised tests to multiple comparisons. This seems to be a very straightforward generalisation, but to our knowledge it has not been suggested before. This approach makes all theory of multiple comparisons developed for continuously distributed statistics automatically applicable to the discrete case. Only two methods: Bonferroni (1935) [4] and Benjamini-Hochberg (1995) [2] were explored in this paper, but it should be possible to similarly generalize other methods, Storey (2002) [17] among others. Secondly, a fuzzy decision procedure ascribing probabilities to rejection of each of multiple hypotheses should appeal to applied scientists given that fuzzy methods are rather popular in contemporary computer-intensive applications, see, for example, Ross (2004) [13].

An evident drawback is the amount of computation required. These procedures should be efficiently programmed if they are to be of practical use. A simple method would be to generate N sets of m p-values from $\prod_{i=1}^{m} \mathrm{Unif}(I_i)$, and to estimate probabilities of rejection τ_i through proportions of rejection out of N. Another attractive option to be explored elsewhere is to use importance sampling methods for more efficient estimation.

FDR control at exact $\frac{m_0}{m}\alpha$ level requires independence of the p-values. But it is worth noting that the calculation of rejection probabilities $\tau(p_i)$ in Sections 3 and 4 holds regardless, due to conditional independence of the randomised p-values. Given that the properties of positive dependence from Benjamini and Yekutieli (2001) [3] between components of the marginally uniform multivariate distribution of the p-values on $[0, 1]^m$ are satisfied, the randomised BH procedure should be conservative.

Interpretation of results of fuzzy multiple comparisons procedures is not straightforward. If a binary decision is required, a simple rule could be adopted, say reject all p-values with probability of rejection above 50%; this would change the FDR level though. We believe that actual probabilities of rejection provide more information, and applied scientists may decide by themselves which hypotheses require further exploration.

Acknowledgments:

This work was in part supported by BBSRC "Exploiting Genomics" grant 28EGM16093'

Appendix

When we examine an interval D_j in the fuzzy subset $\mathcal{F}$ (where $D_j = I_j$ in the non-overlapping case), we need to calculate two quantities, firstly the unconditional probability π_j that a particular hypothesis is rejected, and secondly the probability η_j that no hypotheses in the interval are rejected. Both of these can be calculated from the probabilities $T_{k,l_j}(p_1, p_2)$ (of rejecting exactly k of the hypotheses, for $k = 1, ..., l_j$). Here p_1, p_2 are the boundaries of the interval D_j, $(p_{j-}, p_j$ in the non-overlapping case).

Let the number of randomise p-values in the interval be l_j, and the minimum and maximum ranks be R_{j-} and R_{j+} respectively. For $k = 1, ..., l_j$, let $\alpha_{jk} = (R_{j-} + k - 1)\alpha/m$, $q_{jk} = max(0, (\alpha_{jk} - p_1)/(p_2 - p_1))$ and $t_j = q_{j(k+1)} - q_{jk} = \alpha/m|D_j|$ is independent of k. From now on we suppress the j index on the tie length l_j.

We need to calculate

$$\begin{aligned} T_{k,l}(p_1, p_2) &\equiv P\{P_{jk} < \alpha_{jk}, P_{j(k+1)} > \alpha_{j(k+1)}, ..., P_{jl} > \alpha_{jl}\} \\ &= \frac{l!}{k!} q_{jk}^k P\{P_{j(k+1)} > \alpha_{j(k+1)}, ..., P_{jl} > \alpha_{jl}\} \end{aligned} \tag{3}$$

where $P_{jk}, i = 1, ..., l$ are order statistics from a Uniform on (p_1, p_2).

In order to calculate the probability in Equation 3, the $\{P_{j(k+1)}, ..., P_{jl}\}$ have to be allocated into the intervals defined by $\{\alpha_{j(k+1)}, ..., \alpha_{jl}, p_2\}$ in such a way that the condition in the probability holds. Given such an allocation, the probability is easy to calculate: it is a product of two types of terms:

$$P\{\alpha_{jr} < P_{j(s+1)} < ... < P_{j(s+u)} < \alpha_{j(r+1)}\} = \frac{t_j^u}{u!}$$

$$P\{P_{jl} > ... > P_{j(l-r+1)} > \alpha_{jl}\} = \frac{(1 - q_{jl})^r}{r!}$$

(either u p-values allocated between two adjacent α's or the largest r p-values allocated to the top interval (α_{jl}, p_2)).

The allocations can be labelled uniquely by $l - k$ integers, denoting the number of randomised p-values in the above alpha intervals, eg. $\alpha_1 < P_1 < \alpha_2 < \alpha_3 < P_2 < P_3$ is denoted 102 ($l = 3, k = 0$). If we call these integers $n_{k+1}, ..., n_l$, the probability we need for equation 3 can be written

$$T_{k,l}(p_1, p_2) = \frac{l!}{k!} q_{jk}^k \sum_{\mathcal{Z}_d^{(l-k)}} \frac{t_j^{l-k-n_l^{(d)}} (1 - q_{jl})^{n_l^{(d)}}}{\prod_{i=k+1}^{l} n_i^{(d)}!}$$

where $\mathcal{Z}_d^{(l-k)}$ stands for one of the allocations allowed for $l-k$ intervals. Note that the allocation labels depend only on $l-k$, not j, so can be calculated just once **for each** $l-k$.

The allocations can be calculated in a straightforward way:

```
for $n_1 = 0, 1$ {
  for $n_2 = 0, ..., 2 - n_1$ {
    for $n_3 = 0, ..., 3 - n_1 - n_2$ {
      ...
        for $n_{(l-k)-1} = 0, ..., (l-k) - 1 - \sum_1^{(l-k)-2} n_j$ {
          $n_{(l-k)} = (l-k) - \sum_1^{(l-k)-1} n_j$
          allocation $\mathcal{Z}_d^{l-k} = \{n_1, n_2, ..., n_{l-k}\}$.
},...}
```

We must have $\sum_{i=1}^{r} n_i \leq r$ for each r, since the first r intervals may not contain more than r p-values if the condition in equation 3 is to be satisfied.

References

[1] F. Al-Shahrour, R. Daz-Uriarte, and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20:578–580, 2004.

[2] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *JRSS(B)*, 57(1):289–300, 1995.

[3] Y. Benjamini and D. Yekutieli. The control of the False Discovery Rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[4] C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *In Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Rome: Italy, 1935.

[5] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilit. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[6] D.R. Cox and D.V. Hinkley. *Theoretical Statistics.* Chapman and Hall, London, 1974.

[7] M.B. Dollinger, E. Kulinskaya, and R. G. Staudte. Fuzzy hypothesis tests and confidence intervals. In D.L. Dowe, K.B. Korb, and J.J. Oliver, editors, *Information, Statistics and Induction in Science*, pages 119–128, Singapore, 1996. World Scientific.

[8] C.J. Geyer and Meeden G.D. Fuzzy and randimized confidence intervals and p-values. *www.stat.umn.edu/geyer/fuzz/*, 2004.

[9] C.J. Geyer and Meeden G.D. ump: An R package for UMP and UMPU tests. *www.stat.umn.edu/geyer/fuzz/*, 2004.

[10] P. Gilbert. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):143–158, 2005.

[11] Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.

[12] D.M. Rom. A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77:663–665, 1990.

[13] T.J. Ross. *Fuzzy Logic with Engineering Applications*. John Wiley and Sons Ltd, New York, 2nd edition edition, 2004.

[14] A.J. Roth. Multiple comparison procedures for discrete test statistics. *J. Statist. Plann.Inference*, 82-2(1):101–117, 1999.

[15] R.G. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.

[16] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, 64:479–498, 2002.

[17] Taylor J.E. Storey, J.D. and D. Siegmund. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society: Series B*, 66:187–205, 2004.

[18] R.E. Tarone. A modified Bonferroni method for discrete data. *Biometrics*, 46:515–522, 1990.

[19] V. Tusher, R. Tibshirani, and C. Gilbert. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, USA*, 98:5116–5121, 2001.