On fuzzy FWER and FDR procedures for discrete distributions

Elena Kulinskaya[*], Alex Lewin [†]

October 8, 2007

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 Family Wise Error Rate (Bonferroni, 1935) and False Discovery Rate (Benjamini & Hochberg, 1995) procedures are discussed. An application to linkage disequilibrium testing is given. Software for implementing the procedures is available.

[*]Statistical Advisory Service, 8 Princes Gardens, Imperial College, London SW7 1NA,UK. e-mail: e.kulinskaya@imperial.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

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

1 Introduction

The problem of dealing with multiple comparisons has long been recognised in the statistical literature, starting with Bonferroni (1935, 1936). The Bonferroni correction controls the Family Wise Error Rate (FWER), that is 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), Hochberg (1988) and Rom (1990).

In more recent years, the lack of power of traditional multiple comparisons procedures motivated Benjamini & Hochberg (1995) to introduce a novel class of procedures controlling the False Discovery Rate (FDR). The FDR control is less restrictive than FWER control and admits more powerful procedures. Their procedure is referred to as the BH procedure in what follows. Benjamini & Yekutieli (2001) studied the FDR procedures under dependency. An alternative approach of FDR estimation was introduced in Storey (2002) and Storey et al. (2004).

The controlling procedures cited above were all developed for p-values arising from continuous test statistics. Under appropriate conditions, each procedure will control either the FWER or the FDR at a level α. The proofs for such control use the fact that the p-values have a uniform distribution under the null. For discrete

distributions the level α may not be attainable even for a single test, as the p-values do not come from a uniform distribution under the null. For multiple tests of null hypotheses with different discrete distributions, this problem is exacerbated since even when the desired level is attainable for one test it will not in general be attainable for other tests. The procedures are more conservative, and therefore less powerful. As an example consider a level α Bonferroni procedure for 2 independent tests. Let p_i be the maximum attainable p-value such that $p_i \leq \alpha/2$ for the test i, and $p_1 \neq p_2$ due to different distributions. The attainable level $p_1 + p_2 - p_1 p_2$ is considerably less than α even for the minimum value of $\alpha = 2 \times \max(p_1, p_2)$.

Multiple testing of discrete test statistics is particularly important currently, with the development of novel genomics applications, such as genetics or microarray experiments. Chakraborty et al. (1987) includes a typical genetics example of testing for linkage disequilibrium, i.e. looking at correlation between alleles at pairs of markers. In their example 28 correlation coefficients are calculated. Another example is given in Gilbert (2005), where Fisher's tests are used to identify the positions at which the probability of a non-consensus amino-acid differs between two sequence sets. Other applications include testing gene functional categories for independence with respect to differential gene expression (eg. Al-Shahrour et al., 2004), and association studies in genetics. Here again Fisher's or chi-squared tests are used.

To overcome inherent difficulties in working with discrete distributions, Tarone

(1990) managed to reduce a number of comparisons by disregarding the hypotheses which have no chance of achieving significance after the adjustment. Further improved FWER procedures are given in Roth (1999). Benjamini & Yekutieli (2001) considered a case of discrete test statistics and proved that the BH procedure is then conservative. Gilbert (2005) developed an FDR procedure combining the Tarone (1990) ideas with the BH type procedure.

We use a different approach to multiple comparisons procedures, based on the idea of randomised tests (Cox & Hinkley, 1974). For one test, the test critical function taking on values between 0 and 1 can be used as a fuzzy measure of evidence against the null hypothesis. This quantity can be seen as a fuzzy membership function for the set of rejected tests. (Note that this depends only on the observed p-values and the level α of the test procedure. No randomisation is performed to obtain the fuzzy measure.) The connection between test critical functions taking on values between 0 and 1 and fuzzy quantities was discussed in Dollinger et al. (1996) and applied recently to randomised tests and p-values by Geyer & Meeden (2005).

The purpose of this paper is to show how this idea can be extended to the multiple testing situation. Multiple tests are randomised independently, and the marginal critical function for each test is used when constructing a multiple comparisons procedure. We provide algorithms for the exact calculation of the fuzzy measures, i.e. resulting probabilities of rejection of each test.

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, and Section 5 presents an application to linkage disequilibrium testing (using the data from Chakraborty et al., 1987). Discussion is in Section 6.

An R (R Development Core Team, 2007) package implementing the fuzzy procedures is available from http://www.bgx.org.uk/alex/.

2 Randomised p-values and fuzzy decision rules

Consider a discrete test statistic X, which can take values in $\{x_1, x_2, \cdots, x_n, \cdots\}$. If the observed value of the statistic is x_i, the traditional ('crisp') p-value P for a one-sided test is $p_i \equiv \text{pr}(X \geq x_i)$ calculated under the null hypothesis. Since the set of possible values of X is discrete, the set of possible p-values is also discrete. Under the null the crisp p-value has a discrete uniform distribution, i.e. $\text{pr}(P \leq p_i) = p_i$, as opposed to the continuous Unif$(0, 1)$ distribution for p-values of continuously distributed statistics. Thus in general the exact level-α test is not possible to obtain.

This difficulty may be solved by the introduction of randomised statistical tests. Consider a discrete null distribution of a test statistic X. Let c be the value of the

statistic such that $\text{pr}(X \geq c) > \alpha$ but $\text{pr}(X > c) < \alpha$. Then the exact level-$\alpha$ test can be achieved by using a randomised p-value $P(c) = \text{pr}(X > c) + U\text{pr}(X = c)$ for $U \sim \text{Unif}(0, 1)$ (Cox & Hinkley, 1974). Traditionally this was interpreted as a need for an extra Bernoulli experiment with probability of rejection $(\alpha - \text{pr}(X > c))/P(c)$ when $X = c$. An alternative interpretation is that the p-value is a random variable, uniformly distributed between two discrete consecutive values. Unconditionally, this randomised p-value has a continuous $\text{Unif}(0, 1)$ distribution under the null.

In this work, we denote the crisp p-value for observation x_i by $p_i \equiv \text{pr}(X \geq x_i)$. 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 $p_{i-} \equiv \text{pr}(X > x_i) = \text{pr}(X \geq x_{i+1}), \quad p_{i-} < p_i$. 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. Thus the conditional probability of rejection of the randomised p-value $\text{pr}(P_i \leq \alpha|x_i)$ is

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

It is clear that $\tau(p_i)$ depends only on the observed p-values and the level α.

The idea proposed by Geyer & Meeden (2005) is to use this function $\tau(p_i)$ as a fuzzy measure of evidence against the null hypothesis. We extend this to the multiple comparison situation by calculating the marginal probabilities of rejection for randomised p-values when standard multiple testing procedures are used to control the FWER and FDR. Since the randomised p-values have unconditionally

a Unif$(0,1)$ distribution under the null, 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 by construction. Calculations of rejection probabilities for the p-values in Sections 3 and 4 use this conditional independence. This construction is sufficiently general not to be detrimental for the properties of the resulting procedures as discussed in Section 6.

3 Fuzzy Bonferroni procedure

For continuous p-values, the Bonferroni procedure consists of rejecting each test that has a p-value less than α/m, where m is the number of tests. Thus for the fuzzy Bonferroni procedure we need to calculate $\text{pr}(P_i \leq \alpha/m|x_i)$. As each p-value is compared with the same threshold, the probabilities of rejection are the same as for a single test, with α replaced by α/m.

Definition 3.1. The fuzzy Bonferroni procedure is defined by the marginal 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 of $H_0 : p = 0.5$ vs the 1-sided

alternative $p < 0.5$. The tests reject 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 (p_{i-}, p_i) 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.

4 Controlling FDR for a discrete distribution

As before, we need to calculate the marginal probabilities of rejection for the randomised p-values, this time using an FDR controlling procedure. We choose to use the Benjamini and Hochberg (BH) procedure (Benjamini & Hochberg, 1995). 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 hypotheses corresponding to the largest p-values will be accepted. *As soon as one hypothesis is rejected, all hypotheses with smaller p-values are also rejected.*

The calculation of the probabilities of rejection are more complex than for the Bonferroni procedure, since now the ordering of the p-values must be taken into account. If one imagines generating different sets of randomised p-values, it can

be expected that the order of the randomised p-values will not always be the same as the order of the observed p-values, and will vary from realisation to realisation.

For this reason, it will be useful to think about the support intervals (p_{i-}, p_i) of the randomised p-values. The calculation of probabilities of rejection is much easier when these support intervals do not overlap. This case is considered in Section 4.1. The case of overlapping intervals is presented in Section 4.2.

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. In this case the test statistics have exactly the same discrete null distribution. Therefore the support intervals $(p_{i-}, p_i]$ do not overlap. If there are many tests, it is likely that there will be several observed p-values which are equal. These will have the same probability of rejection. Thus the calculation of the probabilities can be done for each unique support interval ($j = 1, ..., J$), rather than for each p-value ($i = 1, ..., m$), as the number of intervals J can be considerably smaller than m. We denote the probability of rejection for p-values in interval j by π_j. Then the probability of rejection for test i is $\tau_{BH}(p_i) = \pi_j$ where j is the index of the interval to which randomised p-value i belongs.

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. Suppose that there are l observed p-values with the value p_j (a tie of length l). Thus the support interval $(p_{j-}, p_j]$ will always contain l randomised p-values (each uniformly distributed on that interval). Their ranks increase from R_{j-} for the smallest to R_{j+} for the largest. In the BH procedure, the largest rank R_{j+} defines the decision rule. Suppose, without loss of generality, that all hypotheses corresponding to p-values larger than p_j are accepted. Then there are 3 possibilities:

- $p_j \leq \frac{R_{j+}}{m}\alpha$; All randomised p-values are less than $\frac{\alpha}{m}$ multiplied by their respective rank with probability 1, therefore the tie is crisply rejected, i.e. the probability of rejection is $\pi_j = 1$;

- $p_{j-} < \frac{R_{j+}}{m}\alpha < p_j$; The probability of the randomised p-values being less than $\frac{\alpha}{m}$ multiplied by their respective rank is between 0 and 1, thus the tie is fuzzily rejected, $0 < \pi_j < 1$;

- $\frac{R_{j+}}{m}\alpha \leq p_{j-}$; All randomised p-values are greater than $\frac{\alpha}{m}$ multiplied by their respective rank with probability 1, thus the tie is accepted, i.e. $\pi_j = 0$.

Let us look at the fuzzy rejection case in more detail. Consider a particular realisa-

tion of the randomised p-values (of course to calculate the probabilities of rejection we must integrate over all possible realisations). Each realisation of the ordered randomised p-values must be compared with $\alpha_1 = \frac{R_{j-}}{m}\alpha, \cdots, \alpha_l = \frac{R_{j+}}{m}\alpha$. Denote a probability of exactly k randomised p-values rejected out of l by $T_{k,l}(p_{j-}, p_j)$, $0 \leq k \leq l$. Let also $q_k = \max(0, \frac{\alpha_k - p_{j-}}{|p_j - p_{j-}|})$ for $k = 1, \cdots, l$. Then

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

where $U_{(i)}$ are the order statistics from Unif$(0, 1)$. Appendix A gives the details of this calculation. Given k rejections, the probability that a particular hypothesis is rejected is $\binom{l-1}{k-1}/\binom{l}{k} = k/l$. The unconditional probability that any hypothesis out of the l is rejected is the expected proportion of rejections, i.e.

$$\pi_j = l^{-1} \sum_{k=1}^{l} k T_{k,l}(p_{j-}, p_j).$$

We stress that this probability is the exact unconditional probability of rejection for the randomised test. It does not depend on drawing any realisations of randomised p-values.

Next consider decisions about the p-values in previous intervals in each of the 3 above cases.

- If the interval $(p_{j-}, p_j]$ is crisply rejected, all the preceding intervals (those corresponding to smaller p-values) are also crisply rejected. This is due to the aspect of the BH procedure which says that once one p-value is rejected, all smaller ones are also rejected.

- If $(p_{j-}, p_j]$ is a fuzzy interval there are 2 sub-cases to consider.
  - With probability $T_{0l}(p_{j-}, p_j)$ no hypotheses in $(p_{j-}, p_j]$ are rejected, so the preceding interval may be accepted or be crisply/fuzzily rejected on its own merit.
  - With probability $1 - T_{0l}(p_{j-}, p_j)$ at least one hypothesis in $(p_{j-}, p_j]$ is rejected, in which case the preceding interval is crisply rejected.

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

- When the interval $(p_{j-}, p_j]$ is accepted, the previous interval is accepted or crisply/fuzzily rejected on its own merit.

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

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 Benjamini & Yekutieli (2001), applicable to any continuous test statistics. Any m-tuple of randomised p-values 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 2. 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_k values defined before equation 1 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.

Consider now what happens with randomised p-values $\{P_i,\ i = 1, \cdots, m\}$ originating from different distributions. Now the support intervals may overlap, so there is no strict ordering between them. We first partition the unit interval into intervals based on the intersections of the support intervals (so these smaller intervals are non-overlapping). For *each realisation* of m randomised p-values, we can think of allocating these p-values to the non-overlapping intervals. Given a particular allocation, the calculation of π_j for interval j can proceed as in Section 4.1. In order to calculate the $\tau_{BH}(p_i)$ for each test i, we must integrate over the possible allocations of randomised p-values. We stress again that the value of $\tau_{BH}(p_i)$ does not depend on any particular realisation of randomised p-values, but only on the observed discrete p-values.

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

Let each randomised p-value have support in the interval I_i. Partition the support set $\mathcal{I} = \bigcup I_i,\ i = 1, \cdots, m$ into $J \leq 2m$ ordered subintervals $\mathcal{I} = \bigcup D_j,\ j = 1, \cdots, J$, where $D_j = (D_{j-}, D_{j+}]$. Let the probability of randomised p-value P_i belonging to interval D_j be denoted $\phi_{ij} = |D_j \cap I_i|/|I_i|$.

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. Denote by z_i^d the label j of the interval to which randomised p-value i is allocated in allocation d.

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 $s_f = \max\{j : D_{j-} \leq \frac{\mathcal{R}_{j+}}{m}\alpha\}$ and $s_c = \max\{j : \ D_{j+} \leq \frac{\mathcal{R}_{j-}}{m}\alpha\}$, $s_c \leq s_f$. Then all p-values in the interval $D_{reject} = \cup\{D_j,\ j \leq s_c\}$ 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' $\mathcal{F} = \{D_j,\ s_c < j \leq s_f\}$ 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

$$\tau_{BH}(p_i) = \sum_{d=1}^{\Delta} \mathrm{pr}(\mathcal{A}_d)\pi_{z_i^d}^d.$$

where the probability of an allocation $\mathcal{A}_d$ is $\mathrm{pr}(\mathcal{A}_d) = \prod_i \phi_{i,z_i^d}$.

Since we do not need to distinguish between different allocations in subintervals of D_{accept} and D_{reject}, the number of allocations to be considered can be greatly reduced by treating D_{accept} and D_{reject} as one of the subintervals, see Example 3 below.

Lemma 4.2. 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. For any given allocation $\mathcal{A}_d$, the result holds as for Lemma 4.1, with the intervals I_j replaced by D_j. The proof follows by integrating over all possible allocations.

Example 3. Fuzzy BH procedure

Consider the 7 p-values from a mixture of Binomial distributions, given in Table 1. The support set $\mathcal{I} = [0, 0.145]$ is partitioned 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 $36 = 2^2 \times 3^2$: 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}. Allocations of the first three p-values do not change the ranks of the last four p-values within $\mathcal{F}$, 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.

5 Application: testing for linkage disequilibrium

In this section we demonstrate our procedure on a data set used to test linkage disequilibrium (LD), that is the association between alleles at different markers on the same chromosome. Genotype data consist of pairs of alleles at each locus, with no information about which chromosome each allele comes from. Haplotype data include the chromosome information. For example, for a pair of markers, each with two possible alleles (A,a for the first marker and B,b for the second), the possible haplotypes are (A,B), (A,b), (a,B) and (a,b). A pair of markers is in LD in a population if the alleles found at the two markers on the same chromosome are associated in that population.

Linkage disequilibrium data can be presented in the form of 2x2 contingency tables where haplotypes are classified in terms of their alleles at each of the 2 loci of interest. It is usual to use the hypergeometric distribution (as used in the Fisher's exact test) for testing independence between the loci, as there are many tables with low cell counts and thus the approximation used in the chi-squared test is not valid.

The hypergeometric distribution, unlike the chi-squared distribution, can be

used to find significant positive and negative correlations separately. Thus 2-sided tests are used when both positive and negative correlations are of interest. However, there is ongoing controversy about how 2-sided p-values should be constructed for the hypergeometric distribution (Agresti, 2002). Besides, the randomised p-values constructed from 2-sided p-values will not in general be uniform, interfering with our main purpose.

We propose a choice of p-value which does ensure uniformity for the randomised p-values: 1-sided p-values conditioned on the sign of the correlation (Kulinskaya, 2007). These are given by

$$p_i \equiv \begin{cases} \frac{\text{pr}(X \geq x_i)}{\text{pr}(X \geq x_{mode})}, & r \geq 0; \\ \\ \frac{\text{pr}(X \leq x_i)}{\text{pr}(X \leq x_{mode})}, & r < 0; \end{cases}$$

where X is the random variable for one of the cells in the contingency table and follows a hypergeometric distribution conditional on the margins of the table. The quantity x_{mode} is the value of X corresponding to the most probable table under the null, and r is the correlation coefficient (or equivalently the determinant of the 2x2 contingency table). The randomised p-values based on observed p-values constructed as above are Unif$(0, 1)$ under the null.

Often the 1-sided conditional p-values are equal to the usual 2-sided p-values. This happens in two situations. The first case is when the null distribution for a particular table is so skewed that x_{mode} is at one end of the possible range of X (for example if x_{mode} is the minimum possible value for X and so $\text{pr}(X \geq x_{mode}) = 1$). The second case is when the null distribution is symmetric (in this case $\text{pr}(X \geq x_{mode}) = 0.5$, so p_i is twice the usual unconditional 1-sided p-value).

Chakraborty et al. (1987) looked at the relationship between the disease phenylketonuria (PKU) and 8 markers at the human phenylalanine hydroxylase (PAH) locus. As part of this investigation they tested for LD between the markers. For this purpose, haplotypes were divided into cases (with a mutant allele at the PKU locus) and controls (normal allele), since the marker allele frequencies were significantly different for cases and controls. There were 66 case and 66 control haplotypes. Correlation coefficients were calculated for all pairs of markers, 28 in all, and tested for difference from zero (presumably using the chi-squared test, though this is not stated). No multiple testing correction was performed.

Table 3 shows the 1-sided conditional p-values for the controls haplotypes for each pair of markers in the Chakraborty et al. (1987) data set. The markers are given in the table in the same order as they appear on the chromosome, in a similar format as presented in the original paper. As in the original work, the markers which are closest together have the smallest p-values, except for the pairs involving the marker *Hind*III.

Table 4 shows the fuzzy measures τ of evidence against the null of no correlation for each marker pair, using the randomised Benjamini and Hochberg method for controlling FDR at a level of $\alpha = 0.01$. The pairs with $\tau = 1$ here (that is, strong evidence against the null) would also have their null hypotheses rejected in the usual non-fuzzy method. All other null hypotheses would not be rejected, i.e. they would be declared to have no evidence against the null. With our analysis we can show that for the marker *Pvu*II(b), there is evidence for LD with other markers.

6 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) and Benjamini & Hochberg (1995) were explored in this paper, but it should be possible to similarly generalize other methods, Storey et al. (2004) 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).

An evident drawback is the amount of computation required. These procedures should be efficiently programmed if they are to be of practical use. If there are ties in the observed p-values in the general (overlapping intervals) case, the order of computation can be further reduced since we do not have to separately calculate all the different possible allocations of several copies of the same observed p-value (details available on request). Another possibility 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.

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. As long as the properties of positive regression dependence (PRDS) from Benjamini & Yekutieli (2001) between components of the marginally uniform multivariate distribution of the p-values on $[0,1]^m$ are satisfied, the randomised BH procedure is conservative.

A critical feature of the procedures introduced in this paper is the conditional independence of the randomised p-values $P_i|x_i,\ i = 1, \cdots, m$. During the revision stage we found out that this construction is equivalent to a well known technique of embedding a multivariate discrete distribution in a continuous one, termed the standard extension copula by Schweizer & Sklar (1974). Nešlehová (2007) shows that this construction of a continuous joint distribution on $[0,1]^m$ with uniform marginals captures the monotonic dependence between the original random variables. Since the PRDS property of the copula distribution is invariant under comonotone transformations (Benjamini & Yekutieli, 2001, p.1170), we conjecture that it is inherited from the original monotonic dependence between the discrete random variables. Thus our procedure should be general enough not to be unduly conservative. The conjecture requires further work.

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 & Meeden (2005), and is not discussed. Instead we used conditional 1-sided p-values in Section 5, see Kulinskaya (2007).

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 an BBSRC "Exploiting Genomics" grant.

Appendix: Calculation of $T_{k,l}$

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 randomised 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

$$\begin{aligned} T_{k,l}(p_1, p_2) &\equiv P\{P_{jk} < \alpha_{jk}, P_{j(k+1)} > \alpha_{j(k+1)}, ..., P_{jl} > \alpha_{jl}\} \\ &= \frac{l!}{k!} q_{jk}^k P\{P_{j(k+1)} > \alpha_{j(k+1)}, ..., P_{jl} > \alpha_{jl}\} \end{aligned} \quad (2)$$

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 2, 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 2 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 2 is to be satisfied.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: John Wiley and Sons Ltd, 2nd ed.

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

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate:

a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57 289–300.

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

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

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8 3–62.

Chakraborty, R., Lidsky, A. S., Daiger, S. P., Güttler, F., Sullivan, S., Diliella, A. G. & Woo, S. L. C. (1987). Polymorphic DNA haplotypes at the human phenylalanine hydroxylase locus and their relationship with phenylketonuria. *Human Genetics* 76 40–46.

Cox, D. & Hinkley, D. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Dollinger, M., Kulinskaya, E. & Staudte, R. G. (1996). Fuzzy hypothesis tests and confidence intervals. In D. Dowe, K. Korb & J. Oliver, eds., *Information, Statistics and Induction in Science*. Singapore: World Scientific, 119–128.

Geyer, C. & Meeden, G. (2005). Fuzzy and randimized confidence intervals and p-values. *Statistical Science* 20 358–366.

Gilbert, P. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *J. R. Statist. Soc. C* 54 143–158.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75 800–802.

Kulinskaya, E. (2007). On two-sided p-values for non-symmetric distributions. Manuscript. URL `http://www3.imperial.ac.uk/pls/portallive/docs/1/25061696.PDF`.

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis* 98 544–567.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

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

Ross, T. (2004). *Fuzzy Logic with Engineering Applications*. New York: John Wiley and Sons Ltd, 2nd ed.

Roth, A. (1999). Multiple comparison procedures for discrete test statistics. *J. Statist. Plann.Inference* 82 101–117.

Schweizer, B. & Sklar, A. (1974). Operation on distribution functions not derivable from operations on random variables. *Studia Mathematica* 52 43–52.

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73 751–754.

Storey, J. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B* 64 479–498.

Storey, J., Taylor, J. & Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Statist. Soc. B* 66 187–205.

Tarone, R. (1990). A modified Bonferroni method for discrete data. *Biometrics* 46 515–522.

n_i	k_i	p_i	p_{i-}	$p_i - p_{i-}$	$\tau_B(p_i)$
8	0	0.003906	0	0.003906	1
10	1	0.010742	0.000977	0.009766	0.631429
6	0	0.015625	0	0.015625	0.457143
8	1	0.035156	0.003906	0.03125	0.103571
10	2	0.054688	0.010742	0.043945	0
6	1	0.109375	0.015625	0.09375	0
8	2	0.144531	0.035156	0.109375	0

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

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 (overlapping support intervals).** 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, $A_{j\pm} = \mathcal{R}_{j\pm}\alpha/7$

	*Bgl*I	*Pvu*II(a)	*Pvu*II(b)	*Eco*RI	*Msp*I	*Xmn*I	*Hind*III
*Pvu*II(a)	5×10^{-15}	-	-	-	-	-	-
*Pvu*II(b)	1×10^{-5}	1×10^{-5}	-	-	-	-	-
*Eco*RI	2×10^{-4}	2×10^{-4}	3×10^{-2}	-	-	-	-
*Msp*I	1	1	2×10^{-2}	2×10^{-10}	-	-	-
*Xmn*I	1	1	2×10^{-2}	2×10^{-2}	3×10^{-19}	-	-
*Hind*III	7×10^{-4}	7×10^{-4}	7×10^{-2}	1×10^{-3}	3×10^{-7}	3×10^{-7}	-
*Eco*RV	1	1	1×10^{-2}	5×10^{-7}	5×10^{-3}	5×10^{-3}	1×10^{-10}

Table 3: The 1-sided p-values conditional on the sign of the correlation coefficient, for the linkage disequilibrium data set from Chakraborty et al. (1987) Chakraborty et al. (1987). The markers are listed in the order they appear on the chromosome.

	*Bgl*I	*Pvu*II(a)	*Pvu*II(b)	*Eco*RI	*Msp*I	*Xmn*I	*Hind*III
*Pvu*II(a)	1	-	-	-	-	-	-
*Pvu*II(b)	1	1	-	-	-	-	-
*Eco*RI	1	1	0.21	-	-	-	-
*Msp*I	0	0	0.39	1	-	-	-
*Xmn*I	0	0	0.39	1	1	-	-
*Hind*III	1	1	0.10	1	1	1	-
*Eco*RV	0	0	0.62	1	1	1	1

Table 4: Results for the linkage disequilibrium (LD) data set from Chakraborty et al. (1987). The values given are τ, the fuzzy measure of evidence against the null hypothesis of no LD, for the Benjamini and Hochberg FDR method at level $\alpha = 0.01$. The markers are listed in the order they appear on the chromosome.

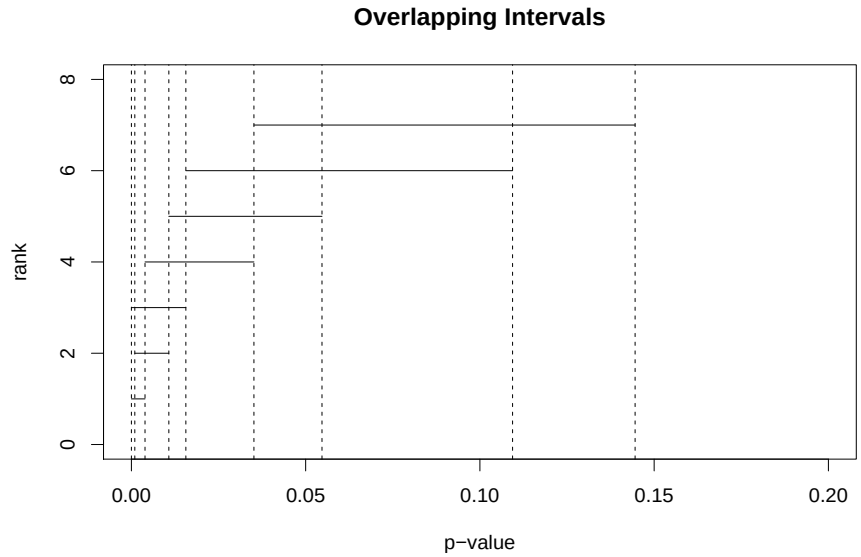

Figure 1: *Plot of p-values and related intersecting support intervals for the data from Example 4, given in Table 2. Support intervals (horizontal segments) are ordered by the ranks of respective p-values on the vertical axis. The support set* $\mathcal{I} = [0, 0.145]$ *is split by vertical dashed lines into 8 subintervals* $D_j,\ j = 1, \cdots, 8$ *on the horizontal axis.*