On fuzzy family wise error rate and false discovery rate procedures for discrete distributions

By ELENA KULINSKAYA

Statistical Advisory Service, Imperial College,
8 Princes Gardens, London SW7 1NA, U.K.
e.kulinskaya@imperial.ac.uk

And ALEX LEWIN

Department of Epidemiology and Public Health, Imperial College,
St Mary's Campus, Norfolk Place, London W2 1PG, U.K.
a.m.lewin@imperial.ac.uk

Summary

Fuzzy multiple comparisons procedures are introduced as a solution to the problem of multiple comparisons for discrete test statistics. The critical function of the randomized p-values is proposed as a measure of evidence against the null hypotheses. The classical concept of randomized 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 family wise error rate and false discovery rate procedures are discussed and an application to linkage disequilibrium testing is given. Software for implementing the procedures is available.

Some keywords: Benjamini-Hochberg procedure; Bonferroni procedure; False discovery rate; Fuzzy decision-making; Multiple comparisons; Randomized tests.

1 Introduction

The Bonferroni correction controls the family wise error rate, that is the probability of committing any type 1 error in families of comparisons under simultaneous consideration. Less conservative family wise error rate procedures using the observed individual p-values were introduced by Simes (1986), Hochberg (1988) and Rom (1990). Benjamini & Hochberg (1995) introduced a novel class of more powerful procedures that control the false discovery rate. Their procedure is referred to as the BH procedure in what follows. Benjamini & Yekutieli (2001) studied the false discovery rate procedures under dependence. An alternative approach, estimating false discovery rate was introduced in Storey (2002) and Storey et al. (2004).

The controlling procedures cited above were developed for p-values arising from continuous test statistics. Under appropriate conditions, each will control either the family wise error rate or the false discovery rate at a level α. The proofs use the fact that the p-values have a $\text{Un}(0,1)$ distribution under the null hypothesis. This will not hold for discrete distributions, even for a single test, and the problem is exacerbated for multiple tests of null hypotheses with different discrete distributions. The procedures are more conservative, and therefore less powerful.

Multiple testing of discrete test statistics is particularly important currently, with the development of novel genomics applications. Chakraborty et al. (1987) includes a typical genetics example of testing for linkage disequilibrium, that is correlation between alleles at pairs of markers. Gilbert (2005) uses Fisher's tests 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 (Al-Shahrour et al., 2004) and association studies in genetics.

To overcome inherent difficulties in working with discrete distributions, Tarone (1990) managed to reduce the number of comparisons by disregarding the hypotheses which have no chance of achieving significance after the adjustment. Further improved family wise error rate 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 a false discovery rate procedure that combines the Tarone (1990) ideas with the BH type procedure.

We use an approach based on the idea of randomized tests (Cox & Hinkley, 1974, pp. 99-101). 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. This depends only on the observed p-values and the level α of the test procedure: no randomization is performed to obtain the fuzzy measure. The connection between test critical functions and fuzzy quantities was discussed in Dollinger et al. (1996) and applied recently to randomized tests and p-values by Geyer & Meeden (2005). We show how this idea can be extended to the multiple testing situation. Multiple tests are randomized independently, and the marginal critical function for each test is used to construct a multiple comparisons procedure. We provide algorithms for exact calculation of the fuzzy measures. An R (R Development Core Team, 2004) package implementing the fuzzy procedures is available from http://www.bgx.org.uk/alex/.

2 Randomized p-values and fuzzy decision rules

Consider a discrete test statistic X, which can take values in $\{x_1 < x_2 < ...\}$. If the observed value of the statistic is x_i, the traditional 'crisp' p-value P for a one-sided test is $p_i \equiv \mathrm{pr}(X \geq x_i)$ 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. $\mathrm{pr}(P \leq p_i) = p_i$, as opposed to the continuous $\mathrm{Un}(0, 1)$ distribution for p-values of continuously distributed statistics. Thus in general it is not possible to obtain the exact level-α test.

This difficulty may be solved by the introduction of randomized tests. For a discrete null distribution of a test statistic X, let c be the value of the statistic such that $\mathrm{pr}(X \geq c) > \alpha$ but $\mathrm{pr}(X > c) < \alpha$. Then the exact level-$\alpha$ test can be achieved by using a randomized p-value $P(c) = \mathrm{pr}(X > c) + U\mathrm{pr}(X = c)$ for $U \sim \mathrm{Un}(0, 1)$ (Cox & Hinkley, 1974, p. 101). Traditionally this was interpreted as a need for an extra Bernoulli experiment with probability of rejection $\{\alpha - \mathrm{pr}(X > c)\}/P(c)$ when $X = c$. An alternative interpretation is that the p-value is a random variable, uniformly distributed between two discrete consecutive values. Unconditionally, this randomized p-value has a continuous $\mathrm{Un}(0, 1)$ distribution under the null.

If $p_{i-} \equiv \mathrm{pr}(X > x_i) = \mathrm{pr}(X \geq x_{i+1})$, then $p_{i-} < p_i$ and the randomized p-value is $P_i|x_i \equiv p_{i-} + U(p_i - p_{i-}) \sim \mathrm{Un}(p_{i-}, p_i)$ conditionally on x_i. Thus the conditional probability of rejection of the null hypothesis based on the randomized p-value $\mathrm{pr}(P_i \leq \alpha|x_i)$ is

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

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

Geyer & Meeden (2005) used $\tau(p_i)$ 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 randomized p-values when standard multiple testing procedures are used to control the family wise error rate and false discovery rate. Since the randomized p-values have unconditionally a $\mathrm{Un}(0, 1)$ distribution under the null hypothesis, all properties of multiple comparison procedures for continuous test statistics are automatically fulfilled. Multiple tests are randomized independently, i.e. conditionally random variables $P_i|x_i,\ i = 1, ..., m$ are independent by construction. Calculations of rejection probabilities for the p-values in §§3 and 4 use this conditional independence. This construction is sufficiently general not to be detrimental to the properties of the resulting procedures, as discussed in §6.

3 Fuzzy Bonferroni procedure

For continuous p-values, the Bonferroni procedure rejects 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 $\mathrm{pr}(P_i \leq \alpha/m|x_i)$.

What we call the fuzzy Bonferroni procedure is defined by the marginal critical functions of the randomized tests:

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

Example 1: Fuzzy Bonferroni procedure for Binomial tests. Consider the results of seven one-sided Binomial tests of $H_0 : p = 0.5$ versus the 1-sided alternative $p < 0.5$. The tests reject for small values of $X_i \sim \mathrm{Bi}(n_i, 0.5), i = 1, ..., 7$. The seven p-values are given in Table 1, and the support intervals (p_{i-}, p_i) are plotted in Fig. 1. The standard level–0.05 Bonferroni procedure compares p-values to $0.05/7 = 0.00714$, thus 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.

4 Controlling false discovery rate for a discrete distribution

4.1 The Benjamini and Hochberg procedure

As before, we calculate the marginal probabilities of rejection for the randomized p-values, this time using the Benjamini & Hochberg (1995) procedure. The continuous BH procedure consists of ordering the p-values, then examining them in turn starting from the largest. Hypothesis i is rejected if the ith p-value is less than $\text{rank}(i)\alpha/m$. As soon as one hypothesis is rejected, all hypotheses with smaller p-values are also rejected.

For the discrete case 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. In general the order of the randomized p-values will vary between realisations. For this reason, it will be useful to think about the support intervals (p_{i-}, p_i) of the randomized p-values. The calculation of probabilities of rejection is much easier when these support intervals do not overlap, e.g. when all test statistics have the same null distribution. This case is considered in §4.2. The case of overlapping intervals is presented in §4.3.

4.2 Nonoverlapping support intervals

If the support intervals $(p_{i-}, p_i]$ do not overlap, and there are many tests, it is likely that several observed p-values will be equal, and these will have the same probability of rejection. We call this subset of equal p-values a tie. The calculation of the probabilities can be done for each of the J distinct support intervals, rather than for each of the m p-values, where J can be considerably smaller than m. Denote the probability of rejection for p-values in interval j by π_j. Then the probability of rejection for test i is $\tau_{\text{BH}}(p_i) = \pi_j$ where j is the index of the interval to which randomized 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. Let the largest p-value rank for interval $(p_{j-}, p_j]$ be R_{j+}. Suppose, without loss of generality, that all hypotheses corresponding to p-values larger than p_j are accepted. Then there are three cases:

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

Case 2: $p_{j-} < R_{j+}\alpha/m < p_j$; The probability of the randomized p-values being less than α/m multiplied by their respective rank is between 0 and 1, thus the tie is fuzzily rejected, $0 < \pi_j < 1$.

Case 3: $p_j \leq R_{j+}\alpha/m$; All randomized p-values are less than α/m multiplied by their respective rank with probability 1, thus the tie is crisply rejected, i.e. $\pi_j = 1$.

Consider the fuzzy rejection case in more detail. Suppose there are l tests with observed p-value p_j, and denote the probability of exactly k randomized p-values out of l being rejected by $T_{k,l}(p_{j-}, p_j)$, $0 \le k \le l$. Calculation of $T_{k,l}$ is given in the Appendix. 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:

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

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

Next consider decisions about the p-values in previous intervals, those corresponding to smaller p-values, in each of the above three cases.

In Case 1, when the interval $(p_{j-}, p_j]$ is accepted, the previous interval is accepted or crisply/fuzzily rejected on its own merit.

In Case 2, if $(p_{j-}, p_j]$ is a fuzzy interval there are 2 sub-cases to be considered. 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. With probability $T_{0l}(p_{j-}, p_j)$ no hypothesis in $(p_{j-}, p_j]$ is rejected, so the preceding interval may be accepted or crisply/fuzzily rejected on its own merit. The probability of rejection for the preceding interval is therefore $\pi_j^{\text{prec}} = \{1 - T_{0l}(p_{j-}, p_j)\} + T_{0l}(p_{j-}, p_j) l^{-1} \sum_{k=1}^{l} k T_{k,l}(p_{j-}^{\text{prec}}, p_j^{\text{prec}})$ and the probability of no rejection in the preceding interval is $T_{0l}(p_{j-}, p_j)T_{0l}(p_{j-}^{\text{prec}}, p_j^{\text{prec}})$.

In Case 3, if the interval $(p_{j-}, p_j]$ is crisply rejected, all preceding intervals are also crisply rejected.

In this paragraph we define what we call the fuzzy BH procedure for ordered nonoverlapping support intervals. Let m ordered p-values have $J \leq m$ distinct values $p_1, ..., p_J$, with ties of length $l_j, j = 1, ..., J$, $\sum l_j = m$. Let each corresponding randomized p-value be uniformly distributed on a support interval $I_j = (p_{j-}, p_j]$, where the intervals $I_j, j = 1, ..., J$ are nonoverlapping and are ordered by value of p_j. Let the ranks of the p-values in the jth tie be from $R_{j-} = \sum_{t<j} l_t + 1$ to $R_{j+} = \sum_{t \leq j} l_t$. Define $s_f = \max\{j : \quad p_{j-} \leq R_{j+}\alpha/m\}$ and $s_c = \max\{j : \quad p_j \leq R_{j+}\alpha/m\}$, $s_c \leq s_f$. Then all p-values in the interval $D_{\text{rej}} = \cup\{I_j, \;\; j \leq s_c\}$ are crisply rejected and all p-values in the interval $D_{\text{acc}} = \cup\{I_j, \;\; j > s_f\}$ are accepted. The fuzzy interval is defined as $\mathcal{F} = \{I_j, \;\; s_c < j \leq s_f\}$. 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, see Algorithm 1.

Algorithm 1. Calculation of rejection probabilities in each interval.

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

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

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

and $\eta_j = \eta_{j+1} T_{0,\,l_j}(p_{j1}, p_{j2})$.

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

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

Lemma 1. *For independent test statistics, and for $m_0 \le m$ true null hypotheses, the above randomized* BH *procedure controls false discovery rate exactly at level $m_0\alpha/m$.*

Proof. This is part of Theorem 5.1 from Benjamini & Yekutieli (2001), applicable to any continuous test statistics. Any m-tuple of randomized 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' $\mathcal{F} = \{I_{s_c}+1, ..., I_{s_f}\}$ are rejected or accepted regardless of their generated values. The false discovery rate is exactly $m_0\alpha/m$, conditional on any generated realisation within the fuzzy subset $\mathcal{F}$. The proof follows by integrating over all possible realisations. □

Example 2: Fuzzy BH *procedure for the same discrete distribution.* Consider $m = 10$ one-sided sign tests for $n = 8$ subjects, $S_i \sim \mathrm{Bi}(8, 0.5)$. Set the false discovery rate level at $\alpha = 0.05$. The p-values are 0.004, 0.035×3, 0.145×2 and 0.363×4. For $p = p_2$ the interval $I_2 = (p_{2-}, p_2] = (0.004,\ 0.035]$ contains $l = 3$ p-values, $R_{2-}\alpha/m = 0.01$ and $R_{2+}\alpha/m = 0.02$. Therefore, $s_c = 1$ and $s_f = 2$.

The q_k values defined in the Appendix are 0.194, 0.355 and 0.516 respectively. We obtain

$$\begin{aligned}
T_{1,3}(p_2) &= 6q_1(q_3 - q_2)(1 - q_3) + 3q_1(1 - q_3)^2 = 0.227, \\
T_{2,3}(p_2) &= 3q_2^2(1 - q_3) = 0.183, \\
T_{3,3}(p_2) &= q_3^3 = 0.137.
\end{aligned}$$

For each of the three hypotheses with p-value of 0.035 the probability of rejection is $\pi_2 = \pi(0.035) = 3^{-1} \sum kT_{k,3}(p_2) = 0.335$ and the probability of rejecting at least one of the three hypotheses is $1 - T_{0,3}(0.035) = 0.547$. The p-value $p_1 = 0.004$ is crisply rejected.

4.3 General case.

When the randomized p-values $\{P_i,\ i = 1, ..., m\}$ originate from different distributions, the support intervals may overlap, so there is no strict ordering between them. We partition the unit interval into intervals based on the intersections of the support intervals, so that these smaller intervals are nonoverlapping. For each realisation of m randomized p-values, we can think of allocating these p-values to the nonoverlapping intervals. Given a particular allocation, the calculation of π_j for interval j can proceed as in §4.2. In order to calculate the $\tau_{\text{BH}}(p_i)$ for each test i, we must integrate over the possible allocations of randomized p-values. We stress again that the value of $\tau_{\text{BH}}(p_i)$ does not depend on any particular realisation of randomized p-values, but only on the observed discrete p-values.

In the next three paragraphs we define what we call the fuzzy BH procedure in the general case of overlapping support intervals. Let each randomized p-value have support in the interval I_i. Partition the support set $\mathcal{I} = \bigcup I_i,\ i = 1, ..., m$ into $J \leq 2m$ ordered subintervals $\mathcal{I} = \bigcup D_j,\ j = 1, ..., J$, where $D_j = (D_{j-}, D_{j+}]$. Let the probability of randomized p-value P_i belonging to interval D_j be denoted by $\phi_{ij} = |D_j \cap I_i|/|I_i|$. Let $\mathcal{A} = \{\mathcal{A}_d, d = 1, ..., \Delta\}$ 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 randomized p-value i is allocated in allocation d.

For each subinterval $D_j, j = 1, ..., J$, denote the maximum and the minimum possible ranks across all allocations $\mathcal{A}_d$ by $\mathcal{R}_{j+}$ and $\mathcal{R}_{j-}$. Define $s_f = \max\{j : \ D_{j-} \leq \mathcal{R}_{j+}\alpha/m\}$ and $s_c = \max\{j : \quad D_{j+} \leq \mathcal{R}_{j-}\alpha/m\}$, $s_c \leq s_f$. Then all p-values in the interval $D_{\text{rej}} = \cup\{D_j, \ j \leq s_c\}$ are crisply rejected, all p-values in the interval $D_{\text{acc}} = \cup\{D_j, \ j > s_f\}$ are accepted, and 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 the ith p-value is

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

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

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

Lemma 2. *For independent test statistics, and for $m_0 \leq m$ true null hypotheses, the above randomized* BH *procedure controls false discovery rate exactly at level $m_0\alpha/m$.*

Proof. For a given allocation $\mathcal{A}_d$, the result holds as for Lemma 2, with the intervals I_j replaced by D_j. The proof follows by integrating over all possible $\mathcal{A}_d$. □

Example 3: Fuzzy BH *procedure.* Consider the seven 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 Fig. 1. Here $s_c = 4,\ s_f = 6$. The first four intervals have $D_{j+} < \mathcal{R}_{j-}\alpha/7$ and therefore constitute D_{rej}; intervals 7 and 8 constitute D_{acc}; intervals 5 and 6 are the fuzzy subset $\mathcal{F}$. The p-values which may end up in the fuzzy subset are p-values 4 to 7. Each can belong to 3 different subintervals, and 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_{rej} and p-value 7 may belong to D_6 or to D_{acc}.

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 $D_{5+} < \mathcal{R}_{5+}\alpha/7$, and every p-value in D_5 is crisply rejected, $\pi_5 = 1$. When there is only one p-value with rank 4 in D_5, it is fuzzily rejected with probability $\pi_5 = 1 - T_{0,l_6}(D_6) + T_{0,l_6}(D_6)\sum_{k=1}^{l} kT_{k,l_5}(D_5)$. This happens only when p-value 4 alone belongs to D_5, with p-value 5 in D_6, and p-values 6 and 7 in D_6 or D_{acc}; this occurs in four possible allocations with l_6 varying from 1 to 3.

Summing up the probabilities of rejection for each p-value we obtain $\tau_{\text{BH}}(P_1) = \tau_{\text{BH}}(P_2) = \tau_{\text{BH}}(P_3) = 1$, $\tau_{\text{BH}}(P_4) = 0.941$, $\tau_{\text{BH}}(P_5) = 0.632$, $\tau_{\text{BH}}(P_6) = 0.281$ and $\tau_{\text{BH}}(P_7) = 0.080$. 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 dataset used to test linkage disequilibrium, 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 the chromosome from which each allele comes. 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 linkage disequilibrium 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 2 x 2 contingency tables in which haplotypes are classified in terms of their alleles at each of the two loci of interest. It is usual to use the hypergeometric distribution, as used in 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 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, p. 93). We propose to use 1-sided p-values conditioned on the sign of the correlation. These are given by

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

where X is the random variable for one of the cell entries 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. The randomized p-values based on observed p-values constructed as above are $\mathrm{Un}(0, 1)$ under the null hypothesis. For symmetric distributions the 1-sided conditional p-values are equal to the usual 2-sided p-values.

Chakraborty et al. (1987) looked at the relationship between the disease phenylketonuria and 8 markers at the human phenylalanine hydroxylase locus. As part of this investigation they tested for linkage disequilibrium between the markers. For this purpose, haplotypes were divided into cases, with a mutant allele at the phenylketonuria 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. No multiple testing correction was performed.

Table 3 shows the 1-sided conditional p-values for the controls haplotypes for each pair of markers. The markers are given in the table in the same order as they appear on the chromosome, in a similar format to that presented in the original paper. As there, the markers which are closest together have the smallest p-values, except for the pairs involving the marker HindIII.

Table 4 shows the fuzzy measures τ of evidence against the null of no correlation for each marker pair, using the randomized Benjamini and Hochberg method for controlling false discovery rate at a level of $\alpha = 0.01$. The pairs with $\tau = 1$ here 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 provide no evidence against the null hypothesis. With our analysis we can show that, for the marker PvuII(b), there is evidence for linkage disequilibrium with other markers.

6 Discussion

We have shown how the classical concept of randomized tests can be extended to multiple comparisons. It should be possible to generalize other methods, such as Storey et al. (2004), along the same lines.

To be of practical use these procedures should be efficiently programmed. 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 calculate separately all the different possible allocations of several copies of the same observed p-value;

details are available on request. Another possibility would be to generate N sets of m p-values from $\prod_{i=1}^{m} \mathrm{Un}(I_i)$, and to estimate probabilities of rejection τ_i through proportions of rejection out of N.

False discovery rate control at exactly level $m_0\alpha/m$ requires independence of the p-values. However, the calculation of rejection probabilities $\tau(p_i)$ in §§3 and 4 holds regardless, because of the conditional independence of the randomized p-values. As long as the properties of positive regression dependence from Benjamini & Yekutieli (2001) between components of the marginally uniform multivariate distribution of the p-values on $[0, 1]^m$ are satisfied, the randomized BH procedure is conservative.

A critical feature of the procedures introduced in this paper is the conditional independence of the randomized p-values $P_i|x_i,\ i = 1, ..., m$. 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 positive regression dependence 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 theory in this paper applies directly only to 1-sided p-values or p-values from symmetric distributions. Treatment of p-values for 2-sided tests with nonsymmetric distributions is technically more involved, see Geyer & Meeden (2005), and is not discussed. Instead we used conditional 1-sided p-values in §5; see also an unpublished Imperial College Technical Report by E. Kulinskaya.

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

Acknowledgement:

This work was in part supported by an 'Exploiting Genomics' grant from the U.K.

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 nonoverlapping case, we need to calculate the unconditional probability π_j that a particular hypothesis is rejected and the probability η_j that no hypothesis in the interval is 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, that is p_{j-}, p_j in the nonoverlapping case.

Let the number of randomized p-values in the interval be l_j, and let 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 index j in the tie length l_j.

We need to calculate

$$\begin{aligned} T_{k,l}(p_1, p_2) &\equiv \mathrm{pr}(P_{jk} < \alpha_{jk}, P_{j(k+1)} > \alpha_{j(k+1)}, ..., P_{jl} > \alpha_{jl}) \\ &= \frac{l!}{k!} q_{jk}^k \mathrm{pr}(P_{j(k+1)} > \alpha_{j(k+1)}, ..., P_{jl} > \alpha_{jl}), \end{aligned} \tag{A1}$$

where $P_{jk}, i = 1, ..., l$ are order statistics from a $\mathrm{Un}(p_1, p_2)$ distribution.

In order to calculate the probability in equation A1, the $\{P_{j(k+1)}, ..., P_{jl}\}$ have to be allocated to 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, it is easy to calculate the probability as a product of two types of term:

$$\begin{aligned} \mathrm{pr}(\alpha_{jr} < P_{j(s+1)} < ... < P_{j(s+u)} < \alpha_{j(r+1)}) &= \frac{t_j^u}{u!}, \\ \mathrm{pr}(P_{jl} > ... > P_{j(l-r+1)} > \alpha_{jl}) &= \frac{(1 - q_{jl})^r}{r!}. \end{aligned}$$

These terms correspond respectively to u p-values being allocated between two adjacent α's and to the largest r p-values being allocated to the top interval (α_{jl}, p_2).

The allocations can be labelled uniquely by $l - k$ integers, denoting the number of randomized p-values in the above alpha intervals; for example, $\alpha_1 < P_1 < \alpha_2 < \alpha_3 < P_2 < P_3$ is denoted by 102 (*i.e.* $l = 3$ and $k = 0$). If we call these integers

$n_{k+1}, ..., n_l$, the probability we need for equation A1 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, and therefore 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 A1 is to be satisfied.

References

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

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** 578–80.

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. *Ann. Statist.* **29** 1165–88.

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. *Hum. Genet.* **76** 40–6.

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. *Information, Statistics and Induction in Science* pp. 119–28. Ed. D. Dowe, K. Korb and J. Oliver, Singapore: World Scientific.

Geyer, C. & Meeden, G. (2005). Fuzzy and randomized confidence intervals and p-values. *Statist. Sci.* **20** 358–66.

Gilbert, P. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Appl. Statist.* **54** 143–58.

HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–02.

NEŠLEHOVÁ, J. (2007). On rank correlation measures for non-continuous random variables. *J. Mult. Anal.* **98** 544-67.

R DEVELOPMENT CORE TEAM (2004). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

ROM, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77** 663–5.

ROTH, A. (1999). Multiple comparison procedures for discrete test statistics. *J. Statist. Plan. Inf.* **82** 101–7.

SCHWEIZER, B. & SKLAR, A. (1974). Operation on distribution functions not derivable from operations on random variables. *Stud. Math.* **52** 43-52.

SIMES, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–4.

STOREY, J. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc.* B **64** 479–98.

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** 515–22.

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. Here $p_i = \mathrm{pr}(X_i \leq k_i)$ is a p-value from a 1-sided binomial test, with $X_i \sim \mathrm{Bi}(n_i; 0.5)$ under the null hypothesis, p_{i-} is the previous attainable p-value and $\tau_B(p_i)$ is the probability of rejection by the fuzzy Bonferroni procedure.

j	D_{j-}	D_{j+}	$\|D_j\|$	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 with overlapping support intervals. Data are given in Table 1. Here 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, and $A_{j\pm} = \mathcal{R}_{j\pm}\alpha/7$.

	BglI	PvuII(a)	PvuII(b)	EcoRI	MspI	XmnI	HindIII
PvuII(a)	5×10^{-15}	-	-	-	-	-	-
PvuII(b)	1×10^{-5}	1×10^{-5}	-	-	-	-	-
EcoRI	2×10^{-4}	2×10^{-4}	3×10^{-2}	-	-	-	-
MspI	1	1	2×10^{-2}	2×10^{-10}	-	-	-
XmnI	1	1	2×10^{-2}	2×10^{-2}	3×10^{-19}	-	-
HindIII	7×10^{-4}	7×10^{-4}	7×10^{-2}	1×10^{-3}	3×10^{-7}	3×10^{-7}	-
EcoRV	1	1	1×10^{-2}	5×10^{-7}	5×10^{-3}	5×10^{-3}	1×10^{-10}

Table 3: Linkage disequilibrium data set. One-sided p-values conditional on the sign of the correlation coefficient. The markers are listed in the order they appear on the chromosome.

	BglI	PvuII(a)	PvuII(b)	EcoRI	MspI	XmnI	HindIII
PvuII(a)	1	-	-	-	-	-	-
PvuII(b)	1	1	-	-	-	-	-
EcoRI	1	1	0.21	-	-	-	-
MspI	0	0	0.39	1	-	-	-
XmnI	0	0	0.39	1	1	-	-
HindIII	1	1	0.10	1	1	1	-
EcoRV	0	0	0.62	1	1	1	1

Table 4: Linkage disequilibrium data set. The values given are τ, the fuzzy measure of evidence against the null hypothesis of no linkage disequilibrium, for the Benjamini and Hochberg false discovery rate method at level $\alpha = 0.01$. The markers are listed in the order they appear on the chromosome.

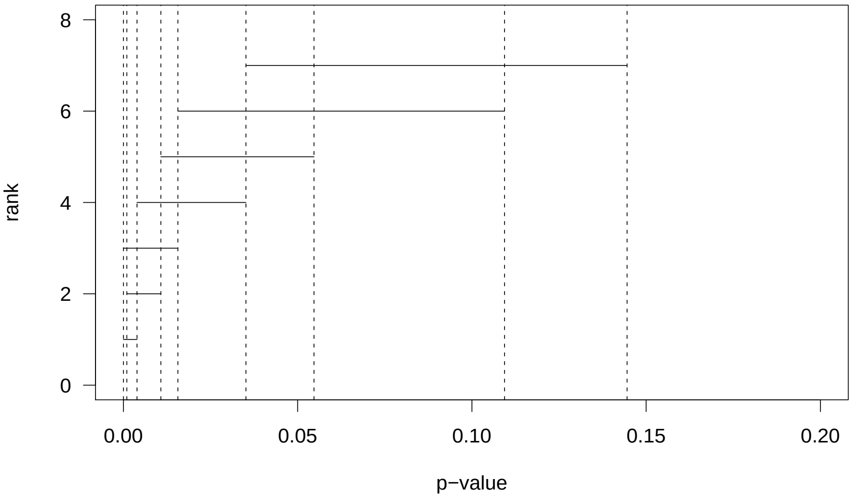

Fig. 1: Plot of p-values and related intersecting support intervals for the data from Example 4, given in Table 2. Support intervals, given by horizontal segments, are ordered by the ranks of respective p-values on the vertical axis. The support set $\mathcal{I} = [0, 0.145]$ is split by vertical dashed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