Estimation of False Discovery Rate Using Permutation P-Values with Different Discrete Null Distributions

Tim Bancroft

Dan Nettleton

Iowa State University

June 16, 2009

Abstract

The false discovery rate (FDR) is a multiple testing error rate which describes the expected proportion of expected type I errors among the total number of rejected hypotheses. Benjamini and Hochberg introduced this quantity and provided an estimator that is conservative when the number of true null hypotheses, m_0, is smaller than the number of tests, m. Replacing m with m_0 in Benjamini and Hochberg's procedure reduces the conservative bias, but requires estimation as m_0 is unknown. Methods exist

to estimate m_0 when each null p-value is distributed as a continuous uniform (0,1) random variable. This paper discusses how to estimate m_0 and therefore FDR when the m p-values are from a mixture of different discrete distributions resulting from permutation testing. The method will be demonstrated through an analysis of proteomics data.

Key Words: Permutation testing; False discovery rate; Multiple testing; Randomized complete block design

1 Introduction

Multiple testing is a growing area of statistics with applications including microarray experiments, proteomics, and quantitative trait loci mapping studies. In some cases, the number of hypotheses tested can be in the thousands, in which case the expected number of type I errors for a typical significance threshold of 0.01 is non-trivial. Many multiple testing error rates exist including the original family wise error rate (FWER) which controls the probability of at least one false positive, i.e. rejecting a hypothesis which is true. Because the FWER is conservative when willing to admit some type I errors, focus has shifted away from FWER control to the pioneering work on the false discovery rate (FDR) of Benjamini and Hochberg (1995), which controls the rate of false findings, i.e. total number of false discoveries to the total number of discoveries. Benjamini and Hochberg (1995) described a procedure to control FDR at a prespecified level α, which is conservative when the number of true null hypotheses, m_0, is smaller than the number of tests, m. The unknown quantity of number of true null hypotheses, m_0, can replace m in their procedure to reduce the conservative bias, but then requires estimation as m_0 is unknown. Many methods exist to estimate m_0 including Benjamini and Hochberg (2000), Storey (2000, 2002), Storey and Tibshirani (2003), Nettleton et al. (2006) among others. Most of the existing methods assume the p-values used to test the m null hypotheses have a continuous uniform distribution on the open interval (0,1) under the null hypothesis. This paper describes how to estimate m_0, and therefore FDR, when the null p-values are from a mixture of discrete non-uniform distributions.

Section 2 describes permutation testing applied to randomized complete block design (RCBD) data with many zero counts that have the potential to create a p-value null distribution where the elements of the discrete support are not equally likely. An-

other example involving next generation sequencing data is also discussed. In Section 3, the proposed procedure is introduced along with some review of existing techniques. The proposed procedure is applied to a proteomics data set in Section 4. A data-based simulation study is presented in Section 5 which compares the estimated FDR to the true false positive fraction (ratio of false positives to total number of rejections) and also evaluates how well the proposed procedure estimates m_0.

2 P-values with Different Discrete Supports

This section describes two scenarios where the analysis results in a collection of p-values from a mixture of discrete distributions, where under the null hypothesis and for any single discrete distribution, the elements of the support are not equally likely.

2.1 A Motivating Example

Some details of a proteomics experiment are given in Section 4 in which over one thousand proteins are tested for differential expression across five genotype groups. Here, we describe the analysis of data from one protein to motivate our proposed approach.

Consider a randomized complete block design (RCBD) with r repetitions and t treatments. Also, suppose the data contain many zero observations. Table 1 shows a possible data set under this scenario with $r = 3$ and $t = 5$.

As is the case in the example depicted in Table 1, a small r and small t result in limited data that does not lend itself well to conventional parametric modeling procedures and therefore would require a non-parametric testing procedure such as

	Treatment				
Repetition	t_1	t_2	t_3	t_4	t_5
1	0	0	0	0	0
2	0	0	0	0	y_1
3	0	0	0	0	y_2

Table 1: Example of a randomized complete block design with $r = 3$ repetitions and $t = 5$ treatments with many observations that are zero.

permutation testing. To employ permutation testing in an RCBD, the data must be permuted within each repetition. Then, the chosen test statistic is computed for each possible permutation and the p-value is determined by the proportion of test statistics that are as extreme or more extreme than the observed test statistic. For this case, we use the traditional F-statistic, which summarizes the amount of within group variability compared to the between group variability. For the data in Table 1, it turns out that there are only two possible values of the test statistic; the observed test statistic (i.e. whenever y_1 and y_2 are in the same column) and the test statistic obtained whenever y_1 and y_2 are in different columns. To see this using permutation testing, note that since repetition 1 is all zero entries, it can stay fixed, which leaves repetition 2 and repetition 3. Now, the p-value obtained when doing all possible permutations of repetition 2 and repetition 3, of which there are $5 \cdot 5$ possible arrangements (five positions for y_1 and five positions for y_2), would be the exact same as if repetition 2 was fixed and only repetition 3 was permuted. This is because the p-value is defined as a fraction, and if both repetition 2 and repetition 3 were permuted, the numerator and the denominator would both be the same multiple larger, thus canceling out. Therefore, when permutation testing is applied to RCBD data, all repetitions where the response is constant, and one repetition of the remaining repetitions can be fixed,

both substantially reducing the number of permutations necessary. Hence, the first repetition may be ignored and the second repetition will be fixed. Then, there are five positions for y_2 which results in five test statistics with only two distinct values. One distinct value corresponds to the arrangement in Table 1. The other distinct value is obtained when y_2 is in any of the other four positions. Hence, the actual test statistic occurs with H_0 probability 1/5 and the other test statistic occurs with H_0 probability 4/5. Note that the test statistic is largest under the arrangement in Table 1 and so the respective p-values are 0.20 and 1, thus, not equally likely under H_0. This distribution is summarized in Table 2. Here, the p-value associated with testing if the 5 treatments all have the same mean response is 0.20.

p-value	.20	1
H_0 probability	.20	.80

Table 2: The null distribution of possible p-values resulting from permutation testing of the randomized complete block design data in Table 1.

Table 3 shows another example that we will briefly go through to illustrate a different p-value support.

	Treatment				
Repetition	t_1	t_2	t_3	t_4	t_5
1	0	y_1	0	0	0
2	y_2	0	0	0	0
3	0	0	y_3	0	0

Table 3: Example of a randomized complete block design with $r = 3$ repetitions and $t = 5$ treatments with sparse data.

Again, we employ permutation testing to test if the 5 treatments all have the same mean response. First, fix repetition 1 and compute test statistics for all possible

permutations of repetition 2 and repetition 3, of which there are 25. In this case, the test statistic can take on one of only five distinct values (if y_1, y_2, and y_3 are distinct) corresponding to the five following arrangements:

1) y_1, y_2, and y_3 are in the same column

2) y_2 and y_3 are in the same column, but not the same column as y_1

3) y_1 and y_2 are in the same column but not the same column as y_3

4) y_1 and y_3 are in the same column but not the same column as y_2

5) y_1, y_2, and y_3 are in different columns.

Note that case 1) occurs with H_0 probability $\frac{1}{25}$, cases 2), 3), and 4) each occur with H_0 probability $\frac{4}{25}$, and case 5) occurs with H_0 probability $\frac{12}{25}$. These five cases result in 5 distinct p-values (again, if y_1, y_2, and y_3 are distinct), and again, these p-values are not equal likely under H_0. Table 2 summarizes this distribution.

p-value	.04	.20	.36	.52	1
H_0 probability	.04	.16	.16	.16	.48

Table 4: The null distribution of possible p-values resulting from permutation testing of the randomized complete block design data in Table 3.

Now suppose that there are many of these RCBD's that comprise one data set and are all to be tested simultaneously via permutation testing. The FDR is to be reported and therefore, m_0 is to be estimated. An example of this is given in the Application Section. The next brief example explains another scenario where an analysis produces p-values with different discrete null distributions.

2.2 Next Generation Sequencing Data

Next generation sequencing is the newest technology for sequencing genomes of organisms. The technology is able to take a sample of DNA, which contains many small subunits of DNA , and for each subunit, determine to which gene along the genome the subunit belongs. This is relevant in the sense that one can take samples from two treatment populations (or two different genotypes) and compare the number of reads to gene A say, between the treatment groups to determine if one treatment population causes more activity in gene A than the other treatment group. Since the whole genome can be sequenced, each subunit of the DNA sample has to match to one of the thousands of genes in the sequence. The data set resulting from these next generation sequencing applications is a breakdown of the number of reads from a DNA sample that matched to each of thousands of genes. In a case with just one experimental unit from each treatment group, the entire data set is comprised of many 2 by 2 tables, one for each gene, where each table breaks down the number of reads to a particular gene and the number of reads to all other genes for each of the two treatment groups. So, for each of m genes, the data is a 2 by 2 table. A hypothetical example is given in Table 5.

	Genotype		
	1	2	Total
Read to Gene A	1	3	4
Not read to Gene A	3453080	3977732	7431812
Total	3453081	3977735	7431816

Table 5: Number of reads to Gene A from two DNA samples from two plants with different genotypes.

The null hypothesis of independence between the genotypes and the number of

(or lack thereof) reads can be tested using a chi square test when the estimated cell frequencies are large, e.g. greater than 5 or 10. When this is not case, as it is with the data in Table 5, the chi-square approximation to the distribution of the X^2 Pearson test statistic is not appropriate and an alternative test is Fisher's exact test which relies on exact small-sample distributions. The null hypothesis is equivalent to testing if the odds ratio is equal to 1 and the p-value is the sum of particular hypergeometric probabilities. Extreme values of n_{11}, the upper left entry in a 2 by 2 table, will make the sample odds ratio different from 1 and provide evidence against independence. Hence, Fisher's exact test defines the two-sided p-value as $\sum_{n_{11} \in S} \mathrm{P}\left[p(n_{11}) \leq p(t_0)\right]$, where $p(t_0)$ denotes the probability of the observed value of n_{11} and the possible values for n_{11} are given by the set $S = \{n_-, \ldots, n_+\}$ for $n_- = \max\{0, n_{1+} + n_{+1} - n\}$ and $n_+ = \min\{n_{1+}, n_{+1}\}$.

For the data in Table 5, $S = \{0, \ldots, 4\}$ and the corresponding hypogeometric probabilities are $\binom{3453081}{0}\binom{3977735}{4}/\binom{7431816}{4} = 0.082$, 0.285, 0.371, 0.215, and 0.047. Since the observed value of n_{11} is 1 and $\mathrm{P}(n_{11} = 1) = 0.285$, then the two sided p-value equals the sum of all the hypergeometric probabilities that are less than or equal to 0.285, i.e. p-value $= 0.082 + 0.285 + 0.215 + 0.047 = 0.629$. Table 6 displays the p-value support for the data given in Table 5.

n_{11}	0	1	2	3	4
p-value	0.129	0.629	1	0.344	0.047

Table 6: Range of possible values for n_{11} and their respective Fisher's Exact Test p-values.

Since the total reads across all genes for each genotype is fixed, the p-value support is simply a function of the possible values for n_{11} given by the set S. There are many genes where the number of reads to that gene between the two genotypes is the same,

and hence, will share the same null distribution. Once all genes are tested, the result will be a collection of m p-values from a mixture of different discrete distributions.

The next section describes how to estimate m_0 for each unique p-value null distribution and how to estimate FDR across all unique p-value null distributions.

3 Proposed Procedure

3.1 Estimating m_0

As mentioned in the Introduction, many methods exist to estimate m_0, the number of true null hypotheses out of m total hypotheses tested, but most are in the framework of parametric testing which results in p-values that are continuous U(0,1) under the null hypothesis. As shown in the above examples, it is possible to obtain a collection of m p-values that do not all follow the same discrete distribution, and further, these discrete distributions do not place equal mass on the elements of their support under the null hypothesis.

Suppose we tested m null hypotheses via permutation testing and obtained $p_1, \ldots, p_m$ each of which belong to one of n unique discrete null distributions, namely $F_1, \ldots, F_n$, with respective support sets $S_1, \ldots, S_n$. Let m_i denote the number of p-values that have null distribution F_i and let m_{0i} denote the number of p-values that correspond to a true null hypothesis with null distribution F_i. Given $\{\widehat{m}_{0i}\}_{i=1}^n$, estimate the experimentwise m_0 as $\widehat{m}_0 = \sum_{i=1}^n \widehat{m}_{0i}$. Therefore, we now focus how to estimate m_{0i} for an arbitrary discrete null distribution, F_i.

3.1.1 Estimating m_{0i}; Iterative Algorithm

Mosig et al. (2001) proposed a histogram based iterative algorithm that estimates the number of true null hypotheses given a collection of p-values all with a continuous U(0,1) null distribution. The basic idea of the iterative algorithm is the following. Given a histogram (with any bin size) of p-values that have a null distribution that is continuous uniform on (0,1), start by assuming that each p-value is truly distributed continuous uniform on (0,1), i.e. all the hypotheses are in fact true null hypotheses. Then, the expectation of the number of p-values in each bin can be calculated. Find the first bin where the observed frequency fails to exceed expectation and for each bin to the left of this bin, count the excess of observed frequencies to expected frequencies. The algorithm declares this total as an estimate of $m - m_0$, and therefore yields an estimate of m_0. Now, recalculate the expected number of p-values that should fall in each bin using the updated estimate of m_0 as the number of true null hypotheses. Again, find the first bin where the observed frequency fails to exceed the updated expectation and for each bin to the left of this bin, find the excess of observed frequencies to the updated expected frequencies. This total replaces the estimate of $m - m_0$, and therefore replaces the estimate of m_0. The algorithm continues in this fashion until convergence.

The objective is to use this algorithm to estimate m_{0i} for each unique p-value distribution $F_i,\ i = 1, \ldots, n$. Even though this algorithm is setup to handle p-values that are distributed uniformly on (0,1) under the null hypothesis, all that the algorithm requires is the null expected frequency for each bin. For arbitrary i, the support S_i is a discrete set and therefore we can allocate a bin for each element of S_i. Hence, to estimate m_{0i} for arbitrary i, gather all p-values with null distribution F_i, and construct a histogram allocating a bin for each element of S_i. Then use the histogram based

estimator to calculate $\widehat{m}_{0i}$. Repeat for all i to obtain $\{\widehat{m}_{0i}\}_{i=1}^{n}$ and calculate $\widehat{m}_0 = \sum_{i=1}^{n} \widehat{m}_{0i}$. Next, the formal definitions of the iterative algorithm are presented for the case of p-values that have any discrete distribution under the null hypothesis.

Again, we are testing m hypotheses based on $p_1, \ldots, p_m$ and each p_i has one of n unique discrete distributions under the null hypothesis. Denote the support of the i^{th} unique discrete distribution, F_i, as S_i and let S_{ij} be the j^{th} smallest element of S_i, where $j = 1, \ldots, n_i$ and where n_i is the number of elements in S_i, i.e. $n_i = |S_i|$. Also, let n_{ij} be the number of observed p-values that have null distribution F_i and are equal to S_{ij}. Let $s_{ij} = S_{ij} - S_{i,j-1}$ for $j = 2, \ldots, n_i$ and $s_{i1} = S_{i1}$, i.e. s_{ij} is the null probability that an arbitrary p-value with distribution F_i equals S_{ij}. Let

$$\tilde{n}_{i,j:n_i} = \frac{\sum_{l=j}^{n_i} n_{il}}{\sum_{l=j}^{n_i} s_{il}}. \tag{1}$$

Let $m_{0i}^{(0)} = \sum_{j=1}^{n_i} n_{ij} = m_i$ and define for any $j \in \{1, \ldots, n_i\}$,

$$m_{0i}^{(k)} = \left(\sum_{l=1}^{j_{ik}-1} s_{il} \right) \cdot m_{0i}^{(k-1)} + \left(1 - \sum_{l=1}^{j_{ik}-1} s_{il} \right) \tilde{n}_{i,j_{ik}:n_i} \text{ for all } k \geq 1, \tag{2}$$

where

$$j_{ik} \equiv \min \left\{ j : n_{ij} \leq s_{ij} \cdot m_{0i}^{(k-1)} \right\}. \tag{3}$$

One can think of j_{ik} as the index of the element of S_i where the frequency of observed p-values that equal S_{ij} does not exceed the null expectation given the estimated number

of p-values with distribution F_i that correspond to true null hypotheses at iteration k. Then, $m_{0i}^{(k)}$ is the estimated number of p-values with distribution F_i that correspond to null hypotheses at iteration k.

3.1.2 A Simple Example

Suppose we have, for some arbitrary i, $m_i = 30$ and $S_i = \{0.04, 0.20, 0.36, 0.52, 1\}$ with corresponding null probabilities $s_i = \{0.04, 0.16, 0.16, 0.16, 0.48\}$. Table 7 displays the observed p-value frequencies and the expected p-value frequencies given $m_{0i}^{(k)}$ for the first $k = 2$ iterations of the algorithm.

Iteration k	$m_{0i}^{(k)}$	p-value	.04	.20	.36	.52	1
0	30	Observed Frequency	4	8	6	5	7
1	22.6	Expected Frequency	1.2	4.8	4.8	4.8	14.4
2	18.752	Expected Frequency	0.904	3.616	3.616	3.616	10.848

Table 7: Observed frequencies and expected frequencies for two iterations of the histogram based estimator for estimating the number of true null hypotheses for a simple example.

Start by constructing a histogram (or table) allocating a bin for each element in S_i and assuming all $m_i = 30$ hypotheses are null, i.e. $m_{0i}^{(0)} = 30$. Given $m_{0i}^{(0)} = 30$ and s_i, find the expected frequency for each bin. Then, find the first bin where the observed frequency does not exceed the expected frequency, which is the bin corresponding to a p-value equal to 1. Next, for all bins to the left of this bin, add up the difference between the observed and expected frequencies. This excess is $(4 - 1.2) + (8 - 4.8) + (6 - 4.8) + (5 - 4.8) = 2.8 + 3.2 + 1.2 + 0.2 = 7.4 \Rightarrow m_{0i}^{(1)} = 30 - 7.4 = 22.6$. Executing the same steps, we estimate $m_{0i}^{(2)} = 30 - 11.248 = 18.752$. Continuing to iterate, we have $\widehat{m}_{0i} = \lim_{k\to\infty} m_{0i}^{(k)} = 14.583$.

In terms of the formal definitions for $k = 1$, $j_{i1} = 5$ since $7 = n_{i5} \leq s_{i5} \cdot m_{0i}^{(0)} =$

$0.48 \cdot 30 = 14.4$. Then, (2) yields

$$\begin{aligned}
m_{0i}^{(1)} &= \left(\sum_{l=1}^{5-1} s_{il}\right) m_{0i}^{(1-1)} + \left(1 - \sum_{l=1}^{5-1} s_{il}\right) \tilde{n}_{i,4:5} \\
&= \left(\sum_{l=1}^{4} s_{il}\right) m_{0i}^{(0)} + \left(1 - \sum_{l=1}^{4} s_{il}\right) \frac{\sum_{l=5}^{5} n_{il}}{\sum_{l=5}^{5} s_{il}} \\
&= (0.04 + 0.16 + 0.16 + 0.16) \cdot 30 + (1 - (0.04 + 0.16 + 0.16 + 0.16)) \frac{7}{0.48} \\
&= (0.52) \cdot 30 + (0.48) \frac{7}{0.48} \\
&= 15.6 + 7 \\
&= 22.6.
\end{aligned}$$

Continuing with the formal definitions yields $m_{0i}^{(2)} = 18.752$, $m_{0i}^{(3)} = 16.75104$, $m_{0i}^{(4)} = 15.71054$, $m_{0i}^{(5)} = 15.16948$, $m_{0i}^{(6)} = 14.88813$, $m_{0i}^{(7)} = 14.74183$, $m_{0i}^{(8)} = 14.66575$, $m_{0i}^{(9)} = 14.62619$, $m_{0i}^{(10)} = 14.60562$, etc.

Although the algorithm is efficient in terms of convergence, there is a limit to this algorithm characterized by Nettleton et al. (2006) for p-values that are continuous U(0,1) under the null hypothesis. The next section extends the limit characterization to handle p-values that have a discrete distribution where the elements in the support have unequal null probabilities.

3.1.3 Limit Characterization

Nettleton et al. (2006) showed the existence of and characterized the limit of the iterative algorithm when the p-values are continuous uniform on (0,1) under the null hypothesis. There is an analogous result to the iterative algorithm when the p-values have a discrete distribution where the elements in the support have unequal null prob-

abilities. Consider an arbitrary discrete distribution F_i with support S_i.

Convergence Result: Let $J_i = \min\{j : n_{ij} \leq \tilde{n}_{i,j:n_i}\}$. Then,

$$\widehat{m}_{0i} = \lim_{k\to\infty} m_{0i}^{(k)} = \frac{\sum_{l=J_i}^{n_i} n_{il}}{\sum_{l=J_i}^{n_i} s_{il}}.$$

The proof of this convergence result for continuous U(0,1) p-values under the null hypothesis can be found in Nettleton et al. (2006) and the proof of this convergence result for discrete non-uniform p-values under the null hypothesis can be found in Bancroft & Nettleton (2009).

Table 8 displays (1) for each j. The limit characterization says to find the first bin (or in this case, the first column) where the observed frequency does not exceed $\tilde{n}_{i,j:n_i}$ which is bin (column) 5, i.e $J_i = 5$. Then,

$$\begin{aligned}
\widehat{m}_{0i} &= \frac{\sum_{l=5}^{5} n_{il}}{\sum_{l=5}^{5} s_{il}} \\
&= \frac{7}{0.48} \\
&= 14.58333.
\end{aligned}$$

j	1	2	3	4	5
p-value (S_{ij})	.04	.20	.36	.52	1
H_0 probability (s_{ij})	.04	.16	.16	.16	.48
Observed Frequency	4	8	6	5	7
$\tilde{n}_{i,j:n_i}$	1.2	4.33	3.6	3	7

Table 8: The p-value support (S_i, i arbitrary), observed frequencies, and tail expectations ($\tilde{n}_{i,j:n_i}$) for a simple example.

Then, for each unique discrete p-value distribution F_i, find $\widehat{m}_{0i}$ via the limit characterization. Then, estimate m_0 by $\widehat{m}_0 = \sum_{i=1}^{n} \widehat{m}_{0i}$. This estimate is used to estimate

FDR which is described in the next section.

3.2 Estimating FDR

Suppose the m ordered p-values $p_{(1)}, \ldots, p_{(m)}$ are to be used to testing the corresponding m hypotheses $H_{(01)}, \ldots, H_{(0m)}$. To control the number of false discoveries made in a collection of discoveries at level α, Benjamini and Hochberg (1995) propose finding the largest integer k such that $\frac{p_{(k)}m}{k} \leq \alpha$. In words, FDR based on a significance cutoff $p_{(k)}$ is the expected number of type I errors divided by the total number of rejections. The numerator in their formulation actually over estimates the number of expected type I errors when m_0 is less than m. Therefore, the actual control of FDR for their procedure is at level $\frac{m_0}{m}\alpha$ and thus is unnecessarily conservative when $m_0 < m$. Replacing m with m_0 will reduce the conservative bias yielding control at the same level α while obtaining a larger list of discoveries.

Instead of specifying a level at which FDR is to be controlled and finding the corresponding p-value cutoff that gives the specified control, one could specify a p-value cutoff c, and find the corresponding FDR for that p-value cutoff. More explicitly, $\widehat{\text{FDR}}(c)$, the estimated false discovery rate for a p-value cutoff c, is

$$\widehat{\text{FDR}}(c) = \min\left\{\frac{p_{(k)}\widehat{m}_0}{k};\ \text{for all } p_{(k)} \geq c\right\}. \tag{4}$$

If (4) is calculated for each value of $c \in \left\{p_{(1)}, \ldots, p_{(m)}\right\}$, the resulting values of $\widehat{\text{FDR}}(c)$ are the q-values of Storey (2003).

As mentioned, the numerator in (4) estimates the expected number of type I errors. If $p_{(1)}, \ldots, p_{(m)}$ all have the same support and possess the property $\text{P}(p-\text{value} \leq p^*) = p^*$ for all p^* in the support under the null hypothesis, then the definition in (4) is valid.

If these conditions do not hold, then the definition in (4) is not valid. To see this, consider the following example.

Suppose we have $m_1 = 50$ p-values with support $S_1 = \{0.20, 1\}$ and we have $m_2 = 50$ p-values with support $S_2 = \{0.04, 0.20, 0.36, 0.52, 1\}$. Table 9 summarizes these supports, null probabilities, and hypothetical observed frequencies.

	S_1		S_2				
p-value	.20	1	.04	.20	.36	.52	1
H_0 probability	.20	.80	.04	.16	.16	.16	.48
Observed Frequency	15	35	4	9	9	8	20

Table 9: Two hypothetical discrete p-value null distributions and hypothetical observed frequencies.

First, using the limit characterization of the histogram based estimator, we have $\widehat{m}_{01} = 43.75$ and $\widehat{m}_{02} = 41.\bar{6}$ and therefore, out of the 100 hypotheses tested, we have an experimentwise estimate of m_0 as $\widehat{m}_0 = 43.75 + 41.\bar{6} = 85.41\bar{6}$. Next, if we ignored that these 100 p-values have two different supports, and treated them as all having support $S_2 = \{0.04, 0.20, 0.36, 0.52, 1\}$, we would estimate, for a p-value cutoff of 0.04, the number of expected type I errors to be $0.04 \cdot 85.41\bar{6} = 3.41\bar{6}$. In reality, the estimated number of expected type I errors made from p-values with support $S_1 = \{0.20, 1\}$ should be $0 \cdot 43.75 = 0$, since 0.20 is the minimum value in the support, and the estimated number of type I errors made from p-values with support $S_2 = \{0.04, 0.20, 0.36, 0.52, 1\}$ should be $0.04 \cdot 41.\bar{6} = 1.\bar{6}$. Hence, the estimate of the total number of type I errors should be $0 + 1.\bar{6} = 1.\bar{6}$, not $3.41\bar{6}$. For this hypothetical example, the estimated FDR would be unnecessarily conservative as we have over estimated the number of expected type I errors by not recognizing the distinction between the two distributions. This motivates that when a collection of m discrete p-values do not all have the same

distribution, the number of expected type I errors should be individually estimated for each unique distribution. Then, retaining the spirit of the definition of FDR, we can estimate FDR(c) across the unique distributions by adding up the estimated number of expected type I errors across the unique distributions and dividing by the total number of rejections. The next section formally describes the proposed procedure of estimating FDR with p-values of different discrete distributions.

3.3 Proposed Procedure

Suppose we have a collection of m p-values, each of which belongs to one of n unique discrete distributions. Denote these n unique distributions as $F_1, \ldots, F_n$ with respective supports $S_1, \ldots, S_n$ and let m_i be the number of p-values that belong to distribution F_i, $i = 1, \ldots, n$. Also, denote the j^{th} smallest element of S_i as S_{ij}. For each i, first calculate $\widehat{m}_{0i}$ via the limit characterization of the histogram based estimator and estimate m_0 as $\widehat{m}_0 = \sum_{i=1}^{n} \widehat{m}_{0i}$. Next, for each i, calculate for a given p-value cutoff c, $\widehat{V}_i(c) = S_{ij_i^*} \cdot m_i \cdot \widehat{\pi}_0$, where $j_i^* = \max\{j : S_{ij} \leq c\}$ and $\widehat{\pi}_0 = \frac{\widehat{m}_0}{m}$. This estimates the expected number of type I errors and note that since c may not be an element of S_i, we use the next smallest element of S_i which will eliminate the conservative bias that would occur if we use c instead of $S_{ij_i^*}$. Also, calculate $R_i(c) = \sum_{j=1}^{n_i} \mathrm{I}\left(p_{ij} \leq S_{ij_i^*}\right)$ where I($\cdot$)=1 if the argument is true and 0 otherwise. Then, the estimated false discovery rate associated with a p-value cutoff c, is defined as $\widehat{\mathrm{FDR}}(c) = \min\left\{ \frac{\sum_{i=1}^{n} \widehat{V}_i(c^*)}{\sum_{i=1}^{n} R_i(c^*)};\ c^* \in \left(\bigcup_i S_i\right) \cap [c, 1] \right\}$.

The next section applies the proposed procedure to a dataset from a Proteomics experiment.

4 Application-Proteomics

Counts of peptides matches to proteins in a database of thousands of proteins were measured on 5 different genotypes over 3 repetitions. Matches were made to 1176 proteins yielding 1176 3x5 matrices of peptide counts. Each of the 15 samples contains a different number of protein pieces and therefore, the counts from the same sample were normalized by the total number of matches for that sample. Table 10 displays the data for a few proteins.

		Genotype				
Protein ID	Repetition	A	B	C	D	E
18	1	0	0	0.0002	0	0.0003
	2	0	0	0	0	0.0003
	3	0	0	0	0	0.0002
39	1	0	0	0	0	0
	2	0.0003	0	0.0005	0	0
	3	0.0003	0.0009	0.0010	0	0
4	1	0.0034	0.0048	0.0042	0.0025	0.0056
	2	0.0049	0.0043	0.0037	0.0047	0.0040
	3	0.0020	0.0032	0.0030	0.0038	0.0028

Table 10: Subset of data (rounded) from a randomized complete block design measuring peptide counts on 1,176 different proteins on five genotypes over 3 repetitions. Data is normalized by dividing the counts for each sample by the sum of the all sample specific counts across proteins.

Note that with just three repetitions and along with the sparse data, these data do not lend itself well to typical parametric mixed model analysis. Therefore, permutation testing is employed to test, for each protein, whether all genotypes have the same abundance of each protein. To fulfill the lone requirement of permutation testing, exchangeability, the data must be permuted within each repetition. As described in Section 2, repetitions with responses equal across genotypes can stay fixed, and one of the remaining repetitions can also stay fixed without changing the proportion of

test statistics as extreme or more extreme than the observed test statistic. The test statistic chosen is the variance of the genotype means, which is rank equivalent to the traditional F-statistic. For protein ID 18, calculating the test statistic for each of the $5! \cdot 5! = 14,400$ permutations (some of which are redundant since only four of the 15 samples matched to protein ID 18 resulting in many 0 normalized counts) results in only 10 unique test statistics which yields a discrete p-value support of $S_{18} = \{0.04, 0.08, 0.12, 0.16, 0.28, 0.40, 0.52, 0.64, 0.76, 1.00\}$. Under the null hypothesis, some of these 10 unique tests statistics occur more often than others, which leads to the unequally spaced elements of S_{18}. To see this, let's look closer at how a p-value of 1 is obtained.

A p-value of 1 corresponds to the smallest possible test statistic. Here, that occurs when the variance of the genotype means is smallest and is when there is only one non-zero observations per column. Again fixing repetition one, this occurs when observation 0.0003 from repetition 2 and observation 0.0002 from repetition 3 are, respectively, in columns corresponding to either genotypes A&B, A&D, B&A, B&D, D&A, or D&B. Under the null hypotheses, each one of these configurations is equally likely. As described in Section 2, since there are only two non-zero observations (for those repetitions that we are permuting), there are only 25 unique configurations of observation 0.0003 from repetition 2 and observation 0.0002 from repetition 3 to the five columns. Hence, a p-value of 1 occurs with null probability $\frac{6}{25}$.

For the actual arrangement for protein 18, the test statistic is the largest possible and therefore, the p-value is $\frac{1}{25} = 0.04$. For protein ID 4, there are no non-zero observations, and each observation is unique. Therefore, there are $5! \cdot 5! = 14,400$ unique test statistics, each occurring with equal probability of $\frac{1}{14,400}$. There were 140

total proteins with a p-value support of 14,400 elements. On the other extreme, there were 94 proteins with a p-value support of just two elements, namely $\{0.20, 1\}$. Also of note, there were 586 proteins with a p-value support of just one element, namely 1. These proteins were ignored in the computation of $\widehat{m}_0$ and estimating FDR as they contained no information. Table 11 displays the breakdown of number of proteins versus number of normalized counts that are 0.

Number of 0 normalized counts	0	1	2	3	4	5	6	7
# proteins	76	45	21	21	14	23	18	37
Number of 0 normalized counts	8	9	10	11	12	13	14	15
# proteins	32	37	38	53	86	139	536	0

Table 11: Breakdown of number of proteins versus number of normalized counts that are 0.

Table 11 shows that there were 536 proteins with 14 of the 15 normalized counts equal to zero, suggesting that each of these proteins will have a p-value singleton support set of $\{1\}$. But, in the previous paragraph, it was stated that there were 586 proteins with a p-value support of one element. The discrepancy is due to the fact that if, for any protein, all nonzero normalized counts are found in only repetition, then the actual number of nonzero normalized counts doesn't influence the p-value support. The p-value distribution in this case will always have a support of just one element, namely 1. Table 11 does not summarize the number of unique p-value null distributions, but rather shows a general structure of the data at hand.

There were 126 unique p-value null distributions (excluding the p-value null distribution with the singleton support $\{1\}$) and the estimated number of true null hypotheses, $\widehat{m}_0$, for each unique p-value distribution was found. For the subset of the 590 informative proteins, $\widehat{\pi}_0 = \frac{\sum_{i=1}^{n} \widehat{m}_{0i}}{m} = \frac{528.6259}{590} = 0.896$ and the estimated FDR for significance cutoffs of 0.001, 0.005, 0.01, and 0.05 are given in Table 12.

c	0.001	0.005	0.01	0.05
$\sum_{i=1}^{n} \widehat{V}_i(c)$	0.1911	1.1792	2.7994	18.5099
$\sum_{i=1}^{n} R_i(c)$	2	5	12	67
$\widehat{\text{FDR}}(c)$	0.0956	0.2358	0.2333	0.2763

Table 12: Estimated number of type I errors and number of rejections across the subset of 590 proteins and the estimated false discovery rate (FDR) for four different significance thresholds.

5 Simulation Study

This section investigates the performance of the proposed procedure via a simulation study that attempts to produce data similar to the data described in the previous section. In words, data for m proteins were simulated, of which, $\pi_0 \cdot m$ ($\pi_0 \in (0, 1)$) were generated under the null hypothesis of equal protein abundance across genotypes and $(1 - \pi_0) \cdot m$ were generated under the alternative hypothesis of not equal protein abundance across genotypes. Once the data were simulated, permutation testing was employed to test the m null hypotheses which resulted in a collection of m permutation p-values that do not all have the same distribution under the null hypothesis. The proposed procedure was employed to estimate the number of true null hypotheses and to estimate the false discovery rate for prespecified significance cutoffs. The estimated number of true null hypotheses was compared to $m_0 = \pi_0 \cdot m$ and the estimated FDR was compared to the false positive fraction. The false positive fraction is the true false discovery rate and is define as $V/(R \wedge 1)$, where V is the true number of type I errors and $R \wedge 1$ is the maximum of total number of rejections and 1. Next, the data simulation is described.

To produce data that are similar to the data in the Application Section, a data-based simulation scheme was employed. When simulating data under the null hypothesis, one of the 61 proteins discussed in Section 4 with a p-value larger than 0.70 was randomly

chosen. The five genotype specific means and standard deviations of the non-zero observations were calculated. One of the five genotype specific means was randomly selected and was used as the mean for each genotype with genotype specific standard deviations remaining the same. These values were used as parameters to simulate data for one protein under the null hypothesis. When simulating data under the alternative hypothesis, one of the 12 proteins discussed in Section 4 with a p-value less than 0.01 was randomly chosen. The five genotype specific means and standard deviations of the non-zero observations were calculated. These values were then used as parameters to simulate data for one protein under the alternative hypothesis. Also, since the data contained many 0 observations, the minimum observation for the randomly chosen protein was chosen as a threshold, that when simulating data, if an observation was smaller than this threshold, that observation became 0. The data were simulated from normal distributions.

Now, when this simulation scheme is employed, the estimates of m_0 are very conservative, e.g. $0.95m$ versus a truth of $0.80m$. Presumably, there is little power to detect the differences in the simulated data between genotypes when the parameters chosen were based on the data of proteins discussed in Section 4 with p-values less than or equal to 0.01. To increase the power when simulating data under the alternative hypothesis, the mean parameters for the five genotypes were simulated from a normal distribution with a mean equal to the average of the five genotype specific means and a standard deviation equal to a multiple, k, larger than the standard deviation of the five genotype specific means. For example, Protein ID 981 had a p-value of 0.00083. Let's say this was randomly chosen to simulate data for a protein under the alternative hypothesis. First, the five genotype specific means of the non-zero observations were

0, 0, 0.00091, 0.00026, and 0.00187, respectively, and the five genotype specific standard deviations of the non-zero observations were 0, 0, 0.00046, 0.00029, and 0.00055, respectively. Next, the average and standard deviation of the five genotype specific means was 0.00061 and 0.00080 respectively. Then, five mean parameters were simulated from N(0.00061, k·0.00080), say $\tilde{\mu}_1, \ldots, \tilde{\mu}_5$. Last, the data for the five genotypes were respectively simulated from N($\tilde{\mu}_1$,0), N($\tilde{\mu}_2$,0), N($\tilde{\mu}_3$,0.00046), N($\tilde{\mu}_4$,0.00029), and N($\tilde{\mu}_5$,0.00055) with appropriate censoring. Three observations were simulated for each genotype.

Once data were simulated for $m = 100$ proteins, permutation testing was employed to test whether the abundance of each protein was the same in all 5 genotypes and the proposed procedure was employed to estimate m_0 and FDR(c), for $c = 0.01, \ldots, 0.05$. The results are summarized below in Table 13 for $N = 100$ simulation runs with a $\pi_0 = 0.80$ and $k = 2$.

	$\widehat{\text{FDR}}(c)$		$V(c)/R(c)$	
c	mean	std. err.	mean	std. err.
0.01	0.0383	0.00017	0.0330	0.00350
0.02	0.0726	0.00044	0.0603	0.00508
0.03	0.1083	0.00077	0.0942	0.00602
0.04	0.1423	0.00113	0.1241	0.00671
0.05	0.1710	0.00142	0.1557	0.00675

Table 13: Average $\widehat{\text{FDR}}(c)$ and $V/(R \wedge 1)$ over $N = 100$ simulation runs where each run consisted of simulating data for $m = 100$ proteins, of which, 80 proteins were simulated under the null hypothesis of equal protein abundance across genotypes.

The mean $\widehat{m}_0$ over $N = 100$ simulation runs was 81.43 with a standard error of 0.1157 which compares very well to the truth of 0.80. Also, the estimated FDR(c) and $V/(R \wedge 1)$ compare very favorably. The estimated FDR(c) is, on average, slightly larger than the true false discovery rate. Since a slight conservative bias is preferred to

a liberal bias, this simulation study suggests that the proposed procedure does in fact control FDR.

6 Summary

This paper proposes how to estimate the number of true null hypothesis, m_0, and the false discovery rate given a collection of m p-values from a mixture of discrete distributions. Further, each discrete null distribution does not have an evenly spaced support resulting in an unequal null probability of observing each element in the support. A simulation study has shown that the proposed procedure does control the false discovery rate and estimates m_0, both with a slight conservative bias.

7 References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, Series B, **57**, 289-300.

[2] Benjamini, Y. and Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics*, **25**, 6083.

[3] Mosig, M.O., Lipkin, E., Galina, K., Tchourzyna, E., Soller, M., and Friedmann, A. (2001). A Whole Genome Scan for Quantitative Trait Loci Affecting Milk Protein Percentage in Israeli-Holstein Cattle by Means of Selective Milk DNA Pooling in a Daughter Design Using an Adjusted False Discovery Rate Criterion. *Genetics*, **151**, 1683-1698.

[4] Nettleton, D., Hwang, J.T.G., Caldo, R.A., and Wise, R.P. (2006). Estimating the Number of True Null Hypotheses From a Histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics*, **Vol. 11, No. 3**, 337-356.

[5] Storey, J.D. (2000). False Discovery Rates: Theory and Applications to DNA Microarrays. Unpublished Ph.D. thesis, Department of Statistics, Stanford University.

[6] Storey, J.D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society*, Series B, **64**, 479-498.