Investigations into refinements of Storey's method of multiple hypothesis testing minimising the FDR, and its application to test binomial data

Author: John H. Nixon
Institution: Agriculture and Agri-Food Canada, Saskatoon Research Centre,107 Science Place, Saskatoon, SK S7N 0X2,Canada

This is a preprint of the article published in the Journal of Computational Statistics and Data Analysis with the following link
http://dx.doi.org/10.1016/j.csda.2012.03.026

Abstract

Storey's method for multiple hypothesis testing "the Optimal Discovery Procedure" (ODP) minimising the false discovery rate (FDR) and giving p-values and q-values (estimates of FDR) for each test, was extended by iteration to enforce consistency between the p-values of the tests and the binary parameters defining which data points contribute to the fitted null hypothesis. These parameters arise when the null hypothesis has to be estimated from the data. The ODP as previously described, is only optimal for fixed values of these parameters. The extension proposed here requires the introduction of a cut-off parameter for the p-values. Motivated by using this method to analyse a set of pairs of frequencies representing gene expression for a set of genes in two libraries, from which it was desired to select those that are most likely to be not following the null hypothesis that the frequency ratio is a fixed unknown number, this method was tested by analysing many similar simulated data sets. The results showed that the ODP modified by iteration could be improved sometimes greatly by a suitable choice of the cut-off parameter, but varying this parameter alone may not lead to the globally optimal solution because statistical testing based on the Binomial distribution is more efficient than using a form of the ODP when the number of non-null hypotheses in the data is small, but the reverse is true when it is large. This may be an effect of using discrete data. Efficiency here is defined in terms of the expected proportion of errors that occur (q-value) when a given proportion of the data is declared "significant" (i.e. the null hypothesis is believed not to hold for them). An improved version of the ODP along these lines is likely to have numerous applications such as in the optimised search for candidate genes that show unusual expression patterns for example when more than two experimental conditions are simultaneously compared and to cases when additional categorical variables or a time series is present in the experimental design.

1 Introduction

In a fundamental advance in the theory of multiple hypothesis testing, Storey [2007] has proposed a general procedure, the Optimal Discovery Procedure (ODP) which, given results for a large number of similar experiments, ranks them in order of significance from the most to the least significant. Storey and Tibshirani [2003] also provide the formula for the calculation of the expected proportion of errors or the False Discovery Rate (FDR) introduced by Benjamini and Hochberg [1995]. This is also known as the q-value and is calculated at each cut-off point determining the tests that are regarded as significant. Storey's method is optimal in the sense that for each cut-off point, the FDR is as small as possible conditional on fixed values for the set of parameters w_i defining the status of each hypothesis. The mathematical proof and background to this work has been described by Storey [2007] and its main application to date was described by Storey Dai and Leek [2005] which is the analysis of gene expression data from microarray experiments [Leung and Cavalieri , 2003].

The quantities w_i [Storey , 2007, Equation 9] arise whenever the null hypothesis contains parameters that must be estimated. They determine which terms are included in the denominator of Storey [2007, Equation 8] and are required for input to Storey's ODP. One way to break the circularity of this is to determine the w_i by a "conventional statistic" , p-value [Storey , 2007], and p-value cut-off λ.

Alternatively a modified ODP could be implemented as an iterated procedure for calculating the p-values of all the tests, with initial estimates either taken from a "conventional statistic" or from the assumption that each hypothesis is null. In one computation cycle, the estimates of the status of each test should be determined by the p-values calculated in the previous cycle using the p-value cut-off. This should be iterated until convergence, i.e. until either no changes or a minimal number of changes occur in the status of any of the hypothesis tests between successive cycles.

It is therefore of interest to compare the performance of different modifications of the ODP namely

1. After one cycle of this procedure when the initial estimate of the status of each hypothesis is 'null hypothesis'
2. After convergence of this procedure with different values of the cut-off parameter λ.
3. Repeating the above with different initial estimates of the status of the hypotheses and/or varying the size of the data set but using the same probability models.

It is also interesting to compare these results with the result of testing the hypotheses with a statistical test not involving the ODP, a "conventional statistic", and the ideal case when all the alternative hypotheses are listed before any of the null hypotheses.

These investigations were initially motivated by trying to use the ODP to find the optimal solution of a multiple hypothesis testing problem arising from a gene expression study. In Section 2 the experiment involving the plant pathogen *Albugo candida* is briefly described (data provided by Dr. H. Borhan at AAFC).

In Section 3 an derivation of the analysis procedure for continuous data is given. This is a modification of Storey's (2007) method of multiple hypothesis testing allowing for iterative refinement. In Section 4 the modification of this method appropriate to the discrete Binomial data set arising from the gene expression study is given including a description of some practical difficulties and proposed solutions. Section 5 summarises the algorithm showing how it can be efficiently implemented as a computer program. The results are subdivided into two parts in sections 6 and 7 concerned with the data analysis and the testing of the method itself using simulated data. Finally some conclusions are given in section 8.

2 Experimental Details

One week old seedlings of the susceptible *Brassica juncea* cultivar “cutlass” were inoculated with the plant pathogenic oomycete *Albugo candida* (race Ac2V) that causes white rust in Brassicaceae. A cDNA library was made from infected cotyledons collected at 10 days after inoculation. Approximately 36000 cDNA clones (average insert size of 800 base pairs) were sequenced from both directions. After trimming with LUCY [Chou and Holmes , 2001] and a quality assessment 50,248 Expressed Sequence Tags (ESTs) were subjected to clustering analysis with TGICL [Pertea *et al.* , 2003] and warehoused using APED (http://sourceforge.net/projects/aped). To differentiate between Ac2V and *B. juncea* cutlass sequences blastn and blastx searches with BLAST [Altchul *et al.* , 1997] were carried out against a comprehensive set of plant databases (unpublished). 14,510 ESTs were identified as putatively pathogen derived during infection of *B. juncea.* We also sequenced clones from an Ac2V spore library (27,547 ESTs). Spores were collected from *B. juncea* cultivar cutlass infected cotyledons at 10-14 days after inoculation.

For each putative distinct transcript, n_1 and n_2 are respectively the number of cloned cDNAs associated with it in the Ac2V spore library and the *Albugo candida* infected *B. juncea* library respectively. The purpose of this analysis is to identify genes that have an abnormal ratio that could indicate that the expression of these genes was significantly different between the spore library and the library from the active infection of *B. juncea.*

This experiment gave rise to the dataset D that consists of $m = 2171$ pairs of frequencies (n_1, n_2) which was summarised (Table 1) as a table of 119 rows, each giving n_1, n_2, and the frequency of this combination in D.

3 A Derivation of the Analysis Procedure for Continuous Data

Suppose there are a set of m hypotheses to be tested with data of the same type, and the ith hypothesis has data x_i associated with it for $1 \leq i \leq m$, where each x_i is a vector of real numbers that are experimental observations or values derived from direct observations. The data D is the set x_i, for $1 \leq i \leq m$.

Suppose that the null (H_0) and alternative (H_1) hypotheses for x are represented by the continuous distributions with densities $f(x)$ and $g(x)$ respectively that may be subject to some other conditions depending on the example, e.g. based on basic probability theory. The unknown parameter π_0 is the prior (before measuring x) probability of the null hypothesis, and $1 - \pi_0$ is the prior probability of the alternative hypothesis. In this model x can have members that are correlated with each other so that some members of x can, when fixed, alter the conditional distribution of other members of x. This allows for covariates or ‘nuisance’ variables whose effect it is desired to eliminate, as in the analysis of covariance. All such variables should be included in x or derivable from x. So for example $\phi(x)$ could be some marginal totals or parameters describing the experiment such as its size. Then the total density can be written as

$$t(x) = \pi_0 f(x) + (1 - \pi_0) g(x). \tag{1}$$

n_1	n_2	freq.	n_1	n_2	freq.	n_1	n_2	freq.
1	1	563	2	5	17	4	1	46
1	2	427	2	6	9	4	2	18
1	3	101	2	7	8	4	3	8
1	4	94	2	8	4	4	4	14
1	5	41	2	9	3	4	5	2
1	6	27	2	10	1	4	6	5
1	7	20	2	12	2	4	7	2
1	8	11	2	13	2	4	8	3
1	9	4	2	15	2	4	9	1
1	10	3	2	17	2	4	12	2
1	11	6	2	18	1	4	13	1
1	12	4	2	21	2	4	14	1
1	13	3	2	22	1	4	16	2
1	14	4	2	25	1	4	18	1
1	15	5	2	27	1	4	19	1
1	16	2	2	28	1	4	87	2
1	17	1	2	31	1	5	1	16
1	18	2	2	36	1	5	2	13
1	19	1	2	44	1	5	3	3
1	20	4	2	65	1	5	4	2
1	21	2	2	87	1	5	5	2
1	23	1	2	109	1	5	6	2
1	27	1	3	1	75	5	7	3
1	28	1	3	2	52	5	9	1
1	34	1	3	3	18	5	13	1
1	36	1	3	4	13	6	1	5
1	38	1	3	5	6	6	2	2
1	40	1	3	6	8	6	3	1
1	41	2	3	7	5	6	8	3
1	50	2	3	9	3	6	19	1
1	65	1	3	10	4	7	1	2
1	86	1	3	11	2	7	2	2
1	88	1	3	12	2	7	8	1
1	99	1	3	13	2	7	14	1
1	111	1	3	17	1	8	7	1
1	276	1	3	18	2	8	15	1
2	1	194	3	19	1	9	4	1
2	2	118	3	20	1	10	1	1
2	3	45	3	24	1	24	26	1
2	4	37	3	87	1			

Table 1: Summary of the data set analysed. The data has been summarised giving the frequency of each occurrence of a pair (n_1, n_2) in the data, and is presented in 3 column format.

The left hand side (LHS) of this equation can be estimated from the data, and if $f(.)$ and $g(.)$ were known, π_0 could be estimated from the data and could take any value satisfying $0 \leq \pi_0 \leq 1$.

Because the prior conditions are the same for every test, what is wanted is a uniform test procedure i.e. one that is applied using the same formula for every test, only the data differ. In general such a procedure can be described as follows where the level of the test is α (the probability of the test rejecting H_0 when it is true) and the Γ_α are a set of nested (if $\alpha < \alpha'$ then $\Gamma_\alpha \subset \Gamma_{\alpha'}$) rejection regions in the space of all possible x values: if $x_i \in \Gamma_\alpha$ then choose the alternative hypothesis H_1 and if $x_i \notin \Gamma_\alpha$ then choose the null hypothesis H_0. The regions Γ_α are defined by a test statistic determined by the following argument. According to the Neymann Pearson lemma stated below [Storey , 2007], the test for H_0 versus H_1 should be done as follows:

$$\text{choose } H_1 \text{ if } \frac{g(x)}{f(x)} > k(\alpha), \text{ and choose } H_0 \text{ if } \frac{g(x)}{f(x)} < k(\alpha)$$

assuming that $\frac{g(x)}{f(x)} = k(\alpha)$ has probability zero. Then, because of the definition of α, the function $k(.)$ is determined by the condition

$$\alpha = \int_{x':g(x') \geq k(\alpha) f(x')} f(x')dx' = P_{x'}\left(\frac{g(x')}{f(x')} \geq k(\alpha)\right) \tag{2}$$

where P_x denotes probability with x chosen according to the null hypothesis. Clearly $k(\alpha)$ must be decreasing as α is increasing. The p-value resulting from the test using data x, $p(x)$, is defined to be the borderline value of α i.e. the smallest value of α for which the test is just significant, therefore $\frac{g(x)}{f(x)} = k\left(p(x)\right)$. Hence from Equation (2),

$$p(x) = P_{x'}\left(\frac{g(x')}{f(x')} \geq \frac{g(x)}{f(x)}\right). \tag{3}$$

This choice of test is a consequence of the Neymann Pearson lemma which states that tests of this form maximises the power ($E[S]/m_1$) i.e. the probability of choosing H_1 when it is true, for a given level of Type 1 error ($\alpha = E[V]/m_0$), i.e. the probability of choosing H_1 when H_0 is true. Precisely stated, this optimality condition is that for any other region Γ' such that $E[V](\Gamma') \leq E[V](\Gamma_\alpha)$, it follows that $E[S](\Gamma') \leq E[S](\Gamma_\alpha)$, in other words the only way that the expected proportion of true positives $E[S](\Gamma')$can exceed $E[S](\Gamma_\alpha)$ is for the expected proportion of false positives $E[V](\Gamma')$ to exceed $E[V](\Gamma_\alpha)$. Here S and V are defined as the frequencies of two of the four possible outcomes for each test for one instance of this procedure, as in Table 2, and the expectation is over many instances of the dataset coming from the stated probability model. Frequently the following terminology is used in the literature: U, V, T, and S are respectively the frequencies of true negatives, false positives, false negatives, true positives, and so $E[V]$ and $E[S]$ are often called expected {false respectively true} positives and abbreviated to EFP and ETP respectively.

In most practical situations, $f(.)$ and $g(.)$ will not be known and so will have to be estimated from the data using their distributional assumptions, along with π_0.

An estimate $\hat{t}(.)$ of $t(.)$ can be made by maximum likelihood using the whole data set D and the distributional constraints on $\hat{t}(.)$ if any. The estimate $\hat{f}(.)$ of $f(.)$, can be made likewise, subject

	Test for null hypothesis		
True situation	Accept	Reject	Total
Null	U	V	m_0
Alternative	T	S	m_1
Total	W	R	m

Table 2: General notation for frequencies of different outcomes of the multiple testing procedure

to its appropriate distributional assumptions (the null hypothesis). Then the question arises as to what subset of the data should be used to get the estimate $\hat{f}(.)$. Either the whole of the data D can be used which has the advantage of simplicity, or a subset of the data D_0 believed to come from the null hypothesis as determined by a simple statistical test can be used. This option is motivated by improving the estimate $\hat{f}(.)$. If the data point x_i is included with weight w_i , where all the weights are 0 or 1, then $D_0 = \{x_i \in D | w_i = 1\}$ where $w_i = \left\{ \begin{array}{l} 0 \text{ if } p(x_i) \leq \lambda \\ 1 \text{ if } p(x_i) > \lambda \end{array} \right\}$ and the function $p(x_i)$ is the p-value for the test with data x_i in the previous cycle of the iterative procedure.

There is a paradoxical property of this statistical method arising from the fact that it gives a statistical test to be applied to each of a set of m sets of data x_i simultaneously, but for each test, the procedure uses information in the entire dataset (because it uses the estimates $\hat{f}(.)$ and $\hat{t}(.)$ that involve the whole dataset or more precisely D_0 and D respectively). This is the reason that the testing procedure can "borrow strength" across tests. The paradox is that the result of an individual test can depend on the dataset it is in. But this is necessary in order that the procedure can pick out repeating patterns and call them significant whereas a single instance of an unusual pattern might be regarded as a random event. This is clearly in accord with intuition, and tests lacking the ability to "borrow strength" across tests would not be able to accomplish this.

From Equation (1) the relationship $g(x) = \frac{t(x) - f(x)\pi_0}{1-\pi_0}$ follows, therefore the statistic for doing the significance test, the estimate of $\frac{g(x)}{f(x)}$, is equivalent to using a statistic which is an estimate of $l(x) = \frac{t(x)}{f(x)}$ because these quantities are related linearly with constant coefficients i.e.

$$\frac{g(x)}{f(x)} = \frac{1}{1-\pi_0} \left(l(x) - \pi_0 \right). \tag{4}$$

Because $t(x)$ is estimated by $\hat{t}(x)$, $l(x)$ is estimated by

$$\hat{l}(x) = \frac{\hat{t}(x)}{\hat{f}(x)}. \tag{5}$$

This test statistic is closely related to the posterior probability of the data x coming from the null hypothesis. From Bayes' theorem $P(H_0|X = x) = P(H_0)P(X = x|H_0)/P(X = x)$ where from above $P(H_0) = \pi_0$, the probability that X is within dx of x under the hypothesis H_0 is $f(x)dx$, and the probability that X is within dx of x is $t(x)dx$, therefore $P(H_0|X = x)$ can be estimated by $\hat{\pi}_0 \hat{f}(x)/\hat{t}(x) = \hat{\pi}_0/\hat{l}(x)$. This statistic is sometimes referred to as the posterior error probability (PEP) [Käll, Storey and Noble , 2008, 2009].

The p-values for all the tests can be obtained using Equation (3), which can be written as

$$p(x) = \int_{x':\hat{l}(x') \geq \hat{l}(x)} \hat{f}(x')dx' \tag{6}$$

where x is the data for the corresponding test, and the condition $\frac{g(x')}{f(x')} \geq \frac{g(x)}{f(x)}$ can be written as $l(x') \geq l(x)$ because the coefficient in (4) is $1/(1-\pi_0) > 0$, and f and l are replaced by their estimates. It follows from Equations (3) and (4) that the larger the test statistic $\hat{l}(x)$ is the more significant the data x is.

4 The Discrete Binomial Data Model

4.1 General Considerations

The model above is a very general model that includes a limiting case of the more complex discrete model defined in this section. The point of introducing the continuous data model was to develop the analysis procedure based on Storey's ideas for a model that approximates a description of the data set to be analysed. This implies that the results of the analysis will only strictly apply to a limiting case of the discrete Binomial model describing the data. This discrete Binomial model will now be described.

In this model each of m hypothesis tests has pair of non-negative integer values (n_1, n_2) so that the whole dataset is $D = \{(n_{1i}, n_{2i})$ for $1 \leq i \leq m\}$. D is a set with multiplicity i.e. the elements can occur multiple times. The null hypothesis is that for each data point (n_1, n_2), a pair of alternative outcomes with probabilities r and $1-r$ have frequencies n_1 and n_2 respectively, and that r is constant for the whole dataset D. The function $\phi(.)$ is given by $\phi(x) = n_1 + n_2$ where $x = (n_1, n_2)$. It is convenient to introduce the notation $j(\phi) = \{\#(n_{1i}, n_{2i}) \in D : n_{1i} + n_{2i} = \phi\}$ for the observed distribution of ϕ in D so that the total number of observations is $m = \sum_{\phi \geq 1} j(\phi)$, and $h(x)$ is the observed distribution of x in D given by $h(n_1, n_2) = \sum_{i=1}^{m} \delta_{n_1 n_{1i}} \delta_{n_2 n_{2i}}$, which is equal to the number of points $(n_{1i}, n_{2i}) \in D$ such that $n_1 = n_{1i}$ and $n_2 = n_{2i}$. Therefore $j(\phi) = \sum_{n_1=0}^{\phi} h(n_1, \phi - n_1)$. For the null hypothesis, because ϕ is uniquely determined by n_1 and n_2,

$$f(x) = P(n_1, n_2) = P(n_1, n_2, \phi) = P(n_1, n_2|\phi)P(\phi) \tag{7}$$

where from the Binomial distribution,

$$P(n_1, n_2|\phi) = \text{Bin}(r, n_1, \phi) = r^{n_1}(1-r)^{\phi - n_1} \begin{pmatrix} \phi \\ n_1 \end{pmatrix}. \tag{8}$$

The test statistics and the p-values are derived for the Binomial model in a manner analogous to the derivation for the continuous model defined above, and this results in Equations (25) and (26) respectively in the Appendix. In these equations the quantities j_0, h_0, m_0 are defined in same way as the corresponding quantities without the subscript zero except that D is replaced by D_0, that is the subset of D believed to come from the null hypothesis. For comparison, the p-values could also be calculated from

$$p_{0\mathrm{Bin}} = \mathrm{Bin}(r, n_1, \phi) + 2.\min\left(\sum_{i=0}^{n_1-1} \mathrm{Bin}(r, i, \phi), \sum_{i=n_1+1}^{\phi} \mathrm{Bin}(r, i, \phi)\right), \tag{9}$$

which is a two tailed test procedure based on the Binomial distribution as null hypothesis and the global alternative hypothesis. Note that the comparison between these two test procedures is valid because both H_0 and H_1 are the same for each test.

4.2 Practical Difficulties with Formulating the Testing Procedure

While formulating the statistical test procedure two problems had to be overcome. The first problem is how to deal with the situation in which some subset of the data had been removed prior to the analysis, which was data with $n_1 = 0$ or $n_2 = 0$. A solution to this is also described in the Appendix and involves making an approximation for the test statistic and the p-value using a modified null hypothesis which significantly affects only a small subset of the tests (those with small values of ϕ). This leads to replacing the p-value p_0 in (9) by the p-value

$$p_{\mathrm{Bin}} = \frac{\mathrm{Bin}(r, n_1, \phi) + 2.\min\left(\sum_{i=1}^{n_1-1} \mathrm{Bin}(r, i, \phi), \sum_{i=n_1+1}^{\phi-1} \mathrm{Bin}(r, i, \phi)\right)}{1 - (1-r)^{\phi} - r^{\phi}} \tag{10}$$

and replacing Equations (25) and (26) by Equations (29) and (30) respectively (see the appendix for details). Equation (10) is the two-sided test for the data (n_1, n_2) using the truncated Binomial null hypothesis.

The second problem is how to evaluate Equation (30) for the p-values considering that it has an infinite number of terms. To solve this, a simulation technique for the null hypothesis based on the fitted values of r and $j_0(.)$ was used. The value $\hat{r}$ (Equation (23)) can be used directly but using $j_0(.)$ directly in the null hypothesis is problematic because it is clearly over-fitted to the data i.e. $j_0(\phi)$ has a lot of random variation because it can be considered to result from a small sample from a data set with a hypothetical smoother distribution which represents the null hypothesis. Below $j_{0c}(\phi)$ is defined as a smoothed [Simonoff , 1996] version of $j_0(\phi)$ to replace it in the null hypothesis. The extent to which these data are smoothed in the definition of the test statistic is of course related to degree to which the method "borrows strength" across tests.

Because of the discussion in the previous paragraph, the sum in (30) was evaluated as an average over N simulated datasets of the proportion of times the test statistic in Equation (13) in the simulated dataset is greater than or equal to the test statistic for a fixed point (n_1, n_2) also calculated from Equation (13).

Each simulated dataset was based on the subset D_0 of data believed to come from the null hypothesis. D_0 has, after arranging its values of ϕ in increasing order, say $\phi = \phi_i$ with frequency q_i for $1 \leq i \leq A$. $F_{0c}(\phi)$ is the cumulative probability distribution representing the null hypothesis for ϕ

where c is an integer parameter. $F_{0c}(\phi)$ was chosen to be constant on each of the intervals between the following points and including these points as lower endpoints of the corresponding interval: $(\phi_{\min}, 0)$, $(\phi_1, q_1/m_0)$, $(\phi_2, (q_1+q_2)/m_0)$, $(\phi_3, (q_1+q_2+q_3)/m_0)$, ...$\left(\phi_c, \frac{1}{m_0}\sum_{j=1}^{c} q_j\right)$ e.g.$F_{0c}(\phi) = 0$ for $\phi_{\min} \leq x \leq \phi_1$, $F_{0c}(\phi) = q_1/m_0$ for $\phi_1 \leq x < \phi_2$ etc., and then piecewise linear joining the points $\left(\phi_c, \frac{1}{m_0}\sum_{j=1}^{c} q_j\right)$, $\left(\phi_{c+1}, \frac{1}{m_0}\left(\frac{1}{2}q_{c+1} + \sum_{j=1}^{c} q_j\right)\right)$, $\left(\phi_{c+2}, \frac{1}{m_0}\left(\frac{1}{2}q_{c+2} + \sum_{j=1}^{c+1} q_j\right)\right)$, ... $\left(\phi_A, \frac{1}{m_0}\left(\frac{1}{2}q_A + \sum_{j=1}^{A-1} q_j\right)\right)$ and $(\phi_{\max}, 1)$ where $m_0 = \sum_{j=1}^{A} q_j$, and $\phi_{\min}$ and $\phi_{\max}$ were chosen as 1 and the largest value of ϕ in the set D respectively. $\phi_{\max}$ is therefore an upper limit to any value of $F_{0c}^{-1}(x)$.

The above definition of $F_{0c}(\phi)$ uses two approaches depending on how large the frequencies q_i are. If they are large, as happens when i is small, the model for D_0 should be such that the probabilities exactly match the observed relative frequencies. However if the frequencies are small, as happens when i is large, some smoothing of the observed frequency distribution should be carried out. The above solution uses both techniques for different ranges of i, and the parameter c determines the crossover between the two different approaches. c is the largest index i of ϕ_i for which a discontinuity occurs in $F_{0c}(.)$ and the largest index i of ϕ_i such that $F_{0c}(.)$ includes its relative frequency exactly.

The simulation was carried out by generating m_0 pseudo-random numbers y from the uniform distribution in the interval $[0, 1]$ and for each of these, ϕ is defined to be the nearest integer to $F_{0c}^{-1}(y)$, which is forced not to be 0 or 1 (another y is generated if this happens), then n_1 is determined from the Binomial distribution $\text{Bin}(r, \phi)$, except that if n_1 is 0 or ϕ , another n_1 is chosen repeatedly until this condition is not satisfied, and finally (n_1, n_2) is recorded as an element of the simulated dataset where $n_2 = \phi - n_1$. This dataset was then summarised by counting the frequency that each pair (n_1, n_2) occurs, as was done with the original dataset D.

The quantity $j_{0c}(\phi)$ is defined in terms of $F_{0c}(.)$ thus:

$$\begin{array}{l} \text{if } \phi > \phi_c \text{ then } j_{0c}(\phi) = \{F_{0c}(\phi + a) - F_{0c}(\phi - a)\}/2a \\ \text{if } \phi \leq \phi_c \text{ then } j_{0c}(\phi) = j_0(\phi) \end{array} \tag{11}$$

so smoothing is done only if $\phi > \phi_c$ as in the definition of $F_{0c}(.)$. Here a is initially calculated as the minimum of (a) the absolute difference between ϕ and its maximum and (b) $\phi/3$, and a is forced to be 1 when it would otherwise be zero. Ideally $F_{0c}(.)$ and $j_{0c}(\phi)$ would be related by

$$F_{0c}(\phi) = \frac{1}{m_0} \sum_{\phi' \leq \phi} j_{0c}(\phi'). \tag{12}$$

The test statistic, modified from Equation (29), is

$$\hat{l}(x) = \frac{m_0 h(n_1, n_2)\left(1 - (1-\hat{r})^{\phi} - \hat{r}^{\phi}\right)}{m\text{Bin}(\hat{r}, n_1, \phi) j_{0c}(\phi)} \tag{13}$$

The formulae used here were chosen based on simple ideas and the need for fast calculation so that Equation (12) is only approximately true. The definitions of $j_{0c}(.)$ and $F_{0c}(.)$ used here are clearly somewhat arbitrary. A better approach to smoothing the distribution of ϕ, especially important for smaller data sets, would probably be based on for example Poisson regression [Zelterman , 2006]. Testing different smoothing strategies is beyond the scope of this paper.

5 Outline of the Algorithm

In this algorithm the data set is partitioned into two parts (those that appear to follow the null hypothesis and those that don't) that can vary during the iterative calculation because λ is fixed while the p-values may change. In this calculation, the subset of data believed to come from the null hypothesis is used to estimate the parameters in the null hypothesis.

To carry out this analysis an efficient algorithm that makes extensive use of the "list" template in the Standard Template Library (STL) in C++ is presented. A data structure based on "list" is very efficient because many of the steps can be performed efficiently after the list has been sorted appropriately, and "list" supports fast sorting. In particular the methods of calculation of the new p-values and $j_{0c}(\phi)$ use this idea when comparing the lists of test statistics from the simulated and observed data sets.

1. Read in the frequency data (n_{1i}, n_{2i}) for $1 \leq i \leq m$ and check for repetition. A repeat count is maintained for each distinct data point. This gives the summary which is a set triplets $(n_1, n_2, h(n_1, n_2))$. These are stored in an STL list of objects q, one object for each repeated data point.

2. Ask the user for a positive integer value N which is the number of simulated datasets to be compared with the observed one when calculating the new p-values, a positive integer value of c used in the smoothing procedure, and a value λ between 0 and 1 used as the p-value cut-off to determine D_0 as the subset of D with p-values greater than λ. D_0 is specified by a slot in q representing the predicate "is in D_0". Initially, D_0 is the whole dataset D unless the following option is used.

3. Another option is to determine D_0 using the p-value cut-off λ by firstly, for each member of q, computing the initial p-values from Equation (10) using $r = \hat{r}$ estimated from Equation (23) using the whole data set D. These p-values are added to q. If this test procedure is to be compared as-is (i.e. without iteration) with the other test procedures, sort q by these p-values then proceed directly to step (6).

4. With ndiff initialised to a non-zero value, repeat the following

   a. If ndiff is not zero the user has the option of continuing the calculation.

   b. Estimate r from Equation (23) using the dataset D_0.

   c. Sort q by increasing ϕ.

   d. Calculate the frequency distribution $j_0(\phi)$ from D_0.

   e. Use the smoothing procedure to calculate $j_{0c}(\phi)$ in Equation (11).

   f. Calculate the test statistic for each element of q from Equation (13).

   g. Sort q by decreasing test statistic.

   h. Initiate the recalculation of the p-value for each member of q as follows:

   i. For N iterations do

        i. Simulate a dataset q_sim based on the null hypothesis defined by $j_{0c}(\phi)$ and $\hat{r}$.
        ii. Calculate the test statistic for each data point in q_sim in the same way i.e. from Equation (13).
        iii. Sort q_sim by decreasing test statistic.
        iv. Use the two sorted lists q and q_sim to increment a variable in q for each datapoint by the frequency that the test statistic for $q\_sim \geq$ the test statistic for this datapoint.

    j. Complete the recalculation of the p-values (and insert them into q) for each test with distinct data as the mean in the above loop of the relative frequency with which the test statistic for the simulated data is $\geq$ the test statistic for the observed data.

    k. Calculate D_0 as $\{x \in D | p(x) > \lambda\}$

    l. Calculate ndiff as the number of tests that have changed status in two consecutive cycles.

5. Estimate $\hat{\pi}_0$ as $\#\{D_0\}/\#\{D\}$.

6. A measure of consistency of this procedure is provided by comparing the observed frequency distribution of ϕ with sum of the smoothed distribution $j_{0c}(\phi)$ fitted to the data believed to come from null hypothesis (D_0) and the frequency distribution of ϕ resulting from the data subset $D \backslash D_0$. This was done with a chi-square test and too small p-values indicate lack of fit and too large p-values indicate over-fitting, both of which must be avoided.

7. Estimated q-values (giving an estimate of the FDR if only that test and all those more significant than it are designated as significant) were obtained from the respective p-values for each test using the algorithm defined by Storey and Tibshirani [2003]. To do this, if the p-values are labelled in increasing order thus $p_{(1)} \leq p_{(3)} \ldots \leq p_{(m)}$, then the q-value for the ith most significant p-value is given by $\hat{q}(p_{(i)})/\hat{\pi}_0 = \min\limits_{t \geq p_{(i)}} \frac{mt}{\#\{p_j \leq t\}}$. Note that this minimum must be where t is one of the p-values. These values should then be calculated in reverse order of i starting at $i = m$ when necessarily $t = p_{(m)}$. Then $\hat{q}(p_{(m)})/\hat{\pi}_0 = p_{(m)}$ and $\hat{q}(p_{(i)})/\hat{\pi}_0 = \min\left\{\frac{mp_{(i)}}{i}, \hat{q}(p_{(i+1)})/\hat{\pi}_0\right\}$ for $m - 1 \geq i \geq 1$. These are multiplied by $\hat{\pi}_0$ to get the estimates of the q-values $\hat{q}(p_{(i)})$ and reported together with the original and final p-values for each test.

8. For testing simulated datasets, for each (n_1, n_2) point in the structure q, the number of null (# null) and alternative hypotheses (# alternative) that gave rise to the total observed number of times (n_1, n_2) occurred is found. Then the running total of the numbers of null and alternative hypotheses is found for the whole structure q, starting with the most significant data point. The actual q-value was calculated as the fraction of null data in the total number of data declared significant at each cut-off i.e.

$$q = \text{sum}(\#\text{null}) / \left(\text{sum}(\#\text{null}) + \text{sum}(\#\text{alternative})\right).$$

    The cumulative probability distributions of the null and alternative p-values were also obtained by plotting sum(#null)$/m$ and sum(#alternative)$/m$ against p-value respectively.

    Also the ideal case can be considered in which all the alternative hypotheses are rejected before any of the null hypotheses. In this case, in the notation of Table 2, $T = 0$ from which follows that

$S = m_1$. At any cut-off determined by a subset of q taken in decreasing order of significance, $R = m_1 + V$ and the q-value is defined as $q\_value = V/R = (R - m_1)/R$. This can be compared with the actual q-value for each proportion of rejections.

Because these loops are basically single pass (excluding the sorting) it is expected that their time complexity will be a linear function of the size of D i.e. m and will therefore be dominated by the time for sorting which is proportional to $Nm\ln(m)$ per iteration of the outer loop.

For analysis of sets of simulated data sets, this algorithm was run in a loop following the generation of each simulated data set and the resulting q-values were linearly interpolated (using nearest points only) onto a grid of values representing the proportion of data declared significant. These interpolated values were averaged and the sample standard deviations were calculated from them.

6 Analysis of the Experimental Data Set and the Power Analysis

The experiment described above gave rise to the dataset D that consists of $m = 2171$ pairs of frequencies (n_1, n_2), which was summarised (Table 1) as a table of 119 rows, each giving n_1,n_2, and the frequency of this combination in D. D was analysed using the above iterative algorithm for the ranges of values of λ and c as indicated in Tables 3 and 4 and with $N = 100$. The algorithm converged (i.e. the number of differences in the status of hypotheses between two consecutive cycles became 0 after a few cycles) except in one case ($c = 18$, $\lambda = 10^{-3}$). For each of the runs, the program calculates the p-values and q-values for each of the tests and reports them in increasing order of p-value (i.e. increasing q-value). Also reported is the fraction y of the data having p-value less than or equal to the current p-value. This takes into account the frequency of occurrence of (n_1, n_2) in the data.

The p-values for the χ^2 test described in the algorithm to check consistency were calculated with separate bins for ϕ in classes $\{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}$, and $\{\geq 9\}$. Bins for which the observed and expected frequencies were forced to be the same (because of the way the smoothing algorithm operates as truncated by c) did not count towards the degrees of freedom. Each of these analyses was done in duplicate to check the algorithm because simulation is involved in the recalculation of the p-values in each cycle. This leads to slightly different results being produced each time the program is run with the same input (Tables 3 and 4). The largest difference between repeat values of $\hat{\pi_0}$ was 0.023 (for the case $\lambda = 10^{-3}$ and $c = 14$) and in all other cases the differences were less than 0.0032. A one-way ANOVA for $\hat{\pi_0}$ gave a standard deviation within treatments (i.e. λ and c) of 0.0022. The repeat values of χ^2 showed little variation with none in about half the cases.

Varying λ varies the thickness of the tail of the smoothed null hypothesis $j_{0c}(\phi)$ with a thicker tail resulting from smaller values of λ. Varying c alters the value of ϕ beyond which the smoothing algorithm actually changes the value of its argument.

The results in Table 4 suggest that λ should be smaller than 10^{-3} and c should be larger than 8. Later results will suggest that when λ is too small the power of the test becomes low so the parameter pair ($\lambda = 10^{-4}$ and $c = 12$) was chosen as a point where the test works close to optimally, and the results are shown in detail for this case, the comparison of the observed and smoothed null distribution of ϕ (Figure 2) and the corresponding cumulative distribution of p-values (Figure 1). The result gave a good fit of the smoothed null hypothesis distribution $j_{0c}(\phi)$ to the data $j(\phi)$ as expected. The high frequencies in Table 1 for small values of n_1 and n_2 explain the large jumps between one y value and the next (Figure 1) because there are large groups of data that are identical and so have the same p-value.

	$\log_{10}(\lambda)$							
c	-30	-10	-7	-6	-5	-4	-3	-2
8	0.0101	0.0106	0.0101	0.0113	0.0113	0.0290	0.0659	0.3192
10	0.0096	0.0111	0.0101	0.0113	0.0124	0.0293	0.0564	0.3192
12	0.0101	0.0106	0.0101	0.0113	0.0124	0.0295	0.0567	0.3192
14	0.0101	0.0101	0.0101	0.0113	0.0124	0.0295	0.0567	0.3192
16	0.0101	0.0106	0.0101	0.0113	0.0124	0.0304	0.0682	0.3192
18	0.0101	0.0101	0.0101	0.0113	0.0124	No Data	0.0682	0.3192
20	0.0101	0.0101	0.0101	0.0111	0.0124	0.0295	0.0682	0.3192

Table 3: 1-mean of the estimates of π_0 obtained from two runs of the analysis for a range of values of the smoothing parameter c and the p-value cut-off parameter λ

	$\log_{10}(\lambda)$							
c	-30	-10	-7	-6	-5	-4	-3	-2
8	0.04	0.04	0.04	0.04	0.04	0.03	0.23	1.00
10	0.50	0.50	0.49	0.52	0.54	0.61	0.89	1.00
12	0.45	0.46	0.45	0.49	0.52	0.55	0.93	1.00
14	0.23	0.23	0.23	0.26	0.29	0.33	0.80	1.00
16	0.33	0.34	0.33	0.38	0.41	0.77	1.00	1.00
18	0.30	0.30	0.30	0.36	0.41	No Data	0.99	1.00
20	0.27	0.27	0.27	0.29	0.33	0.88	1.00	1.00

Table 4: p-values for the mean consistency check χ^2 obtained from two runs of the analysis for a range of values of the smoothing parameter c and the p-value cut-off parameter λ

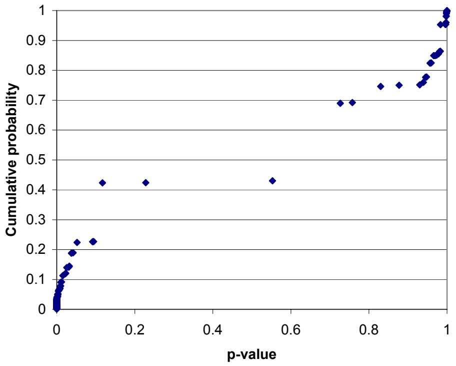
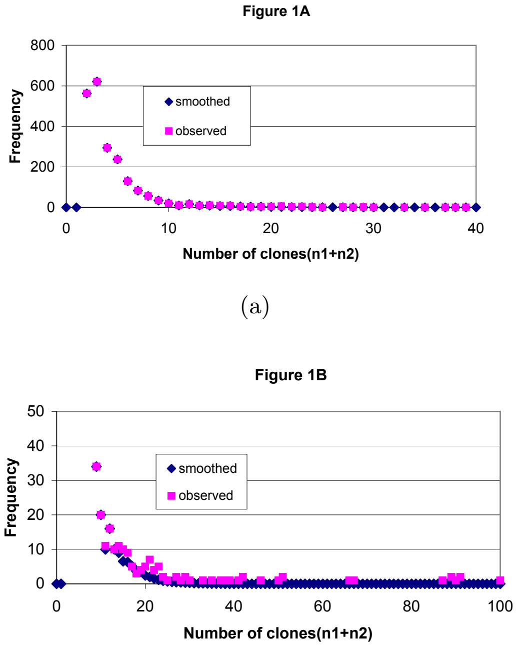


Figure 1: The p-value distribution for the analysis of the observed data with parameters $\lambda = 10^{-4}$ and $c = 12$.



(a)

(b)

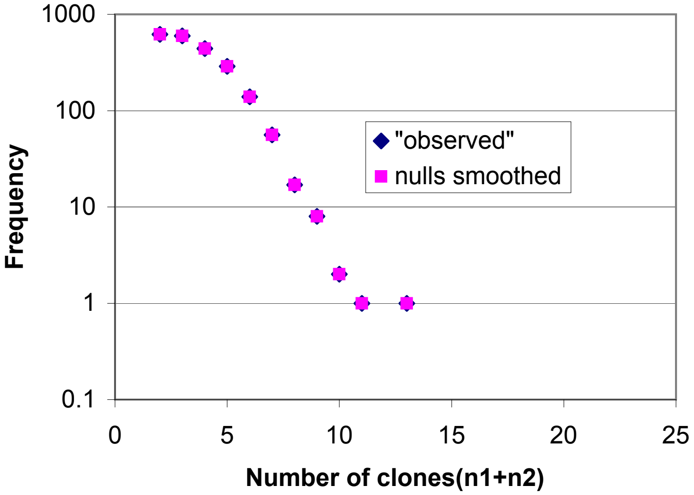
Figure 2: Comparison between the observed and null hypothesis fitted frequency distributions of $n_1 + n_2$ for the experimental data (Table 1) using parameters $\lambda = 10^{-4}$ and $c = 12$.

In this analysis it is not clear how well the statistical technique is performing because the actual status of each data point is not known. Therefore two sets of simulations (scenarios 1 and 2) were carried out as in Table 5 to generate data sets somewhat similar to the observed data set. In

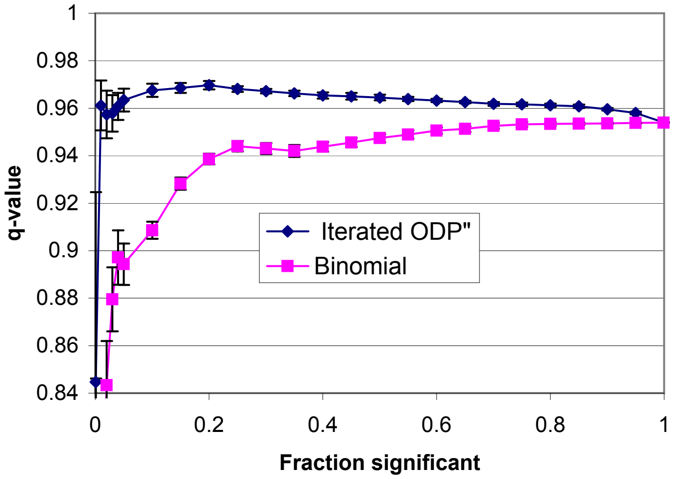
	Scenario 1		Scenario 2	
	$n_1 + n_2$	$n_1 \vert n_1 + n_2$	$n_1 + n_2$	$n_1 \vert n_1 + n_2$
Null data	Poisson^T(3)	Binom^T(0.33)	Poisson^T(3)	Binom^T(0.33)
Alternative data	Poisson^T(3)	Uniform[1,$n_1 + n_2 - 1$]	Poisson^T(15)	Uniform[1,$n_1 + n_2 - 1$]

Table 5: Probability models used for two sets of simulations. Both the probability distributions of $n_1 + n_2$ and of n_1 given $n_1 + n_2$ are specified for the null and alternative subsets of the data in each simulation. Poisson$^T(\mu)$ is the Poisson distribution with mean μ and truncated such that values 0 or 1 cannot be generated (i.e. whenever they occur another value is generated from the Poisson distribution), where μ was chosen to agree approximately with the truncated mean of the observed data distribution. Binom$^T(r)$ is the truncated Binomial distribution with n specified by the distribution in the cell to its immediate left, and r is the other Binomial parameter. The result cannot be 0 or n, otherwise another value is chosen. Uniform[a,b] is the uniform distribution with non-zero values from $X = a$ to b inclusive.

both sets of simulated data sets the number of null and alternative data points was 2071 and 100 respectively to agree with the observed total of 2171 data points, and assuming approximately 5% of the data are not from the null hypothesis and have very small p-values (Figure 1). The truncated mean of ϕ for ϕ up to 5 is 3.12, and this accounts for 79% of the data (the mean of all values of ϕ is 5.18 but this includes the effect of a long tail in the distribution) and so the mean was taken as 3 for the simulated null data in both sets of simulations. All the simulated data sets were analysed with the algorithm described here based on the ODP, and the truncated Binomial test i.e. Equation (10). The parameter values $N = 100$, $\lambda = 10^{-4}$ and $c = 12$ were used as was done for the experimental data. This value of c almost prevents smoothing (exactly for Simulation 1), which is not required anyway because the simulated data does not have the problems the observed data had namely the long tail with sparsely distributed data in the frequency distribution of ϕ. This makes the χ^2 statistic now meaningless.

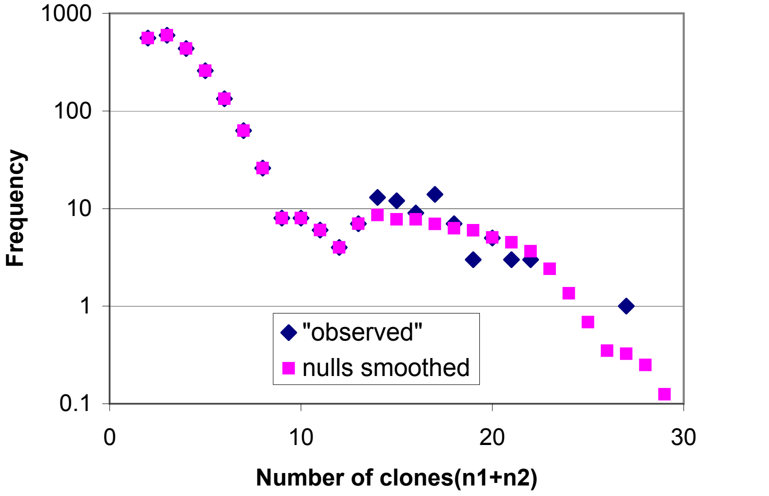


(a) Comparison between the "observed" and null hypothesis fitted frequency distributions of $n_1 + n_2$ for instance 1 of the simulated data.

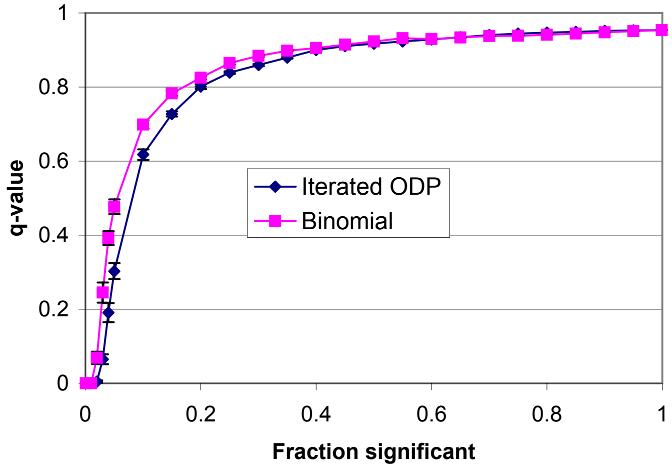


(b) Comparison between the power of the iterated ODP and the truncated Binomial test. Means of 10 simulations are plotted with errorbars representing one standard deviation of the mean.

Figure 3: Results for scenario 1 of Table 5 using parameters $\lambda = 10^{-4}$ and $c = 12$ with 100 alternative data points out of a total of 2171 data points.



(a) Comparison between the "observed" and null hypothesis fitted frequency distributions of $n_1 + n_2$ for instance 1 of the simulated data.



(b) Comparison between the power of the iterated ODP with the truncated Binomial test. Means of 10 simulations are plotted with errorbars representing one standard deviation of the mean.

Figure 4: Results for scenario 2 of Table 5 using parameters $\lambda = 10^{-4}$ and $c = 12$ with 100 alternative data points out of a total of 2171 data points.

For the two scenarios described in Table 5, the exact q-value (fraction of errors in the subset of D selected by the cut-off) for each cut-off point is related to the fraction of the data selected by the cut-off (Figures 3b and 4b). This plot seems to be a good way to assess the relative efficiency of different multiple hypothesis testing procedures. If the status of each hypothesis is known because the data are simulated, the most important point on the x axis is where x equals the actual proportion of alternative hypotheses in the data. Comparisons between different procedures should be made at this point if it is known. For scenario 1, Figure 3 shows that better discrimination is obtained by using the truncated Binomial test (10) rather than the iterated ODP i.e. the q-value at this point is lower than for the ODP. For scenario 2, the result is reversed, with this iterated ODP analysis having the advantage. The results for scenario 2 are very different from the results for scenario 1, because in scenario 2, some data points having the larger values of ϕ are significant with a q-value estimated as zero (also evident in Figure 4(a) because the difference between the two plots represents alternative hypotheses). This is expected because many of these data points will be likely to be non-null on the basis of ϕ alone because of the difference between the distributions of ϕ in this scenario.

7 Further investigations

Because the iterated ODP may or may not be superior to the truncated Binomial test, which was not an expected result, further results are needed to verify that the statistical methods described here are performing correctly and verify the relative efficiencies of the different procedures in different scenarios. Independent sets of 10 simulated datasets of integer pairs (n_1, n_2) were generated with $\pi_0 = 0.5$ and $m = 10^4$ and were analysed using the above algorithm with $\lambda = 10^{-3}$ with two different initial assumptions of the status of the hypotheses (1) that all the hypotheses are null (i.e. $w = 1$) and (2) the hypotheses have their initial status determined by the truncated Binomial test using the p-value cut-off equal to λ. As before the mean fraction of errors in those hypotheses chosen as significant (q-value) in the 10 datasets was plotted against the fraction of data declared significant. In addition the corresponding results were obtained from sets of 10 independent datasets after one and two cycles of the algorithm, with both the above initial assumptions, and compared with the result of the truncated Binomial test itself. In these analyses, smoothing was disabled because it is unnecessary for these large data sets and it can introduce artificial discontinuities in $j_{0c}(\phi)$. This makes the parameter c irrelevant (it is effectively infinite).

The progress of the iterative algorithm described here in the two cases and the independence of the final result on the initial assignment of hypotheses to the individual tests are clearly shown in Figure 5. Two-sample t-tests calculated from the means and variances of both the fully converged results showed that they were not significantly different from each other (smallest p-value in 998 tests was 0.023 and the distribution of p-values was approximately uniform). Moreover a general result to help validate the iterative procedure (3) was that if the algorithm was initialised with the p-values from the Binomial test, the result was almost the same as when it was initialised by assuming that all the hypotheses are null.

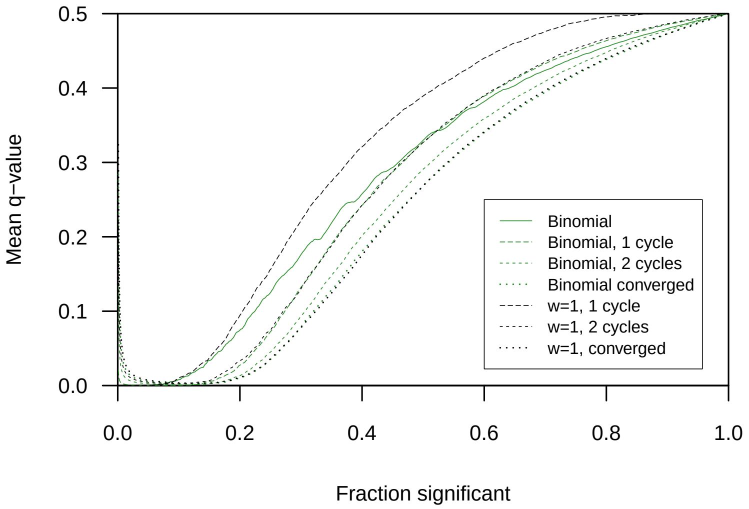


Figure 5: Progress of the convergence of the algorithm described here based on the ODP, initialised by the truncated Binomial test and by the assumption that all the hypotheses are null.

To study further out how well the multiple hypothesis testing procedures perform and how they compare with each other (including the comparison between the ODP and the iterated ODP which is hardly possible in the results of section 6 because very little iteration was needed in these cases), sets of 10 simulated data sets of integer pairs (n_1, n_2) were generated, each set using a combination of $\pi_0 \in \{0.9, 0.5\}$ and $m \in \{10^3, 10^4, 10^5\}$ with the assignment of the hypotheses done without randomization. Four replicates of this was done independently and each replicate was analysed with a different value of λ. The simulations were all done according to probability models in Table 6 and were analysed (Figures 6 and 7) with λ taking each of the values $10^{-5}, 10^{-4}, 10^{-3}$, and 3×10^{-3}, and with $N = 100$. The ideal case where all the true alternative hypotheses are ranked more significant than any of the true null hypotheses was included for comparison.

	$n_1 + n_2$	$n_1 \mid n_1 + n_2$
Null data	Poisson$^T(20)$	Bin$^T(0.3)$
Alternative data	Poisson$^T(20)$	Uniform$[1,n_1 + n_2 - 1]$

Table 6: Probability models used for a set of simulations. The notation is exactly the same as for Table 5.

The following multiple hypothesis testing procedures were compared using the simulated data sets described above:

1. the truncated Binomial test according to Equation (10) (This is actually a simple case where each test is done independently).
2. the ODP with all the hypotheses initially assumed to be null ($w = 1$, 1 cycle). This is the same as running the algorithm in Section 5 for one cycle.

3. the iterated ODP run to convergence with different values of λ as described above.

The means and variances of estimates of π_0 were obtained (Tables 7 and 8) from the iterated ODP. The results were very poor for $m = 10^3$ and for $\lambda = 10^{-5}$ but improved as each of m and λ increased leading to reasonable estimates when $m = 10^5$ and $\lambda = 3 \times 10^{-3}$. These are consistent with earlier results for single estimates of π_0 for $\pi_0 \in \{0.5, 0.7, 0.9, 0.99\}$ and $m = 10^4$ (data not shown). These overestimates are obviously biased upward because the bias exceeds the standard deviation at least 10 fold in many cases. The bias and the complexity of the problem makes it probably impractical to use the Cramer Rao lower bound on the variance of the estimator to objectively assess it, except possibly using simulations. All that can be said with certainty is that this estimator of π_0 is capable of improvement. As suggested below it may be better to try to optimise the statistical power of the multiple testing procedure as defined by the q-value as a function of the fraction of data declared significant, which is far more informative than just the estimate of π_0.

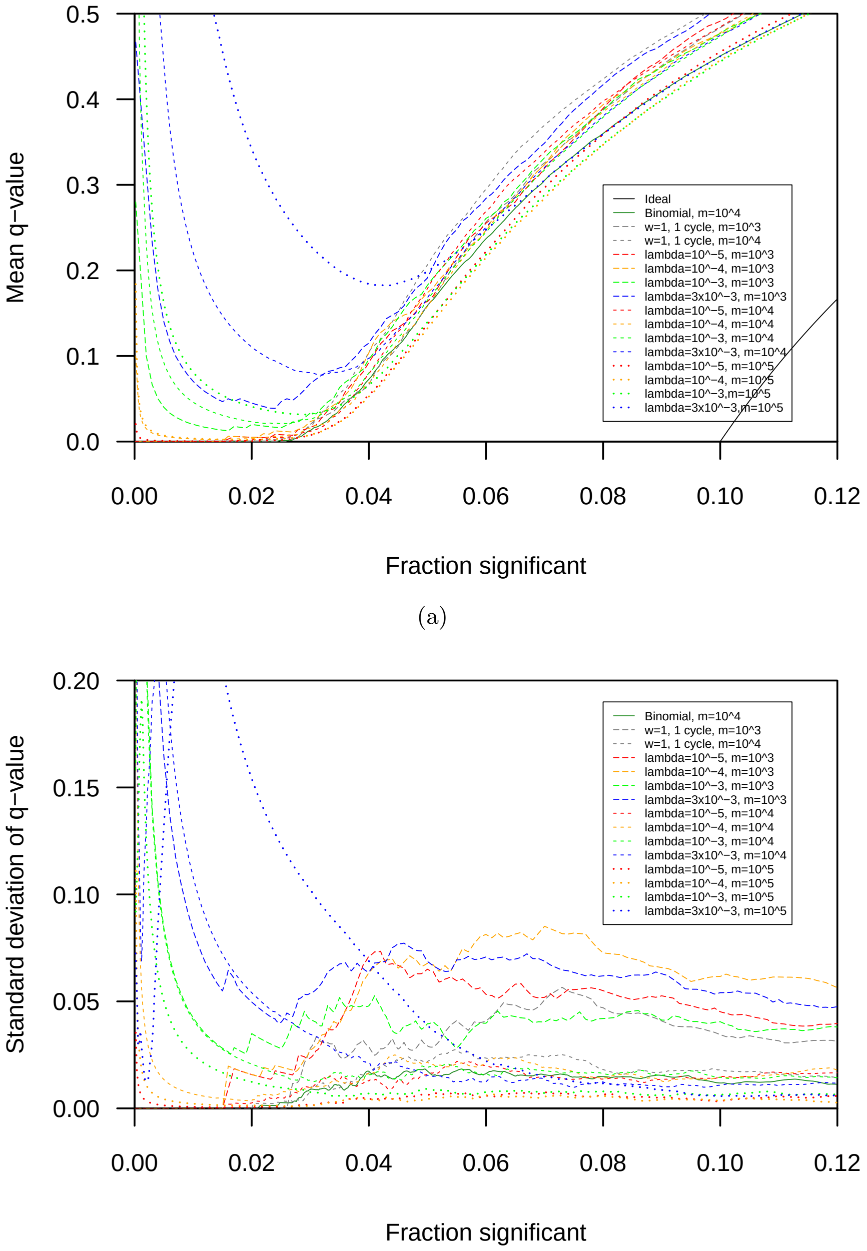
	λ values for mean				λ values for s.d.			
m	10^{-5}	10^{-4}	10^{-3}	3×10^{-3}	10^{-5}	10^{-4}	10^{-3}	3×10^{-3}
10^3	0.982	0.975	0.966	0.961	0.0042	0.0038	0.0051	0.0068
10^4	0.979	0.971	0.963	0.951	0.0012	0.0017	0.0024	0.0025
10^5	0.9733	0.9639	0.947	0.924	0.00078	0.00082	0.0026	0.0044

Table 7: Mean and standard deviation of estimates of π_0 from analyses of 10 simulated data sets with $\pi_0 = 0.9$ for each value of m. Independent simulations were used for the analyses with different values of λ.

	λ values for mean				λ values for s.d.			
m	10^{-5}	10^{-4}	10^{-3}	3×10^{-3}	10^{-5}	10^{-4}	10^{-3}	3×10^{-3}
10^3	0.910	0.875	0.830	0.798	0.013	0.013	0.010	0.011
10^4	0.891	0.846	0.779	0.742	0.0051	0.0059	0.0074	0.0067
10^5	0.844	0.872	0.708	0.526	0.0057	0.0036	0.0044	0.0077

Table 8: Mean and standard deviation of estimates of π_0 from analyses of 10 simulated data sets with $\pi_0 = 0.5$ for each value of m. Independent simulations were used for the analyses with different values of λ.

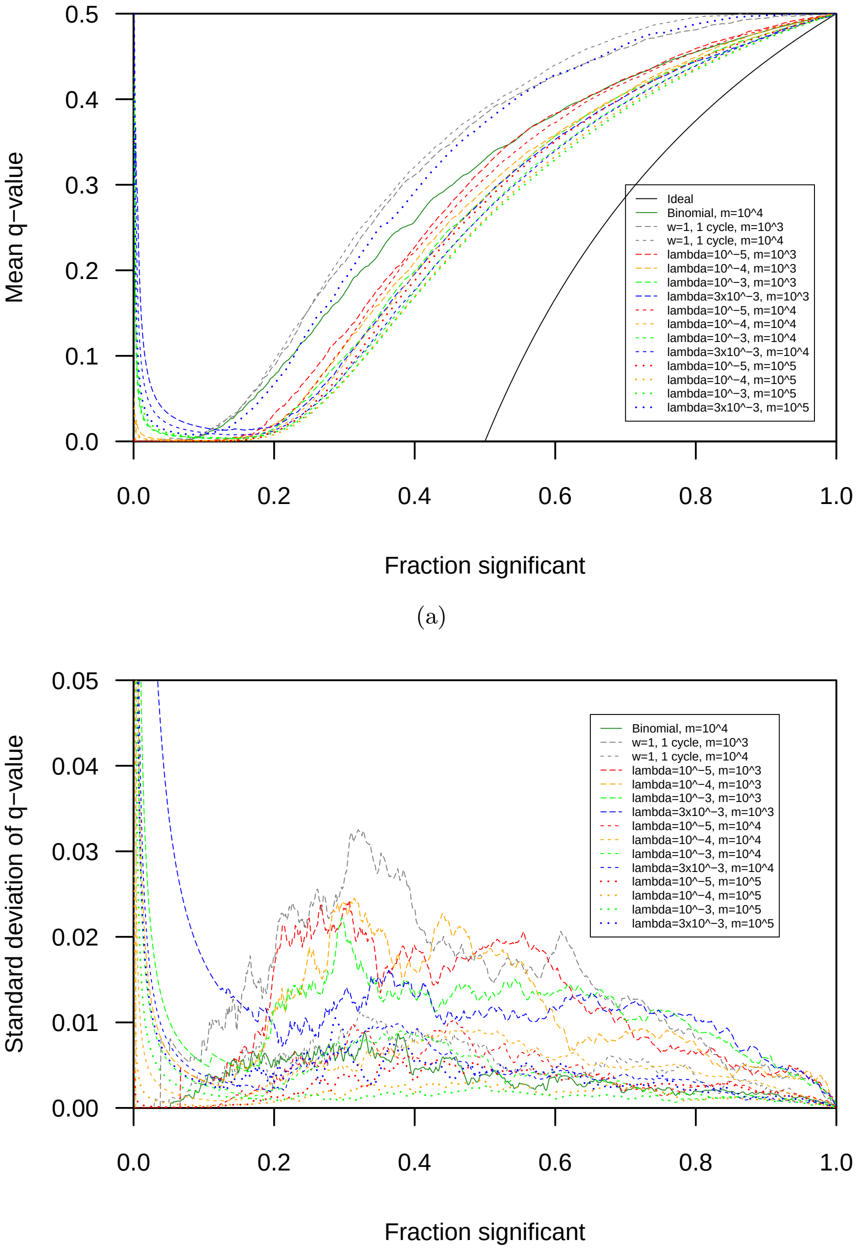
As was done in the earlier results, the actual q-value was plotted against the fraction of the data declared significant. The larger the fraction of the data declared significant is for a given q-value, the better the statistical procedure is, so the better procedures have the curve further to the right. These curves describing the efficiency (i.e. statistical power) of the multiple testing procedures (Figures 6 and 7) allow many comparisons to be made.



(a)

(b)

Figure 6: Comparison of the power of the ODP and its iterative extension for different values of the p-value cut-off parameter λ. Different sizes of the dataset(m) are compared and the proportion of null hypotheses is π_0=0.9 (without randomization). The ideal case and the truncated Binomial test are included for comparison.



(a)

(b)

Figure 7: Comparison of the power of the ODP and its iterative extension for different values of the p-value cut-off parameter λ. Different sizes of the dataset(m) are compared and the proportion of null hypotheses is π_0=0.5 (without randomization). The ideal case and the truncated Binomial test are included for comparison.

A complex picture emerges on examining these results for the iterated ODP. For $\pi_0 = 0.5$ there is a trend that the power increases as λ increases from 10^{-5} to 3×10^{-3}. The only exception to this is for $m = 10^5$ when the power for $\lambda = 3 \times 10^{-3}$ is less than that of all the other cases. When $\pi_0 = 0.9$ the picture is similar this time with exceptions for $\lambda = 3 \times 10^{-3}$ and for $\lambda = 10^{-3}$ with $m = 10^5$. In addition to this many of the curves cross making the comparisons dependent on the error rate acceptable (if comparison is made according to the fraction of data that can then be declared significant) or on the fraction of the data declared significant (if comparison is made according to the expected fraction of errors in those that are declared significant). The best value of λ to be used in the procedures (3) is probably close to 10^{-3} or 3×10^{-3} and is certainly dependent on m. Too small a value of λ leads to loss of efficiency in the ascending parts of the curves and too large a value leads to large q-values for small values of the fraction of significant data chosen by the cut-off (Figures 6 and 7). The latter is a consequence of the statistical analysis erroneously choosing some null data points in the first few results listed in order of increasing p-value and seems to be an artefact of the iterative procedure. The results can also be compared for the same value of λ for different sizes of the data set (m). In this case there is the trend that larger values of m lead to greater statistical power but there are exceptions here too particularly with the larger values of λ. It is interesting to note that the near optimal values of λ from Figures 6 and 7 correspond to the results with the most accurate estimates of π_0. However the characteristic curves of the tests are far more informative than just the estimates $\hat{\pi_0}$ because it could happen that $\hat{\pi_0}$ is accurate if an equal number of type I and type II errors were made by the test for a particular value of the p-value cut-off λ, regardless of how large these error rates are.

It is also interesting to compare these results with the one cycle ODP which shows that only in one case ($m = 10^3$, $\pi_0 = 0.9$) was the efficiency of this method (ignoring the descending parts of the curves that occur for the larger values of λ that are clearly artefacts mentioned above) comparable with the efficiency of the iterated ODP, in all other cases the one-cycle ODP was outperformed by the iterated ODP for all the parameter values tested.

The efficiency of the truncated Binomial test is interesting because it is independent of the other parameters (λ, m) unlike the different versions of the iterated ODP. Therefore the comparison between the iterated ODP and the truncated Binomial tests does depend on these parameters. Generally larger m favours the use of the iterated ODP test with this being true for smaller m as π_0 decreases (this conclusion is also supported by single analyses of datasets with $\pi_0 = 0.99$ and $m = 10^5$, data not shown). Therefore it is an interesting and important problem to generally determine the relative efficiencies of the different procedures so that the best one can be chosen (i.e. what is the best value of λ to take in the iterated ODP or should the truncated Binomial test be used instead) in the given circumstances (m and π_0).

8 Discussion

The iterative modification of Storey's ODP proposed here was designed to be an extension of the ODP to cases where the null hypothesis is not completely specified i.e. it is a distribution containing unknown parameters. In this case another set of binary parameters arise, that represent whether or not each data point contributes to the estimation of the null hypothesis, and the iterative extension to the ODP enforces some consistency between them and the p-values of the tests. This however involves another unknown parameter λ used as a cut-off for the p-values.

The iterated ODP was applied to the problem of testing a set of genes each having frequencies of occurrence of transcripts n_1 and n_2 in a pair of libraries against the null hypothesis that these frequencies are in proportion to the sizes of the libraries. The data with n_1 or n_2 equal to 0 was removed prior to this analysis.

The result of this analysis is the p-values and estimated q-values of all the tests listed in increasing order of q-value (and p-value). The q-value of a test is the expected frequency of errors, if the current test and all the more significant tests are assumed to be cases where the null hypothesis does not hold. In this example, the result is the list of genes in decreasing order of significance against the common expected ratio r amongst the null hypotheses, without regard to the direction in which the estimate of r for the gene differs from the overall estimate of r.

The results with simulated data show that definite improvements in power can be obtained when compared with Storey's ODP i.e. this algorithm with one cycle and with initial estimates of the status of the hypotheses being either that they are all (1) null hypotheses (Figures 6 and 7) or (2) taken from the truncated Binomial test (Figure 5). The extent of the improvement in the power observed here in case (1) is dependent on λ and the number of data points generated from the alternative hypothesis. However sometimes, particularly with small datasets, the most powerful procedure found here is the test based on the truncated Binomial distribution, but for the largest datasets used here ($m = 10^5$) the Binomial test was bettered by the iterated ODP for some values of the fraction of significant data, especially when π_0 was 0.5. This is contrary to what is expected based on the general arguments used here which are a modification of Storey's derivation to the case when two distributions are not known and so have to be estimated from the data. These arguments assume continuous data and the result seems to show an effect of using discrete data that is converted to frequency data that is not adequately dealt with in the derivation.

There still remains the problem of determining the optimum value of the cut-off parameter λ, which was roughly estimated in this study by trying a few values. This parameter also affects the accuracy with which the proportion of null hypotheses is estimated, with more accurate values resulting from the values of λ that give the greatest statistical power (proportion of results declared significant for a fixed proportion of errors in those declared significant).

Because of the simplicity of the essential idea behind the statistical technique being developed, it is anticipated that there will be many applications and extensions of the method to other areas that involve dichotomous classification of entities having discrete or continuous data associated with them (for example genes associated or not associated with a disease state) such that one or both classes has a distribution with a known functional form but with unknown parameters.

However it is now clear that much more work has to be done to optimise the algorithm for multiple hypothesis testing. This was illustrated by the example of binomial data but the result is obviously true much more generally. The optimising strategy is also likely to be dependent on the size of the dataset and the fraction of data arising from the alternative hypothesis. In this paper only strategies involving w_i equal to 0 or 1 were considered, leading in an obvious way to the p-value cut-off parameter λ. More generally, strategies involving fractional values of w dependent on the data point or upon the p-value of the test could be considered, which would be interpreted as probabilities or weight factors. It is believed that future improvements to this statistical technique along these lines can also be made within the efficient computational scheme described above. Similar general conclusions would be applicable to more complex data analyses such as when, for each test, more than two comparisons are being made for each gene i.e. there would be say frequencies $(n_1, n_2 \ldots n_j)$ representing a gene's expression level across j experimental conditions, or when in addition there are data for more than one categorical variable or replication. Therefore extensions of the study in this paper are likely to be increasingly important as more complex and large data sets are generated and analysed to find for example candidate genes responsible for complex traits such as human diseases.

9 Appendix: derivation of the test statistic and p-value for each test, including the case when for a subset of (n_1, n_2) the data should not be used

For the null hypothesis, $f(x) = \mathrm{Bin}(r, n_1, \phi)P(\phi)$ where $x = (n_1, n_2)$ and $\phi = n_1 + n_2$. This derivation of the test statistic involves separately estimating the numerator and denominator of Equation (5) by maximum likelihood. Under the null hypothesis, the likelihood of the subset of data D_0 (the subset of D believed to come from cases where the null hypothesis is true) can be written as

$$L_0 = \prod_{n_1, n_2 \geq 0} \left[(\mathrm{Bin}(r, n_1, \phi)P(\phi))^{h_0(n_1,n_2)} \right] \tag{14}$$

where $D_0 = \{(n_1, n_2)$ with frequency $h_0(n_1, n_2)$ for all n_1 and $n_2\}$. L_0 must be maximised by choosing r and $P(.)$ such that $P(\phi) \geq 0$ for all $\phi > 0$ with the constraint

$$\sum_{\phi \geq 0} P(\phi) = 1. \tag{15}$$

The likelihood can be written as $L_0 = L_1 \times L_2$ where

$$L_1 = \prod_{n_1 \geq 0} \prod_{n_2 \geq 0} (\mathrm{Bin}(r, n_1, \phi))^{h_0(n_1,n_2)} \tag{16}$$

and

$$L_2 = \prod_{n_1 \geq 0} \prod_{n_2 \geq 0} P(\phi)^{h_0(n_1,n_2)} \tag{17}$$

which can be simplified to

$$L_2 = \prod_{\phi \geq 0} \prod_{n_1 = 0}^{\phi} P(\phi)^{h_0(n_1, \phi - n_1)} = \prod_{\phi \geq 0} P(\phi)^{j_0(\phi)}. \tag{18}$$

Further, because L_1 is independent of $P(\phi)$ and L_2 is independent of r, the maximum likelihood is the product of the maximum of L_1 obtained by varying r within $[0,1]$ and the maximum of L_2 obtained by varying $P(\phi)$ subject to (15).

The second of these problems is, apart from a constant factor, the problem of maximising the likelihood of the multinomial model, and the result is known [Santner and Duffy , 1989, pp 44-45] to be unique and determined by the relative frequencies of the data i.e.

$$P(\phi) = j_0(\phi)/m_0. \tag{19}$$

For maximising L_1 by varying r, a maximum of L_1 (there could be more than one) can be at either end of the interval i.e. 0 or 1, or at an interior point. If a maximum is at $r = 0$, then at this point

$$L_1 = \prod_{n_1 \geq 0} \prod_{n_2 \geq 0} \left[0^{n_1} 1^{n_2} \left(\begin{array}{c} n_1 + n_2 \\ n_1 \end{array} \right) \right]^{h_0(n_1,n_2)}. \tag{20}$$

This is 0 if $n_1 > 0$ and $h_0(n_1, n_2) > 0$ for any point (n_1, n_2). Therefore to avoid this contradiction, for all (n_1, n_2) either $h_0(n_1, n_2) = 0$ or $n_1 = 0$. In other words, for every point (n_1, n_2), $h_0(n_1, n_2) \neq 0 \implies n_1 = 0$. This only applies to trivial data sets that only have data for $n_1 = 0$ which I will exclude. Likewise a maximum is at $r = 1$ only occurs in trivial data sets and can be excluded. Local maxima in the interior i.e. in $(0, 1)$ satisfy $0 = \frac{\partial ln(L_1)}{\partial r}$ i.e.

$$\sum_{n_1 \geq 0} \sum_{n_2 \geq 0} h_0(n_1, n_2) \left(\frac{n_1}{r} - \frac{n_2}{1-r} \right) = 0. \tag{21}$$

The value of r that satisfies this is unique and given by

$$\hat{r} = \frac{\sum_{n_1 \geq 0} \sum_{n_2 \geq 0} h_0(n_1, n_2) n_1}{\sum_{n_1 \geq 0} \sum_{n_2 \geq 0} h_0(n_1, n_2)(n_1 + n_2)}, \tag{22}$$

or equivalently

$$\hat{r} = \frac{\sum_{i \in T_0} n_{1i}}{\sum_{i \in T_0} (n_{1i} + n_{2i})} \tag{23}$$

where T_0 is an index set determining the subset of m_0 data
$D_0 = \{(n_{1i}, n_{2i}) : i \in T_0\}$. Therefore the optimum point is unique so the maximum likelihood estimate of $f(x)$ is given by

$$\hat{f}(x) = \text{Bin}(\hat{r}, n_1, \phi) j_0(\phi)/m_0. \tag{24}$$

There is no distributional assumption associated with the maximum likelihood estimate of the distribution of x, $t(x)$ for the whole dataset D so this should be given by the maximum likelihood without the above constraint on the form of $P(n_1, n_2)$. In this case the likelihood $L = \prod_{i=1}^{m} P(n_{1i}, n_{2i})$ is to be maximised subject only to the constraint $\sum_{n_1 \geq 0} \sum_{n_2 \geq 0} P(n_1, n_2) = 1$. It is easy to show that this leads to $\hat{t}(x) = h(x)/m$ by the above quoted theorem. Therefore the test statistic is

$$\hat{l}(x) = \frac{m_0 h(n_1, n_2)}{m \text{Bin}(\hat{r}, n_1, \phi) j_0(\phi)}. \tag{25}$$

In this discrete case, the expression (6) for the p-value becomes

$$p(n_1, n_2) = \frac{1}{m_0} \sum_{\phi' \geq 0} j_0(\phi') \times \sum_{n_1' : \hat{l}(n_1', \phi' - n_1') \geq \hat{l}(n_1, n_2)}^{\phi'} \mathrm{Bin}(\hat{r}, n_1', \phi'). \tag{26}$$

This procedure should be modified when for the subset of pairs $(n_1, n_2) \notin S$, the data $h(n_1, n_2)$ is not going to be used, for example because this subset of data has already been removed prior to the analysis for any reason or because it is believed to be probably in error. This leaves only the subset of m_{0s} elements from D_0 such that $(n_1, n_2) \in S$, and the whole dataset D is likewise reduced. Then a null hypothesis should be set up that refers only to the subset of reliable data defined by the condition $(n_1, n_2) \in S$ and the test statistic must be re-derived.

Now null hypothesis is $f(x) = \mathrm{Bin}(r, n_1, \phi) P(\phi) H(n_1, n_2)/k$ and the normalisation constant is

$$k = \sum_{(n_1, n_2) \in S} \mathrm{Bin}(r, n_1, \phi) P(\phi) \tag{27}$$

where H is the indicator function for S i.e.

$$H(n_1, n_2) = \left\{ \begin{array}{cc} 1 & (n_1, n_2) \in S \\ 0 & \text{otherwise} \end{array} \right. .$$

Now

$$L_0 = \prod_{n_1 \geq 0, n_2 \geq 0} \left[(\mathrm{Bin}(r, n_1, \phi) P(\phi))^{h_0(n_1, n_2)} \right] \times \prod_{n_1 \geq 0, n_2 \geq 0} \left[\left(\frac{H(n_1, n_2)}{k} \right)^{h_0(n_1, n_2)} \right]$$

and the second factor is either the constant $k^{-m_{0s}}$ or 0. Maximising the likelihood obviously requires that this factor is non-zero and a necessary and sufficient condition for this is that $h_0(n_1, n_2) = 0$ whenever $(n_1, n_2) \notin S$. Then the likelihood is

$$L_0 = k^{-m_{0s}} \prod_{(n_1, n_2) \in S} \left[(\mathrm{Bin}(r, n_1, \phi) P(\phi))^{h_0(n_1, n_2)} \right]$$

which must be maximised with the constraints (15) and (27). This is equivalent to maximising $\ln(L_0)$ subject to these constraints and requires that
$\frac{\partial Q}{\partial P(\phi)} = \frac{\partial Q}{\partial r} = 0$ where

$$Q = \sum_{(n_1, n_2) \in S} h_0(n_1, n_2) \left(\ln(P(\phi)) + n_1 \ln r + n_2 \ln(1 - r) + \ln \left(\begin{array}{c} \phi \\ n_1 \end{array} \right) \right) -$$

$$m_{0S} \ln(k) - \mu \left(\sum_{\phi \geq 1} P(\phi) - 1 \right) - \sigma \left(k - \sum_{(n_1, n_2) \in S} \mathrm{Bin}(r, n_1, \phi) P(\phi) \right). \tag{28}$$

The difficulty here is to solve the Equation $\frac{\partial Q}{\partial r} = 0$ for r which has a degree related to the size of S, so there seem to be no simple formulae for the ML estimates of r and $P(\phi)$, which makes the method very difficult to implement in practice. However for the case in which S is defined by $n_1 \neq 0$ and $n_2 \neq 0$ as applies for the data analysed in this paper, replacing the Binomial distribution in Equation (26) by the Binomial distribution conditional on $n_1 \neq 0$ and $n_2 \neq 0$, shows that Equation (25) should be replaced by

$$\hat{l}(x) = \frac{m_0 h(n_1, n_2)\left(1 - (1-\hat{r})^{\phi} - \hat{r}^{\phi}\right)}{m.\mathrm{Bin}(\hat{r}, n_1, \phi) j_0(\phi)} \tag{29}$$

and Equation (26) should be replaced by

$$p(n_1, n_2) = \frac{1}{m_0} \sum_{\phi' \geq 0} \left\{ j_0(\phi') \times \frac{\sum_{\substack{n_1' = 1 \\ n_1' : \hat{l}(n_1', \phi' - n_1') \geq \hat{l}(n_1, n_2)}}^{\phi' - 1} \mathrm{Bin}(\hat{r}, n_1', \phi')}{1 - (1-\hat{r})^{\phi} - \hat{r}^{\phi}} \right\} \tag{30}$$

This gives a good approximation when this S is appropriate and the main interest is in the tests where $\phi = n_1 + n_2$ is not small. This is because then the probabilities of generating data (n_1, n_2) from the original model involving the unmodified Binomial distribution and such that $n_1 = 0$ and $n_2 = 0$, are then small. For example if $r = 0.5$ the probability that $n_1 = 0$ or $n_2 = 0$ is $2^{1-\phi}$ when $\phi \geq 1$, which rapidly decreases with ϕ.

10 Acknowledgements

I would like to thank Dr H. Borhan for providing me with the data prior to publication of his paper and for contributions to the manuscript, Mr. M.G. Links for doing preliminary analysis of the data and both these colleagues and Dr S. Karcz for comments resulting in considerable improvements to the manuscript. I am also grateful to the anonymous referees who made suggestions that greatly improved the manuscript.

11 References

References

Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. ,1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25, 3389–3402.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*,57, 289–300.

Chou, H.H., and Holmes M.H. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics.* 17, 1093–1104.

Käll, L., Storey, J.D., and Noble, W.S. 2008. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* 24, i42–i48.

Käll, L., Storey, J.D., and Noble, W.S. 2009. QVALITY: non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics* 25, 964–966.

Leung, Y.F., and Cavalieri, D. 2003. Fundamentals of cDNA microarray data analysis *Trends in Genetics* 19, 649–659.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., and Parvizi, B. *and others* 2003. TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics.* 19, 651–652.

Santner, T.J. and Duffy, D.E.1989. *The Statistical Analysis of Discrete Data* Springer-Verlag.

Simonoff, J. S. 1996. *Smoothing methods in statistics* New York: Springer-Verlag.

Storey, J.D., and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natn. Acad. Sci. USA* 100, 9440–9445.

Storey, J.D. 2007. The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc. B* ,69, 347–368.

Storey, J.D., Dai J.Y., and Leek J.T. 2005. The Optimal Discovery Procedure for Large-Scale Significance Testing, with Applications to Comparative Microarray Experiments. *University of Washington Biostatistics Working Paper Series. Paper* 260.

Zelterman D. 2006. *Models for Discrete Data: revised edition* Oxford University Press.