Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 68(2) 2014 303-314

Multiple Hypothesis Testing: A Review

Stefanie R. Austin[1], Isaac Dialsingh[2] and Naomi S. Altman[1]

[1]*Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111, USA*

[2]*Department of Mathematics and Statistics, University of the West Indies, St. Augustine Campus, Trinidad and Tobago*

Received 04 June 2014; Revised 21 June 2014; Accepted 22 June 2014

SUMMARY

Simultaneous inference was introduced as a statistical problem as early as the mid-twentieth century, and it has been recently revived due to advancements in technology that result in the increasing availability of data sets containing a high number of variables. This paper provides a review of some of the significant contributions made to the field of multiple hypothesis testing, and includes a discussion of some of the more recent issues being studied.

Keywords: Family-wise error rate, FWER, False discovery rate, FDR, Discrete test, Adaptive FDR, Simultaneous testing, Simultaneous inference.

1. INTRODUCTION

Data sets containing a high number of variables, notably those generated by high-throughput experiments in fields such as genomics and image analysis, have been becoming increasingly available as technology and research advances. For this reason multiple hypothesis testing remains an area of great interest. This review covers some of the major contributions to multiple hypothesis testing and provides a brief discussion on other issues surrounding the standard assumptions of simultaneous inference. This is not meant to be a comprehensive report but rather a history and overview of the topic.

1.1 Single Hypothesis

In the case of a single hypothesis, we typically test the null hypothesis H_0 versus an alternative hypothesis H_1 based on some statistic. We reject H_0 in favor of H_1 whenever the test statistic lies in the rejection region specified by some rejection rule. Here it is possible to make one of two types of errors: Type I and Type II. A Type I error, or false positive, occurs when we decide to reject the null hypothesis when it is in fact true. A Type II error, or false negative, occurs when we do not reject the null hypothesis when the alternative hypothesis is true. Table 1 summarizes the error possibilities.

Table 1. Possible outcomes for a single hypothesis test

	Declared True	Declared False
True Null	Correct ($1 - \alpha$)	Type I Error (α)
False Null	Type II Error (β)	Correct ($1 - \beta$)

Typically, a rejection region is chosen so as to limit the probability of a Type I error to some level α. Ideally, we also choose a test that offers the lowest probability of committing a Type II error, β, while still controlling

Corresponding author: Naomi S. Altman
E-mail address: nsa1@psu.edu

α at or below a certain level. In other words, we maximize power $(1 - \beta)$ while maintaining the Type I error probability at a desired level.

1.2 Multiple Hypotheses

When conducting multiple hypothesis tests, if we follow the same rejection rule independently for each test, the resulting probability of making at least one Type I error is substantially higher than the nominal level used for each test, particularly when the number of total tests m is large. This can be easily seen when considering the probability of making zero Type I errors. For m independent tests, if α is the rejection level for each p-value, then this probability becomes $(1 - \alpha)^m$. Because $0 < \alpha < 1$, it follows that

$$(1 - \alpha)^m < (1 - \alpha)$$

and so the probability of making no Type I errors in $m > 1$ tests is much smaller than in the case of one test. Consequently, the probability of making at least one such error in m tests is higher than in the case of one test. For example, we use a rejection rule of $p < .05$ for each of 100 total independent tests, the probability of making at least one Type I error is about 0.99.

To address this issue, multiple testing procedures seek to make the individual tests more conservative so as to minimize the number of Type I errors while maintaining an overall error rate, which we denote q. The cost of these procedures is often a reduction in the power of the individual tests. Tests are typically assumed to be independent, although there do exist methods in cases of dependency, which is discussed briefly in Section 4.1.

We assume that we are testing m independent null hypotheses, H_{01}, H_{02}, ..., H_{0m}, with corresponding p-values p_1, p_2, ..., p_m, and we call the i^{th} hypothesis "significant" if we reject the null hypothesis H_{0i}. In Table 2 we summarize the possible configurations when testing m hypotheses simultaneously. We see that V is the number of false rejections (or false discoveries), U is the number of true non-rejections (or true acceptances), S is the number of true rejections, and T is the number of false non-rejections. Here m_0, the total number of true null hypotheses, is fixed but unknown. Though random variables V, S, U, and T are not observable, the random variables $R = S + V$ and $W = U + T$, the number of significant and insignificant tests, respectively, are observable. The proportion of false rejections is V/R when $R > 0$ and the proportion of false acceptances is T/W when $W > 0$.

Table 2. Possible outcomes for m hypothesis tests

	Significant	Not Significant	Total
True Null	V	U	m_0
False Null	S	T	m_1
Total	R	W	m

The Type I error rates most discussed in the literature are:

1. Family-wise error rate (FWER): Probability of at least one Type I error,

   FWER = Prob($V \geq 1$)

2. False discovery rate (FDR): Expected proportion of false rejections,

   FDR = $E(Q)$, where $Q = \begin{cases} V/R & R > 0 \\ 0 & R = 0 \end{cases}$

2. CONTROLLING FAMILY-WISE ERROR RATE

The earliest multiple hypothesis adjustment methods focused on controlling the family-wise error rate (FWER), and these are still commonly used today. The FWER is defined as the probability of making at least one false rejection when all null hypotheses are true. Instead of controlling the probability of a Type I error at a set level for each test, these methods control the overall FWER at level q. The trade-off, however, is that they are often overly conservative, resulting in low-power tests.

Many of the methods in this class are based on the idea of ordered p-values. That is, prior to performing any adjustments, we first order the m p-values as $p_{(1)}$, $p_{(2)}$ ..., $p_{(m)}$ such that $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$, with corresponding null hypotheses $H_{0(1)}$, $H_{0(2)}$, ..., $H_{0(m)}$. Most procedures are then developed using either the first-order Bonferroni inequality or the Simes inequality (Shaffer 1995). The inequalities are very similar and can even be viewed as different formulations of the same concept.

2.1 Bonferroni Inequality

The first-order Bonferroni inequality states that, given any set of events E_1, E_2, ..., E_m, the probability

of at least one of the events occurring is less than or equal to the sum of their marginal probabilities (Shaffer 1995). In the context of multiple hypothesis testing, the event of interest is the rejection of a null hypothesis. The applicable form of the inequality then, for $0 \le \alpha \le 1$, is

$$\text{Prob}\left(\bigcup_{i=1}^{m} \left(p_i \le \frac{\alpha}{m} \right) \right) \le \alpha$$

The primary method based on this concept was proposed by Bonferroni, and it also happens to be the most popular among all procedures for controlling FWER. In its simplest form, to maintain the FWER at level q, set the nominal significance level for each test at $\alpha = q/m$ (Shaffer 1995). That is, for test i, if the corresponding p-value is $p_i < q/m$, we reject null hypothesis H_{0i}.

Others have also developed procedures around this idea. One such method includes a sequential, step-down algorithm proposed by Holm (1979), shown to be uniformly more powerful than Bonferroni's simple procedure. To maintain an error rate at level q, reject all null hypotheses in the set

$$\left\{ H_{0(i)} : i < \min\left(k : p_{(k)} > \frac{q}{m+1-k} \right) \right\}$$

Another suggestion for improvement is to replace the quantity α/m with $[1 - (1 - \alpha)^{1/m}]$, which is always a larger value (Shaffer 1995). This is a common idea used when developing procedures to control the false discovery rate.

2.2 Simes Inequality

Simes (1986) extended Bonferroni's inequality; in the context of multiple hypothesis testing, the Simes inequality can be stated the following way: for ordered p-values, $p_{(1)}$, $p_{(2)}$, …, $p_{(m)}$, corresponding to independent, continuous tests (so that the p-values are Uniform(0, 1)), then assuming all hypotheses are true:

$$\text{Prob}\left(p_{(i)} \ge i\frac{\alpha}{m} \right) = 1 - \alpha$$

where $0 \le \alpha \le 1$. Using this inequality, Simes created a simple multiple testing rule: to maintain an error rate at level q, reject all null hypotheses in the set

$$\left\{ H_{0(i)} : p_{(i)} \le \frac{i\alpha}{m} \right\}.$$

Two common methods that also utilize the Simes inequality were developed by Hochberg (1988) and Hommel (1988).

Hochberg's procedure is very similar to Holm's proposed method from Section 2.1, except it was formulated as a step-up procedure. It has also been shown to be more powerful than Holm's procedure. Again using the ordered p-values and maintaining the error rate level at q, reject all null hypotheses in the set

$$\left\{ H_{0(i)} : i \le \max\left(k : p_{(k)} \le \frac{q}{m+1-k} \right) \right\}$$

More powerful, and only marginally more difficult to execute, Hommel's (1988) procedure is an alternative, yet less popular, option. Under the same conditions as discussed in this section, to control at level q reject all null hypotheses:

1. Compute $k = \max \{ i \in \{1, ..., m\} : p_{(m-i+j)} > \frac{j\alpha}{i}$ for $j = 1, \cdots, i \}$.
2. If no maximum exists, then reject all null hypotheses. Else, reject $\{H_{0i} : p_i \le \alpha/k\}$.

3. CONTROLLING FALSE DISCOVERY RATE

More modern approaches in multiple hypothesis testing focus on controlling the false discovery rate (FDR). The FDR is defined as the expected percentage or proportion of rejected hypotheses that have been wrongly rejected (Benjamini and Hochberg 1995).

Instead of controlling the probability of a Type I error at a set level for each test, these methods control the overall FDR at level q. When all null hypotheses are actually true, the FDR is equivalent to the FWER. If, however, the number of true null hypotheses is less than the number of total hypotheses-that is, when $m_0 < m$ - the FDR is smaller than or equal to the FWER (Benjamini and Hochberg 1995). Thus, methods that control FWER will also control the FDR. We see, then, that controlling the FDR is a less stringent condition than controlling the FWER, and consequently FDR procedures are more powerful.

Controlling the FDR was made popular by Benjamini and Hochberg (1995), who developed a simple step-up procedure performed on the ordered p-

values of the tests (Benjamini and Hochberg 1995). Since then there have been several other proposed FDR procedures. These are summarized in this section.

3.1 Continuous Tests

The density of the p-values can be expressed as

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p)$$

where $f_0(p)$ and $f_1(p)$ are the densities of the p-values under the null and alternative hypotheses, respectively (Dialsingh 2012). For continuous tests, p-values are uniformly distributed on (0, 1) when the null is true. However, the distribution under the alternative hypothesis is unknown. Methods for estimating π_0 when the test statistics are continuous have been developed by coupling the mixture model with the assumption that either $f(p)$, the density of marginal p-values, or $f_1(p)$, the density of p-values under the alternative, is non-increasing.

The following is a summary of commonly-used methods for controlling FDR when the p-values are continuous. For all procedures, we assume that we are testing m independent null hypotheses, H_{01}, H_{02}, ..., H_{0m}, of which m_0 are truly null, with corresponding p-values, p_1, p_2, ..., p_m. Additionally, all methods here are based on ordered p-values. That is, instead of using the original, unordered p-values, we consider instead the ordered values, $p_{(1)}, p_{(2)}, ..., p_{(m)}$, such that $p_{(1)} \leq p_{(2)} \leq ... \leq p(m)$, with corresponding null hypotheses $H_{0(1)}$, $H_{0(2)}$, ..., $H_{0(m)}$.

3.1.1 Benjamini and Hochberg Procedure

Benjamini and Hochberg (1995) presented the first procedure for controlling FDR in their 1995 paper, and it still remains the most common procedure to date (the BH algorithm). To control FDR at level q, reject all null hypotheses where

$$\left\{ H_{0(i)} : i \leq \max\left(k : p_{(k)} \leq \frac{iq}{m} \right) \right\}$$

It has been shown that when the test statistics are continuous and independent, this procedure controls the FDR at level π_{0q} (Benjamini and Hochberg 1995), where π_0 is the proportion of true null hypotheses. Ferreira and Zwinderman (2006) later developed some exact and asymptotic properties of the rejection behavior of the BH algorithm.

3.1.2 Benjamini and Liu Procedure

While the BH algorithm is a step-up procedure, Benjamini and Liu (1999) suggested an alternative step-down procedure for controlling FDR (the BL algorithm). To control FDR at level q, the procedure is conducted as follows:

1. Calculate the critical values, $\delta_i = 1 - \left[1 - \min\left(\frac{mq}{m-i+1} \right) \right]^{1/(m-i+1)}$ for $i = 1, ..., m$.
2. Let k be the value such that $k = min\{i : p(i) > \delta_i\}$.
3. Reject the null hypotheses $H_{0(1)}$, $H_{0(2)}$, ..., $H_{0(k-1)}$.

They demonstrated that this procedure neither dominates nor is dominated by the step-up procedure of Benjamini and Hochberg.

3.1.3 Storey's Procedure

Storey (2002) suggests a different approach to adjusting for multiple hypotheses. While the previous methods involved fixing an FDR level q and determining from there which tests to reject, Storey uses the opposite approach: he fixes which tests are rejected (in a sequential way) and then estimates the corresponding false discovery rate. The basic idea of Storey's procedure is as follows:

1. Define a set of rejection regions, $\{[0, \gamma_j]\}$. One easy way to do this is to let $\gamma_i = p_{(i)}$, or the series of ordered p-values. Then, for γ_i, the rejection region is $\{p_{(1)}, ..., p_{(i)}\}$.
2. For each rejection region, estimate the FDR. This will lead to a series of FDR estimates, $\{\widehat{FDR_j}\}$.
3. Choose the rejection region that provides an acceptable estimate of FDR.

Storey's approach can also be used by estimating a variation on the FDR: the positive FDR (pFDR), the false discovery rate conditional on nonzero rejections: $E(V/R|R > 0)$ (Storey 2003). This is often a more interpretable and easily estimable value. In his paper, Storey proposed that an estimate of pFDR for a given γ_j is $(\gamma_j \hat{m}_0) / R(\gamma_j)$, where $\hat{m}_0$ is an estimate of the true number of null hypotheses, and $R(\gamma_j) = \#(p \leq \gamma_j)$ is the number of tests that would be rejected for the given rejection region. Further discussion on using m_0 instead of m is given in Section 3.3.

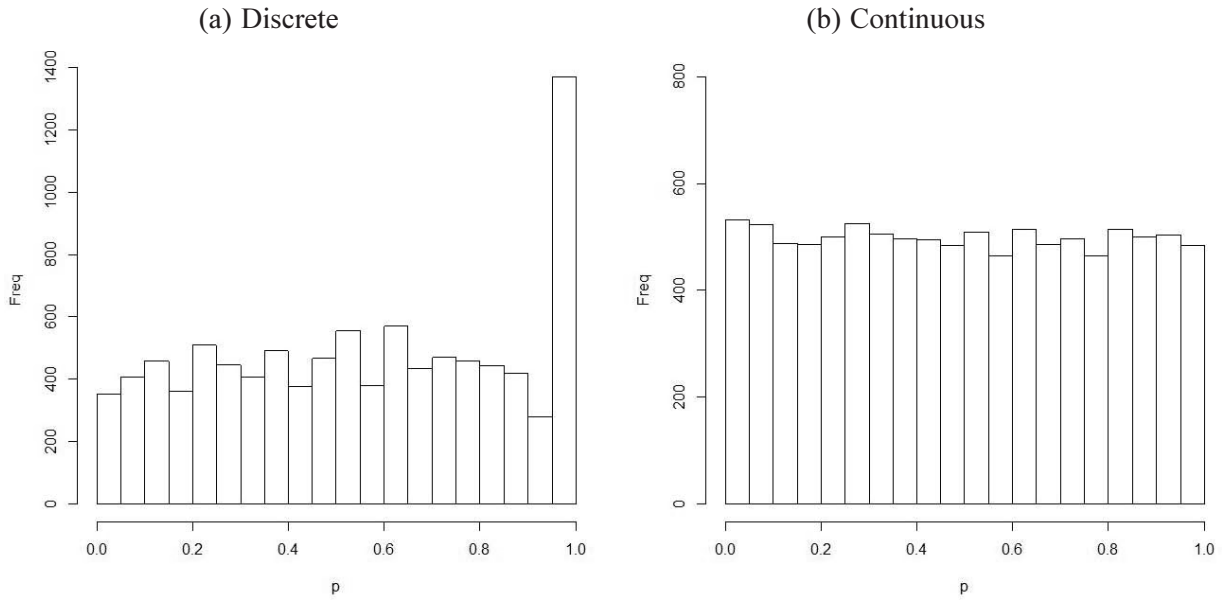

Fig. 1. Histograms of p-values coming from m = 10,000 tests, all of which correspond to true null hypotheses.

3.2 Discrete Tests

Most research to date has been dedicated to the case of continuous data. In these situations, the resulting test statistics are continuous with known distributions when the null hypothesis is true. As well, the p-values are continuous and known to follow a Uniform (0, 1) distribution under the null hypothesis. For discrete data, however, this is no longer the case. Non parametric tests, such as Fisher's exact tests, lead to p-values that are discrete and non-uniform. To illustrate this point, we create histograms of p-values that come from m = 10,000 tests, all of which correspond to a true null hypothesis, as shown in Fig. 1. Note that in the continuous case, the observed p-values form a near-uniform distribution. However, in the case of discrete data, we are far from uniform and in fact see a peak at $p = 1$.

Furthermore, the distribution of achievable p-values of a given discrete test is dependent on the ancillary statistic. As a result, in the case of multiple hypotheses, the distribution of p-values will vary by test. Consequently, the use of a subscript becomes necessary in the mixture model from Section 3.1 to highlight this difference. The model can be rewritten as

$$f_i(p_i) = \pi_0 f_{0i}(p_i) + (1 - \pi_0) f_{1i}(p_i)$$

where $f_i(p_i)$ is the density of the i^{th} observed p-value and $f_{0i}(p_i)$ and $f_{1i}(p_i)$ are the null and alternative densities of the i^{th} p-value, respectively. One can immediately see the potential problems with having unique distributions for each test. Recently more focus has been given to the situation of discrete testing, though the topic has yet to be as extensively explored.

Using *midP*-values instead of p-values in the BH algorithm was the first suggestion for addressing multiple testing for discrete data. Lancaster (1961) defined the *midP*-values as the average of the observed and the next smallest possible p-value. The distribution of the *midP*-value is more uniform under the null hypothesis than is the p-value, and using the *midP*-value should lead to results that are at least as powerful as when using the p-value, but also may exceed the nominal FDR level. This idea is discussed further by Routledge (1994), Berry and Armitage (1995), and Fellows (2010).

3.3 Adaptive Procedures

Estimating $m_0 = \pi_0 m$, the number of true null hypotheses, can improve FDR procedures by making them more powerful. When replacing m by m_0 in the BH or the BL algorithm we can control the FDR at exactly the level of q. When using $\hat{m}_0$ instead of m in an FDR procedure, we call these "adaptive" methods. The value of π_0, and thus of m_0, can be estimated using a variety of methods. The idea behind many of the proposed methods were first introduced by Mantel (1980), and Black (2004) provides a nice discussion of the benefits of using an adaptive method.

3.3.1 Estimating m_0 for Continuous Tests

The following is a summary of commonly-used methods used for estimating m_0 (or π_0) when the p-values are continuous. In addition to those given below, continuous estimators have also been proposed in the form of adaptive algorithms developed by Benjamini and Hochberg (2000), Storey *et. al*. (2004), Benjamini *et. al*. (2006), Gavrilov *et. al*. (2009), Blanchard and Roquain (2009), or Liu and Sarkar (2011).

Storey's Method: Storey's (2002) method is one of the most popular methods used today, and has been shown to estimate π_0 reasonably well for continuous p-values. The estimator is given by

$$\hat{\pi}_0 = \frac{\#(p_i > \lambda)}{m(1-\lambda)}$$

where $\lambda \in [0, 1]$ is a tuning parameter and #(S) is the number of elements in *S*. Storey also offers an adaptive method for selecting an optimal λ (Storey 2002). This procedure typically provides a conservative estimate of π_0.

Pounds and Cheng Method: Pounds and Cheng (2006) proposed an estimator of π_0 when the test statistics are continuous. The estimator is given by

$$\hat{\pi}_0 = \begin{cases} \min(1, 2\bar{p}) & \text{for two-sided tests} \\ \min(1, 2\bar{t}) & \text{for one-sided tests} \end{cases}$$

where $\bar{p} = \frac{1}{m}\sum_{i=1}^{m} p_i$ is the average p-value for all m hypothesis tests, $\bar{t} = \frac{1}{m}\sum_{i=1}^{m}[2.\min(p_i, 1-p_i)]$, and $\min(a, b)$ is the minimum of a and b. In general, the Pounds and Cheng estimator is biased upward, but the bias is small when $pf_1(p)$ is small or when π_0 is close to 1.

Location Based Estimator: Dalmasso *et. al*. (2005) proposed an estimator for π_0 for continuous and independent tests, which they coined the Location Based Estimator (LBE). The LBE is simple estimator that is obtained from the expectation of transformed p-values, using the transformation $\psi(p) = [-\log(1 - p)]^n$, where log() is the natural logarithm function and $n \geq 0$ is an integer tuning parameter. Taking the ratio of the expected value of ψ under the alternative and null hypotheses gives us the following estimator:

$$\hat{\pi}_0 = \frac{(1/m)\sum_{i=1}^{m}[-\log(1-p_i)]^n}{n!}$$

where $n!$ is the factorial of n. In their paper, Dalmasso *et. al*. does provide one example of how to select the tuning parameter n but notes that other criteria could be considered. The LBE provides a bias-variance balance and, because of its relatively low variance, it often performs better in terms of mean-squared error than π_0 estimators that had been developed by that time, including the Storey and Pounds and Cheng methods.

Nettleton's Method: Nettleton *et. al*. (2006) presents an algorithm for estimating m_0 by estimating the proportion of observed p-values that follow the uniform distribution. The algorithm used in Nettleton's method is as follows:

1. Partition the interval [0, 1] in B bins of equal width.
2. Assume all null hypotheses are true, and set $m_0^{(0)} = \pi_0^{(0)} m = m.$
3. Calculate the expected number of p-values for each bin given the current estimate of the number of true null hypotheses.
4. Beginning with the leftmost bin, sum the number of p-values in excess of the expected until a bin with no excess is reached.
5. Use the excess sum as an updated estimate of m_1, and then use that to update the estimate of $m_0 = m - m_1$.
6. Return to Step 3 and repeat the procedure until convergence is reached.

The number of bins is a tuning parameter, and using $B = 20$ has been recommended (Nettleton *et al*. 2006).

3.3.2 Estimating m_0 for Discrete Tests

There do not exist FDR procedures designed specifically for discrete data, other than the use of the *midP*-values. However, the non-uniformity of the p-values may be addressed in the first step of estimating m_0 and then utilizing an adaptive FDR method. We

maintain the assumptions and notations used thus far. However, now we further assume that π_0 does not depend on the set of null distributions, of which there are d unique discrete distributions ($d \le m$), f_{01}, f_{02}, ..., f_{0d}, which are known. These distributions correspond to the d unique ancillary statistics A_1, A_2, ..., A_d (each of which can be viewed as fixed or random). For a null distribution f_{0j} there is a finite set of achievable p-values S_{j1}, S_{j2}, ..., S_{jT_j} with $S_{j1} < S_{j2} < \cdots S_{jT_j} = 1$ and with corresponding probabilities s_{j1}, s_{j2}, ..., s_{jT_j}. Note that set $S_j = \{S_{j1}, S_{j2}, ..., S_{jT_j}\}$, is the support of f_{0j}, with $s_{j1} = S_{j1}$ and $s_{jk} = S_{jk} - S_{j,\,k-1}$ for $2 \le k \le T_j$. The hypotheses are partitioned into sets so that if the null distribution of the i^{th} p-value is known to be f_{0j}, then the corresponding support is S_j. However, the distribution of p_i is not known when the corresponding null hypothesis is false.

The following is a summary of commonly-used methods used for estimating the number or proportion of true null hypotheses when the p-values are discrete. The performances of continuous and discrete m_0 estimators in the presence of discrete data are explored in Dialsingh (2012) and Austin (2014).

Pounds and Cheng Method: In the same paper as their 2006 continuous estimator, Pounds and Cheng (2006) also proposed an estimator of π_0 for the discrete case. Similar to the continuous estimator, the discrete estimator is given by

$$\hat{\pi}_0 = \begin{cases} \min(1, 2\bar{p}) & \text{For two-sided tests} \\ \min(1, 8\bar{t}) & \text{For one-sided tests} \end{cases}$$

where $\bar{p} = \frac{1}{m}\sum_{i=1}^{m} p_i$ is the average p-value for all m hypothesis tests, $\bar{t} = \frac{1}{m}\sum_{i=1}^{m}[2\min(p_i, 1-p_i)]$, and $\min(a, b)$ is the minimum of a and b. Simulations show that this estimator is conservative but robust for discrete tests.

Regression Method: Proposed by Dialsingh (2012), the regression method can be used when the mixture distribution, $\text{Prob}(p_i = S_{it} | A_i = a)$, can be estimated from the data. For each of the d unique null distributions, there exists a finite set of achievable p-values. So, for null distribution f_{0j}, there is a known support S_j. With a slight abuse of notation, we say that $H_{0i} \in f_{0j}$ if the i^{th} null hypothesis is assumed to have distribution f_{0j}. Then we have ($p_i = S_{jt} | H_{0i} \in f_{0j}$, H_{0i} true) = ϕ_{0jt}, which is known. However, the distribution of p_i is not known when the null hypothesis is false; we denote this unknown probability as Prob ($p_i = S_{jt} | H_{0i} \in f_{0j}$, H_{0i} false) = ϕ_{1jt}. We assume that $\phi_{0jt} < \phi_{1jt}$ for small S_{jt} and $\phi_{0jt} > \phi_{1jt}$ for large S_{jt}. Then we have

$$\text{Prob}\,(p_i = S_{jt} | H_{0j} \in f_{0j}) = \phi_{jt} = \pi_0\phi_{0jt} + (1 - \pi_0)\phi_{1jt}$$

When the set $D_j = \{H_{0i} : H_{0i} \in f_{0j}\}$ is significantly large, ϕ_{jt} can be estimated from the data as

$$\hat{\phi}_{jt} = \frac{K_{jt}}{M_j}$$

where M_j is the cardinality of set D_j and K_{jt} is the number of hypotheses in D_j that have p-value S_{jt}. We know that $E(\hat{\phi}_{jt}) = \phi_{jt} = \pi_0\phi_{0jt} + (1 - \pi_0)\phi_{1jt}$ and that ϕ_{0jt} is known. The regression method estimates π_0 by regressing $\hat{\phi}_{jt}$ on ϕ_{0jt} by assuming $(1 - \pi_0)\phi_{1jt}$ is the constant intercept. Thus the slope of the resulting regression equation is an estimator of π_0. To obtain reasonable estimates of $\hat{\phi}_{jt}$ it is preferred that each M_j is sufficiently large.

Bancroft Method: The method developed by Bancroft *et. al*. (2013) is an adaptation of Nettleton's method for continuous tests to discrete cases. Similar to Nettleton's method, the idea is to create bins in the interval [0, 1] and to use the excess of expected versus observed p-values in those bins to iteratively update the estimate of m_0. However, because the possible attainable p-values depend on the null distribution, we no longer look at entire set of m p-values but rather at each of the d sets of p-values corresponding to the d unique null distributions, f_{01}, f_{02}, ..., f_{0d}. Nettleton's algorithm can be applied to each set of p-values separately to come up with an estimate of m_{0j}, the number of tests corresponding to true null hypotheses in set D_j. Note that the initial estimate of m_{0j} would be M_j, the total number of tests in D_j.

Furthermore, because the p-values of discrete tests are not uniformly distributed over [0, 1], bins need not and should not be of equal width. Instead, since the support S_j is known for each unique null distribution f_{0j}, bins can be created such that each bin houses a single value from S_j. Then, the same algorithm from Nettleton's method is applied to these bins.

The algorithm is run d instances to find estimates of m_{01}, m_{02}, ..., m_{0j}. Then, the estimate of π_0 becomes

$$\hat{\pi}_0 = \frac{\hat{m}_{01} + \hat{m}_{02} + ... + \hat{m}_{0j}}{m}$$

T-Methods: The T-methods, proposed by Dialsingh (2012), are based on Tarone's (1990) idea of removing hypotheses for which there is no power prior to performing any analyses. For certain values of the ancillary statistic, the number of achievable p-values is small, yielding a component of the null distribution that is far from uniform. Often the corresponding hypothesis tests have zero power because the minimum achievable p-value is larger than the boundary of the rejection region, say $\alpha = .05$. Because filtering out these tests improves the uniformity of the p-values, Tarone suggested removing these tests to improve the power of multiple comparison adjustments. The remaining tests are then used to estimate π_0 developed for continuous p-values.

3.3.3 Gilbert's Procedure

Analogous to T-methods for estimating π_0, Gilbert (2005) developed a procedure using the idea that multiplicity adjustments do not need to account for hypothesis tests that have no power. Gilbert's procedure uses the BH algorithm on only a subset of the tests by removing the tests whose minimum achievable p-value is less than q. To control FDR at level q, Gilbert's procedure is conducted as follows:

1. Let $m(I)$ be the number of tests with power, with corresponding or-dered p-values

   $P_{(1)}, p_{(2)}, ..., p_{(m(I))}$ and null hypotheses $H_{0(1)}, H_{0(2)}, ..., H_{0(m(I))}$.

2. Apply the BH algorithm on only these $m(I)$ tests.

One can perform an adaptive version of Gilbert's procedure as well, as suggested in Dialsingh (2012). More recently, Heyse (2011) contributed an alternative multiple-testing procedure for categorical data that uses the exact conditional distribution of potential outcomes.

3.4 Variations on the FDR

There exists a number of variations on the false discovery rate as defined by Benjamini and Hochberg. In Section 3.1.3 we discussed the positive FDR (*p*FDR) as used by Storey. In the same year Tsai *et. al*. (2003) discussed the properties and relationships of several variations on the FDR. Including the *p*FDR, these alternatives included:

1. Conditional FDR (*c*FDR): *c*FDR = $E(V/R|R = r)$
2. Marginal FDR (*m*FDR): *m*FDR = $E(V)/E(R)$
3. Empirical FDR (*e*FDR): *e*FDR = $E(V)/r$

Pounds and Cheng (2004) developed a method, coined the spacings LOESS histogram (SPLOSH), for estimating the *c*FDR, the expected pro-portion of false rejections given that there are r total rejections. The method was applied to independent, continuous tests; however, there is no model assumed on the observed p-values and thus this method may be applicable to other situations.

The local FDR (*l*FDR), coined by Efron (2005), estimates the probability of the null model conditional on the observed test statistic. He applied it only to continuous tests. It is based on empirical Bayes analysis of the mixture model of the null and alternative hypothesis distributions. It has been studied by Efron *et. al*. (2001), Efron and Tibshirani (2002), Efron (2004), and Strimmer (2008).

4. OTHER CONSIDERATIONS

4.1 Dependent Tests

In all methods discussed thus far there is an assumption of independence among the tests. However, in many cases, such as when comparing several treatments with a single control, this assumption is not valid. In fact, in practice, dependent tests statistics are encountered more often than independent test statistics. Therefore, multiple testing adjustment when the hypotheses are dependent remains an open area of research. Schwartzman and Lin (2011) discussed the effect of dependency and derived approximation for the mean, variance, and distribution of the false discovery rate when the data are correlated. They showed that correlation may increase the bias and variance of the false discovery estimator.

The common BH and BL algorithms can still be used for dependent tests under certain situations. Benjamini and Yekutieli (2001) proved that the BH algorithm controls FDR when the test statistics are positive dependent (Two random variables X_1 and X_2 are positive dependent if $P(X_1 \cap X_2) > P(X_1)P(X_2)$) under the null hypotheses. They furthermore showed that under other kinds of dependency, a small

conservative modification can be be made to maintain the nominal error rate. The following year Sarkar (2002) showed that the FDR is also controlled in the BL procedure if the test statistics are positive dependent (Sarkar 2002). In 2008 Sarkar further investigated the performance of common FDR procedures under positive dependence (Sarkar 2008).

Most of the FWER-controlling procedures can be used in certain dependency situations. Sarkar and Chang (1997) showed that the Simes method still controls the FWER under positive dependency of the test statistics. Later, Guo *et al*. (2009) developed FWER-controlling methods based on adaptive Holm and Hochberg procedures and proved that they control the FWER under positive dependence. That same year Sun and Cai (2009) developed a multiple-testing procedure that exploits the dependence structure among hypotheses assuming that the data were generated from a two-stage hidden Markov model.

Friguet *et. al.* (2009) utilized factor analysis to model the dependence structure of the responses given the predictors. They used this to develop modified t-tests that take advantage of the common factor structure to reduce the error rate variance. They showed that this procedure is more powerful and results in more precise FDR estimates than the traditional BH algorithm when used on dependent tests.

4.1.1 Pairwise Comparisons

A more specific dependence issue in multiple testing arises when each hypothesis is composite and one tests for differences across all or many parameter pairs. There has been extensive study on these types of pairwise procedures for a single response variable, and Jaccard *et al.* (1984) provides a nice overview of procedures developed until that time. They found that Tukey (1953) is one of the best methods when the assumptions of equal sample sizes, homogeneous variances, and normality hold. When one or more assumptions fail, they recommend procedures by Kramer (1956) and Games *et al*. (1981). All of these are considered one-step procedures that control for multiple testing when there is a single response variable. Jiang and Doerge (2006) proposed a two-step procedure when there are many composite hypotheses, with many pairwise comparisons of parameters within each, that controlled the FDR and was more powerful than traditional one-step procedures.

4.1.2 Multivariate Test Statistics

When handling multiple hypotheses, the first major step is to rank the tests in order of perceived significance. This is typically done in a univariate matter, through ordering the p-values of the individual tests. However, as shown by Storey (2007), overall performance can be improved by borrowing information across all tests to perform the ranking. Storey referred to this as the "optimal discovery procedure," which maximizes the expected number of true rejections, which in turn results in a reduced FDR. A generalized Bayesian discovery procedure was later developed by Guindani *et al*. (2009).

Chi (2008) also considered the idea of multivariate p-values. He created a procedure that uses an arbitrary family of nested regions of the p-values' domain and showed that this controlled the FDR and *p*FDR and was relatively powerful.

4.2 Weighted P-Values

The idea of weighted p-values was first presented by Genovese *et al*. (2006) as a way to incorporate prior information about the hypotheses. If the assignment of weights to p-values is positively associated with the non-null hypotheses, power is increased, except in cases where the power is already near one. Wasserman and Roeder (2009) derived methods for choosing optimal weights and showed that power is robust to misspecification of the p-value weights. Roquain and van de Wiel (2009) further discussed the weighting of p-values in the Benjamini and Hochberg procedure.

4.3 Power

As discussed in Section 1.2, error can be defined in a number of ways, though most modern procedures focus on controlling the false discovery rate. Power, too, can be defined in more than one way. Most researchers defined power as the average probability of rejecting a non-null hypothesis, otherwise known as the per pair power (Einot and Gabriel 1975). However, other power definitions include the probability of rejecting at least one non-null hypothesis, or any-pair power, and the probability of rejecting all false hypotheses, or all-pairs power (Ramsey 1978).

5. CONCLUSION

This review serves as a brief introduction to the topic and its many issues. Though the problem of simultaneous inference has been recognized for many years, multiple hypothesis testing has recently became a more intense area of research due to the greater access and availability of data. The methods focus on controlling some Type I error rate while maintaining power of the individual tests. There remain many potential areas of research in this field, including the problems surrounding discrete data or dependent tests.

5.1 Topics Not Covered

The topic of multiple hypothesis testing is too far-reaching to be covered in its entirety in this review, and there are many researchers whose contributions have not been acknowledged here. The following is a list of some of the multiple testing topics not covered in this review: simultaneous confidence intervals; binary data; one-sided tests; generalized linear and mixture models; survival analysis; linear contrasts; bootstrapping and resampling; Bayesian methods; decision theory methods; methods based on Central Limit Theorem; projection methods; and significant analysis of microarrays (SAM) and other nonparametric methods. Many of these topics are covered in the texts listed in the next section.

ACKNOWLEDGEMENTS

Naomi S. Altman acknowledges partial support from NSF DMS 1007801 and from NIH UL1RR033184. Naomi Altman and Isaac Dialsingh also acknowledge partial support from NSF IOS 0820729.

REFERENCES

Austin, S.R. (2014). Multiple hypothesis testing of discrete data. Master's thesis, The Pennsylvania State University.

Bancroft, T., Du, C. and Nettleton, D. (2013). Estimation of false discovery rate using sequential permutation p-values. *Biometrics,* **69**, 1-7.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.*, **B57**, 289-300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Behavioral Edu. Statist.,* **25,** 60-83.

Benjamini, Y., Krieger, A.M. and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, **93,** 491-507.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inf.*, **82,** 163-170.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Ann. Statist.*, **29,** 1165-1188.

Berry, G. and Armitage, P. (1995). Mid-p confidence intervals: A brief review. *J. Roy. Statist. Soc.,* **D44**, 417-423.

Black, M.A. (2004). A note on the adaptive control of false discovery rates. *J. Roy. Statist. Soc.,* **B66,** 297-304.

Blanchard, G. and Roquian, E. (2009). Adaptive false discovery rate control under independence and dependence. *J. Machine Learning Res.,* **10,** 2837-2871.

Bretz, F., Hothorn, T. and Westfall, P. (2010). *CRC Press*, 2 ed. Springer, New York.

Chi, Z. (2008). False discovery rate control with multivariate p-values. *Elect. J. Statist.,* **2,** 368-411.

Curran-Everett, D. (2001). Multiple comparisons: philosophies and illustrations. *Amer. J. Physiol. -Regulatory, Integrative and Comparative Physiology,* **279**, R1-R8.

Dalmasso, C., Broet, P. and Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics,* **21,** 660-668.

Dialsingh, I. (2012). False discovery rates when the statistics are discrete. PhD thesis, The Pennsylvania State University.

Dickhaus, T. (2014). *Simultaneous Statistic Inference: with Applications in the Life Sciences*. Springer, Berlin Heidelberg.

Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.,* **18**, 71-103.

Dudoit, S. and van der Laan, M.J. (2007). *Multiple Testing Procedures with Applications to Genomics.* Springer.

Efron, B. (2004). Large-scale simultaneous hypothesis testing. *J. Amer. Statist. Assoc.*, **99,** 96-104.

Efron, B. (2005). Local false discovery rates. Technical Report, Stanford University.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.,* **23**, 70-86.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, **96,** 1151-1160.

Einot, I. and Gabriel, K.R. (1975). A study of the powers of several methods of multiple comparisons. *J. Amer. Statist. Assoc.*, **70,** 574-583.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Statist. *Methods Medical Res.,* **17,** 347-388.

Fellows, I. (2010). The minimaxity of the mid p-value under linear and squared loss functions. *Comm. Statist. -Theory Methods,* **40,** 244-254.

Ferreira, J.A. and Zwinderman, A.H. (2006). On the Benjamini-Hochberg method. *The Ann. Statist.,* **34**, 1827-1849.

Friguet, C., Kloareg, M. and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.*, **104**, 1406-1415.

Games, P.A., Keselman, H.J. and Rogan, J.C. (1981). Simultaneous pairwise multiple comparison procedures for means when sample sizes are unequal. *Psychological Bull.,* **90**, 594-598.

Gavrilov, Y., Benjamini, Y. and Sarkar, S.K. (2009). An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.,* **37,** 619-629.

Genovese, C.R., Roeder, K. and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika,* **93,** 509-524.

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

Guindani, M., Muller, P. and Zhang, S. (2009). A Bayesian discovery procedure. *J. Roy. Statist. Soc.,* **B71,** 905-925.

Guo, W., Sarkar, S.K. and Peddada, S.D. (2009). Adaptive multiple testing procedures under positive dependence. Technical Report., Temple University.

Heyse, J.F. (2001). A false discovery rate procedure for categorical data. In: *Recent Advancements in Biostatistics*, vol. 4. World Scientific Publishing Co., **3,** 43-58.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika,* **75,** 800-803.

Hochberg, Y. and Tamhane, A.C. (2009). *Multiple Comparison Procedures*. Wiley.

Holm, S. (1997). A simple sequentially rejective multiple test procedure. *Scandinavian J. Statist.,* **6**, 65-70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika,* **75,** 383-386.

Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*. CRC Press.

Jaccard, J., Becker, M.A. and Wood, G. (1984). Pairwise multiple comparison procedures: a review. *Psychological Bull.,* **96,** 589-596.

Jiang, H. and Doerge, R.W. (2006). A two-step multiple comparison procedure for a large number of tests and multiple treatments. *Statist. Appl. Genet. Mole. Biol.,* **5**.

Kramer, C.Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics,* **12,** 307-310.

Lancaster, H. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.,* **56,** 223-234.

Liu, F. and Sarkar, S.K. (2011). A new adaptive method to control the false discovery rate. Technical Report, Temple University.

Liu, W. (2010). *Simultaneous Inference in Regression*. CRC Press.

Mantel, N. (1980). Assessing laboratory evidence for neoplastic activity. *Biometrics,* **36,** 381-399.

Miller, R.G.J. (2011). S*imultaneous Statistical Inference*, 2 ed. Springer, New York.

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. *J. Agric. Biol. Environ. Statist.,* **11,** 337-356.

Pan, W. (2001). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18,** 546-554.

Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics,* **20,** 1737-1745.

Pounds, S. and Cheng, C. (2006). Robust estimation of the false discovery rate. *Bioinformatics,* **22,** 1979-1987.

Ramsey, P.H. (1978). Power differences between pairwise multiple comparisons. *J. Amer. Statist. Assoc.,* **73**, 479- 485.

Roquain, E. (2009). Optimal weighting for false discovery rate control. *Elect. J. Statist.,* **3,** 678-711.

Routledge, R. (1994). Practicing safe statistics with the *mid*-p. *Canad. J. Statist.,* **22,** 103-110.

Sarkar, S.K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Ann. Statist.,* **30,** 239-257.

Sarkar, S.K. (2008). On methods controlling the false discovery rate. *The Ind. J. Statist.,* **A70**, 135-168.

Sarkar, S.K. and Chang, C. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.,* **92,** 1601-1608.

Schwartzman, A. and Lin, X. (2001). The effect of correlation in false discovery rate estimation. *Biometrika,* **98,** 199-214.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Rev. Psychol.*, **46,** 561-584.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. 751-754.

Storey, J.D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc.,* **B64**, 479-498.

Storey, J.D. (2003). The positive false discovery rate. *Ann. Statist.,* **31,** 2013-2035.

Storey, J.D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. Roy. Statist. Soc.,* **B69**, 347-368.

Storey, J.D., Taylor, J.E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist., Soc.,* **B66,** 187-205.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, **9,** 303.

Sun, W. and Cai, T.T. (2009). Large-scale multiple testing under dependence. *J. Roy. Statist. Soc.,* **B71**, 393-424.

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

Toothaker, L.E. (1993). *Multiple Comparison Procedures.* SAGE.

Tsai, C., Hsueh, H. and Chen, J.J. (2003). Estimation of false discovery rates in multiple testing: Application to gene microarray data. *Biometrics,* **59**, 1071-1081.

Tukey, J.W. (1953). The problem of multiple comparisons. Technical Report, Princeton University.

Wasserman, L. and Roeder, K. (2009). Genome-side significance levels and weighted hypothesis testing. *Statist. Sci.,* **24,** 398-413.

Westfall, P.H., Tobias, R.D. and Wolfinger, R.D. (2011). *Multiple Comparisons and Multiple Tests Using SAS*, 2 ed. SAS Institute.