Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

MULTIPLE COMPARISONS WITH A STANDARD USING FALSE DISCOVERY RATES

Dashi I. Singham
Roberto Szechtman

Department of Operations Research
Naval Postgraduate School
1 University Circle
Monterey, CA 93943, USA

ABSTRACT

We introduce a new framework for performing multiple comparisons with a standard when simulation models are available to estimate the performance of many different systems. In this setting, a large proportion of the systems have mean performance from some known null distribution, and the goal is to select alternative systems whose means are different from that of the null distribution. We employ empirical Bayes ideas to achieve a bound on the false discovery rate (proportion of selected systems from the null distribution) and a desired probability an alternate type system is selected.

1 INTRODUCTION

One objective of simulation optimization procedures is to select the best system when simulation models are used to estimate performance of different systems. Similar procedures exist for choosing the best subset of systems, or choosing systems that are better than some standard. We consider a large number of systems, where the goal is to choose a subset of systems that have particularly low (or high) mean performance relative to some standard. The majority of systems have output which can be categorized according to some null probability distribution with a known mean. An alternate distribution models the output from systems with smaller means, and we attempt to select these systems and classify them as non-null, or "alternate". The goal is to divide the systems into "null" and "alternate" to further study the systems deemed alternate.

This work is inspired by recent literature on large scale inference using empirical Bayes methods in Efron (2010). The methods are motivated by situations where one wishes to test a large number of hypotheses, and it is presumed that some proportion of these hypotheses are "null" while the remaining hypotheses are non-null. In Efron (2010), this setup is explored with applications to biological systems in microarray testing, where there are a large number of gene expression levels to be tested. The goal is to isolate the few likely to be significant for further investigation. We parallel these ideas to the problem of multiple comparisons with a standard when the data for many different systems is simulated. This problem involves selecting the systems that are different than the standard according to some probability guarantees on the proportion of null systems included in the selection, and the probability that an alternate is included.

One method for large scale testing of hypotheses is the Benjamini-Hochberg algorithm (Benjamini and Hochberg 1995). This algorithm sorts the data for each system according to p-values associated with the hypothesis test, and determines a threshold for selecting systems as alternate. The false discovery rate (FDR) is the probability that a selected system is actually null. This is in contrast to the typical Type I error, which is the probability that a null system is selected as an alternate. The algorithm ensures that the expected false discovery rate (Efdr) is below some chosen value. The idea is that controlling the Efdr helps select a subset where most of the systems are alternate, so that future effort investigating these systems is not wasted on many null systems. Generally, choosing the subset according to Efdr can lead to a larger

subset and more alternates included than rules relying on Bonferroni bounds to control the probability that any null system is included. By allowing some proportion of selected systems to be null, more alternates can be identified.

We develop a new algorithm for studying the problem of multiple comparisons with a standard that can be applied in simulation settings. We assume i.i.d. normally distributed data where the null systems have a known mean but unknown variance. This algorithm is a threshold policy that selects systems with p-values smaller than some critical value. In addition to controlling the Efdr as the Benjamini-Hochberg algorithm does, the algorithm controls the probability that an alternate system will be correctly selected (analogous to the power of a test). We exploit the simulation environment and determine the number of samples that must be collected for each system to achieve these metrics. These sample sizes depend on the variance of the systems and the difference in means needed to distinguish alternates from null systems.

Section 2 briefly reviews the simulation literature dealing with subset selection problems and Section 3 summarizes the Benjamini-Hochberg algorithm. Section 4 describes the framework and derivation of sample sizes needed for our algorithm and Section 5 details the algorithm implementation and numerical results. Finally, Section 6 concludes.

2 LITERATURE REVIEW

Our motivation parallels that of the classical problem of subset selection and comparison with a standard, and we briefly review past work in these areas. In the subset selection problem the objective is to select multiple systems with the best performance means, or select a subset that includes the best system. In the multiple comparisons with a standard problem, the goal is to identify systems that are better than some standard, or choose the best system where special consideration is given to one system that is considered standard.

Algorithms for choosing a subset of systems have different objectives. One is to maximize the probability of choosing the top m systems (Chen, He, Fu, and Lee 2008), while others wish to choose a subset of size m that contains some number of the best systems (Koenig and Law 1985). In Ryzhov and Powell (2009), possible subsets are treated as individual alternatives and sampling rules applied to choose the best subset. Chingcuanco and Osorio (2013) consider an opportunity cost function as an objective rather than a zero-one function for assessing whether the correct systems have been included. A common objective is to identify all systems whose means are within a specified distance from the mean of the best system. Corlu and Biller (2013) consider this problem while using historical data to quantify input uncertainty associated with each system.

Many algorithms exist for performing multiple comparisons with a standard to determine the best system, where special consideration is given to some standard system. The first method for comparing multiple systems with a standard is in Paulson (1952). The algorithm guarantees with some specified probability that the standard is selected if it is actually the best, and delivers a sample size required to determine that another system is the best with respect to the standard. This work, as well as many recent algorithms, assume i.i.d. normal observations from each system. Nelson and Goldsman (2001) handle different variances across systems using a two-stage procedure where the first stage estimates the variance and determines the sampling effort required in the second stage. Kim (2005) employs a fully sequential procedure and is able to accommodate a much larger number of systems. Xie and Frazier (2013) employ a dynamic programming solution to sequentially decide where to sample based on different sampling costs, and accommodate a general class of payoff functions for correct and incorrect selections.

3 BENJAMINI-HOCHBERG ALGORITHM

This section outlines the Benjamini-Hochberg (BH) algorithm as an example of empirical Bayes methods. These methods involve testing a large number of hypotheses by comparing p-values calculated for each test. Let $p_i, i \in 1,\ldots N$ be the p-values from each of N tests. The null hypothesis says that the p-values

should have a $U(0,1)$ distribution if the null is true for all tests. The p-values are then sorted in increasing order $p_{(1)}, p_{(2)}, \ldots, p_{(i)}, \ldots, p_{(N)}$, and those with the smallest values signal a likely alternate result. The BH algorithm offers a method for selecting a p-value threshold for choosing alternate systems by controlling the Efdr. Let $q \in (0,1)$ be the desired upper bound on the Efdr. Then, define $i_{\max}$ as the largest index for which

$$p_{(i)} \leq \frac{i}{N} q.$$

The BH algorithm rejects the null (selects as alternate) all systems with $p_{(i)}$ such that $i \leq i_{\max}$. Let π_0 be the proportion of systems that are null. The following theorem then applies.

Theorem 1 Assume that the p-values are uniformly distributed if the null distribution applies for all systems. If the p-values corresponding to the correct null hypothesis are independent of each other, then

$$\text{Efdr}(q) = \pi_0 q \leq q$$

where π_0 is typically assumed to be unknown, but close to 1. Proofs appear in Benjamini and Hochberg (1995) and Efron (2010).

4 PROCEDURE

Inspired by Theorem 1, we develop a new procedure to classify a large number of simulated systems by controlling the expected false discovery rate. Each system will be classified as either null or alternate. The objective is to select a subset of systems that include a large number of alternate systems that are likely to be better than the null standard. Selecting as many alternate systems as possible is desirable, but controlling the proportion of null systems included in the selection (the false discovery rate) is also critical. We develop an algorithm to choose a threshold for selecting a subset of systems that will balance these two objectives.

We allow for an arbitrarily large number of systems, N. For simplicity in exposition, we assume the system output is independent and normally distributed. These are not necessary requirements for using empirical Bayes methods generally, but these assumptions are available in simulation through replications and batched means methods. These assumptions allow us to explicitly calculate probabilities that null and alternate systems are selected. Without loss in generality, we say that null system i is normally distributed $\mathcal{N}(\mu, \sigma_i^2)$, with each system having a different unknown variance. This value of μ is analogous to the standard in multiple comparisons problems, and we attempt to identify systems that have means smaller than $\mu - \varepsilon$ for some $\varepsilon > 0$. The parameter ε is analogous to an indifference-zone parameter in ranking and selection methods (see Kim and Nelson (2001)). Our algorithm would still work (although it would be conservative) if some of these null systems actually had means greater than the standard μ and some of the alternates had means smaller than $\mu - \varepsilon$.

We have a belief that $N_0 = \pi_0 N$ of these systems are null, where $0 \leq \pi_0 \leq 1$, and $N_1 = N - N_0 = \pi_1 N$ of these systems are from an alternate configuration with a mean smaller than μ, where $\pi_0 + \pi_1 = 1$. We will discuss in Section 4.4 how to estimate π_0, though there may be some prior belief on this value. We derive exact results for known π_0, and numerical experiments show that estimated values of π_0 can be used with encouraging results. Additionally, we derive properties of our procedure under worst-case conditions, where each alternate system i has output that is distributed as $\mathcal{N}(\mu - \varepsilon, \sigma_i^2)$, where the σ_i values unknown. To further simplify the analysis, let $\mu = 0$ so that the distribution of the null systems are $\mathcal{N}(0, \sigma_i^2)$ and the alternates are $\mathcal{N}(-\varepsilon, \sigma_i^2)$.

4.1 Classification method

The two desired output metrics from our algorithm are an Efdr of $q \in (0,1)$ and a probability that an alternate is correctly selected as $1 - \beta \in (0,1)$ (analagous to the power of a test). We note that there is a tradeoff between these two output measures. Selecting only a small number of systems (with the most

extreme p-values) will have a smaller Efdr, but fewer of the actual alternates are selected for future study. Selecting more systems will increase the probability an alternate is correctly selected, but risks including more null systems, which will reduce the Efdr. In order to obtain these desired output metrics, two inputs to our procedure must be carefully chosen. The first input is a threshold $u \in (0,1)$ that classifies system i as alternate if $p_i \leq u$. The second input is a sample size n_i for each system.

The procedure consists of two main steps. In the first step, sample from each system and construct p-values p_i from each system's output. The second step selects systems with p-values less than a critical threshold u^* and classifies them as alternate. The remainder of this section describes the overall framework and how to choose the threshold u^* to obtain the desired Efdr. Choosing the sample sizes n_i^* when σ_i is known is discussed in Section 4.2 and the unknown σ_i case is discussed in 4.3. We assume π_0 is known in both these cases, and Section 4.4 describes a method for estimating π_0.

The result of the algorithm's classification method is displayed in Table 1. This table parallels one from Efron (2010) and shows the four possible categories for the N systems. Each system is either null or alternate in reality, and is either selected or not selected as alternate in our algorithm. The values in the table are the expected number of systems that will fall into each of the four categories. Because null systems have uniformly distributed p-values, on average uN_0 of them will be selected as alternate. Define a function $\gamma(u;\varepsilon,\sigma_i,n_i)$ to be the probability that alternate system i with variance σ_i^2 and n_i samples delivers a p-value smaller than or equal to u. The second row of Table 1 show the expected number of alternate systems nonselected and selected using this function γ because for alternate systems, $P(p_i \leq u) = \gamma(u;\varepsilon,\sigma_i,n_i)$.

Table 1: Expected number of null and alternate systems nonselected and selected.

	Nonselection	Select as Alternate	Sum
Null	$(1-u)N_0$	uN_0	N_0
Alternate	$\sum_{i=1}^{N_1} 1-\gamma(u;\varepsilon,\sigma_i,n_i)$	$\sum_{i=1}^{N_1} \gamma(u;\varepsilon,\sigma_i,n_i)$	N_1
Sum	# nonselected	# selected	N

Many algorithms attempt to control the Type I error u, which is the probability that a null is classified as alternate. In some cases, a Bonferonni bound is used to control the probably that any of the selected systems are actually null. Here, our two metrics are the Efdr,

$$q := \text{Efdr} = \frac{uN_0}{uN_0 + \sum_{i=1}^{N_1} \gamma(u;\varepsilon,\sigma_i,n_i)}, \tag{1}$$

and the expected proportion of alternates selected for future study,

$$1-\beta := \frac{\sum_{i=1}^{N_1} \gamma(u;\varepsilon,\sigma_i,n_i)}{N_1}. \tag{2}$$

We need to choose inputs u^* and n_i^* so that we can jointly obtain desired values of q and $1-\beta$. Denote the expected number of alternate systems selected as $K := \sum_{i=1}^{N_1} \gamma(u;\varepsilon,\sigma_i,n_i)$. Solving both (1) and (2) for K implies $uN_0(\frac{1}{q}-1) = (1-\beta)N_1$. Further solving for the value of u^* yields

$$u^* = \frac{(1-\beta)N_1}{N_0}\frac{q}{1-q} = \frac{(1-\beta)\pi_1}{\pi_0}\frac{q}{1-q}. \tag{3}$$

Thus, we can control the Efdr using u^* based on the choice of $1-\beta$ and knowledge of π_0. Choosing n_i carefully for each system helps select any individual alternate system with probability $1-\beta$. Note that $(1-\beta)N_1 = K$, and K is the sum of N_1 terms. Choosing n_i such that each term within the sum is equal to $1-\beta$ will satisfy our requirements. Let p_i^+ be p-values calculated for alternate systems. We need to solve the following for n_i:

$$1-\beta = \gamma(u^*;\varepsilon,\sigma_i,n_i) = P(p_i^+ \leq u). \tag{4}$$

Assuming that as n_i increases $\gamma(u^*;\varepsilon,\sigma_i,n_i)$ increases for alternate systems, we find n_i^* as the solution to (4) using root-finding methods. We explore cases where this calculation may become tractable in the next sections.

4.2 Known Variance

This section describes how to choose n_i^* when the variance of each system is known. Given n_i samples of a system with mean $\overline{X}_i$, calculate p_i as

$$p_i := \Phi\left(\frac{\overline{X}_i-\mu}{\sigma_i/\sqrt{n_i}}\right) = \Phi\left(\frac{\overline{X}_i-0}{\sigma_i/\sqrt{n_i}}\right). \tag{5}$$

Null systems deliver uniform p-values using (5). We next derive $\gamma(u;\varepsilon,\sigma_i,n_i)$ for alternate systems. Let Z be a standard normal random variable. Then,

$$P(p_i^+ \le u) = P(\Phi(Z-\varepsilon\sqrt{n_i}/\sigma_i) \le u) = P(Z \le \Phi^{-1}(u)+\varepsilon\sqrt{n_i}/\sigma_i) = \Phi(\Phi^{-1}(u)+\varepsilon\sqrt{n_i}/\sigma_i). \tag{6}$$

As suggested in Section 4.1, we can control the FDR using u^*, and obtain a probability that the alternate will be selected as $1-\beta$ by solving (4) for n_i and rounding up to the nearest integer to obtain the sample sizes

$$n_i^* = \left\lceil \left(\frac{\sigma_i}{\varepsilon}\right)^2 (\Phi^{-1}(1-\beta)-\Phi^{-1}(u^*))^2 \right\rceil. \tag{7}$$

The sample size increases as σ_i increases and ε decreases, as expected.

4.3 Unknown Variance

This section describes choosing n_i^* when the variance of each system is unknown. Let $\overline{X}_i$ and $\hat{\sigma}_i^2$ be the mean and sample variance of n_i samples from system i. Calculate a t-statistic for null systems as

$$\frac{\overline{X}_i-\mu}{\hat{\sigma}_i/\sqrt{n_i}} = \frac{\overline{X}_i-0}{\hat{\sigma}_i/\sqrt{n_i}} \sim t_{n_i-1} \tag{8}$$

where t_{n_i-1} has a t-distribution with n_i-1 degrees of freedom. The cumulative distribution function of this t-distribution is denoted $F_{t_{n_i-1}}(\cdot)$. Calculate the p-value for system i as

$$p_i := F_{t_{n_i-1}}\left(\frac{\overline{X}_i}{\hat{\sigma}_i/\sqrt{n_i}}\right). \tag{9}$$

Under the null hypothesis that all system means are 0, the distribution of the calculated p_i values will be uniform. If the system output is generated independently across systems, using our threshold algorithm with u^* given in (3) will select the systems with the smallest p-values while controlling the expected false discovery rate. Next, we derive the distribution of p_i^+ values for alternate systems. Paralleling (8) and (9), calculate

$$t_{n_i-1}^+ \stackrel{D}{=} \frac{\overline{X}_i+\varepsilon-\varepsilon}{\hat{\sigma}_i/\sqrt{n_i}} \stackrel{D}{=} \frac{\sigma_i n_i^{-1/2}Z-\varepsilon}{\hat{\sigma}_i/\sqrt{n_i}} \stackrel{D}{=} \frac{Z-\varepsilon\sqrt{n_i}/\sigma_i}{\chi_{n_i-1}/\sqrt{n_i-1}}, \tag{10}$$

where Z is a standard normal random variable, $\overline{X}_i$ has distribution $\mathcal{N}(-\varepsilon,\sigma_i^2/n_i)$, $\chi^2_{n_i-1} \stackrel{D}{=} (n_i-1)\hat{\sigma}_i^2/\sigma_i^2$, and $\stackrel{D}{=}$ denotes equality in distribution. Hence, the distribution of $t_{n_i-1}^+$ is a noncentral t-distribution. The corresponding p-value is

$$p_i^+ := F_{t_{n_i-1}}(t_{n_i-1}^+).$$

If $\varepsilon > 0$, then there is more weight in the distribution of p_i^+ towards zero (values of p_i for the null system stochastically dominate the values of p_i^+ for the alternates), suggesting that systems with smaller p-values are more likely to be alternates. We continue by deriving the distribution of p_i^+ values. It follows that

$$\gamma(u;\varepsilon,\sigma_i,n_i) = P(p_i^+ \le u) = P(t_{n_i-1}^+ \le F_{t_{n_i-1}}^{-1}(u)). \tag{11}$$

This distribution function can be calculated using properties of the noncentral t-distribution, and thus we can solve (4) numerically using (11) to determine the samples sizes for each system, and compare it to the optimal value of n_i^* in the known variance case (7). Letting $\hat{n}_i^* = (\hat{\sigma}_i/\varepsilon)^2(\Phi^{-1}(1-\beta) - \Phi^{-1}(u^*))^2$, Eq. (7) leads to

$$\hat{n}_i^* - n_i^* = \left(\frac{(\Phi^{-1}(1-\beta) - \Phi^{-1}(u^*))}{\varepsilon}\right)^2 (\hat{\sigma}_i^2 - \sigma_i^2),$$

where

$$(n_0-1)^{1/2}(\hat{\sigma}_i^2 - \sigma_i^2) \overset{D}{\approx} \sigma_i^2 N(0,2),$$

and n_0 is the sample size used to calculate variance estimates. Hence, for n_0 sufficiently large, $\hat{\sigma}_i^2 \approx \sigma_i^2$, meaning that $\hat{n}_i^* \approx n_i^*$. This and Eqs. (6)–(7) in turn suggest that $P(p_i^+ \le u) \approx 1-\beta$ when the variance estimate is used. We quantify this difference in a more rigorous manner next.

We use the following standard ordering notation in the remainder of the paper. For a positive sequence $\{\eta_n\}_{n\in\mathbb{N}}$, a sequence of random variables $\{\xi_n\}_{n\in\mathbb{N}}$ is $O_p(\eta_n)$ if for all $\zeta > 0$ there exists a constant M such that $P(|\xi_n/\eta_n| > M) < \zeta$ for all n sufficiently large. The sequence is $o_p(\eta_n)$ if $P(|\xi_n/\eta_n| > M) \to 0$ for $M > 0$ arbitrary, as $n \to \infty$.

Equation (10) leads to

$$t_{n_i-1}^+ \le F_{t_{n_i-1}}^{-1}(u) \iff Z - F_{t_{n_i-1}}^{-1}(u)\chi_{n_i-1}/\sqrt{n_i-1} \le \varepsilon\sqrt{n_i}/\sigma_i,$$

so that

$$P(t_{n_i-1}^+ \le F_{t_{n_i-1}}^{-1}(u)) = P(Z - F_{t_{n_i-1}}^{-1}(u)\chi_{n_i-1}/\sqrt{n_i-1} \le \varepsilon\sqrt{n_i}/\sigma_i). \tag{12}$$

Writing $\chi_{n_i-1}^2 \overset{D}{=} \sum_{k=1}^{n_i-1} Z_k^2$ and using the Central Limit Theorem lead to

$$\sqrt{n_i-1}\left(\frac{\chi_{n_i-1}^2}{n_i-1} - 1\right) = N(0,2) + o_p(1).$$

Appealing to the Delta method (Theorem 3.1 of (van der Vaart 1998)) results in

$$(\chi_{n_i-1} - \sqrt{n_i-1}) = N(0,1/2) + o_p(1),$$

Hence,

$$Z - F_{t_{n_i-1}}^{-1}(u)\chi_{n_i-1}/\sqrt{n_i-1} - N\left(-F_{t_{n_i-1}}^{-1}(u), 1 + \frac{(F_{t_{n_i-1}}^{-1}(u))^2}{2(n_i-1)}\right) = o_p(n_i^{-1/2}),$$

Since $F_{t_{n_i-1}}^{-1}(\cdot) = \Phi^{-1}(\cdot) + O(n_i^{-1/2})$ (see p. 81 of (Serfling 1980)), Slutsky's Lemma leads to

$$Z - F_{t_{n_i-1}}^{-1}(u)\chi_{n_i-1}/\sqrt{n_i-1} = N\left(-\Phi^{-1}(u), 1 + \frac{(\Phi^{-1}(u))^2}{2(n_i-1)}\right) + o_p(n_i^{-1/2}).$$

It follows from (11) and (12) that

$$\gamma(u;\varepsilon,\sigma_i,n_i) = \Phi\left(\frac{\varepsilon\sqrt{n_i}/\sigma_i + \Phi^{-1}(u)}{\left(1+(\Phi^{-1}(u))^2/(2n_i)\right)^{1/2}}\right) + o(n_i^{-1/2}).$$

Then, as in the known variance scenario, we solve for n_i such that

$$1-\beta = \Phi\left(\frac{\varepsilon\sqrt{n_i}/\sigma_i + \Phi^{-1}(u^*)}{\left(1+(\Phi^{-1}(u^*))^2/(2n_i)\right)^{1/2}}\right) + o(n_i^{-1/2}),$$

where u^* is as in (3). Since $\Phi^{-1}(1-\beta+o(n_i^{-1/2})) = \Phi^{-1}(1-\beta)+o(n_i^{-1/2})$,

$$\left(1+(\Phi^{-1}(u^*))^2/(2n_i)\right)^{1/2}\Phi^{-1}(1-\beta) = \varepsilon\sqrt{n_i}/\sigma_i + \Phi^{-1}(u^*) + o(n_i^{-1/2}).$$

The function

$$g(n_i) = \left(\frac{\sigma_i}{\varepsilon}\right)^2\left(\Phi^{-1}(1-\beta)\left(1+\frac{(\Phi^{-1}(u^*))^2}{2n_i}\right)^{1/2} - \Phi^{-1}(u^*)\right)^2$$

is a contraction and has negative first derivative. Hence, starting with a guess $\tilde{n}_i = (\sigma_i/\varepsilon)^2\left(\Phi^{-1}(1-\beta)-\Phi^{-1}(u^*)\right)^2$ (cf. Eq. (7)) and applying $g(\tilde{n}_i)$, one can sandwich n_i with a lower bound

$$\left(\frac{\sigma_i}{\varepsilon}\right)^2\left(\Phi^{-1}(1-\beta)-\Phi^{-1}(u^*)\right)^2 \le n_i$$

and an upper bound

$$n_i \le \left(\frac{\sigma_i}{\varepsilon}\right)^2\left(\Phi^{-1}(1-\beta)\left(1+\left(\frac{\varepsilon}{\sigma_i}\right)^2\frac{(\Phi^{-1}(u^*))^2}{2\left(\Phi^{-1}(1-\beta)-\Phi^{-1}(u^*)\right)^2}\right)^{1/2} - \Phi^{-1}(u^*)\right)^2,$$

for ε sufficiently small that the $o(n_i^{-1/2})$ term can be dropped.

A conservative approach is to set

$$n_i^* = \left\lceil\left(\frac{\sigma_i}{\varepsilon}\right)^2\left(\Phi^{-1}(1-\beta)\left(1+\left(\frac{\varepsilon}{\sigma_i}\right)^2\frac{(\Phi^{-1}(u^*))^2}{2\left(\Phi^{-1}(1-\beta)-\Phi^{-1}(u^*)\right)^2}\right)^{1/2} - \Phi^{-1}(u^*)\right)^2\right\rceil. \tag{13}$$

Replacing σ_i^2 by its estimate $\hat{\sigma}_i^2(n_0)$ over n_0 trial samples produces the estimator

$$\bar{n}_i^* = \left\lceil\left(\frac{\hat{\sigma}_i(n_0)}{\varepsilon}\right)^2\left(\Phi^{-1}(1-\beta)\left(1+\left(\frac{\varepsilon}{\hat{\sigma}_i(n_0)}\right)^2\frac{(\Phi^{-1}(u^*))^2}{2\left(\Phi^{-1}(1-\beta)-\Phi^{-1}(u^*)\right)^2}\right)^{1/2} - \Phi^{-1}(u^*)\right)^2\right\rceil,$$

with $\bar{n}_i^* - n_i^* = O_p(n_0^{-1/2}\varepsilon^{-2})$. Set $n_0 = O(\varepsilon^{-\delta})$, for $\delta > 0$, so that $\bar{n}_i^* - n_i^* = O_p(\varepsilon^{\delta/2-2})$. Taylor expanding the right-hand side of Eq. (6) about n_i^*,

$$\begin{aligned}\Phi(\Phi^{-1}(u)+\varepsilon\sqrt{\bar{n}_i^*}/\sigma_i) - \Phi(\Phi^{-1}(u)+\varepsilon\sqrt{n_i^*}/\sigma_i) &= \frac{\phi(\Phi^{-1}(u)+\varepsilon\sqrt{n_i^*}/\sigma_i)}{2\sigma_i}\frac{\varepsilon}{\sqrt{n_i^*}}O_p(\varepsilon^{\delta/2-2}) \\ &= O_p(\varepsilon^{\delta/2}).\end{aligned}$$

Hence, we conclude that the error in Eq. (6) induced by estimating the variance is of order $O_p(\varepsilon^{\delta/2})$. For instance, setting $\varepsilon = 0.05$ and $\delta = 2$ results in $n_0 = 400$ trial samples, and an error in Eq. (6) of order $O(\varepsilon)$.

4.4 Estimating π_0

The value of u^* requires knowledge of π_0. To estimate the proportion of null systems π_0, we employ a method from Efron (2010). Define the z-values for each system as

$$z_i = \Phi^{-1}(p_i). \tag{14}$$

The "zero assumption" says that the density of alternate system z-values is zero on some range, because the peak of the distribution of z-values is assumed to contain mostly null systems. This range for the zero assumption could be $\mathscr{A} := [0,\infty)$, or could include negative values depending on the values of ε and σ_i^2. Let $F_0(\mathscr{A})$ be the probability that a z-value from a null system lies in range $\mathscr{A}$. Then, the expected number z-values observed in $\mathscr{A}$ is

$$E[\#z_i \in \mathscr{A}] = \pi_0 N F_0(\mathscr{A}).$$

Efron (2010) suggests using the resulting estimator

$$\hat{\pi}_0 = \frac{\#z_i \in \mathscr{A}}{NF_0(\mathscr{A})} \tag{15}$$

based on the observed z-values. This estimator can sometimes yield an estimate of $\pi_0 \geq 1$ depending on the range of $\mathscr{A}$ used, which leads to $u^* \leq 0$. For computational purposes we assume that at least one system is an alternate, so that $\pi_0 \leq (N-1)/N$ and $u^* > 0$. More complex methods for estimating the proportion of null systems are discussed in Efron (2007).

5 IMPLEMENTATION

We summarize the results from Section 4 into an algorithmic implementation. First, we outline the algorithm when the values of σ_i and π_0 are known:

1. Calculate u^* using (3), and n_i^* using (7).
2. Sample each system i using n_i^* samples, and calculate p_i according to (5).
3. Select systems with $p_i \leq u^*$ as alternate.

By construction, the above method will select a subset with an Efdr of q and with an expected proportion $1-\beta$ of alternates included in the subset. When σ_i and π_0 must be estimated, we propose the following method which we call Algorithm 1. This algorithm has a first stage which generates estimates $\hat{\sigma}_i$ and $\hat{\pi}_0$.

Algorithm 1

1. For each system $i \in 1,\ldots N$ collect an initial n_0 samples (n_0 should be sufficient to estimate σ_i^2) and calculate p_i according to (9) using estimates $\hat{\sigma}_i$.
2. Estimate π_0 using (14) and (15). Use this to calculate u^* using (3).
3. Calculate the sample sizes n_i^* from (13) using $\hat{\sigma}_i$.
4. Resample each system using n_i^* new samples and calculate updated $\hat{\sigma}_i$ and p_i values.
5. Repeat Step 2 with the recalculated p_i values to obtain an updated value of $\hat{\pi}_0$ and u^*.
6. Select systems with $p_i \leq u^*$ as alternate.

We describe the metrics collected to evaluate the performance of the algorithm. The first two are the Efdr and $1-\beta$. These two should be controlled by the algorithm by design, and numerical experiments will show the effects of using estimates of σ_i^2 and π_0. Additionally, we measure the Type I error (the proportion of null systems selected), and the total proportion of systems selected as alternate to measure

the selection size. Finally, to measure the sampling effort of Algorithm 1, we record the average sampling ratio $\frac{1}{N}\sum_{i=1}^{N} n_i^*/(\sigma_i/\varepsilon)^2$ to estimate the effort needed for each system relative to σ_i and ε.

Each table reports results from an experiment consisting of 1000 replications to estimate the mean and standard deviation of each output metric. Each replication simulates $N = 1000$ systems with $\pi_0 = 90\%$ of them as null. Then, π_0 is estimated in the algorithm and we choose the zero-assumption range $\mathscr{A} = [\Phi^{-1}(1/2 - \varepsilon/4), \infty)$. We generate the value of σ_i for each system from $2 + Exp(3)$ where $Exp(x)$ is an exponential random variable with mean x, and choose initial sample sizes for variance estimation to be $n_0 = 1000$.

We compare the results of Algorithm 1 to the Benjamini-Hochberg algorithm when similar sampling effort is applied to each algorithm. The BH algorithm can be applied to the second stage after n_i^* samples are collected with results very close to those of Algorithm 1. We choose the average n_i^* value (from the results of Algorithm 1) as the sample size for each of the N systems and apply the BH algorithm. The overall number of samples is the same across both algorithms, but the BH algorithm applies effort equally while Algorithm 1 uses n_i^* for system i.

Both algorithms perform favorably, with the BH algorithm typically yielding a smaller Efdr and smaller $1-\beta$ than Algorithm 1. Table 2 compares the BH algorithm and Algorithm 1 when $q = 0.1$, $1-\beta = 0.9$ and $\varepsilon = 0.1$. Both algorithms appear able to control the Efdr to be smaller than 0.1. Algorithm 1 using n_i^* does a slightly better job selecting more alternate systems (93.2%), though this is conservative relative to the desired 90% selection rate. We will see in all the tables that the value of $1-\beta$ delivered by Algorithm 1 is noticeably larger than that for the BH algorithm. The Type I error is small for both algorithms, and the proportion of systems selected is approximately 10%. We note that the expected number of systems selected will be $\pi_1(1-\beta)/(1-q)$ and in the forthcoming examples should be equal to $\pi_1 = 10\%$ because we choose $\beta = q$.

Table 2: Parameters $q = 0.1$, $1-\beta = 0.9$, $\varepsilon = 0.1$, the mean and std. deviation of $\frac{1}{N}\sum_{i=1}^{N} n_i^*/(\sigma_i/\varepsilon)^2$ are 15.14 and 2.76.

	BH-Mean	BH-Std. Dev.	Algorithm 1-Mean	Algorithm 1-Std. Dev.
Efdr	0.090	0.030	0.096	0.039
$1-\beta$	0.896	0.036	0.932	0.043
Type I Error	0.010	0.004	0.011	0.005
Proportion selected	0.099	0.005	0.103	0.007

Table 3 reduces ε to 0.05, and the results are similar to those for $\varepsilon = 0.1$, but the mean sampling ratio increases as expected from 15.14 to 17.62. Again, we note that both algorithms achieve desired performance across the four metrics, though we note that our implementation of the BH algorithm benefits from a sample size selection achieved by using the average of the system sample sizes delivered by Algorithm 1.

Table 3: Parameters $q = 0.1$, $1-\beta = 0.9$, $\varepsilon = 0.05$, the mean and std. deviation of $\frac{1}{N}\sum_{i=1}^{N} n_i^*/(\sigma_i/\varepsilon)^2$ are 17.62 and 4.01.

	BH-Mean	BH-Std. Dev.	Algorithm 1-Mean	Algorithm 1-Std. Dev.
Efdr	0.090	0.030	0.094	0.039
$1-\beta$	0.912	0.036	0.954	0.040
Type I Error	0.010	0.004	0.011	0.005
Proportion selected	0.100	0.005	0.106	0.007

Table 4 reports results for $\varepsilon = 0.1$, but has $q = 0.05$ and $1-\beta = 0.95$. Both algorithms are able to deliver an Efdr smaller than q, and sampling management in Algorithm 1 delivers an estimated $1-\beta$ of 96.6%. The mean relative samples needed increases to 20.14 to obtain the smaller q and higher $1-\beta$ than in Table 2.

Table 4: Parameters $q = 0.05$, $1-\beta = 0.95$, $\varepsilon = 0.1$, the mean and std. deviation of $\frac{1}{N}\sum_{i=1}^{N} n_i^*/(\sigma_i/\varepsilon)^2$ are 20.14 and 2.83.

	BH-Mean	BH-Std. Dev.	Algorithm 1-Mean	Algorithm 1-Std. Dev.
Efdr	0.046	0.021	0.050	0.027
$1-\beta$	0.910	0.032	0.966	0.026
Type I Error	0.005	0.002	0.006	0.003
Proportion selected	0.095	0.004	0.102	0.004

We conclude with a table showing the results when the BH algorithm is applied without sample size management. Table 5 reports performance of the BH algorithm when all systems receive $n_0 = 1000$ samples. The same input parameters as those in Table 4 are used. We observe that Edfr is still controlled, but the values of $1-\beta$, Type I error, and proportion of selected systems are close to zero due to so few systems being selected. The benefit of Algorithm 1 is that it guides the sample size selection to encourage selection of more alternates.

Table 5: BH algorithm with $n_0 = 1000$ samples for each system. Parameters $q = 0.05$, $\beta = 1-0.95$, $\varepsilon = 0.1$.

	BH-Mean	BH-Std. Dev.
Efdr	0.048	0.196
$1-\beta$	0.002	0.006
Type I Error	0.000	0.000
Proportion selected	0.000	0.001

6 CONCLUSIONS

We present an algorithm for performing multiple comparisons with a standard under a framework where most of the systems are from a "null" distribution, and the remainder are from an "alternate" distribution. The intent is to isolate the subset of systems that are likely from the alternate distribution. The algorithm selects a subset of systems while controlling the expected false discovery rate and the probability that an alternate system is selected. The framework is similar to the Benjamini-Hochberg algorithm in that it uses a threshold policy on the p-values for each system while controlling the Efdr. We derive expressions for the sample sizes needed for each system to obtain a desired probability that an alternate is correctly selected, and assess the effect of variance estimation on these sample sizes. Future work will consider uncertainty in the mean of the alternate distribution, and relax the normality assumption.

REFERENCES

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* 15 (1): 289–300.

Chen, C.-H., D. He, M. Fu, and L. H. Lee. 2008. "Efficient Simulation Budget Allocation for Selecting an Optimal Subset". *INFORMS Journal on Computing* 20 (4): 579–595.

Chingcuanco, F., and C. Osorio. 2013. "A Procedure to Select the Best Subset Among Simulated Systems Using Economic Opportunity Cost". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 452–562. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Corlu, C. G., and B. Biller. 2013. "A Subset Selection Procedure Under Input Parameter Uncertainty". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk,

R. Hill, and M. E. Kuhl, 463–473. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Efron, B. 2007. "Size, Power and False Discovery Rates". *The Annals of Statistics* 35 (4): 1351–1377.

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

Kim, S.-H. 2005. "Comparison with a Standard via Fully Sequential Procedures". *ACM Transactions on Modeling and Computer Simulation* 15 (2): 155–174.

Kim, S.-H., and B. L. Nelson. 2001. "A Fully Sequential Procedure for Indifference-Zone Selection in Simulation". *ACM Transactions on Modeling and Computer Simulation* 11 (3): 251–273.

Koenig, L. W., and A. M. Law. 1985. "A Procedure for Selecting a Subset of Size m Containing the l Best of k Independent Normal Populations, with Applications to Simulation". *Communications in Statistics-Simulation and Computation* 14 (3): 719–734.

Nelson, B. L., and D. Goldsman. 2001. "Comparisons with a Standard in Simulation Experiments". *Management Science* 47 (3): 449–463.

Paulson, E. 1952. "On the Comparison of Several Experimental Categories with a Control". *The Annals of Mathematical Statistics* 23 (2): 239–246.

Ryzhov, I. O., and W. Powell. 2009. "The Knowledge Gradient Algorithm for Online Subset Selection". In *Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, 137–144. Nashville, Tennessee.

Serfling, R. 1980. *Approximation Theorems of Mathematical Statistics*. Wiley.

van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Xie, J., and P. I. Frazier. 2013. "Sequential Bayes-Optimal Policies for Multiple Comparisons with a Known Standard". *Operations Research* 61 (5): 1174–1189.

AUTHOR BIOGRAPHIES

DASHI SINGHAM received her Ph.D. from the University of California at Berkeley, and currently is a Research Assistant Professor of Operations Research at the Naval Postgraduate School. Her research interests include simulation analysis and applied statistics. Her email address is dsingham@nps.edu.

ROBERTO SZECHTMAN received his Ph.D. from Stanford University, and currently is an Associate Professor in the Operations Research Department at the Naval Postgraduate School. His research interests include applied probability and military operations research. His email address is rszechtm@nps.edu.