Published in final edited form as:

J Nonparametr Stat. 2011 ; 23(3): 583–604. doi:10.1080/10485252.2010.482154.

Randomised P -values and nonparametric procedures in multiple testing

Joshua D. Habiger^{*} and Edsel A. Peña

Department of Statistics, 216 LeConte College, University of South Carolina, Columbia, SC 29208, USA

Abstract

The validity of many multiple hypothesis testing procedures for false discovery rate (FDR) control relies on the assumption that P -value statistics are uniformly distributed under the null hypotheses. However, this assumption fails if the test statistics have discrete distributions or if the distributional model for the observables is misspecified. A stochastic process framework is introduced that, with the aid of a uniform variate, admits P -value statistics to satisfy the uniformity condition even when test statistics have discrete distributions. This allows nonparametric tests to be used to generate P -value statistics satisfying the uniformity condition. The resulting multiple testing procedures are therefore endowed with robustness properties. Simulation studies suggest that nonparametric randomised test P -values allow for these FDR methods to perform better when the model for the observables is nonparametric or misspecified.

Keywords

multiple testing; randomisation; false discovery rate; microarray; P -value statistics

1. Introduction

Consider the problem of testing a pair of hypotheses $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ based on the realisation x , called data, of a random observable X whose distribution F is assumed, possibly incorrectly, to belong to a known class of distributions $\mathcal{F} = \{F(x; \theta); \theta \in \Theta\}$. The class $\mathcal{F}$ is said to be a model for F . Assume further that the test rejects H_0 for large values of a statistic $S(X)$. Given data $X = x$, the P -value or significance value is defined by

$$P(x) = \mathbf{P}\{S(X) \geq S(x) | X \sim F(\cdot; \theta_0)\}.$$

The P -value statistic is $P(X)$. Given an $\alpha \in [0, 1]$, the test which rejects H_0 whenever $P(X) \leq \alpha$, which we shall denote by $\hat{\alpha}(X; \alpha)$, is called an α -level test. The size of $\hat{\alpha}(X; \alpha)$ is the probability of rejecting H_0 when in fact $\theta = \theta_0$.

For the size of $\mathcal{X}(X; \alpha)$ to be α for any $\alpha \in [0, 1]$, it is necessary and sufficient that $P(X)$ have a uniform distribution under $\theta = \theta_0$. However, this condition will only be true when $S(X)$ has a continuous distribution and the assumed model $\mathcal{F}$ is correct. If $S(X)$ has a discrete or mixed-type distribution or the model $\mathcal{F}$ is misspecified, i.e. $F \notin \mathcal{F}$, then this uniformity condition does not hold and the size of $\mathcal{X}(X; \alpha)$ need not be α . The impact of the breakdown of this uniformity condition may be deemed acceptable for a single hypothesis test. However, when testing many hypotheses simultaneously, the breakdown of this uniformity condition may cause multiple testing procedures to perform poorly.

Benjamini and Hochberg (1995) introduced a method for controlling the false discovery rate (FDR), which is the expected proportion of erroneously rejected null hypotheses (false discoveries) among rejected null hypotheses (discoveries). FDR methods are useful in situations when it is acceptable for a small proportion of discoveries to be false discoveries. Many other methods have been developed to estimate the FDR for specified rejection regions or observed P -values (see Efron, Tibshirani, Storey, and Tusher 2001; Storey 2002, 2003; Efron 2004; Genovese and Wasserman 2004; Jin and Cai 2007; Ruppert, Nettleton, and Hwang 2007; Sun and Cai 2007). However, many FDR methods are developed under the assumption that each P -value statistic $P_i(X_i)$ for testing the i th pair of hypotheses H_{0i} versus H_{1i} is uniformly distributed under H_{0i} . But in many practical situations calling for FDR control or estimation, this uniformity condition does not hold. For example, in a standard microarray analysis, T -tests or Wilcoxon rank-sum tests are employed to test the null hypotheses that gene expression levels are not differentially expressed across two treatment groups. However, the Wilcoxon rank-sum test statistic has a discrete distribution, while the T -test is valid only if a normal model is correct for each of the F_i s.

The fact that the uniformity condition may not be satisfied has been partly addressed. Some FDR methods allow for the estimation of the distribution of the P -value statistics under the null hypotheses (see, e.g. Yekutieli and Benjamini 1999; Korn, Troendle, McShane, and Simon 2004; Pollard and van der Laan 2004; van der Laan and Hubbard 2006; Efron 2007; Dudoit and van der Laan 2008; Efron 2008), and hence do not require that P -value statistics satisfy the uniformity condition. However, as noted in Efron (2008), these methods will lead to FDR estimators with a larger standard error. In Section 5, we will see that these methods may also lead to biased FDR estimators. Other FDR methods have been developed for P -value statistics with discrete distributions (see, e.g. Gilbert 2005; Kulinskaya and Lewin 2009).

In this paper, rather than modifying or developing an FDR method, we focus on the P -value statistics themselves that are being used in many of these FDR methods. The idea is to augment the observable x by an independently generated uniform random variate u , and to view the doubleton (x, u) as full data. The collection of decision functions given by $\mathcal{X} = \{\mathcal{X}(X, U; \alpha): \alpha \in [0, 1]\}$ is viewed as a stochastic process. The decision process will then induce a well-defined P -value statistic, denoted $P(X, U)$, which is uniformly distributed, under H_0 , for a model $\mathcal{F}$ if and only if the size of $\mathcal{X}(X, U; \alpha)$ is α under model $\mathcal{F}$ for each α in $[0, 1]$. This framework provides an avenue for generating P -value statistics that satisfy the uniformity condition when test statistics have discrete distributions. These P -value statistics are then the statistics on which existing FDR methods will operate. If $P(X, U)$ depends in a

nontrivial manner on U , it will be said to be a randomised P -value statistic. This framework will allow for robust nonparametric procedures to be used to generate P -value statistics that are uniformly distributed under the null hypotheses. Hence, those FDR methods which assume the uniformity condition are endowed with robustness properties.

In Section 2, the definitions of a valid decision process and a randomised P -value statistic are given. A discussion regarding randomisation in multiple testing is in Section 3. In Section 4, the randomised Wilcoxon rank-sum test P -value statistic is defined and compared with the nonrandomised Wilcoxon P -value and T -test P -value statistics. In Section 5, the P -values discussed in Section 4 are applied to an FDR-controlling and an FDR-estimating procedure. Simulation studies suggest that if F is non-normal, these procedures perform much better when using the randomised Wilcoxon P -values rather than the T -test P -values and comparable when F is normal. Procedures always perform better when using the randomised Wilcoxon P -values as compared with the nonrandomised Wilcoxon P -values. In Section 6, a real microarray data set is used to illustrate the methods in this paper and to demonstrate that the set of null hypotheses that are declared false will depend on the type of P -value and FDR procedure used to test the null hypotheses. Concluding remarks are in Section 7.

2. Randomised P -value statistics

2.1. Decision processes and P -values

In this section, we introduce the notion of a randomised P -value statistic associated with a randomised decision or test function. Recall that a randomised test function for testing H_0 versus H_1 based on $X \in \mathcal{X}$ is a function $\varphi: \mathcal{X} \rightarrow [0, 1]$, with the interpretation that $\varphi(x)$ is the probability of rejecting H_0 when $X = x$. Let U be uniformly distributed over the unit interval and independent of X . We define a decision function $\delta: \mathcal{X} \times [0, 1] \rightarrow \{0, 1\}$ via $\delta(x, u) = I\{\varphi(x) \leq u\}$. The important point is that this δ is a nonrandomised decision function with respect to the augmented random observable (X, U) . The decision function δ has the interpretation that if $\delta = 1$ (0), then H_0 is rejected (not rejected).

Let us suppose that $X \sim F$ with $F \in \mathcal{F}$, and consider the problem of testing $H_0: F \in \mathcal{F}_0 \subset \mathcal{F}$ versus $H_1: F \in \mathcal{F} \setminus \mathcal{F}_0 = \mathcal{F}_1$ based on X and U . For the decision function $\delta(X, U)$, its size is defined by $\eta = \sup_{F \in \mathcal{F}_0} E\{\delta(X, U) | (X, U) \sim F \otimes \lambda\}$, where $(X, U) \sim F \otimes \lambda$ means that $X \sim F$, $U \sim \lambda$ with λ being Lebesgue or uniform probability measure on $[0, 1]$, and with X and U independent. For brevity, we write $\sup_{F \in \mathcal{F}_0} E_F\{\delta(X, U)\} \equiv \sup_{F \in \mathcal{F}_0} E\{\delta(X, U) | (X, U) \sim F \otimes \lambda\}$ for the remainder of the paper. We could specify the size of η to determine the complete specification of $\delta(X, U)$. Thus it will be beneficial to indicate this symbiotic interplay by writing $\delta(X, U; \eta)$. The stochastic process indexed by η and defined by $\delta = \{\delta(X, U; \eta): \eta \in [0, 1]\}$ will be called a decision process provided it satisfies the condition that for every $F \in \mathcal{F}$, the mapping $\eta \mapsto \delta(x, u; \eta)$ is nondecreasing, right-continuous, and with $\delta(x, u; 0) = 0$ and $\delta(x, u; 1) = 1$ for almost all $(x, u) \in \mathcal{X} \times [0, 1]$ with respect to $F \otimes \lambda$. The index η will be referred to as the size index.

For a given decision process δ , we define its induced P -value statistic by

$$P_{\Delta}(X, U) = \inf\{\eta \in [0, 1] : \delta(X, U; \eta) = 1\}. \quad (1)$$

This will be referred to as a randomised P -value statistic when it depends nontrivially on U . A realised P -value $P(x, u)$ can be interpreted as the smallest size index η allowing for the rejection of H_0 given the augmented data (x, u) using the decision process.

2.2. Examples of decision processes and P -value statistics

The definition of the P -value statistic in Equation (1) is quite general since the form of the decision process could be quite general. Therefore, before proceeding we demonstrate this definition more concretely by obtaining the specific forms of $P(X, U)$ for three commonly encountered decision processes. For each decision process, suppose that $S(X)$ is a one-dimensional test statistic. Let Q be a distribution function on the range space of $S(X)$. Define $q(s) = Q(s) - Q(s-)$ and the quantile function $Q^{-1}(u) = \inf\{s : Q(s) \geq u\}$.

Let us first consider a ‘lower-tailed’ decision or test function. This type of decision function is used when small values of $S(x)$ are evidence against the null hypothesis. The decision function has the form, for $\eta \in [0, 1]$,

$$\delta^-(X, U; \eta) = I\{S(X) < Q^{-1}(\eta)\} + I\{S(X) = Q^{-1}(\eta)\} I\{U \leq \gamma^-(\eta)\},$$

where

$$\gamma^-(\eta) = \frac{\eta - Q(Q^{-1}(\eta)-)}{q(Q^{-1}(\eta))},$$

with the convention that $0/0 = 0$. Observe that $E\{\delta^-(X, U; \eta) | S(X) \sim Q\} = \eta$.

Form the decision process $\delta^- = \{\delta^-(X, U; \eta) : \eta \in [0, 1]\}$. Note that $\{\eta : \delta^-(X, U; \eta) = 1\} = \{\eta : \eta \geq Q(S(x)-) + uq(S(x))\}$. Thus, by Equation (1), the associated P -value statistic for the decision process δ^- is

$$P_{\Delta^-}(X, U) = \inf\{\eta \in [0, 1] : \delta^-(X, U; \eta) = 1\} = Q(S(X)-) + Uq(S(X)). \quad (2)$$

Analogously, consider an ‘upper-tailed’ decision function. This decision function is used when large values of $S(x)$ are evidence against a null hypothesis. The decision function has the form, for $\eta \in [0, 1]$,

$$\delta^+(X, U; \eta) = I\{S(X) > Q^{-1}(1-\eta)\} + I\{S(X) = Q^{-1}(1-\eta)\} I\{U \leq \gamma^+(\eta)\},$$

where

$$\gamma^+(\eta) = \frac{\eta - [1 - Q(Q^{-1}(1 - \eta))]}{q(Q^{-1}(1 - \eta))}.$$

Again, note that $E\{\delta^+(X, U; \eta) | S(X) \sim Q\} = \eta$.

Form the upper-tailed decision process $\delta^+ = \{\delta^+(X, U; \eta) : \eta \in [0, 1]\}$ and observe that $\{\eta : \delta^+(x, u; \eta) = 1\} = \{\eta : \eta \geq 1 - Q(S(x)) + uq(S(x))\}$. It follows from Equation (1) that the P -value statistic for δ^+ is

$$P_{\Delta^+}(X, U) = \inf\{\eta \in [0, 1] : \delta^+(X, U; \eta) = 1\} = 1 - Q(S(X)) + U_q(S(X)). \quad (3)$$

The P -value expressions in Equations (2) and (3) are the forms obtained according to the P -value definition of Cox and Hinkley (1974), so their definition is subsumed by the P -value definition in Equation (1).

The ‘two-sided’ decision function has the form given by, for $\eta \in [0, 1]$ and for a fixed $c \in [0, 1]$,

$$\begin{aligned} \delta_c^0(X, U; \eta) &= \max\{\delta^-(X, U; (1-c)\eta), \delta^+(X, U; c\eta)\} \\ &= \delta^-(X, U; (1-c)\eta) + \delta^+(X, U; c\eta), \end{aligned}$$

with the second form following since it is easy to see that the probability that $\delta^-(X, U; (1-c)\eta)$ and $\delta^+(X, U; c\eta)$ simultaneously take the value of one is zero. Note that this decision function is a combination of the lower- and upper-tailed test functions where the lower-tailed test is allocated a size of $(1-c)\eta$ and the upper-tailed test is allocated a size of $c\eta$.

Form the decision process $\Delta_c^0 = \{\delta_c^0(X, U; \eta) : \eta \in [0, 1]\}$. It is important to point out that c should not depend on η and (X, U) , otherwise Δ_c^0 need not satisfy some of the conditions of a decision process.

As to be expected, the P -value statistic induced by Δ_c^0 is closely related to the P -value statistics for δ^+ and δ^- . We present the form of this P -value statistic in Theorem 2.1 and then present some distributional equivalences in Corollary 2.2.

Theorem 2.1—Fix $a, c \in [0, 1]$ and let δ^- , δ^+ , and Δ_c^0 be defined as above. Then,

$$P_{\Delta_c^0}(X, U) = \min \left\{ \frac{P_{\Delta^-}(X, U)}{1-c}, \frac{P_{\Delta^+}(X, U)}{c} \right\}. \quad (4)$$

Proof: Note that

$$\begin{aligned}
& \{\eta: \delta_c^0(x, u; \eta) = 1\} = \{\eta: \delta^+(x, u; c\eta) = 1\} \cup \{\eta: \delta^-(x, u; (1-c)\eta) = 1\} \\
& = \{\eta: c\eta \geq 1 - Q(S(x)) + uq(S(x))\} \cup \{\eta: (1-c)\eta \geq Q(S(x)) - uq(S(x))\} \\
& = \left\{ \eta: \eta \geq \frac{P_{\Delta^+}(x, u)}{c} \right\} \cup \left\{ \eta: \eta \geq \frac{P_{\Delta^-}(x, u)}{1-c} \right\} \\
& = \left\{ \eta: \eta \geq \min \left[\frac{P_{\Delta^+}(x, u)}{c}, \frac{P_{\Delta^-}(x, u)}{1-c} \right] \right\}.
\end{aligned}$$

Hence, the result follows from Equation (1).

Corollary 2.2—For δ^+ , δ^- , and Δ_c^0 defined as above,

$$P_{\Delta_c^0}(X, U) \stackrel{d}{=} \min \left\{ \frac{1 - P_{\Delta^+}(X, U)}{1-c}, \frac{P_{\Delta^+}(X, U)}{c} \right\} \stackrel{d}{=} \min \left\{ \frac{P_{\Delta^-}(X, U)}{1-c}, \frac{1 - P_{\Delta^-}(X, U)}{c} \right\},$$

where ‘ $\stackrel{d}{=}$ ’ means equal in distribution.

Proof: It suffices to show that $1 - P_{\Delta^-}(X, U) \stackrel{d}{=} P_{\Delta^+}(X, U)$ since this would imply that $1 - P_{\Delta^+}(X, U) \stackrel{d}{=} P_{\Delta^-}(X, U)$. But since $U \stackrel{d}{=} 1 - U$ and $Q(s-) = Q(s) - q(s)$, we have

$$\begin{aligned}
1 - P_{\Delta^-}(X, U) &= 1 - \{Q(S(X)) - Uq(S(X))\} \\
&= 1 - \{Q(S(X)) - Q(S(X)) + Uq(S(X))\} \\
&= 1 - Q(S(X)) + \{1 - U\}q(S(X)) \\
&\stackrel{d}{=} 1 - Q(S(X)) + Uq(S(X)) = P_{\Delta^+}(X, U).
\end{aligned}$$

The constant c can be viewed as a weight between the lower- and upper-tailed tests. If it is more desirable, or perhaps optimal, to reject the null hypothesis for large values of $S(x)$ rather than small values of $S(x)$, then c can be chosen to be close to 1. Observe, also, that when $c = 1/2$, which may be preferable under symmetry considerations, then we obtain from expression (4) the usual P -value statistic in two-sided hypotheses testing, given by

$$P_{\Delta_{1/2}^0}(X, U) = 2 \min\{P_{\Delta^+}(X, U), P_{\Delta^-}(X, U)\}.$$

2.3. Properties of P-value statistics

It is desirable to have a decision process in which the size index η coincides with the size of $\delta(X, U; \eta)$ for any η . Since the size of $\delta(X, U; \eta)$ will depend on $\mathcal{F}_0$, it is important to emphasise that the validity of a test is relative to the sub-model $\mathcal{F}_0$. We shall therefore say that the decision function $\delta(X, U; \eta)$ is $\mathcal{F}_0$ -size-valid at η if $\sup_{F \in \mathcal{F}_0} E_F \delta(X, U; \eta) = \eta$, and the decision process is $\mathcal{F}_0$ -size-valid if, for every $\eta \in [0, 1]$, $\delta(X, U; \eta)$ is $\mathcal{F}_0$ -size-valid at η . Likewise, we say that δ is $\mathcal{F}_0$ -level-valid at η if $\sup_{F \in \mathcal{F}_0} E_F \delta(X, U; \eta) \leq \eta$, and is $\mathcal{F}_0$ -level-valid if $\delta(X, U; \eta)$ is $\mathcal{F}_0$ -level valid at η for every $\eta \in [0, 1]$.

Next we consider the distribution of the P -value statistic under $\mathcal{F}_0$. We say that $P(X, U)$ is $\mathcal{F}_0$ -uniform if $\sup_{F \in \mathcal{F}_0} \mathbf{P}_F \{P(X, U) \leq t\} = t$ for every $t \in [0, 1]$. Likewise, we say that $P(X, U)$ is $\mathcal{F}_0$ -uniformly bounded if $\sup_{F \in \mathcal{F}_0} \mathbf{P}_F \{P(X, U) \leq t\} \leq t$ for every $t \in [0, 1]$.

The size and level validity of Δ_c^0 are tied-in to each other. In Theorem 2.3, we see that $P(X, U)$ is $\mathcal{F}_0$ -uniform if and only if Δ_c^0 is $\mathcal{F}_0$ -size-valid and $P(X, U)$ is $\mathcal{F}_0$ -uniformly bounded if and only if Δ_c^0 is $\mathcal{F}_0$ -level-valid.

Theorem 2.3—Let $\mathcal{D} = \{\delta(X, U; \eta) : \eta \in [0, 1]\}$ be a decision process for $H_0 : F \in \mathcal{F}_0$ versus $H_1 : F \in \mathcal{F}_1$. Then $P(X, U)$ is $\mathcal{F}_0$ -uniformly bounded if and only if Δ_c^0 is $\mathcal{F}_0$ -level-valid, and $P(X, U)$ is $\mathcal{F}_0$ -uniform if and only if Δ_c^0 is $\mathcal{F}_0$ -size-valid.

Proof: Let $F \in \mathcal{F}$. Then there exists an $N \subset \mathcal{X} \times [0, 1]$, dependent on F , with $w \mapsto \delta(x, u; w)$ right-continuous and nondecreasing for $(x, u) \in N^c$ and with $(F \otimes \lambda)(N) = 0$. We first show that

$$\{(x, u) : P_\Delta(x, u) \leq t\} = \{(x, u) \in N^c : \delta(x, u; t) = 1\} \cup \{(x, u) \in N : P_\Delta(x, u) \leq t\}. \quad (5)$$

Consider an $(x, u) \in N^c$. If $w < P(x, u)$, then $w < \inf\{w' : \delta(x, u; w') = 1\}$ implying that $\delta(x, u; w) = 0$ since $\delta(x, u; \cdot)$ is right-continuous and nondecreasing for $(x, u) \in N^c$. Conversely, if $\delta(x, u; w) = 1$, then since $\delta(x, u; \cdot)$ is right-continuous and nondecreasing for $(x, u) \in N^c$ and takes values in $\{0, 1\}$, then $w \geq \inf\{w' : \delta(x, u; w') = 1\} = P(x, u)$. This proves Equation (5). From Equation (5) and since $(F \otimes \lambda)(N) = 0$, it follows that $\mathbf{P}_F\{P_\Delta(X, U) \leq t\} = \mathbf{P}_F\{\delta(X, U; t) = 1\}$ for every $F \in \mathcal{F}$, so that

$$\begin{aligned} \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{P_\Delta(X, U) \leq t\} &= \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{\delta(X, U; t) = 1\} \\ &= \sup_{F \in \mathcal{F}_0} E_F\{\delta(X, U; t)\} \leq t \end{aligned}$$

for every $t \in [0, 1]$ if and only if Δ_c^0 is $\mathcal{F}_0$ -level-valid, and with equality if and only if Δ_c^0 is $\mathcal{F}_0$ -size-valid.

Theorem 2.3 suggests that the P -value uniformity condition can be verified by checking that Δ_c^0 is $\mathcal{F}_0$ -size-valid, which in some situations can be verified by verifying that the model $\mathcal{F}$ is valid. For example, Δ_c^0 in Section 5 is $\mathcal{F}_0$ -size-valid if a normal model for F is appropriate. However, if the sample size of the data for testing a null hypothesis is small, model validation techniques may be considered unpalatable, thus warranting the use of a randomised nonparametric type test to ensure that the P -value statistic satisfies the uniformity condition. Such types of P -values will be studied in greater detail in Section 4. For now, we use Theorem 2.3 to examine the relationship among decision processes Δ_c^+ , Δ_c^- , and Δ_c^0 and P -values P^+ , P^- , and $P_{\Delta_c^0}$.

Corollary 2.4—Let Δ_c^+ and Δ_c^- be defined as in Section 2.2 and let $c \in [0, 1]$ be fixed. If Δ_c^+ and Δ_c^- are both $\mathcal{F}_0$ -level-valid, then Δ_c^0 is also $\mathcal{F}_0$ -level-valid, hence $P_{\Delta_c^0}(X, U)$ is $\mathcal{F}_0$ -uniformly bounded.

Proof: The second implication follows from the first assertion. The first assertion follows immediately by noting that

$$\begin{aligned}
\sup_{F \in \mathcal{F}_0} E_F \delta_c^0(X, U; \eta) &= \sup_{F \in \mathcal{F}_0} \{E_F \delta^+(X, U; c\eta) + E_F \delta^-(X, U; (1-c)\eta)\} \\
&\leq \sup_{F \in \mathcal{F}_0} \{E_F \delta^+(X, U; c\eta) + \sup_{F \in \mathcal{F}_0} E_F \delta^-(X, U; (1-c)\eta)\} \\
&\leq c\eta + (1-c)\eta = \eta,
\end{aligned}$$

with the last inequality arising from the $\mathcal{F}_0$ -level-validity of δ^+ and δ^- .

Corollary 2.5—Assume the conditions of Corollary 2.4. If, furthermore, there exists an $F_0 \in \mathcal{F}_0$ such that $S(X) \sim Q$ when $X \sim F_0$, then Δ_c^0 is $\mathcal{F}_0$ -size-valid, hence $P_{\Delta_c^0}(X, U)$ is $\mathcal{F}_0$ -uniform.

Proof: Using the additional condition, we have that

$$E_{F_0} \delta_c^0(X, U; \eta) = E_{F_0} \delta^-(X, U; (1-c)\eta) + E_{F_0} \delta^+(X, U; c\eta) = \eta.$$

Coupling this with the result of Corollary 2.4, we obtain that $\sup_{F \in \mathcal{F}_0} E_F \delta_c^0(X, U; \eta) = \eta$, thereby proving that Δ_c^0 is $\mathcal{F}_0$ -size-valid.

3. Discussion on randomised testing

Before considering more concrete examples, we discuss the use of a randomiser U in multiple testing. Such a device is usually considered unpalatable or undesirable (cf. Randles and Wolfe 1979, p. 34) in the single-pair hypothesis testing problem. There are two major arguments against randomised tests or decision functions. The first is our uneasy feeling that we are relinquishing the decision-making to a ‘biased coin’ when we randomise, while the second is the common-sensical objection that, when using a randomised multiple testing procedure, we may end up rejecting one null hypothesis and not another *even* if the observed data for these tests are identical. Certainly, both of these concerns are quite compelling, but let us try to clarify the relevant issues.

3.1. An implementation perspective

The notion of randomised testing or decision-making is clearly not new as this is critical in the Neyman and Pearson (1933) hypothesis testing framework and is also central in decision and game theory. In fact, optimal strategies are usually randomised, also called mixed, strategies. Let us examine how we operationally implement a randomised test function, $\varphi(x)$, which takes values in $[0, 1]$. Given data x , H_0 is rejected with probability $\varphi(x)$. If $\varphi(x) \in \{0, 1\}$, then our decision is clear. However, if $\varphi(x) \in (0, 1)$, we then generate a $U \sim U[0, 1]$, and if $U \leq \varphi(x)$ we reject H_0 , otherwise we fail to reject H_0 . This protocol, however, leads to a very-human psychological difficulty since the ‘coin flip’ (U) acquires a very central role in the decision-making process, eliciting the strident cry ‘Why should I let a coin make my final decision?’

However, suppose that for every experiment leading to data x we *simultaneously* generate a uniform variate u independently of x , and call (x, u) the full data. We then let the decision function δ depend immediately on (x, u) and make it nonrandomised, that is, taking $\{0, 1\}$ -values only. In this paper, this is done by defining $\delta(x, u) = I\{u \leq \varphi(x)\}$ where the $\varphi(\cdot)$ is the usual randomised test function depending only on data x . Then, the psychological difficulty mentioned above disappears, or is at least diminished, since with the full data (x, u) at hand, an unequivocal decision can be made. Since the u -component of the data was obtained together with the experimental component x , we may view the randomisation experiment as simply a data-gathering experiment. In essence, at the moment of generating the randomiser u , no *Sword of Damocles* is hanging over our heads to make an immediate decision. This is a crucial difference with the standard protocol, so the ‘coin flip’ in this revised protocol does not acquire a ‘door-die’ quality. In essence, defining a decision function in such a manner emphasises that any randomised decision function could be viewed as nonrandomised, so long as we view (x, u) as the full experimental data.

3.2. Whither common sense or mathematical metric?

As mentioned earlier, another objection to randomised multiple testing procedures is the potential to end up with two conflicting conclusions for two genes even when they possess the same data or rank vectors. This clearly runs counter to our intuition. However, our intuition is based on our common sense, whereas a mathematical metric of assessing such procedures should be utilised. To elaborate on this matter in a concrete manner, let us suppose that we have 10,000 genes and for the m th gene we observe the value of a random variable X_m which has a binomial distribution with parameters 10 and θ_m , and suppose that we want to test for gene m the pair of hypothesis $H_{0m} : \theta_m = 0.5$ versus $H_{1m} : \theta_m = 0.7$. We may assume that the X_m s are independent of each other. For the m th pair of hypotheses, the most powerful Neyman–Pearson test of size $\alpha = 0.05$ (of course, in a multiple testing setting we actually need to adjust this value, e.g. using the Sidak sizes, but for illustrative purpose, let us stick with 0.05) is a randomised test which rejects H_{0m} if $X_m > 7$, while if $X_m = 7$, we reject H_{0m} with a probability of 0.89. The difficulty alluded to above is the possibility that, for example, genes 1 and 2 may have observed values $x_1 = 7$ and $x_2 = 7$. When using this multiple testing procedure there is a probability of $(2)(0.89)(0.11) = 0.1958$ that we will end up with different decisions for these genes, even though they have exactly the same data. We could certainly have used the nonrandomised 0.05-level tests for each of these genes where H_{0m} is rejected if and only if $X_m > 7$. If this multiple testing procedure is used, then we will never encounter the common-sensical impasse described above.

But, the question is, which of these two multiple testing procedures is better, mathematically speaking? Let us now assume that among these 10,000 genes, 20% have their alternative hypotheses being correct. Since the probability of observing the value of 7 for a $\text{Bin}(10, 0.7)$ distribution is 0.2668, then we would expect that among the 2000 genes with $\theta_m = 0.7$, about 534 of them will get a value of $x_m = 7$. Among these 534 cases, if we use the randomised multiple testing procedure, we expect to find 475 of them to have correct alternative hypotheses, whereas if we used the nonrandomised multiple testing procedure, in all of these 534 cases we will fail to reject their null hypotheses. Thus, many more real discoveries will be achieved by using the randomised multiple decision function, with potential practical,

important, and even ethical implications such as improving the odds of discovering new and effective drugs. But then, one may still wonder how we could justify making different decisions for these genes where the observed values of X_m s are equal to 7. However, statistics, at least from the frequentist viewpoint, is never about *individual* cases ('genes'), but rather it is always about the *ensemble* of cases. Thus, from this perspective, we feel that the afore-mentioned common-sensical objection to randomised multiple testing procedures is more of a mirage.

3.3. Alternatives to randomised testing

The goal of this paper is to provide a general definition of a P -value statistic, and to further demonstrate that the type of P -value statistic being used in a multiple testing method will have an impact on the performance of the method, especially if the procedure is developed under the uniformity condition. However, not all methods are developed under this uniformity condition. Let us pause to discuss some of these alternative methods.

Others have considered modifying the nonrandomised (crisp) P -value, defined by $P(x) = \mathbf{P}_{H_0} \{S(X) \leq S(x)\}$, but without making use of a randomiser. Lancaster (1961) defined the mid- P -value via $P_{\text{mid}} = \mathbf{P}_{H_0} \{S(X) > S(x)\} + (1/2)\mathbf{P}_{H_0} \{S(X) = S(x)\}$. Westfall and Soper (1998) considered weighting the nonrandomised P -values, while Westfall and Soper (2001) considered constructing P -values with weighted test statistics. Both methods were designed to control the family-wise error rate (the probability of making one or more false discoveries). Geyer and Meeden (2005) used fuzzy set theory to develop the 'fuzzy' P -value, and show that it can be uniformly distributed under the null hypothesis. Here, the test function $\phi(x)$ is viewed as a fuzzy set membership function. Berger and Casella (2005) show that, for a realisation of the data x , the corresponding fuzzy P -value can be viewed as a uniform random variable over the interval $(p_l(x), p_u(x)]$ where $p_l(x) = \mathbf{P}_{H_0} \{S(X) > S(x)\}$ and $p_u(x) = \mathbf{P}_{H_0} \{S(X) \leq S(x)\}$. That is, the fuzzy P -value (which we should point out is not actually a 'value' in the traditional sense) is $P_{\text{fuzzy}} \sim U(p_l(x), p_u(x))$. Hence, fuzzy P -values are not directly applicable to multiple testing methods that require P -values to be realisations of a random variable. For example, Kulinskaya and Lewin (2009) modify the BH procedure, to allow for fuzzy P -values.

As pointed out by a referee, some multiple testing methods (see Westfall and Young 1989; Tarone 1990; Westfall and Wolfinger 1997; Hommel and Krummenauer 1998; Roth 1999; Gilbert 2005; Gutman and Hochberg 2007; Westfall and Troendle 2008) exploit the discreteness of the nonrandomised P -values. For example, consider testing the pairs of hypotheses $H_{0m} : p_{1,m} = p_{2,m}$ versus $H_{1m} : p_{1,m} < p_{2,m}$ for $m = 1, 2, \dots, M$ with data $X_{1,m} \sim \text{Bin}(10, p_{1,m}), X_{2,m} \sim \text{Bin}(10, p_{2,m})$ using Fisher's exact tests. Recall that the P -value for Fisher's exact test is based on the conditional distribution $P(X_{1,m} = x_{1,m} | X_m = x_m, p_{1,m} = p_{2,m})$, where $X_m = X_{1,m} + X_{2,m}$. For brevity, we write $P_0(X_{1,m} = x_{1,m} | X_m = x_m) = P(X_{1,m} = x_{1,m} | X_m = x_m, p_{1,m} = p_{2,m})$. Note that if $X_m = 0$, for example, then $X_{1,m} = 0$. Hence, $P_0(X_{1,m} = 0 | X_m = 0) = 1$, so the (nonrandomised) P -value for this test is 1. Even if $X_m = 1$, then $P_0(X_{1,m} = 0 | X_m = 1) = P_0(X_{1,m} = 1 | X_m = 1) = 1/2$ which again results in a (nonrandomised) P -value of 1. In Tarone (1990) (other procedures mentioned operate in a similar manner), the standard Bonferroni (1936) procedure is adapted by automatically accepting those null

hypotheses whose X_m is small (or large), and testing the remaining $M^* - M$ null hypotheses at level α/M^* . It is shown that the procedure controls the family-wise error rate at level α .

Now, let us consider an interesting example posed by a referee, and briefly compare two multiple testing methods. Method 1 is the method described above, which automatically accepts some null hypotheses and tests the remaining M^* null hypotheses at level α/M^* using nonrandomised P -values from Fisher's exact test. Method 2 simply tests all M null hypotheses at level α/M using randomised P -values from Fisher's exact test. Now, suppose that $p_{1,m} = p_{2,m} = 0.0001$ for $m = 1, 2, \dots, 4999$, $p_{1,5000} = 0.25$ and $p_{2,5000} = 2p_{1,5000}$. Then method 1 will likely automatically accept the first 4999 true null hypotheses (note that $P(X_m = 1 | p_{1,m} = p_{2,m} = 0.0001)^{4999} = 0.99$) and it will test $H_{0,5000}$ at level $\alpha/1$. On the other hand, method 2 tests all 4999 true null hypotheses and the false null hypotheses at level $\alpha/5000$. Clearly, in this scenario, method 1 is superior to method 2 since it will likely test the false null hypothesis at level α rather than $\alpha/5000$, and automatically accept the true null hypotheses. However, let us consider the other side of the coin. Suppose that, for $m = 1, 2, \dots, 4999$, $p_{1,m} = 0.0001$ and $p_{2,m} = 2p_{1,m}$, while for $m = 5000$, $p_{1,m} = p_{2,m} = 0.25$. Method 1 will likely erroneously accept all of the false null hypotheses (note $P(X_m = 1 | p_{1,m} = 0.0001, p_{2,m} = 0.0002)^{4999} = 0.98$), and test the true null hypothesis at level α . On the other hand, method 2 will test the 4999 false null hypotheses at level $\alpha/5000$, rather than automatically (erroneously) accepting them. Further, it will test the true null hypothesis at level $\alpha/5000$ rather than α , so that the probability of making a false discovery is also much lower for method 2. Thus, each of these methods have certain situations in which they would be expected to perform well. It would be interesting to compare these two methods in more detail, but our goal in this present paper is limited to just comparing P -value statistics rather than different multiple testing methods.

4. Concrete examples

Observe that the formulation in the preceding section is general in that the observable X may reside in a complicated sample space $\mathcal{X}$. In particular, it could be a discrete, continuous, or mixed-type random vector. It could even be a collection of stochastic processes. Thus, the notion of a randomised P -value statistic carries over to these complicated data structures.

In our concrete examples, we consider the two-group location-shift model. Let

$\mathbf{X} = (X_1, \dots, X_{n_1}) \stackrel{\text{i.i.d.}}{\sim} F(x)$, and $\mathbf{Y} = (Y_1, \dots, Y_{n_2}) \stackrel{\text{i.i.d.}}{\sim} F(y - \theta)$ where $F(\cdot) \in \mathcal{F} = \{H : H \text{ is a continuous distribution function on } \mathbb{R}\}$, $\theta \in \mathbb{R}$, and $\mathbf{X}$ and $\mathbf{Y}$ are independent. We also let $U \sim U[0, 1]$ which is independent of $(\mathbf{X}, \mathbf{Y})$. The problem is to test $H_0 : (\theta, F) \in \mathcal{F}_0$ versus $H_1 : (\theta, F) \in \mathcal{F}_1$ where $\mathcal{F}_0 = \{(\theta, F) : \theta \geq 0, F \in \mathcal{F}\}$ and $\mathcal{F}_1 = \{(\theta, F) : \theta < 0, F \in \mathcal{F}\}$. The decision functions will be of form $\delta : \mathbb{R}_{n_1} \times \mathbb{R}_{n_2} \times [0, 1] \times [0, 1] \rightarrow \{0, 1\}$, where $\delta(\mathbf{x}, \mathbf{y}, u; \eta) = 1(0)$ means reject (do not reject) H_0 at size index η with augmented data $(\mathbf{x}, \mathbf{y}, u)$. We now describe the specific decision processes of interest.

4.1. Decision processes

The standard two-sample T -test for a shift in location could be invoked to test H_0 versus H_1 . Define the decision function δ_T by

$$\delta_T(\mathbf{x}, \mathbf{y}, u; \eta) = I\{T(\mathbf{x}, \mathbf{y}) \leq \mathcal{T}_{n_1+n_2-2}^{-1}(\eta)\}$$

with

$$T(\mathbf{x}, \mathbf{y}) = \frac{\bar{y} - \bar{x}}{s_p \sqrt{1/n_1 + 1/n_2}}; \quad s_p = \sqrt{\frac{(n_1-1)s_x^2 + (n_2-1)s_y^2}{(n_1+n_2-2)}},$$

where $\bar{x}$, $\bar{y}$, s_x , and s_y are the sample means and standard deviations of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\mathcal{T}_k(\cdot)$ is the cumulative distribution function of Student's central T -distribution with k degrees-of-freedom. Form the decision process $T = \{\delta_T(\mathbf{X}, \mathbf{Y}, U; \eta) : \eta \in [0, 1]\}$. The well-known two-sample T -test P -value can be recovered from Equation (1) since

$$\begin{aligned} P_{\Delta_T}(\mathbf{X}, \mathbf{Y}, U) &= \inf\{\eta \in [0, 1] : \delta_T(\mathbf{X}, \mathbf{Y}, U; \eta) = 1\} \\ &= \inf\{\eta \in [0, 1] : I[T(\mathbf{X}, \mathbf{Y}) \leq \mathcal{T}_{n_1+n_2-2}^{-1}(\eta)] = 1\} \\ &= \inf\{\eta \in [0, 1] : I[\mathcal{T}_{n_1+n_2-2}(T(\mathbf{X}, \mathbf{Y})) \leq \eta] = 1\} \\ &= \mathcal{T}_{n_1+n_2-2}(T(\mathbf{X}, \mathbf{Y})). \end{aligned}$$

Observe that since $T(\mathbf{X}, \mathbf{Y})$ has a continuous distribution, there is no need for randomisation. Hence T and P_T could be viewed as nonrandomised.

Next we consider the validity of the T -test under a normal model and the more general model $\mathcal{F}$. Denote the family $\mathcal{F}_0^{\text{Norm}} = \{(\theta, F) : \theta \geq 0, F = N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ where $N(\mu, \sigma^2)$ is a normal distribution function with mean μ and variance σ^2 . It is well-known that $T(\mathbf{X}, \mathbf{Y})$ has distribution function $\mathcal{T}_{n_1+n_2-2}$ if and only if $(\theta, F) \in \mathcal{F}_0^{\text{Norm}}$ and $\theta = 0$. Further, $\mathbf{P}_{(\theta, F)}\{T(\mathbf{X}, \mathbf{Y}) \leq t\} < \mathcal{T}_{n_1+n_2-2}(t)$ for every $\theta > 0$, $t \in \mathbb{R}$, and $F \in \mathcal{F}_0^{\text{Norm}}$. Hence

$$\begin{aligned} \sup_{(\theta, F) \in \mathcal{F}_0^{\text{Norm}}} E_{(\theta, F)}\{\delta_T(\mathbf{X}, \mathbf{Y}, U; \eta)\} &= \sup_{(\theta, F) \in \mathcal{F}_0^{\text{Norm}}} \mathbf{P}_{(\theta, F)}\{T(\mathbf{X}, \mathbf{Y}) \leq \mathcal{T}_{n_1+n_2-2}^{-1}(\eta)\} \\ &= \mathcal{T}_{n_1+n_2-2}(\mathcal{T}_{n_1+n_2-2}^{-1}(\eta)) = \eta \end{aligned}$$

for every $\eta \in [0, 1]$. Thus, T is $\mathcal{F}_0^{\text{Norm}}$ -size-valid, and by Theorem 2.3, $P_T(\mathbf{X}, \mathbf{Y}, U)$ is $\mathcal{F}_0^{\text{Norm}}$ -uniform. On the other hand, $T(\mathbf{X}, \mathbf{Y})$ is not distributed according to a $\mathcal{T}_{n_1+n_2-2}$ distribution for many $(\theta, F) \in \mathcal{F}_0$ even when $\theta = 0$ (for instance, take F to be an exponential cumulative distribution function $F(x) = (1 - e^{-x})I(x > 0)$). Hence, T need not be $\mathcal{F}_0$ -size-valid and, by Theorem 2.3, P_T need not be $\mathcal{F}_0$ -uniform.

Next we consider a nonrandomised Wilcoxon decision process. Define

$$\delta_W(\mathbf{x}, \mathbf{y}, u; \eta) = I\{W(\mathbf{x}, \mathbf{y}) < \mathcal{W}_{n_1, n_2}^{-1}(\eta)\},$$

where $W(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{n_2} R_j(\mathbf{x}, \mathbf{y})$ and $R_j(\mathbf{x}, \mathbf{y}) = \text{rank of } y_j \text{ among } (\mathbf{x}, \mathbf{y})$, and $\mathcal{W}_{n_1, n_2}(\cdot)$ is the cumulative distribution function of a Wilcoxon rank-sum statistic for vectors of length n_1 and n_2 and $\theta = 0$. The decision process corresponding to δ_W is given by $\delta_W(\mathbf{X}, \mathbf{Y}, U; \eta) : \eta \in [0, 1]$ and the associated P -value statistic is

$$\begin{aligned} P_{\Delta_W}(\mathbf{X}, \mathbf{Y}, U) &= \inf\{\eta \in [0, 1] : \delta_W(\mathbf{X}, \mathbf{Y}, U; \eta) = 1\} \\ &= \inf\{\eta \in [0, 1] : I[W(\mathbf{X}, \mathbf{Y}) < \mathcal{W}_{n_1, n_2}^{-1}(\eta)] = 1\} \\ &= \mathcal{W}_{n_1, n_2}(W(\mathbf{X}, \mathbf{Y})). \end{aligned}$$

Again, δ_W and P_{Δ_W} do not depend on u and could therefore be considered to be nonrandomised.

We can again use Theorem 2.3 to examine the distribution of P_{Δ_W} . By Theorem 4.2.13 in Randles and Wolfe (1979), for every $F \in \mathcal{F}$ and $\theta > \theta'$, $\mathbf{P}_{(\theta, F)}\{W(\mathbf{X}, \mathbf{Y}) \leq w\} \leq \mathbf{P}_{(\theta', F)}\{W(\mathbf{X}, \mathbf{Y}) \leq w\}$ for every $w \in \{n_2(n_2 + 1)/2, \dots, n_2(2n_1 + n_2 + 1)/2\}$, the range of $W(\mathbf{X}, \mathbf{Y})$. Hence,

$$\begin{aligned} \sup_{(\theta, F) \in \mathcal{F}_0} E_{(\theta, F)}\{\delta_W(\mathbf{X}, \mathbf{Y}, U; \eta)\} &= \sup_{(\theta, F) \in \mathcal{F}_0} \mathbf{P}\{W(\mathbf{X}, \mathbf{Y}) < \mathcal{W}_{n_1, n_2}^{-1}(\eta)\} \\ &= \mathcal{W}_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta) -) \\ &\leq \eta \end{aligned}$$

for every $\eta \in [0, 1]$. Therefore, δ_W is $\mathcal{F}_0$ -level-valid. Thus, by Theorem 2.3, $P_{\Delta_W}(\mathbf{X}, \mathbf{Y}, U)$ is $\mathcal{F}_0$ -uniformly bounded. However, by the discreteness of $\mathcal{W}_{n_1, n_2}$ we have

$\mathcal{W}_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta) -) < \eta$ for some η . For example, take $\eta \in (\mathcal{W}_{n_1, n_2}(w), \mathcal{W}_{n_1, n_2}(w + 1))$ with w and $w + 1$ being possible values of $W(\mathbf{X}, \mathbf{Y})$. Thus, δ_W is not $\mathcal{F}_0$ -size-valid, and again by Theorem 2.3, $P_{\Delta_W}(\mathbf{X}, \mathbf{Y}, U)$ is not $\mathcal{F}_0$ -uniform.

Now let us consider the randomised Wilcoxon rank-sum test. Define decision function δ_{WR} by

$$\delta_{WR}(\mathbf{x}, \mathbf{y}, u; \eta) = I\{W(\mathbf{x}, \mathbf{y}) < \mathcal{W}_{n_1, n_2}^{-1}(\eta)\} + I\{u \leq \gamma(\eta)\} I\{W(\mathbf{x}, \mathbf{y}) = \mathcal{W}_{n_1, n_2}^{-1}(\eta)\},$$

where $\gamma(\eta) = [\eta - \mathcal{W}_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta) -)] / [w_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta))]$ and $w_{n_1, n_2}(\cdot)$ is the probability mass function of a Wilcoxon rank-sum statistic for vectors of length n_1 and n_2 when $\theta = 0$ and $F \in \mathcal{F}$. Form the decision process $\delta_{WR} = \{\delta_{WR}(\mathbf{X}, \mathbf{Y}, U; \eta) : \eta \in [0, 1]\}$. It can be verified that $\{(\mathbf{x}, \mathbf{y}, u, \eta) : \delta_{WR}(\mathbf{x}, \mathbf{y}, u; \eta) = 1\} = \{(\mathbf{x}, \mathbf{y}, u, \eta) : \eta - \mathcal{W}_{n_1, n_2}(W(\mathbf{x}, \mathbf{y}) -) + uw_{n_1, n_2}(W(\mathbf{x}, \mathbf{y}))\} = 1\}$. Hence

$$\begin{aligned} P_{\Delta_{WR}}(\mathbf{X}, \mathbf{Y}, U) &= \inf\{\eta \in [0, 1] : \delta_{WR}(\mathbf{X}, \mathbf{Y}, U; \eta) = 1\} \\ &= \inf[\eta \in [0, 1] : I\{W(\mathbf{X}, \mathbf{Y}) < \mathcal{W}_{n_1, n_2}^{-1}(\eta)\} + I\{U \leq \gamma(\eta)\} I\{W(\mathbf{X}, \mathbf{Y}) = \mathcal{W}_{n_1, n_2}^{-1}(\eta)\} = 1] \\ &= \mathcal{W}_{n_1, n_2}(W(\mathbf{X}, \mathbf{Y}) -) + U w_{n_1, n_2}(W(\mathbf{X}, \mathbf{Y})). \end{aligned}$$

Though the null hypotheses in this section are stated in terms of the location-shift null model $\mathcal{F}_0$, it can be shown that P_{WR} is size-valid under a less restrictive null model, $\mathcal{H}_0$, and hence P_{WR} is $\mathcal{H}_0$ -size-valid. Similar arguments can be used to show that P_{WR} is $\mathcal{F}_0$ -size-valid and that P_{WR} is $\mathcal{F}_0$ -uniform.

Theorem 4.1—Suppose that $X_1, \dots, X_{n_1} \stackrel{\text{i.i.d.}}{\sim} F$ and $Y_1, \dots, Y_{n_2} \stackrel{\text{i.i.d.}}{\sim} G$ and $F, G \in \mathcal{F}$. For testing $H_0: (F, G) \in \mathcal{H}_0 \equiv \{(F, G) \in \mathcal{F}: F(t) \leq G(t), \forall t \in \mathbb{R}\}$ versus $H_1: (F, G) \in \mathcal{H}_1 \equiv \{(F, G) \in \mathcal{F}: F(t) < G(t), \forall t \in \mathbb{R}\}$, the decision process P_{WR} is $\mathcal{H}_0$ -size-valid and $P_{WR}(\mathbf{X}, \mathbf{Y}, U)$ is $\mathcal{H}_0$ -uniform.

Proof: It follows from Theorem 2.3 that $P_{WR}(\mathbf{X}, \mathbf{Y}, U)$ is $\mathcal{H}_0$ -uniform if P_{WR} is $\mathcal{H}_0$ -size-valid. Hence, it suffices to show that $\sup_{(F, G) \in \mathcal{H}_0} E_{F, G}\{\delta_{WR}(\mathbf{X}, \mathbf{Y}, U; \eta)\} = \eta$ for any $\eta \in [0, 1]$. For any F and G and for any $F' \in \mathcal{F}$ such that $F'(t) \leq F(t), \forall t \in \mathbb{R}$, we have, by Theorem 4.3.3 in Randles and Wolfe (1979), that $\mathbf{P}_{(F', G)}\{W(\mathbf{X}, \mathbf{Y}) \leq w\} \leq \mathbf{P}_{(F, G)}\{W(\mathbf{X}, \mathbf{Y}) \leq w\}$ for any $w \in \mathbb{R}$. Further, $\mathbf{P}_{(G, G)}\{W(\mathbf{X}, \mathbf{Y}) \leq w\}$ does not depend on G for $G \in \mathcal{F}$. Hence, $\sup_{(F, G) \in \mathcal{H}_0} \mathbf{P}_{(F, G)}\{W(\mathbf{X}, \mathbf{Y}) \leq w\} = \mathbf{P}_{(G, G)}\{W(\mathbf{X}, \mathbf{Y}) \leq w\} = \mathcal{W}_{n_1, n_2}(w)$ for any $G \in \mathcal{F}$. Hence,

$$\begin{aligned} \sup_{(F, G) \in \mathcal{H}_0} E_{(F, G)}\{\delta_{WR}(\mathbf{X}, \mathbf{Y}, U; \eta)\} &= \mathcal{W}_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta) -) + \gamma(\eta) w_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta)) \\ &= \mathcal{W}_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta) -) + \left\{ \frac{\eta - \mathcal{W}_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta) -)}{w_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta))} \right\} w_{n_1, n_2}(\mathcal{W}_{n_1, n_2}^{-1}(\eta)) = \eta \end{aligned}$$

for any $\eta \in [0, 1]$. Thus, P_{WR} is $\mathcal{H}_0$ -size-valid and $P_{WR}(\mathbf{X}, \mathbf{Y}, U)$ is $\mathcal{H}_0$ -uniform.

Let us also examine the densities of the three P -value statistics for several different data generating distributions F . Figure 1 displays the simulated densities of $P_T(\mathbf{X}, \mathbf{Y}, U)$, $P_W(\mathbf{X}, \mathbf{Y}, U)$, and $P_{WR}(\mathbf{X}, \mathbf{Y}, U)$ when F is a normal, a Cauchy, or a Laplace distribution function and $n_1 = n_2 = 5$. Notice that P_T is uniformly distributed only when F is normal, while P_{WR} is uniformly distributed for every F considered. Observe that P_W is not uniformly distributed for any F considered, though the distribution-free property of P_W is manifested by noting that the density plot for P_W is invariant with respect to F .

5. Application to FDR-controlling procedures

The framework from Section 2 easily extends to the multiple testing setting. However, it now becomes possible to define decision functions and P -values in more general ways. We briefly present a framework for this multiple testing setting for clarity.

For each $m \in \mathcal{M} = \{1, 2, \dots, M\}$, let observable random entity $X_m \in \mathcal{X}_m$ have (marginal) distribution function F_m . Further, let $\mathcal{F}_m$ be a model for F_m and $\mathcal{F}_{0m}$ a nonempty subset of $\mathcal{F}_m$. Consider testing $H_{0m}: F_m \in \mathcal{F}_{0m}$ with observable random entities X_m and U_m , where X_m is independent of U_m , and the U_m s are independent standard uniform random variables. Define decision functions $\delta_m: \mathcal{X}_m \times [0, 1] \rightarrow \{0, 1\}$ by $\delta_m(X_m, U_m)$. Again we write $\delta_m(X_m, U_m; \eta)$ for $\delta_m(X_m, U_m)$ when δ_m has size index η . The size of δ_m is $\eta = \sup_{F_m \in \mathcal{F}_{0m}} E_{F_m}\{\delta(X_m, U_m)\}$. Assume that for every $F_m \in \mathcal{F}_m$, $\delta_m(x_m, u_m; \eta)$ is nondecreasing and right-continuous in η with $\delta_m(x_m, u_m; 0) = 0$ and $\delta_m(x_m, u_m; 1) = 1$ for almost all $(x_m, u_m) \in \mathcal{X}_m \times [0, 1]$ with

respect to $F_m \otimes \lambda_m$, where λ_m is a uniform probability measure on $[0, 1]$. Define decision processes $\delta_m = \{\delta_m(X_m, U_m; \eta) : \eta \in [0, 1]\}$. The P -value statistic for δ_m is defined via

$$P_{\Delta_m}(X_m, U_m) = \inf\{\eta \in [0, 1] : \delta_m(X_m, U_m; \eta) = 1\}.$$

There are currently many FDR methods that assume P -values from true null hypotheses are uniformly distributed. However, we restrict our study of the impact of the use of randomised P -values on just two of these methods. The methods considered are the BH procedure developed in Benjamini and Hochberg (1995) and Efron's empirical Bayes procedure developed in Efron et al. (2001) and Efron (2004, 2007, 2008).

5.1. Simulation setup

The simulation study is designed to mimic a small-scale microarray analysis. That is, the number of hypotheses to be tested is large, the sample size for each test is small, and most null hypotheses are true. Specifically, we consider the two-sample location shift model for each $m \in \mathcal{M} = \{1, 2, \dots, 1000\}$. Assume that $\mathbf{X}_m = (X_{1m}, \dots, X_{n_1m}) \stackrel{\text{i.i.d.}}{\sim} F_m(x)$; and $\mathbf{Y}_m = (Y_1, \dots, Y_{n_2m}) \stackrel{\text{i.i.d.}}{\sim} F_m(y - \theta_m)$, where $F_m \in \mathcal{F}$, $\theta_m \in \mathbb{R}$, and $\mathbf{X}_m$, $\mathbf{Y}_m$, and U_m are independent. Define decision processes $\delta_{T,m} = \delta_T$, $\delta_{W,m} = \delta_W$, and $\delta_{WR,m} = \delta_{WR}$ for $m \in \mathcal{M}$, where δ_T , δ_W , and δ_{WR} are defined as in Section 4. The P -values for each of the decision processes are then defined via $P_T(\mathbf{X}_m, \mathbf{Y}_m, U_m)$, $P_W(\mathbf{X}_m, \mathbf{Y}_m, U_m)$, and $P_{WR}(\mathbf{X}_m, \mathbf{Y}_m, U_m)$. The problem is to test $H_{0m} : (\theta_m, F_m) \in \mathcal{F}_0$ versus $H_{1m} : (\theta_m, F_m) \in \mathcal{F}_1$, where $\mathcal{F}_0$ and $\mathcal{F}_1$ are defined as in Section 4.

Data are generated as follows. For a single data set, H_{0m} is tested with independent $X_{1m}, \dots, X_{5m}, Y_{1m}, \dots, Y_{5m}$ and U_m for $m \in \{1, 2, \dots, 1000\}$. For $m \in \{1, 2, \dots, 900\}$ $X_{1m}, \dots, X_{5m}, Y_{1m}, \dots, Y_{5m} \stackrel{\text{i.i.d.}}{\sim} F_m(\cdot - 0)$. For $m = 901, 902, \dots, 1000$, $X_{1m}, \dots, X_{5m} \stackrel{\text{i.i.d.}}{\sim} F_m(\cdot - 0)$, while $Y_{1m}, \dots, Y_{5m} \stackrel{\text{i.i.d.}}{\sim} F_m(\cdot - (-2))$. We consider the same three different data generating distributions from the last section. That is, F_m is a normal, a Cauchy, or a Laplace distribution with scale 1 and location -2 or 0 . The number of simulation replications is 5000.

5.2. Benjamini and Hochberg (1995) procedure

The BH procedure is designed to control the FDR. Let random variable V be the total number of falsely rejected null hypotheses (false discoveries) and let random variable R be the total number of rejected null hypotheses (discoveries). In our notation,

$V = \sum_{m=1}^M I(H_{0m} \text{ is true}) \delta_m$ and $R = \sum_{m=1}^M \delta_m$. Let random variable Q be defined by $Q = V / \max\{R, 1\}$. The FDR is defined via $\text{FDR} = E[Q]$, where the expectation operator $E(\cdot)$ is with respect to the (true) joint distribution governing the random observables. The BH procedure is defined as follows. Let p_m be realised P -values for testing null hypotheses H_{0m} , $m \in \mathcal{M}$, and let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(M)}$ be the ordered P -values with $H_{0(j)}$ being the null hypothesis associated with $p_{(j)}$. The procedure rejects $H_{0(1)}, \dots, H_{0(k)}$, where k is defined via

$$k = \max \left\{ m \in \mathcal{M} : p_{(m)} \leq \frac{m\alpha}{M} \right\}.$$

Benjamini and Hochberg (1995) showed that if the P -value statistics from the true null hypothesis are uniformly distributed and independent, then the FDR for this procedure is

less than or equal to $M_0\alpha/M$, where $M_0 = \sum_{m=1}^M I(H_{0m} \text{ is true})$, the number of ‘true’ null hypotheses. Benjamini and Yekutieli (2001) and Sarkar (2002) showed that the procedure controls the FDR under certain dependency structures as well.

P -values for the j th replicated data set are computed via $P_T(\mathbf{x}_{mj}, \mathbf{y}_{mj}, u_{mj})$, $P_W(\mathbf{x}_{mj}, \mathbf{y}_{mj}, u_{mj})$, and $P_{WR}(\mathbf{x}_{mj}, \mathbf{y}_{mj}, u_{mj})$ for $j = 1, 2, \dots, 5000$. The procedure is applied at $\alpha = (0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5)$. For a given α and for data set j , the true false discovery proportion $q_j(\alpha)$ and the number of discoveries $r_j(\alpha)$ are recorded. Plots of $(\alpha, \overline{q(\alpha)})$ and $(\alpha, \overline{r(\alpha)})$, where $\overline{q(\alpha)} = \sum_{j=1}^{5000} q_j(\alpha)/5000$ and $\overline{r(\alpha)} = \sum_{j=1}^{5000} r_j(\alpha)/5000$ are the average false discovery proportion and the average number of discoveries, respectively, are presented in Figure 2.

Simulation studies suggest that the BH procedure controls the FDR at level $M_0\alpha/M = 0.9\alpha$ when invoking randomised Wilcoxon P -values when F_m is a normal, Laplace, or Cauchy distribution function. When making use of the T -test P -values, however, the FDR is not necessarily controlled at 0.9α . The procedure is conservative in that the FDR is often less than 0.9α when using nonrandomised Wilcoxon P -values. The randomised Wilcoxon P -values always allow for more discoveries than the nonrandomised Wilcoxon P -values, and typically allow for more discoveries than the T -test P -values, the exception being when F is normal, which is to be expected since both tests are valid but with the T -test being more powerful than the Wilcoxon test.

5.3. Efron’s empirical Bayes procedure

The procedure studied in this section, which was introduced in Efron et al. (2001) and further developed in Efron (2004, 2007, 2008), makes use of Z -values, which are typically computed via $z_m = \Phi^{-1}(Q(s_m))$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function and $s_m \equiv s_m(x_m)$ is a test statistic with distribution Q under H_{0m} . It is then assumed that the Z -values have marginal mixture density

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

so that Z_m has density f_0 under H_{0m} and density f_1 under H_{1m} . The weight π_0 can be viewed as the proportion of true null hypotheses or as the prior probability that H_{0m} is true. Notice that f_0 is a standard normal distribution by the probability integral transformation if the statistic $Q(S_m)$ has a uniform distribution under H_{0m} . However, as noted in the previous section, $Q(S_m)$ will not have a uniform distribution when Q is discrete or when $S_m \stackrel{d}{\neq} Q$ under H_{0m} . We will consider computing Z -value statistics via

$$Z_m(X_m, U_m) = \Phi^{-1}(Q(S_m(X_m) - 1) - U_m \Delta Q(S_m(X_m))),$$

where $Q(s) = Q(s) - Q(s-)$. Observe that $Z_m(X_m, U_m)$ can be equivalently expressed $Z_m(X_m, U_m) = \Phi^{-1}(P_{-}(X_m, U_m))$. Note that even if Q is a discrete distribution, Z_m has a standard normal density so long as $P_{-}(X_m, U_m)$ has a uniform distribution under H_{0m} .

The local fdr at $Z_m = z$ (Efron 2004) is defined via

$$\text{fdr}(z) = \frac{\pi_0 f_0(z)}{f(z)} = \mathbf{P}(H_{0m} \text{ is true} | Z_m = z).$$

Sun and Cai (2007) have shown that decision rules of form $\delta_m(z_m) = I(\text{fdr}(z_m) < \gamma)$ possess certain optimality properties. However, since the local fdr depends on the unknown quantities $f_0(z)$, $f_1(z)$, and π_0 , these quantities need to be estimated in order to implement such a decision rule.

If a local fdr estimator is developed under the assumption that f_0 is a standard normal density, it shall be referred to as a type 1 estimator, while if f_0 is estimated by $\hat{f}_0$ (Efron 2004 calls $\hat{f}_0$ the empirical null distribution), the local fdr estimator is called a type 2 estimator. These estimators are denoted by

$$\text{fdr}_1^{\hat{}}(z) = \frac{\hat{\pi}_0 f_0(z)}{\hat{f}(z)} \quad \text{and} \quad \text{fdr}_2^{\hat{}}(z) = \frac{\hat{\pi}_0 \hat{f}_0(z)}{\hat{f}(z)}.$$

It is important to note that though it can be assumed that Z_m has a standard normal density under H_{0m} (and hence estimate $\text{fdr}(z)$ by $\text{fdr}_1^{\hat{}}(z)$), it is not necessary to make use of such an assumption. One may make use of the empirical null distribution and estimate $\text{fdr}(z)$ with $\text{fdr}_2^{\hat{}}(z)$. However, as stated in Efron (2008), ‘the empirical null is an expensive luxury from the point of view of estimation efficiency’. That is, it is desirable to use type 1 estimators if indeed f_0 is a standard normal density since type 2 local fdr estimators tend to have a larger standard error.

We investigate both the bias and variability of these local fdr estimators relative to the type of Z-value considered using the simulated data sets described in Section 5.1. For each set of 1000 hypotheses tests, Z-values are computed $z_{\text{WR}_m} = \Phi^{-1}(P_{\text{WR}}(\mathbf{x}_m, \mathbf{y}_m, u_m))$ and $z_{T_m} = \Phi^{-1}(P_T(\mathbf{x}_m, \mathbf{y}_m, u_m))$. The two types of estimates of the local fdr are computed for $z = (-4, -3.8, \dots, -2.2, -2)$ using the software locfdr available in R (Efron, Turnbull, and Narasimhan 2009). The discrete Z-values $\Phi^{-1}(P_{\text{W}}(\mathbf{x}_m, \mathbf{y}_m))$ are omitted from consideration because the estimation algorithms fail to converge. The true local fdr at z for simulated data set j is approximated by

$$\text{fdr}_j(z) = \frac{\sum_{m=1}^M I(\theta_{mj}=0) I(z-0.1 < z_{mj} \leq z+0.1)}{\max\{\sum_{m=1}^M I(z-0.1 < z_{mj} \leq z+0.1), 1\}},$$

where z_{mj} is the Z -value for testing H_{0m} from simulated data set j . The type i local fdr estimate for data set j is denoted $\hat{\text{fdr}}_{ij}(z)$. All default parameters were used in the R function `locfdr` (Efron et al. 2009) for computing $\hat{\text{fdr}}_1(z)$ and $\hat{\text{fdr}}_2(z)$. The maximum likelihood method was used to estimate f_0 for type 2 local fdr estimators. The average bias and standard deviation of the local fdr estimators over the 5000 replications are computed via

$$\text{bias}(\hat{\text{fdr}}_i(z)) = \overline{\hat{\text{fdr}}_i(z)} - \overline{\text{fdr}(z)} \quad \text{and} \quad s(\hat{\text{fdr}}_i(z)) = \sqrt{\frac{\sum_{j=1}^{5000} [\hat{\text{fdr}}_{ij}(z) - \overline{\hat{\text{fdr}}_i(z)}]^2}{4999}}$$

where $\overline{\hat{\text{fdr}}_i(z)} = \sum_{j=1}^{5000} \hat{\text{fdr}}_{ij}(z) / 5000$ and $\overline{\text{fdr}(z)} = \sum_{j=1}^{5000} \text{fdr}_j(z) / 5000$.

Figure 3 has plots of $(\overline{\text{fdr}(z)}, \text{bias}(\hat{\text{fdr}}_1(z)))$, and $(\overline{\text{fdr}(z)}, s(\hat{\text{fdr}}_1(z)))$ when Z -values are z_{WR_m} and z_{T_m} . We see that $\hat{\text{fdr}}_1(z)$ performs well in terms of bias and variability when considering the randomised Wilcoxon-type Z -values for each data-generating distribution considered. On the other hand, this estimator tends to be biased when F_m is non-normal, when considering T -test-type Z -values. As for $\hat{\text{fdr}}_2(z)$, we see that it is more biased and variable than $\hat{\text{fdr}}_1(z)$ regardless of which Z -value statistic is considered. However, when using the randomised Wilcoxon Z -values, recall that f_0 is standard normal so long as the F_m are continuous since P_{WR} is $\mathcal{F}_0$ -uniform. Hence, the less biased and less variable local fdr estimator can be used to define a decision rule if making use of the randomised Wilcoxon Z -values regardless of F_m . On the other hand, when computing Z -values with a T -test statistic, if F_m is not normal then f_0 need not be a standard normal density since P_T need not be uniformly distributed when $\theta_m = 0$. In short, the simulation study suggests that the type 2 local fdr estimator is indeed an ‘expensive luxury’, and also demonstrates that it is unnecessary if we make use of the randomised Wilcoxon Z -values.

6. Application to a microarray data set

Methods in Sections 4 and 5 are applied to the microarray data set in Timmons et al. (2007), where five Affymetrix chips from brown fat cells and eight Affymetrix chips from white fat cells yielded a total of 13 gene expression measurements for each of $M = 12,488$ genes. The goal is to determine, for each gene, if a gene’s expression level is related to fat cell type. The form of the data for testing H_{0m} is in Table 1.

We assume that $X_{1m}, \dots, X_{5m} \stackrel{\text{i.i.d.}}{\sim} F_m(x)$ and $Y_{1m}, \dots, Y_{8m} \stackrel{\text{i.i.d.}}{\sim} F_m(y - \theta_m)$ for $\theta_m \in \mathbb{R}$ and $F_m \in \mathcal{F}$ for each $m \in \{1, 2, \dots, M\}$. The goal is to test $H_{0m} : \theta_m = 0, F_m \in \mathcal{F}$ against $H_{1m} : \theta_m \neq 0, F_m \in \mathcal{F}$ for each m .

We first test the null hypotheses with Efron's empirical Bayes procedure using Z -values from T -tests and randomised Wilcoxon tests, which we again denote by $z_{Tm} = \Phi^{-1}(P_T(\mathbf{x}_m, \mathbf{y}_m, u_m))$ and $z_{WRm} = \Phi^{-1}(P_{WR}(\mathbf{x}_m, \mathbf{y}_m, u_m))$, respectively. The software locfdr is used to perform the analysis (Efron et al. 2009). The decision rule

$$\text{Reject } H_{0m} \text{ if } \hat{\text{fdr}}_i(z_m) < 0.2$$

is used to determine which H_{0m} s are declared false. Note that, as mentioned in Efron (2004) and Sun and Cai (2007), the FDR defined in Section 5, which is referred to as the 'tail area FDR' in Efron (2004, 2007), can be considerably less than 0.2 for this decision rule. In

Figure 4, we see that $\hat{\text{fdr}}_1(z) < \hat{\text{fdr}}_2(z)$ whether Z -values are computed from T -test or the randomised Wilcoxon test. In fact $\hat{\text{fdr}}_2(z_{Tm}) > 0.2$ and $\hat{\text{fdr}}_2(z_{WRm}) > 0.2$ for every m allowing for no null hypotheses to be rejected (genes to be 'discovered'). On the other hand,

$\hat{\text{fdr}}_1(z_{Tm}) < 0.2$ for 1881 z_{Tm} s and $\hat{\text{fdr}}_1(z_{WRm}) < 0.2$ for 2026 z_{WRm} s. Clearly, whether or not f_0 is assumed to be a standard normal distribution under H_{0m} , and which type of Z -value is used in the analysis, has a major impact on the alternative hypotheses that are 'discovered'.

Next we analyse the data with the BH procedure at FDR levels of $\alpha = 0.05, 0.1$, and 0.2 . This procedure requires that P -values for testing $H_{0m} : \theta_m = 0, F_m \in \mathcal{F}$ against $H_{1m} : \theta_m \neq 0, F_m \in \mathcal{F}$ be calculated so that small P -values are evidence against H_{0m} . Note that in the empirical Bayes analysis above, Z -values are computed using lower-tailed P -value statistics even though the alternative hypotheses are two sided. The form of the two-sided P -value for testing H_{0m} with a T -test or Wilcoxon rank-sum test is well known. From Theorem (2.1) with $c = 1/2$, we define the two-sided randomised Wilcoxon P -value via

$$P_{\Delta_{WR}^2}(\mathbf{x}_m, \mathbf{y}_m, u_m) = 2 \min\{P_{\Delta_{WR}}(\mathbf{x}_m, \mathbf{y}_m, u_m), 1 - P_{\Delta_{WR}}(\mathbf{x}_m, \mathbf{y}_m, u_m)\},$$

where Δ_{WR} is defined as in Section 4 with $n_1 = 5$ and $n_2 = 8$. It can be verified that this P -value statistic is $\mathcal{F}_0^0$ -uniform where $\mathcal{F}_0^0 = \{(\theta, F) : \theta = 0, F \in \mathcal{F}\}$ using arguments in Section 4 and Corollary 2.5. The BH procedure was applied at alpha levels of 0.05, 0.1, and 0.2 using the two-sided T -test P -values, Wilcoxon P -values, and randomised Wilcoxon P -values. Again, we see that the set of genes declared differentially expressed will depend on how the P -value for each test is calculated, as can be seen in Table 2. For $\alpha = 0.05$, the BH procedure allows for 812, 863, and 908 discoveries when using P -values from T -tests, Wilcoxon tests, and randomised Wilcoxon tests, respectively (columns 2–4). Only 637 hypotheses were rejected by the BH procedure at $\alpha = 0.05$ for all three P -value types (row 2, column 5) even though 1065 null hypotheses are rejected at $\alpha = 0.05$ for at least one of the P -value types (row 2, column 6). Of course, we cannot know the true false discovery proportion for any of the three sets of rejected null hypotheses. However, the uniformity condition will always be satisfied for the randomised Wilcoxon P -values if it is satisfied for

the T -test P -values. On the other hand, it may be the case that the uniformity condition is satisfied for the randomised Wilcoxon P -values but is not satisfied for the T -test P -values.

7. Concluding remarks

Many multiple testing procedures, especially those designed to control or estimate the FDR, are defined in terms of the P -values of the individual tests. We have provided a framework for generating a P -value that is uniformly distributed under the null hypothesis even in situations when the test statistic is discrete. This framework views a decision function $\delta(x, u; \alpha)$ as an element of the decision process $\{\delta(x, u; \alpha) : \alpha \in [0, 1]\}$. Though we call a P -value depending on uniform variate u as randomised, we believe this terminology may be slightly misleading, for when viewing u as part of data, the P -value is in fact not randomised with respect to augmented data (x, u) .

Methods in this paper make nonparametric distribution-free test procedures usable for FDR methods, providing an alternative to a parametric model for the random observables. As an example, we defined a nonparametric randomised Wilcoxon decision process to test for a shift in location across two treatment groups. It was demonstrated that the nonparametric randomised P -value statistic generated from this process is uniformly distributed under the null hypothesis so long as the data generating distribution is continuous, while the T -test P -value statistic is uniformly distributed under the null hypothesis provided the data generating distribution is normal. If data are non-normal, a situation common in microarray analysis, we have illustrated that those FDR methods requiring the uniformity assumption perform better when using a nonparametric randomised Wilcoxon P -value rather than a T -test P -value or nonrandomised Wilcoxon P -value. Methods in this paper could also be used to construct P -values for FDR methods when random observables have discrete distributions, such as when data are dichotomous.

Not all FDR methods assume the uniform model for P -value statistics from true null hypothesis. In the future, it would be interesting to compare some of these methods with some FDR methods requiring the uniform assumption. The goal of this paper, however, was to compare P -value statistics, and to demonstrate that the manner in which P -value statistics are calculated needs to be carefully considered, especially when applied to FDR methods that require P -value statistics to have uniform distributions under the null hypotheses.

Acknowledgments

The authors wish to thank the reviewers for their helpful and constructive comments. They also wish to thank Wensong Wu, Professor Joshua Tebbs, and Professor Phillip Dixon for helpful comments and discussions, and also thank Professor Peter Westfall for alerting them and providing them with relevant references. The authors also acknowledge NSF Grant DMS0805809; National Institutes of Health (NIH) Grant RR17698; and the Environmental Protection Agency (EPA) Grant RD-83241902-0 to the University of Arizona with subaward number Y481344 to the University of South Carolina. These grants partially supported this work.

References

- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*. 1995; 57:289–300.

- Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annals of Statistics*. 2001; 29:1165–1188.
- Berger RL, Casella G. Comment: Fuzzy and Randomized Confidence Intervals and P -Values. *Statistical Science*. 2005; 20:372–374.
- Bonferroni C. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936; 8:3–62.
- Cox, DR.; Hinkley, DV. *Theoretical Statistics*. London: Chapman and Hall; 1974.
- Dudoit, S.; van der Laan, MJ. *Springer Series in Statistics*. New York: Springer; 2008. *Multiple Testing Procedures with Applications to Genomics*.
- Efron B. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*. 2004; 99:96–104.
- Efron B. Size, Power and False Discovery Rates. *The Annals of Statistics*. 2007; 35:1351–1377.
- Efron B. Microarrays, Empirical Bayes and the Two-Group Model. *Statistical Science*. 2008; 23:1–22.
- Efron B, Tibshirani R, Storey JD, Tusher P. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*. 2001; 96:1151–1160.
- Efron, B.; Turnbull, B.; Narasimhan, B. R Package Version 1.1-6. CRAN; 2009. *Locfdr: Computes Local False Discovery Rates*. <http://cran.r-project.org/web/packages/locfdr/>
- Genovese C, Wasserman L. A Stochastic Process Approach to False Discovery Rate Control. *Annals of Statistics*. 2004; 32:1035–1061.
- Geyer CJ, Meeden GD. Fuzzy and Randomized Confidence Intervals and P -Values' (with comments and a rejoinder by the authors). *Statistical Science*. 2005; 20:358–387.
- Gilbert PB. A Modified False Discovery Rate Multiple-Comparisons Procedure for Discrete Data, Applied to Human Immunodeficiency Virus Genetics. *Journal of the Royal Statistical Society Series C*. 2005; 54:143–158.
- Gutman R, Hochberg Y. Improved Multiple Test Procedures for Discrete Distributions: New Ideas and Analytical Review. *Journal of Statistical Planning and Inference*. 2007; 137:2380–2393.
- Hommel G, Krummenauer F. Improvements and Modifications of Tarone's Multiple Test Procedure for Discrete Data. *Biometrics*. 1998; 54:673–681.
- Jin J, Cai TT. Estimating the Null and the Proportion of Nonnull Effects in Large-Scale Multiple Comparisons. *Journal of the American Statistical Association*. 2007; 102:495–506.
- Korn EL, Troendle JF, McShane LM, Simon R. Controlling the Number of False Discoveries: Application to High-Dimensional Genomic Data. *Journal of Statistical Planning and Inference*. 2004; 124:379–398.
- Kulinskaya E, Lewin A. On Fuzzy Familywise Error Rate and False Discovery Rate Procedures for Discrete Distributions. *Biometrika*. 2009; 96:201–211.
- Lancaster HO. Significance Tests in Discrete Distributions. *Journal of the American Statistical Association*. 1961; 56:223–234.
- Neyman J, Pearson E. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*. 1933; 231:289–337.
- Pollard KS, van der Laan MJ. Choice of a Null Distribution in Resampling-Based Multiple Testing. *Journal of Statistical Planning and Inference*. 2004; 125:85–100.
- Randles, RH.; Wolfe, DA. *Wiley Series in Probability and Mathematical Statistics*. New York–Chichester–Brisbane: John Wiley & Sons; 1979. *Introduction to the Theory of Nonparametric Statistics*.
- Roth AJ. Multiple Comparison Procedures for Discrete Test Statistics. *Journal of Statistical Planning and Inference (Multiple Comparisons; Tel Aviv, 1996)*. 1999; 82:101–117.
- Ruppert D, Nettleton D, Hwang JT. Exploring the Information in p -Values for the Analysis and Planning of Multiple-Test Experiments. *Biometrics*. 2007; 63:483–495. [PubMed: 17715492]
- Sarkar SK. Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures. *The Annals of Statistics*. 2002; 30:239–257.
- Storey J. A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B*. 2002; 64:479–498.

- Storey J. The Positive False Discovery Rate: A Bayesian Interpretation and the q -Value. *The Annals of Statistics*. 2003; 31:2012–2035.
- Sun W, Cai TT. Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control. *Journal of the American Statistical Association*. 2007; 102:901–912.
- Tarone RE. A Modified Bonferroni Method for Discrete Data. *Biometrics*. 1990; 46:515–522. [PubMed: 2364136]
- Timmons J, Wennmalm K, Larsson O, Walden T, Lassmann T, Petrovic N, Hamilton D, Gimeno R, Wahlestedt C, Baar K, Nedergaard J, Cannon B. Myogenic Gene Expression Signature Establishes that Brown and White Adipocytes Originate from Distinct Cell Lineages. *Proceedings of the National Academy of Sciences of the United States of America*. 2007; 104(11):4401–4406. [PubMed: 17360536]
- van der Laan MJ, Hubbard AE. Quantile-Function Based Null Distribution in Resampling Based Multiple Testing. *Statistical Applications in Genetics and Molecular Biology* (electronic). 2006; 5:Art. 14, 30.
- Westfall PH, Soper KA. Weighted Multiplicity Adjustments for Animal Carcinogenicity Tests. *Journal of Biopharmaceutical Statistics*. 1998; 8:23–44.
- Westfall PH, Soper KA. Using Priors to Improve Multiple Animal Carcinogenicity Tests. *Journal of the American Statistical Association*. 2001; 96:827–834.
- Westfall PH, Troendle JH. Multiple Testing with Minimal Assumptions. *Biometrical Journal*. 2008; 50:745–755. [PubMed: 18932134]
- Westfall PH, Wolfinger RD. Multiple Tests with Discrete Distributions. *The American Statistician*. 1997; 51:3–8.
- Westfall PH, Young SS. p Value Adjustments for Multiple Tests in Multivariate Binomial Models. *Journal of the American Statistical Association*. 1989; 84:780–786.
- Yekutieli D, Benjamini Y. Resampling-Based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics. *Journal of Statistical Planning and Inference*. 1999; 82:171–196.

Figure 1.

Density plots of 100,000 simulated P -values for each of $P_T(\mathbf{X}, \mathbf{Y}, U)$, $P_W(\mathbf{X}, \mathbf{Y}, U)$, and $P_{WR}(\mathbf{X}, \mathbf{Y}, U)$ when $X_1, \dots, X_5, Y_1, \dots, Y_5 \stackrel{\text{i.i.d.}}{\sim} F$, and F (location, scale) are normal(0, 1), Cauchy(0, 1), and Laplace(0, 1) distribution functions.

Figure 2.

The average FDR and average number of discoveries over 5000 replications for the BH procedure with $\alpha \in [0.01, 0.5]$. $F_m(\text{location}, \text{scale})$ are $N(\theta_m, 1)$, $\text{Cauchy}(\theta_m, 1)$, and $\text{Laplace}(\theta_m, 1)$ distributions, where $\theta_m = 0$ for $m = 1, \dots, 900$ and $\theta_m = 2$ for $m = 901, \dots, 1000$. P -values P_T , P_W , and P_{WR} are denoted by '+', ' $\triangle$ ', and ' $\circ$ ', respectively.

Figure 3.

Plots of average bias and standard deviation of

$\text{fdr}_1(\hat{z}_{T_m})$ (+), $\text{fdr}_1(\hat{z}_{W_{R_m}})$ (o), $\text{fdr}_2(\hat{z}_{T_m})$ (x), and $\text{fdr}_2(\hat{z}_{W_{R_m}})$ versus $\overline{\text{fdr}(z)}$. z -values are computed $z_{W_{R_m}} = \Phi^{-1}(P_{W_R}(\mathbf{x}_m, \mathbf{y}_m, u_m))$ and $z_{T_m} = \Phi^{-1}(P_T(\mathbf{x}_m, \mathbf{y}_m, u_m))$. Data are generated as in Section 5.1.

Local fdr analysis

Figure 4.

Plots of the local fdr estimates arising from microarray data in Timmons et al. (2007). Notes:

$\hat{fdr}_1(z_T) = (\dots)$, $\hat{fdr}_1(z_{WR}) = (-)$, $\hat{fdr}_2(z_T) = (-.-.-)$, $\hat{fdr}_2(z_{WR}) = (---)$.

Table 1
Illustration of the observed gene expression measurements along with uniform variates and P -values.

m	x_{1m}	x_{2m}	...	x_{5m}	y_{1m}	y_{2m}	...	y_{8m}	u_m	P_T	P_W	P_{WR}
1	1.22	1.66	...	2.33	5.64	1.79	...	4.05	0.210	0.317	0.177	0.150
2	3.57	19.22	...	11.89	5.17	29.49	...	11.26	0.766	0.193	0.362	0.350
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
12,488	2.52	10.91	...	22.67	10.70	7.35	...	12.81	0.872	0.505	0.528	0.521

Table 2

The number of discoveries when using T -test P -value (T), Wilcoxon P -value (W), and randomised Wilcoxon P -value (WR) for FDR control at level α are displayed for $\alpha \in \{0.05, 0.1, 0.2\}$.

α	T	W	WR	$T \cap W \cap WR$	$T \cup W \cup WR$
0.05	812	863	908	637	1065
0.10	1483	1499	1609	1200	1834
0.20	2758	2304	2749	2141	3121

The number of null hypotheses that were rejected using T -test, Wilcoxon, and randomised Wilcoxon P -values are in column 4 and the number of null hypotheses that were rejected when using at least one of the three P -values are in column 5.