Check for updates

SOFTWARE TOOL ARTICLE

[REVISED] FDRestimation: Flexible False Discovery Rate Computation in R [version 2; peer review: 2 approved]

Megan H. Murray [ID]1, Jeffrey D. Blume[1,2]

[1]Biostatistics, Vanderbilt University Medical Center, Nashville, TN, 37235, USA
[2]School of Data Science, University of Virginia, Charlottesville, VA, 22904, USA

Abstract

False discovery rates (FDR) are an essential component of statistical inference, representing the propensity for an observed result to be mistaken. FDR estimates should accompany observed results to help the user contextualize the relevance and potential impact of findings. This paper introduces a new user-friendly R pack-age for estimating FDRs and computing adjusted p-values for FDR control. The roles of these two quantities are often confused in practice and some software packages even report the adjusted p-values as the estimated FDRs. A key contribution of this package is that it distinguishes between these two quantities while also offering a broad array of refined algorithms for estimating them. For example, included are newly augmented methods for estimating the null proportion of findings - an important part of the FDR estimation procedure. The package is broad, encompassing a variety of adjustment methods for FDR estimation and FDR control, and includes plotting functions for easy display of results. Through extensive illustrations, we strongly encourage wider reporting of false discovery rates for observed findings.

Keywords

false discovery rate, multiple comparisons, adjusted p-value, null proportion estimation, R Package

Open Peer Review

Reviewer Status ✔ ✔

	Invited Reviewers	
	1	**2**
version 2 (revision) 19 Oct 2021	✔ report	✔ report
version 1 03 Jun 2021	? report	

1. **Ted Westling** [ID], University of Massachusetts Amherst, Amherst, USA

   Yujian Wu [ID], University of Massachusetts Amherst, Amherst, USA

2. **David Colquhoun** [ID], University College London, London, UK

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the RPackage gateway.

Corresponding author: Megan H. Murray (murraymegan25@gmail.com)

Author roles: Murray MH: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Blume JD**: Conceptualization, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Murray MH and Blume JD. **FDRestimation: Flexible False Discovery Rate Computation in R [version 2; peer review: 2 approved]** F1000Research 2021, **10**:441 https://doi.org/10.12688/f1000research.52999.2

First published: 03 Jun 2021, **10**:441 https://doi.org/10.12688/f1000research.52999.1

> **REVISED** **Amendments from Version 1**
>
> The article was revised in response to the review by Ted Westling and Yujian Wu. No main ideas or content were changed. Figures 3, 5, 6, 7 were updated. Specifically Figure 7 is new and was added to provide evidence of the simulated MSE of the different null proportion methods.
>
> **Any further responses from the reviewers can be found at the end of the article**

Introduction

The reporting of observed results is not without controversy when multiple comparisons or multiple testing is involved. Classically, p-values were adjusted to maintain control of the family-wise error rate (FWER). However, this control can come at the cost of substantial Type II Error rate inflation, especially in large-scale inference settings where the number of comparisons is several orders of magnitude greater than the sample size. Large scale inference settings occur frequently in the analysis of genomic, imaging, and proteomic data, for example. Recently, it has become popular to control the false discovery rate (FDR) instead of the FWER in these settings because its Type II Error rate inflation is much less severe. The FDR is essentially the propensity for a finding to be mistaken i.e., the propensity for a non-null claim to be, in fact, wrong.

Controlling the FDR at or below a specific level, say γ, does *not* imply that the Type I Error rate, per-comparison or family-wise, is also controlled at the same level. The increase in Type I Errors that is allowed by FDR control is accompanied by fewer Type II errors. Moreover, different approaches to controlling the FDR allow for different degrees of error trade-off. And software for implementing these approaches vary widely in their scope, options, and accessibility. In addition, methods for controlling the FDR, which use the classical rejection-testing framework, are often confused with the methods used to provide an estimate of the FDR for a particular result.

The `FDRestimation` package distinguishes between methods for FDR control and methods for FDR estimation, and it allows the user to easily access complex statistical routines for computing the desired quantity. The plotting functions allow users to visually assess results and differences between methods. We should note that the base package function `stats::p.adjust` is now frequently used to compute the estimated FDR, however `stats::p.adjust` actually reports the adjusted p-values for FDR control, and these are not always the same thing. More on this important distinction later. Our package also provides a wide range of methods for estimating the FDR, estimating the proportion of null results, and computing the adjusted p-values. We hope by clearly illustrating the usage of our package in routine settings that these FDR methods will become more accessible and gain even more popularity in routine practice.

Simple motivating example

We begin with a simple example to fix ideas. Table 1 shows five unadjusted (raw) p-values for experimental features along with their corresponding Z-values. The third column lists the Benjamini-Hochberg adjusted p-values to be used for FDR control (Benjamini and Hochberg, 1995). Controlling the FDR at level γ amounts to selecting all of the adjusted p-values in column 3 that are below γ. Note here that the adjusted p-values are monotonically increasing, just like the raw p-values, but inflated.

If the goal is to control the FDR at 5%, then only the first feature would be declared interesting and selected. Throughout the paper, we use the term "interesting" to describe features that are selected by a procedure with FDR control. We do not use the term "significant" in order to avoid confusion with those features that would have been selected from by a procedure with strict Type I Error control.

The fourth column presents FDR estimates for each feature. As we show later, there are several ways to invert the FDR control procedures to yield an estimate of the FDR. Our package performs this inversion for most popular methods.

Table 1. Example with 5 features using the Benjamini-Hochberg adjustment and assuming a two-sided normal distribution.

Feature	Raw p-value	Z-value	Adjusted p-value	FDR	Lower Bound FDR
Feature 1	0.005	2.807	0.025	0.025	0.019
Feature 2	0.049	1.969	0.064	0.122	0.126
Feature 3	0.050	1.960	0.064	0.083	0.128
Feature 4	0.051	1.951	0.064	0.064	0.130
Feature 5	0.700	0.385	0.700	0.700	0.481

The FDRs here were obtained by inverting the Benjamini-Hochberg FDR control procedure, and so we will refer to them as the BH FDRs (Benjamini and Hochberg, 1995). In practice we find these estimates to be the most context useful when making scientific decisions about which findings to pursue.

Importantly, these are clearly not identical to the BH adjusted p-values nor are they even monotone. The non-monotonicity results from the group-wise p-value adjustment procedure ("step-up") and the non-smooth estimate of the p-value mixture distribution, which is needed for FDR estimation. The important insight is that the set of features that are selected by the FDR control procedure is not equivalent to the set of feature whose individual FDR is less than the control threshold. For example, if the FDR threshold was $\gamma = 0.07$, then the first four features would be selected by BH to control the group-wise FDR at 7%. However, only the first and fourth features have estimated false discovery rates below 0.07, and thus only these two features would be reported as having a false discovery propensity less than 7%. Note that both approaches come from the same Benjamini-Hochberg machinery, and thus have the same distributional assumptions. The distinction between adjusted p-values and estimated FDRs are critical here.

Because FDRs are only estimates, and because there are a variety of estimation approaches, it helps to have a feature-specific benchmark for each FDR. The fifth column provides such a benchmark; it displays a well-known lower bound on the FDR assuming a Gaussian posterior and a null proportion of 50%. These assumptions are relatively benign for reasons we discuss later and represent a "best-case" scenario. This benchmark shows two things: (1) the adjusted p-values are a poor substitute for the FDRs, and (2) the smoothness of the FDR estimation approach is important.

Methods
FDR methods
p-value based approaches

Let $p_1,\ldots,p_m$ be the individual unadjusted p-values derived for each of m different features or tests. For clarity, the ith p-value is for the ith feature and has not been adjusted for any multiple testing. It is sometimes referred to as the "univariate" p-value. The sorted or ranked p-values are represented by $p_{(1)},\ldots,p_{(m)}$ where $p_{(1)}$ is the smallest, $p_{(m)}$ is the largest and with $p_{(k)}$ is the kth ranked p-value.

Let γ be the false discovery rate threshold for interesting findings. This threshold is context specific, and is either set by the researcher or according to a community standard. This threshold is specified *a priori* when performing FDR control procedures, but it need not be specified for FDR estimation procedures. The Benjamini-Hochberg algorithm for FDR control is to find the largest index, say k, such that

$$p_{(i)} \leq \gamma \frac{i}{m} \text{ for } i \in \{1,2,\ldots,m\} \tag{1}$$

This can be written compactly $k = \max\left[i : p_{(i)} \leq \gamma i/m\right]$. Then all features with $p_{(1)},\ldots,p_{(k)}$ are deemed interesting at the FDR γ threshold and considered "findings". This is called a "step-up" procedure because not all of the rejected features will have unadjusted p-values that meet the above criterion. Only the largest of them must meet that criterion. Because this is a "step-up" procedure, the adjusted p-values will depend on the raw p-values from other features. The Benjamini-Hochberg adjusted p-value for the ith feature is notated in this paper by $\tilde{p}_i$ and defined in Equation (2), where := means "is defined as".

$$\tilde{p}_{(i)} := \min_{j \geq i}\left(\frac{p_{(j)}m}{j}\right) \leq \gamma \tag{2}$$

These adjusted p-values are monotone increasing in raw p-value ranking, so one can directly compare $\tilde{p}_i$ to γ to see if a particular feature would be rejected as null for the FDR threshold γ. Importantly, the feature specific FDR estimates need not be monotone. To see this, re-arrange Equation (1) as follows in Equation (3).

$$FDR_i := \frac{p_i m}{\text{rank}(p_i)} \cdot \hat{\pi}_0 \tag{3}$$

The derivation of FDR is described in the following section. This shows that the BH procedure is, in effect, estimating the feature specific FDR as FDR_i. See also Efron LSI for motivation for this definition (Efron, 2013). Because estimation of the feature specific FDR does not include group-wise control of the FDR, the "step-up" monotonicity condition does not apply. Thus, feature specific FDR estimates such as FDR_i are not always monotone in raw p-value ranking.

A consequence of this dichotomy is that an individual feature may be rejected at FDR γ level by the BH algorithm even though its feature specific FDR estimate is actually greater than γ. This is largely a consequence of the smoothness of the FDR estimates and the fact that they can have substantial variability. Note that there are several methods for estimating the FDR, and some methods may be better suited to certain contexts. Our package offers several methods for FDR estimation, as described in later sections of this paper.

To illustrate we simulated real data from 100 hypothesis tests and captured the 100 raw p-values. For context, 80 of these p-values were generated from a uniform distribution (and hence under the null) while the other 20 were generated from a skewed distribution representing the alternative. Results are computed using our `p.fdr` function, which we detail later. The raw p-values are displayed in Figure 1 as black points; Figure 2 shows only the 20 features with the smallest ranked
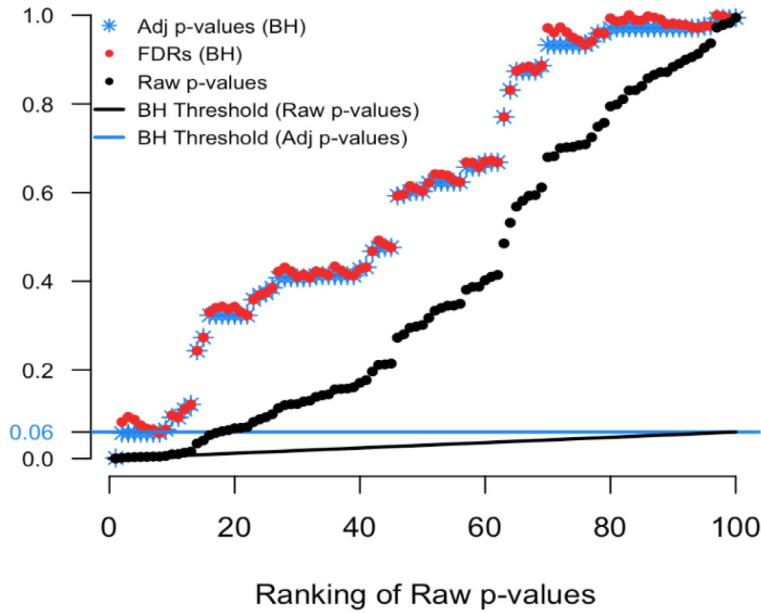

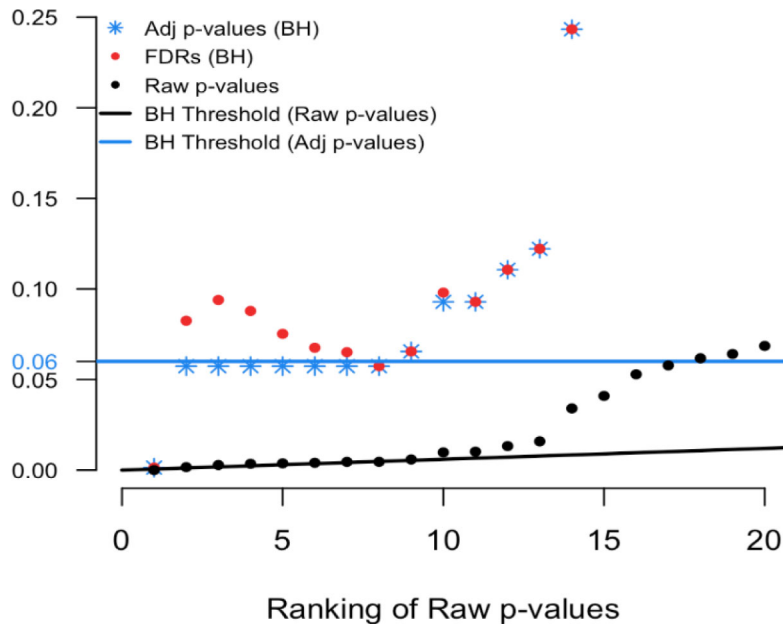**Figure 1. Simulated example comparing raw p-values and the threshold of interest.**

Figure 2. Magnified section of Figure 1.

raw p-values. The black sloped line is the BH rejection line from Equation (1). Also included in the plot are the BH adjusted p-values (blue stars), the BH FDR threshold for interesting findings (blue horizontal line), and the BH FDR estimates (red points).

In Figure 2 we see that exactly 8 of the adjusted p-values fall below our threshold of interest (blue line, set here to 0.06). Therefore, the BH FDR procedure that controls the group-wise FDR identifies the 8 smallest p-values as interesting findings. However, notice the non-monotonicity of the individual FDRs. Only the first and last of the 8 lowest FDRs are less than 0.06.

From these results it should be clear that the feature-specific FDRs and the BH adjusted p-values have different purposes and interpretations. To emphasize, when a feature is identified as 'interesting' by an FDR control procedure, it does not always follow that the feature's individual propensity to be a false discovery is less than the desired threshold. Both quantities must be computed, as the tasks are not always exchangeable.

Z-value based approaches

For FDR estimation, it is often helpful to transform the p-values $p_1,\ldots,p_m$ to into Z-values $z_1,\ldots,z_m$ using the standard normal quantiles. For example, $z_i = \Phi^{-1}(1 - p_i)$ for one-sided p-values or $z_i = \Phi^{-1}(1 - p_i/2)$ for two-sided p-values. Efron explains the rationale as an attempt to leverage the distributional properties of a set of Gaussian random variables (Efron, 2013). Note that these Z-values are not intended to be the original test statistics. We will adopt Dr. Bradley Efron's formulation as described here (Efron, 2013).

We begin with the classic two-group model, which assumes each of the m features is either null (distribution known) or alternative (distribution unspecified), but that this status is unknown. As a group the combined data can be used to provide an estimate of the mixture distribution, where the mixing proportion (π_0) is also unknown. Let $f_0(z)$ be the probability density function of the z-values when they come from the true null distribution and $f_1(z)$ be the probability density function of the z-values when they come from the alternative distribution. Then $F_0(\cdot)$ and $F_1(\cdot)$ denote the probability of rejection for any subset $\mathcal{Z}$ of the real line such that,

$$F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)dz \text{ and } F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z)dz \tag{4}$$

With mixing or null proportion π_0, the proportion of non-null features is simply $\pi_1 = 1 - \pi_0$. The mixing distribution function is

$$F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z}) \tag{5}$$

When working with Z-values, it is reasonable to use a gaussian distribution for the theoretical null probability density function, so that $f_0(z) \sim N(0,1)$ (Efron, 2013). When estimating the FDR, is it also common to assume that $\pi_0 = 1$ because doing so results in a conservative estimate of the FDR. Then, an application of Bayes famous theorem yields:

$$FDR(\mathcal{Z}) := Pr\{null|z \in \mathcal{Z}\} = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})} \tag{6}$$

Substituting the natural empirical estimate of the mixture distribution $F(\mathcal{Z})$ results in empirical Bayes estimates the global FDR Equation (6) (Benjamini and Hochberg, 1995) (Efron, 2013). For example, the obvious empirical estimate of the mixing distribution function is the step function $\hat{F}(\mathcal{Z}_i) = \text{rank}(p_i)/m$. Notice that the right hand side of Equation (1) then looks like $\gamma \cdot \hat{F}(\mathcal{Z}_i)$ or γ times the step function. In some settings smoothing $\hat{F}(\mathcal{Z}_i)$ can be beneficial. Very often it is assumed $\pi_0 = 1$ and $F_0(\mathcal{Z}) = 1 - \Phi(\mathcal{Z})$ for one-sided tests. An advantage of estimating the FDR from the right hand side of Equation (6) is that one only needs to accurately estimate the mixture distribution function to get good estimates of the FDR and this does not require the independence of the z-values.

Figure 3 and Figure 4 show the application of this framework using the same simulated data as in the last example (100 tests, 80 truly null). In the z-space, the null distribution is now the standard normal and the alternative distribution was set to $N(2,1)$ (of course this is unknown, in practice). Figure 3 shows these densities overlaid on a histogram of the raw data. The blue curve indicates the null density, the red curve indicates the alternative density, and the black curve is the mixture density with $\pi_0 = 0.8$. Clearly, the blue curve does not fit the histogram well, with a much shorter right tail than the histogram shows. So, assuming all 100 tests come from the null distribution does come with a penalty.
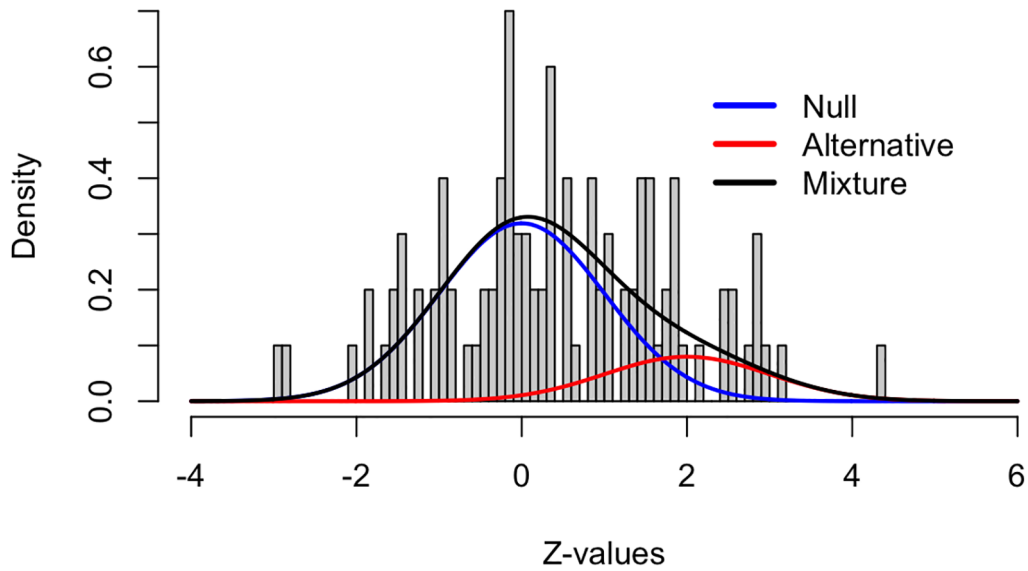
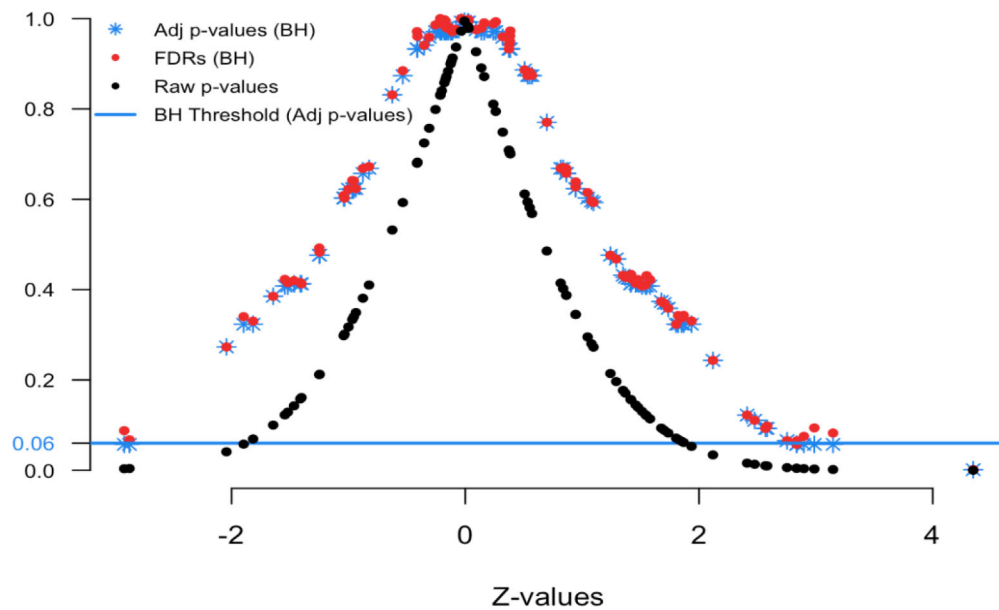**Figure 3. Density histogram of the simulated example.**

Figure 4. FDR simulated Z-values plot.

Figure 4 displays the relationship between the Z-values and various FDR quantities. The black dots show the raw p-values (y-axis) versus their Z-value (x-axis); the red dots show the estimated FDRs (y-axis) versus their Z-value (x-axis); and the blue stars show the BH adjusted p-values (y-axis) versus their Z-value (x-axis). This is the comparable plot to Figure 1, where the x-axis has been changed from p-value ranking to z-scale. The usefulness of this plot is that is shows what the desired FDR quantity is for a given Z-value. This provides context for our FDRs and adjusted p-values.

Here we see that Z-values greater than 2.85 and less than -2.5 have adjusted p-values less than 0.06 (blue threshold line, horizontal). This means in order to control the group-wise FDR, one would identify features with these Z-values as "interesting". Notice that the Z-value above 4 has a FDR less than 0.06. Also the Z-value of 2.9 has a FDR less than 0.06. In practice, we find that the display in Figure 1 is more intuitive for non-statisticians, but that Figure 4 provides some essential insight into the stability and smoothness of the estimation procedure.

Lower bound on the FDR

The previous section introduced an empirical Bayes estimator for the FDR, which has become one of the most popular estimates. However, there are many different approaches for estimating the FDR. We have found it helpful in practice to be able to benchmark the magnitude of the FDR under known conditions in order to provide a contrast for estimators that rely heavily on distributional assumptions. This lower bound can help to contextualize findings and illuminate differences masked by empirical assumptions.

Our preferred benchmark is a well-known lower bound on the posterior probability of the null (hypothesis) under a gaussian model. This lower bound depends only on the data for the feature or test of interest and it does not borrow strength across features (for better or worse). Hence, it can also be used when only a single test is performed, i.e. when only a single p-value is available. In our experience, the gaussian assumption tends to have minimal influence because sampling distributions tend to be symmetric.

The lower bound arises as follows. Let the joint density of data from a single feature be $g(X_1,…,X_n|\theta)$ where θ is a parameter of interest. The likelihood function is $L_n(\theta) \propto g(x_1,…,x_n|\theta)$ and denote the maximum likelihood estimator as $\hat{\theta}_n$. Recall that the null hypothesis is $H_0 : \theta = \theta_0$. Let $\pi_0 = P(null)$ be the prior probability of the null and let z be the observed test statistic of the null hypothesis. Then, a lower bound on the posterior probability, $P(null|x_1,…,x_n)$, which is effectively the FDR, is given by

$$P(null|x_1,…x_n) \geq \left(1 + \frac{L(\hat{\theta}_n)\pi_1}{L(\theta_0)\pi_0}\right)^{-1} \approx \left(1 + \exp\left(z^2/2\right)\frac{\pi_1}{\pi_0}\right)^{-1} \tag{7}$$

The first inequality holds because $\int g(x_1,…x_n|\theta_1)h(\theta_1)d\theta_1 \, leq g(x_1,…x_n|\hat{\theta}_n)$ for all $\theta_1 \sim h(\theta_1)$. Note that $\int h(\theta_1)d\theta_1 = \pi_1$ by definition. The second approximation comes from the general asymptotic behavior of a classical likelihood ratio test, where $-2\log\frac{L(\theta_0)}{L(\hat{\theta}_n)} \sim \chi_1^2 = [N(0,1)]^2$ for one-dimensional parameters. This lower bound is similar to that derived and explored by Berger (1985). Our function uses default odds, $\pi_1/\pi_0 = 1$, reasonable in many circumstances, which easily can be changed. As the z-statistic approaches zero, the lower bound approaches 1/2, as would be expected.

For illustration, consider feature 4 in Table 1. Feature 4 has an observed p-value of 0.051, but has a univariate gaussian lower bound on the FDR of $0.13 = (1 + \exp(1.951^2/2))^{-1}$. In this case the BH estimated FDR is 0.064, substantially below the lower bound. This discrepancy in estimates is due to differing underlying assumptions. In contrast, feature 2 has a p-value of 0.049 and FDR of 0.122, very close to its lower bound. Although feature 4 has nearly the same p-value as feature 2, its BH FDR is nearly half that of feature 2. The univariate gaussian lower bound is helpful for identifying when FDR estimates may be optimistic, as in the case above. Similarly, we see that the adjusted p-values can be much less than the lower bound, which is another reason why they should not be mistaken for FDR estimates.

Adjustment methods

The computation of adjusted p-values and FDRs for each method follows a similar intuitive approach. First, estimates of the FDR for each feature are obtained using the preferred method, e.g. Benjamini-Hochberg or Benjamini-Yekutieli. Step-up or step-down adjustments are not applied at this stage. Next, adjusted p-values are obtained from the estimated FDRs by applying the step-up or step-down adjustment that is associated with the method. The step adjustment is necessary for error control but not for FDR estimation. For methods that do not have a step-up or step-down component, e.g. Bonferroni, the adjusted p-values and FDRs will be the same. The distinction between the estimated FDRs and the adjusted p-values is an important one that is routinely confused in practice.

Note that all estimates of adjusted p-values and FDRs are forced to be 1 or less. Also, when ranks are used in our package the `ties.method = "random"`. This means for example that if the 4 smallest p-values in a vector tie in value then they will be assigned ranks 1,2,3,4 randomly. The user can change the ties method in the input to the function.

Below we illustrate this with the remaining five methods (BH is discussed above).

Benjamini-Yekutieli

Benjamini-Yekutieli (BY) is a step-up method for controlling the false discovery rate under arbitrary dependence (Benjamini and Yekutieli, 2001). For a pre-specified dependence structure, there exits an adjustment function called $c(m)$ that is used to modify the Benjamini-Hochberg estimate of the FDR. For example, in the case of flexible positive

dependence, the function $c(m) = \sum_{j=1}^{m} \frac{1}{j}$ is used. Then, the threshold criteria is to find the largest index i such that Equation (8) holds, which is a scaled version of the BH criterion given in Equation (1).

$$p_{(i)} \leq \gamma \frac{i}{m} \frac{1}{c(m)} \tag{8}$$

This can be written compactly $k = \max \left[i : p_{(i)} \leq \gamma \cdot i/(m \cdot c(m)) \right]$ or for non-ordered vectors of p-values $k = \max$ $\left[\text{rank}(p_i) : p_i \leq \gamma \cdot \text{rank}(p_i)/(m \cdot c(m)) \right]$. Then all features with $p_{(1)}, \ldots, p_{(k)}$ are deemed interesting at the FDR γ threshold and considered "findings". Recall that Benjamini-Hochberg procedure uses the step function ($F(p_{(i)}) = i/m$) as its implicit empirical estimate of the mixing distribution function (CDF) Check this notation. The Benjamini-Yekutieli procedure amounts to simply using a modified estimate for the CDF, namely ($F(p_{(i)}) = i/(m \cdot c(m))$).

Mathematically, the adjusted p-values and estimated FDRs are

$$\tilde{p}_{(i)}^{BY} := \min_{j \geq i} \left(p_{(j)} \frac{m \cdot c(m)}{j} \right) \leq \gamma \tag{9}$$

$$FDR_i^{BY} := p_i \frac{m \cdot c(m)}{\text{rank}(p_i)} \cdot \hat{\pi}_0 \tag{10}$$

Comparing this form to the general formula for the FDR in Equation (6), we see that the BY correction amounts to changing the estimate of the mixture distribution $F(\mathcal{Z})$ from $[\text{rank}(p_i)/m]$ to $[\text{rank}(p_i)/(m \cdot c(m))]$ to account for dependence. Note that we have avoided using the ordered notation for False discovery rate estimates, say $FDR_{(i)}$, because although those estimates are dependent on ordered p-values the FDR estimates themselves do not have to be monotonic.

Here we see the BY FDRs, or red points, jump above and below the 0.06 threshold in ranks 1 to 6. Then in ranks 7 and greater the red dots remain above the threshold and quickly are adjusted to the value of 1. The positive dependence correction causes these BY FDRs to be closer to 1, or more conservative.

Bonferroni

The Bonferroni correction controls the family wise error rate (FWER) (Bonferroni, 1936). We include it in our function because of its popularity in multiple adjustments even though it is not directly related to FDR. For this method we would reject the null hypothesis for each $p_i \leq \frac{\gamma}{m}$ in order to control the FWER at $\leq \gamma$ level. In our functions the adjusted p-values and adjusted FDRs will always be identical for this method.

$$\tilde{p}_i^{Bon} := p_i m \leq \gamma \tag{11}$$

$$FDR_i^{Bon} := p_i m \cdot \hat{\pi}_0 \tag{12}$$

From this form we see that the Bonferroni correction amounts to changing the estimate of the mixture distribution $F(\mathcal{Z})$ to $[1/m]$.

Sidak

The Sidak or Dunn-Sidak correction controls the family wise error rate (FWER) (Šidák, 1967). This correction method is exact for tests that are independent, it is conservative for tests that are positively dependent, and it is liberal for tests that are negatively dependent. For this method is slightly less strict then the traditional Bonferroni method. For each $p_i \leq \gamma_{Sid} = 1 - (1-\gamma)^{\frac{1}{m}}$ reject the null hypothesis in order to control the FWER at $\leq \gamma$ level. In our functions the adjusted p-values and adjusted FDRs will always be identical for this method.

$$\tilde{p}_i^{Sid} := 1 - (1 - p_i)^m \leq \gamma \tag{13}$$

$$FDR_i^{Sid} := 1 - (1 - p_i)^m \cdot \hat{\pi}_0 \tag{14}$$

From this form we see that the Sidak correction amounts to changing the estimate of the mixture distribution $F(\mathcal{Z})$ to $[p_i/(1 - (1-p_i)^m)]$ assuming $F_0(Z) = p_i$.

Holm

The Holm method, also known as the Holm-Bonferroni method, controls the FWER and is less conservative and therefore uniformly more powerful than the Bonferroni correction (Holm, 1979). For this method we use the step-down procedure which would reject the null for those rankings $1,\ldots,(k-1)$ such that k is the smallest ranking where:

$$p_{(k)} \leq \frac{\gamma}{m+1-k} \tag{15}$$

From the above equation we see that it relies on the ranking or j that means our function's outputted adjusted p-value and FDR can be different.

$$\tilde{p}_{(i)}^{Holm} := \max_{j \leq i} \left(p_{(j)}(m+1-j) \right) \leq \gamma \tag{16}$$

$$FDR_i^{Holm} := p_i(m+1-\text{rank}(p_i)) \cdot \hat{\pi}_0 \tag{17}$$

From this form we see that the Holm correction amounts to changing the estimate of the mixture distribution $F(\mathcal{Z})$ to $[1/(m+1-\text{rank}(p_i))]$.

Hochberg

The Hochberg method uses the same equation as the Holm method, Equation (15) (Hochberg, 1988). However for this method we use the step-up procedure. This means we would reject the null for those rankings $1,\ldots,j$ such that j is the largest ranking where:

$$p_{(j)} \leq \frac{\gamma}{m+1-j} \tag{18}$$

This change from the step-down to the step-up procedure results in the Hochberg correction being more powerful than the Holm method.

$$\tilde{p}_{(i)}^{Hoch} := \min_{j \geq i} \left(p_{(j)}(m+1-j) \right) \leq \gamma \tag{19}$$

$$FDR_i^{Hoch} := p_i(m+1-\text{rank}(p_i)) \cdot \hat{\pi}_0 \tag{20}$$

From this form we see that the Hochberg correction is the same as the Holm and amounts to changing the estimate of the mixture distribution $F(\mathcal{Z})$ to $[1/(m+1-\text{rank}(p_i))]$.

Null proportion (π_0) estimation

The proportion of truly null features (π_0), also known as the mixing proportion, is an important component of the FDR estimate that can be a strong driver of the estimate. While generally not identifiable, reasonable estimates of π_0 can be obtained under certain conditions. Many of the popular FDR estimation routines take a conservative approach by setting $\pi_0 = 1$, which results in a larger, i.e. conservative, FDR estimates.

The default in p.fdr is to assume that $\pi_0 = 1$. However, users are able to set the null proportion to a particular value or specify an estimation routine to estimate π_0 from the data. Many methods have been proposed for estimating the mixing proportion π_0 in a two-component mixture. p.fdr includes several of these methods such as Storey, Meinshausen, Jiang, Nettleton, and Pounds (Storey and Tibshirani, 2003; Meinshausen, Rice, et al., 2006; Jiang and Doerge, 2008; Nettleton, Hwang, Caldo, and Wise, 2006; Pounds and Morris, 2003). In next section, we propose a new approach that we call "Last Histogram Height". This new approach is simple, appears to have excellent performance over a wide range of scenarios, and less computationally intensive that Storey's approach, which is quite popular. An evaluation and comparison to existing approaches is described in the subsequent subsection.

Last Histogram Height

Under the null, a test statistic for a feature, say a Z-value, is standard normal. As such, the corresponding p-value has a uniform distribution over the unit interval. Therefore, if all the features were null, we would expect an empirical histogram

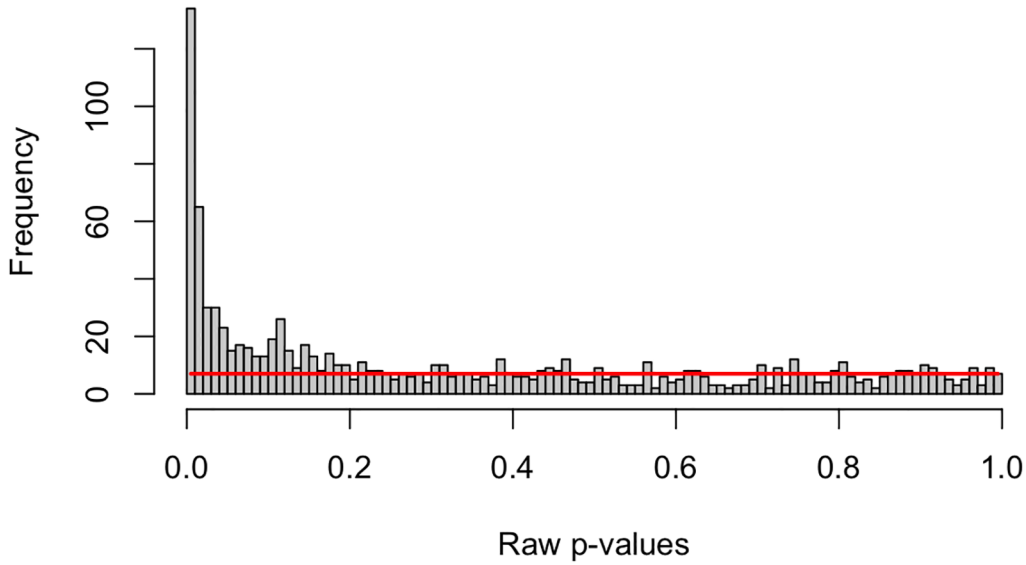## Last Histogram Height Estimation Method

Figure 5. Simulated histogram of p-values with horizontal line at the last bin height.

of the observed p-values to be approximately flat. Moreover, we see that the distribution of non-null p-values tends to be shifted toward zero.

The "Last Histogram Height" method uses the bin height of p-values near 1 to estimate the true proportion of null features. We rely on the assumption that larger p-values are more likely to be come from null features. Let bin heights be $H_1, H_2, \ldots, H_B$, where B is the total number of bins. When $B = m$ (m is the number of features) and all features are null, we would expect $H_i \approx 1$ for all $i = 1, \ldots, B$. The caveat is estimating bin height is sensitive to the choice of bin width. However, we have found that Scott's normal reference rule tends to work very well for this method (Scott, 1979).

When $\pi_0 < 1$, the empirical distribution of the p-values (as shown by the histogram) will not be uniform over the unit interval. Departure from the uniform becomes easier to detect as π_0 moves further from 1 because the histogram shape quickly deviates from a uniform appearance. An example is presented in Figure 5, which shows a histogram of raw p-values from our simulated example of $m = 1000$ features. The red horizontal line is drawn at the height of the last bin, H_B. In this approach H_B is our "null height" and $H_B \cdot B$ is an estimate of the total number of null features. We then divide that by the number of total features (m) to estimate the null proportion (see Equation (21)):

$$\hat{\pi}_0 = \frac{H_B B}{m} \qquad (21)$$

The approach works because we would expect $\pi_0 * m/B$ null p-values to be in each bin. This simple method performed well over many different simulation settings, as we described in the next section. It is also relatively free of constraining assumptions on the alternative distribution. We note that this approach can also be viewed as a form of central matching, as discussed by Efron (Efron, 2013), with center mass $\frac{1}{B}$ and a very small bin width. The "Last Histogram Height" algorithm is as follows:

Algorithm 1: Last Histogram Height Method

Result: Null proportion estimate

1. Plot a histogram of the raw p-values, $p_1, p_2, \ldots p_m$, with B number of bins, where $B < m$

   - The most consistent bin method is `scott`, according to our simulations

2. Store the histogram bin heights H_b for each bin $b = 1, 2, \ldots, B$

3. Call the height of last bin H_B the "null height"

4. Set the estimate of π_0 to be

$$\hat{\pi}_0 = \frac{H_B B}{m}$$

Storey

Storey et al. (2003) propose an iterative procedure for estimating π_0. This procedure is popular and tends to have good performance characteristics over a wide range of scenarios. Storey's method relies on the fact that null p-values are uniformly distributed. As such, the bin height of p-values greater than 1/2 should give a conservative estimate of the null proportion. But there is nothing magical about 1/2, so Storey uses a tuning parameter. Let λ identify "large" p-values, e.g., $\#p_i > \lambda$ where $i = 1,\ldots,m$, such that the estimate of the null proportion $\hat{\pi}_0(\lambda)$, can be tuned by λ to yield a desirable bias-variance trade-off. Storey smoothes $\hat{\pi}_0(\lambda)$ before tuning, which provides some numerical stability. Note that for the "Last Histogram Height" approach, the bin height closest to one is used to estimate the null proportion, which is conceptually similar to using $\lim_{\lambda} \hat{\pi}_0(\lambda)$ as Storey does. Storey's algorithm for estimating π_0 is as follows:

Algorithm 2: Storey's Method

Result: Null proportion estimate

1. Let $p_{(1)}, p_{(2)}, \ldots p_{(m)}$ be the ordered p-values. This also denotes the ordering of the features in terms of their evidence against the null hypothesis.

2. For a range of λ, say $\lambda = 0, 0.05, 0.10, \ldots, 0.95$, and $i = 1, \ldots, m$, calculate

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

3. Let $\hat{f}$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ

4. Set the estimate of π_0 to be when $\lambda = 1$:
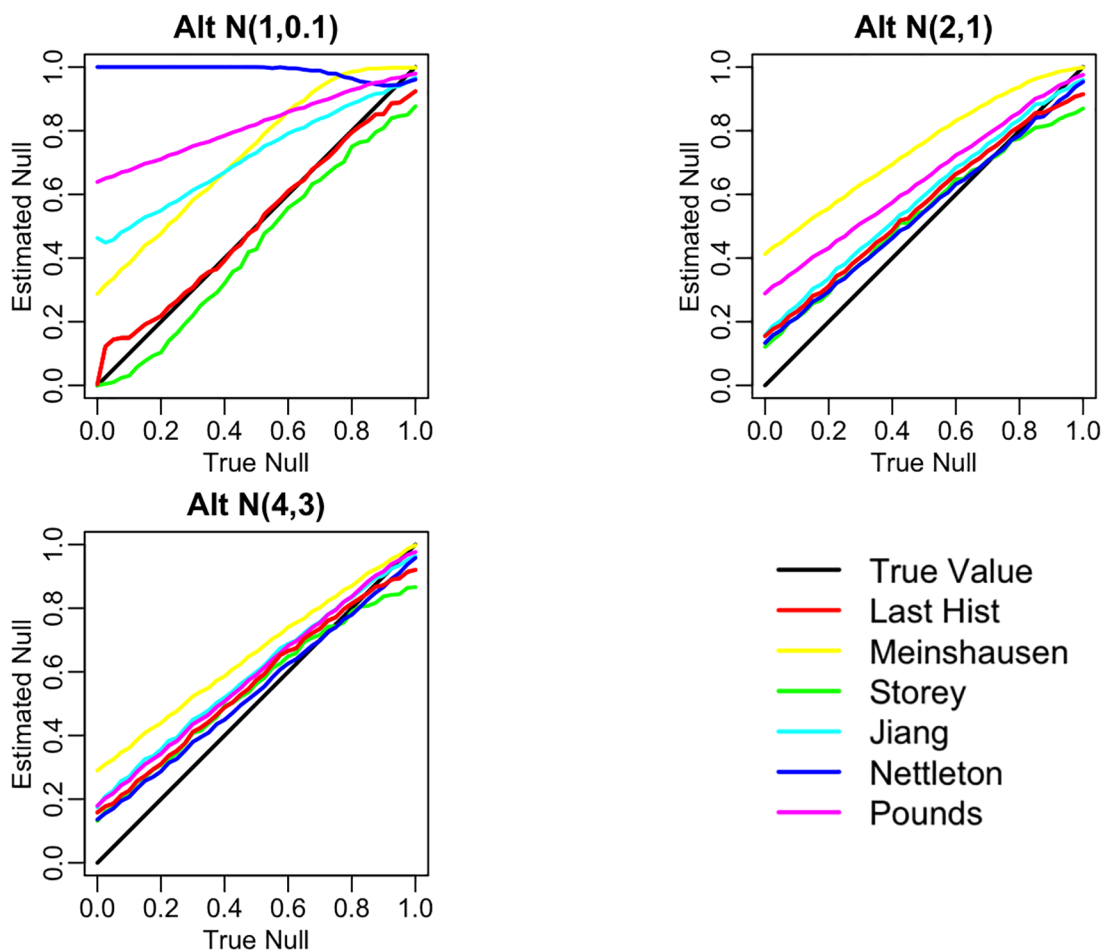
$$\hat{\pi}_0 = \hat{f}(1)$$

Figure 6. Comparison of null proportion estimation methods performance.

Comparison

Below in Figure 6 are three plots showing the range of behavior of the six methods for estimating the null proportion that are included in our R package. These plots show the arrogate behavior of each method for estimating π_0 over 1000 simulations where the methods are used on a set of 100 features. A standard normal distribution was used for null features and three different alternative distributions were examined for alternative features (three different plots). The x-axis represents the true π_0 used to generate data and ranges from 0 to 1. The y-axis represents the average estimate π_0 (over the 1,000) simulations for each of the six methods. Figure 7 shows the corresponding mean squared error (MSE) for these simulations.

"Last Histogram Height" and Storey's method preformed the best across these scenarios (and others not shown here). They routinely produce the closest estimates of the true null proportion and have some of the lowest MSEs. Although we only display three different mixture distributions for a set of 100 features here, we tested 12 different mixture distributions over three different features set sizes to confirm our results. Our recommendation is to use the default of setting $\pi_0 = 1$ when the majority of features are expected to be null or nearly null. But in cases where the null proportion is likely to be different from one (say less than 0.95 or 0.9), then the "Last Histogram Height" algorithm tends to perform the best.

Implementation

`FDRestimation` is a user-friendly R package that directly computes and displays false discovery rates from p-values or z-scores under a variety of assumptions. The following sections will explain the primary functions in this package and illustrate how to implement them.
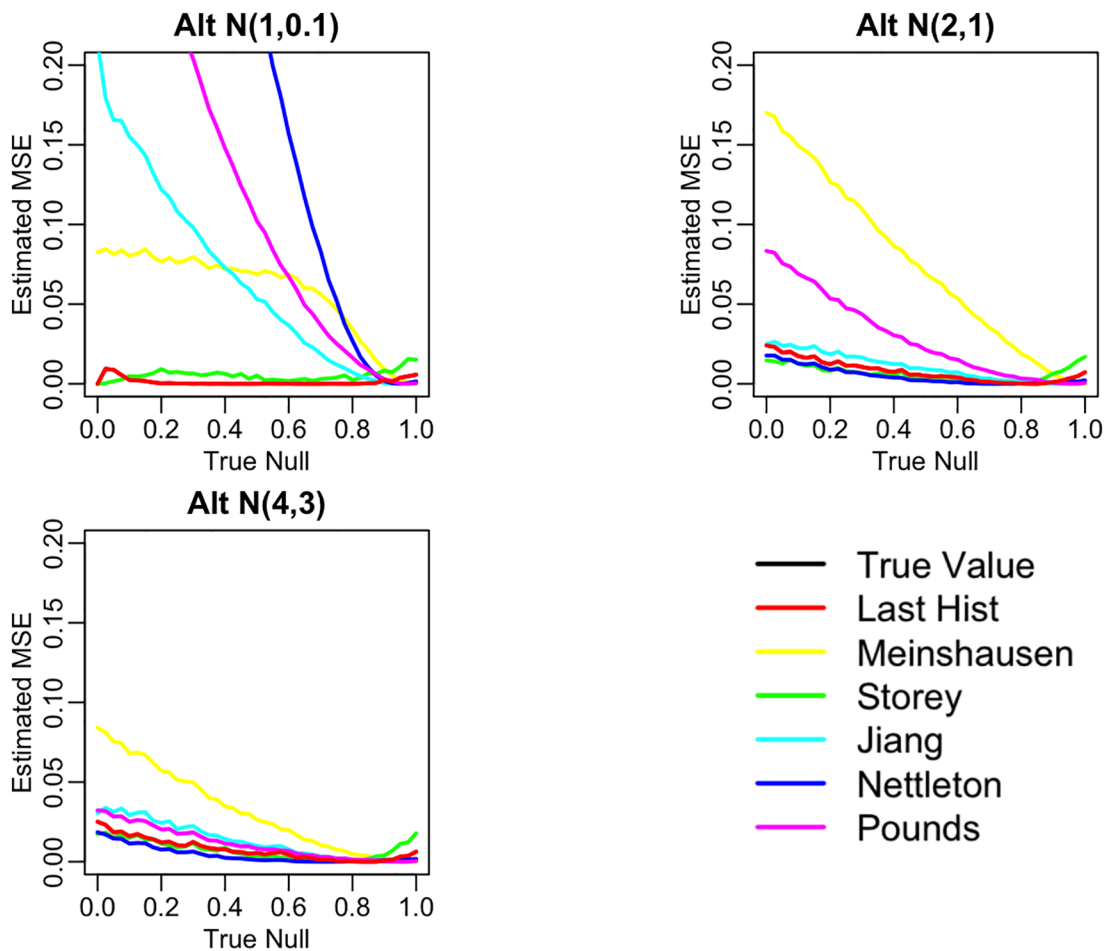

Figure 7. Comparison of null proportion estimation methods MSE.

`p.fdr` Function

This `p.fdr` function is used to compute FDRs and multiple-comparison adjusted p-values from a vector of raw p-values. The stats package function `stats::p.adjust` is similar in that it will produce multiple-comparison adjusted p-values. However, `stats::p.adjust` returns the BH adjusted p-value labeled as the FDR estimate. Strictly speaking this is inaccurate, because the BH FDR estimate should not have the forced monotonicity that its adjusted p-values must have. In addition, when estimating the FDR, our `FDRestimation::p.fdr` function allows adjustments of key assumptions that are not adjustable in the `stats::p.adjust` implementation (they are set to the simplest, most popular options).

This `FDRestimation::p.fdr` function allows for the following adjustment methods: Benjamini-Hochberg, Benjamini-Yekutieli (with both positive and negative correlation), Bonferroni, Holm, Hochberg, and Sidak (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Bonferroni, 1936; Holm, 1979; Hochberg, 1988; Šidák, 1967). It also allows the user to specify the threshold for important findings, the assumed pi0 value, the desired pi0 estimation method, whether to sort the results, and whether to remove NAs in the imputed raw p-value vector count (`stats::p.adjust` actually counts NAs as viable features in its Bonferroni adjustment). Table 2 shows all of the inputs for this function and their descriptions.

The underlying methods for estimating the null proportion can be set by using the "estim.method" and "set.pi0" arguments. The default value of "set.pi0" is 1, meaning it assumes that all features are null features. Accordingly, this approach will yield conservative estimates of the FDR. Alternatively, and less conservatively, one can attempt to estimate

Table 2. Inputs to the `p.fdr` function taken directly from the R documentation (Murray, 2020).

Arguments	Description
p values	A numeric vector of raw p-values.
z values	A numeric vector of Z-values to be used in pi0 estimation or a string with options "two.sided", "greater" or "less". Defaults to "two.sided".
threshold	A numeric value in the interval [0,1] used in a multiple comparisons hypothesis tests to determine significance from the null. Defaults to 0.05.
adjust.method	A string used to identify the adjustment method. Defaults to *BH.* Options are *BH*, *BY*, *Bon*, *Holm*, *Hoch*, and *Sidak*.
BY.corr	A string of either "positive" or "negative" to determine which correlation is used in the BY method. Defaults to *positive.*
just.fdr	A Boolean TRUE or FALSE value which output only the FDR vector instead of the list output. Defaults to FALSE.
default.odds	A numeric value determining the ratio of pi1/pi0 used in the computation of single lower bound FDR. Defaults to 1.
estim.method	A string used to determine which method is used to estimate the null proportion or pi0 value. Defaults to set.pi0.
set.pi0	A numeric value to specify a known or assumed pi0 value in the interval [0,1]. Defaults to 1. Which means the assumption is that all inputted raw p-values come from the null distribution.
hist.breaks	A numeric or string variable representing how many breaks are used in the pi0 estimation histogram methods. Defaults to "scott".
ties.method	A string a character string specifying how ties are treated. Options are "first", "last", "average", "min", "max", or "random". Defaults to "random".
sort.results	A Boolean TRUE or FALSE value which sorts the output in either increasing or non-increasing order dependent on the FDR vector. Defaults to FALSE.
na.rm	A Boolean TRUE or FALSE value indicating whether NA's should be removed from the inputted raw p-value vector before further computation. Defaults to TRUE.

the null proportion from the data. To do this, we recommend using "Last Histogram Height", as it was the simplest routine and one of the most accurate in our simulations (Murray, 2020).

Here we see an example of how to use this `FDRestimation::p.fdr` function in R. We simulate 100 features with a true null proportion of 80%.

```
set.seed(88888)

# Simulate Data
sim.data.p= c(runif(80),runif(20, min=0, max=0.01))

# Full set
p.fdr(p=sim.data.p, threshold=0.05, adjust.method="BH")

# First 5 p-values for Figure 8
p.fdr(p=sim.data.p[ 1:5], threshold=0.05, adjust.method="BH")
```

The function will return a list object of the `p.fdr` class. In Figure 8 we see this list object from the first five p-values for with the following components (Murray, 2020).

- **fdrs** A numeric vector of method adjusted FDRs.

- **Results Matrix** A numeric matrix of method adjusted FDRs, method adjusted p-values, and raw p-values.

- **Reject Vector** A vector containing Reject.H0 and/or FTR.H0 based off of the threshold value and hypothesis test on the adjusted p-values.

```
$fdrs
[1] 1.0000000 0.9567112 0.9680525 0.8344546 0.6980411

$`Results Matrix`
    BH FDRs Adjusted p-values Raw p-values
1 1.0000000          0.6980411    0.2393817
2 0.9567112          0.8344546    0.5740267
3 0.9680525          0.8344546    0.7744420
4 0.8344546          0.8344546    0.8344546
5 0.6980411          0.6980411    0.2792164

$`Reject Vector`
[1] "FTR.H0" "FTR.H0" "FTR.H0" "FTR.H0" "FTR.H0"

$pi0
[1] 1

$threshold
[1] 0.05

$`Adjustment Method`
[1] "BH"

$Call
p.fdr(pvalues = sim.data.p[1:5], threshold = 0.05, adjust.method = "BH")

attr(,"class")
[1] "p.fdr"
```

Figure 8. Example of output produced with p.fdr code.

- **pi0** A numeric value for the pi0 value used in the computations.

- **threshold** A numeric value for the threshold value used in the hypothesis tests.

- **Adjustment Method** The string with the method name used in computation(needed for the plot.fdr function).

`get.pi0` Function

The `get.pi0` function is used to estimate the null proportion from the raw p-values. The user can choose one of six different methods included in our function: Last Histogram Height, Storey, Meinshausen, Jiang, Nettleton, and Pounds (Storey and Tibshirani, 2003; Meinshausen et al., 2006; Jiang and Doerge, 2008; Nettleton et al., 2006; Pounds and Morris, 2003). The user may also change the methods of determining the number of histogram breaks, which is an essential component for many of the methods implemented here. Table 3 shows function arguments and their descriptions.

Here we see an example of how to use this `get.pi0` function in R. We used the simulated data from above `sim.data.p` where the true null proportion was set to 80%. In the first example, for the purposes of the estimation routine, π_0 was set to a single value with the `set.pi0=0.8` argument (1 is the default). Alternatively, we can use one of the six estimation methods in `get.pi0` instead of specifying π_0 *a priori.* Below is an example where we set the estimation method to `"last.hist"` (i.e., "Last Histogram Height"). In that case, the get.pi routine returned an estimate of null proportion of 0.95.

```
set.seed(88888)

# Set null proportion with known value
get.pi0(sim.data.p, estim.method="set.pi0", set.pi0=0.8)

[1] 0.8

# Get null proportion with last histogram height method
get.pi0(sim.data.p, estim.method="last.hist")

[1] 0.85
```

Table 3. Inputs for the `get.pi0` function taken directly from the R documentation (Murray, 2020).

Arguments	Description
p values	A numeric vector of raw p-values.
set.pi0	A numeric value to specify a known or assumed pi0 value in the interval [0,1]. Defaults to 1. Which means the assumption is that all inputted raw p-values come from the null distribution.
estim.method	A string used to determine which method is used to estimate the null proportion or pi0 value. Defaults to set.pi0.
z values	A numeric vector of Z-values to be used in pi0 estimation or a string with options "two.sided", "greater" or "less". Defaults to "two.sided".
threshold	A numeric value in the interval [0,1] used in a multiple comparisons hypothesis tests to determine significance from the null. Defaults to 0.05.
default.odds	A numeric value determining the ratio of pi1/pi0 used in the computation of single lower bound FDR. Defaults to 1.
hist.breaks	A numeric or string variable representing how many breaks are used in the pi0 estimation histogram methods. Defaults to "scott".
na.rm	A Boolean TRUE or FALSE value indicating whether NA's should be removed from the inputted raw p-value vector before further computation. Defaults to TRUE.

`plot.p.fdr` Function

This `plot.p.fdr` function is used to plot the results of `p.fdr`. By default, the adjusted FDRs, adjusted p-values and raw p-values are plotted along with two threshold lines to help contextualize the points. Any combination of p-values and thresholds can be removed from the plot. The user can set the axis limits, the location of the legend, the title of the plot and the plotting symbols and colors. Table 4 shows all the function arguments and their descriptions.

Here we see an example of the `plot.p.fdr` function in R. We used our simulated data `sim.data.p`, where the a true null proportion was 80%, for illustration. Figure 9 show the default plot, and Figure 10 zooms in on an interesting subset of findings.

Table 4. Inputs for the `plot.p.fdr` function taken directly from the R documentation (Murray, 2020).

Arguments	Description
p.fdr.object	A p.fdr object that contains the list of output.
raw.pvalues	A Boolean TRUE or FALSE value to indicate whether or not to plot the raw p-value points. Defaults to TRUE.
adj.pvalues	A Boolean TRUE or FALSE value to indicate whether or not to plot the adjusted p-value points. Defaults to TRUE.
sig.line	A Boolean TRUE or FALSE value to indicate whether or not to plot the raw p-value significance line. Defaults to TRUE.
adj.sig.line	A Boolean TRUE or FALSE value to indicate whether or not to plot the adjusted significance threshold. Defaults to TRUE.
threshold	A numeric value in the interval [0,1] used in a multiple comparisons hypothesis tests to determine significance from the null. Defaults to 0.05.
x.axis	A string variable to indicate what to plot on the x-axis. Can either be "Rank" or "Zvalues". Defaults to "Rank".
xlim	A numeric interval for x-axis limits.
ylim	A numeric interval for y-axis limits. Defaults to c(0,1).
zvalues	A numeric vector of Z-values to be used in pi0 estimation or a string with options "two.sided", "greater" or "less". Defaults to "two.sided".
legend.where	A string "bottomright", "bottomleft", "topleft", "topright". Defaults to "topleft" is x.axis="Rank" and "topright" if x.axis="Zvalues".
main	A string variable for the title of the plot.
pch.adj.p	A plotting 'character', or symbol to use for the adjusted p-value points. This can either be a single character or an integer code for one of a set of graphics symbols. Defaults to 17.
pch.raw.p	A plotting 'character', or symbol to use for the raw p-value points. This can either be a single character or an integer code for one of a set of graphics symbols. Defaults to 20.
pch.adj.fdr	A plotting 'character', or symbol to use for the adjusted FDR points. This can either be a single character or an integer code for one of a set of graphics symbols. Defaults to 20.
col	A vector of colors for the points and lines in the plot. If the input has 1 value all points and lines will be that same color. If the input has length of 3 then col.adj.fdr will be the first value, col.adj.p will be the second, and col.raw.p is the third. Defaults to c("dodgerblue","firebrick2", "black").

```
# Figure 9
plot(p.fdr(p=sim.data.p))

# Figure 10
plot(p.fdr(p=sim.data.p), xlim=c(0,25), ylim=c(0,0.25))
```
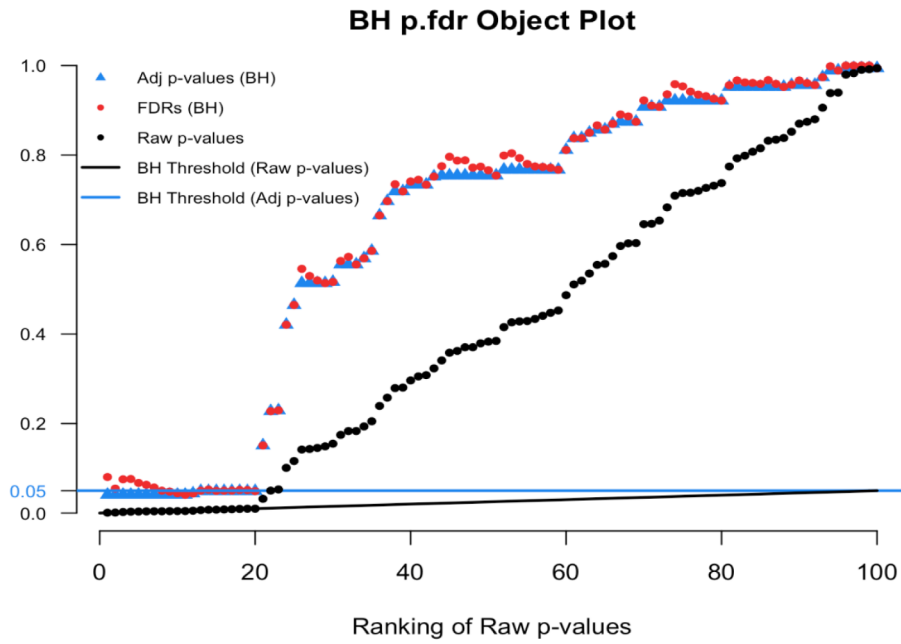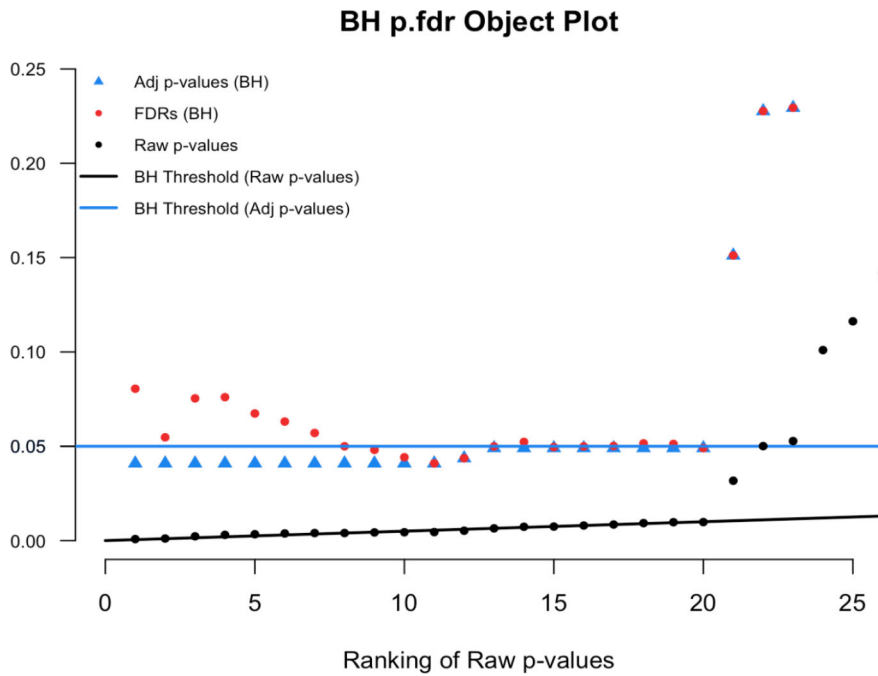

Figure 9. Benjamini-Hochberg p.fdr plot.

Figure 10. Magnified section of Figure 9.

Operation

This article was written using R version 4.0.3 (2020-10-10 on https://cran.r-project.org/bin/windows/base/old/4.0.3/) and `FDRestimation` version 1.0.0. The `FDRestimation` R package is available from CRAN and works on R versions 3.4 and above.

The package can be installed from CRAN using the following code:

```
# Install from CRAN
install.packages("FDRestimation")

# Load the package
library(FDRestimation)
```

Conclusions

We encourage the use of FDR methods and desire to illuminate the importance of contextualizing important findings. Our package provides useful and easy tool for those want to compute the false discovery rate, analogous to the role that `stats::p.adjust` plays for multiple comparison adjustments in everyday practice. Importantly, we hope it is now clear that p-value adjustments are not interchangeable with FDRs. In addition, `FDRestimation` package clearly delineates between methods for FDR control and methods for FDR estimation, while still allowing the user to choose from many different inputs and assumptions for their data. The more flexibility the user has at their disposal with these methods, better interpretations and applications will result.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

R package `FDRestimation` is available from CRAN: https://cran.r-project.org/package=FDRestimation.

Source code available from: https://github.com/murraymegan/FDRestimation.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.4684221.

License: MIT + file LICENSE

Acknowledgements

We would like to thank the Vanderbilt University Biostatistics SEDS Lab for help discussing these ideas.

References

Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J Royal Statistical Society.* 1995; **57**(1): 289–300.
Reference Source

Efron B: *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press; 2013. ISBN 9780511761362.

Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Annals Stat.* 2001; pages 1165–1188.

Bonferroni C: **Teoria statistica delle classi e calcolo delle probabilita.** *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze.* 1936; **8**: 3–62.

Šidák Zbyněk: **Rectangular confidence regions for the means of multivariate normal distributions.** *J Am Stat Assoc.* 1967; **62**(318): 626–633.

Holm S: **A simple sequentially rejective multiple test procedure.** *Scandinavian J Stat.* 1979; pages 65–70.

Hochberg Y: **A sharper bonferroni procedure for multiple tests of significance.** *Biometrika.* 1988; **75**(4): 800–802.
Publisher Full Text

Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A.* 2003; **100**(16): 9440–9445.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Meinshausen N, Rice J, *et al.*: **Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses.** *Annals Stat.* 2006; **34**(1): 373–393.
Publisher Full Text

Jiang H, Doerge RW: **Estimating the proportion of true null hypotheses for multiple comparisons.** *Cancer informatics.* 2008; **6**: 117693510800600001.
PubMed Abstract | **Free Full Text**

Nettleton D, Hwang JTG, Caldo RA, *et al.*: **Estimating the number of true null hypotheses from a histogram of p values.** *J Agri Biol Environmental Stat.* 2006; **11**(3): 337.
Publisher Full Text

Pounds S, Morris SW: **Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values.** *Bioinformatics.* 2003; **19**(10): 1236–1242.
Publisher Full Text

Scott DW: **On optimal and data-based histograms.** *Biometrika.* 1979; **66**(3): 605–610.
Publisher Full Text

Murray M: **Fdrestimation.** Nov 2020.
Reference Source

Open Peer Review

Current Peer Review Status:

David Colquhoun (iD)

Department of Neuroscience, Physiology and Pharmacology, University College London, London, UK

First, I must say that my knowledge about methods used for multiple comparisons is not great. That being said, I find the paper entirely convincing.

My comments will be restricted to matters of clarity and notation, apart from one technical point.

I must confess to having played a small part in the confusion about notation. In 2014 I wrote a paper about FDR but I was talking about the interpretation of p values from single tests, not multiple comparisons[1]. I said in that paper "IR should clarify that this paper is not about multiple comparisons. It is well known that high false discovery rates occur when many outcomes of a single intervention are tested." This was a mistake because up to that point, the term FDR had been used almost exclusively for the correction of p values for multiple comparisons.

I therefore added to the confusion by changing my notation from FDR to False positive risk (FPR) in my 2017[2] and 2019[3] papers.

In the 2017 paper, I wrote
"It is worth noting that all the methods for correcting for multiple comparisons aim to correct only the type 1 error. The result is, therefore, a (corrected) *p*-value, so it will still underestimate the false positive risk, for the reasons to be described."[2]

I was therefor very happy to see in your paper the section labelled "Lower bound on the FDR". There you use a lower bound which you attribute to Berger (1985) - a reference that is missing from the reference list. Your eqn 7 is the same as eqn 1.2 in Berger & Sellke JASA,1987), which they attribute to, Edwards, Lindeman and Savage (1963).

In my 2019 paper[3], I use a similar approach (which I attribute to Goodman,1999b). I compare it with 2 other methods in Fig. 3 of that paper. As you can see, it agrees will with my own approach

which differs only by using the *t* distribution rather than normal. But in order to get this agreement it is necessary to multiply by 2 the exponential term in the likelihood ratio exp($-z$^2/2) (see appendix A3, eqn A9). It's pointed out in Held and Ott,2018,eq.13 that this factor of two is necessary if one is talking about 2-tailed tests. Since you are, perhaps you should add it. This is confirmed by the fact that when I run your p.fdr with a single p value of 0;05 it gives what you call FDR (and I call FPR_50) as 0.128, whereas with the factor of 2 included I get FPR of 0.227 corresponding to a likelihood ration of 3.4 -and it's a well-known result that p=0.05 gives a likelihood ratio of around 3 by a variety of arguments.

In summary, I think that the paper would be improved by two things.
1. A discussion of the nomenclature used in this area. Is it sensible to use FDR only for corrected p values, or should the same term be used when you are referring to P(H0 | data)? I'd argue that FPR is more appropriate and that "risk" is more appropriate than "rate", given that the likelihood ratio is calculated by the p-equals approach, so we are considering only evidence from the one experiment m that we're analysing rather than long term error rates.

2. Inclusion of the factor of 2 in the calculation of the likelihood ratio (or at least a mention of the fact that your expression is not right for 2 tail tests).

References

1. Colquhoun D: An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci*. 2014; **1** (3): 140216 PubMed Abstract | Publisher Full Text
2. Arandjelović O: A more principled use of the p-value? Not so fast: a critique of Colquhoun's argument.*R Soc Open Sci*. 2019; **6** (5): 181519 PubMed Abstract | Publisher Full Text
3. Colquhoun D: The False Positive Risk: A Proposal Concerning What to Do Aboutp -Values. *The American Statistician*. 2019; **73** (sup1): 192-201 Publisher Full Text

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am not a professional statistician. I'm a biophysicist with a strong interest in statistics. In 1971 I wrote a textbook, Lectures on Biostatistics, and have recently written three papers about the (mis)interpretation of p values

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 November 2021

https://doi.org/10.5256/f1000research.78269.r97323

✔ **Ted Westling** (iD)
Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, USA
Yujian Wu (iD)
Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, USA

The authors have clearly answered all the questions, updated the figures and code as suggested.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Non-parametric estimation and inferrence

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 07 July 2021

? **Ted Westling** 🆔

Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, USA

Yujian Wu 🆔

Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, USA

This article emphasizes the necessity of distinguishing between the computation of adjusted p-values for FDR control and FDR estimation. A variety of commonly used methods for FDR estimation are introduced in the paper and the implemented package makes the methods easily accessible. The authors also provide a new method for estimation of the proportion for the null hypothesis. A simulation study verifies the new method works well when compared with other existing methods. The authors also have detailed interpretations of the functions in the package, together with some example code in the following section. Overall, the paper develops in a logical way with replicable results.

We have several comments and questions about the article and package:

1. All three functions have one common argument, "z values", which can either be a vector or a string. When we tried to pass only one vector to the argument (without providing p-values) to compute the FDR, the function didn't work. we had to pass the p-values corresponding to the z-values to make it work. Given that there's a z-value based approach for FDR estimation, is it possible to make the function work whenever it takes a numerical vector for the argument "z values"?

For example, if we run the following code:
```
z <- runif(10)
```

p.fdr(zvalues = z)
```

It will give me an error message "Error in TRUE %in% (pvalues > 1 | pvalues < 0) : argument "pvalues" is missing, with no default", unless we run it this way:
```

p.fdr(pvalues = pnorm(z))
```


2. The lower bound on FDR is introduced in the paper as one way of identifying optimistic FDR estimations. Is it reasonable to incorporate this lower bound into the output of the function "p.fdr"?

3. In terms of algorithm 1, is it necessary to store the heights for every bin? Is the height of the last bin alone enough?

4. In figure 6, when the true null proportion is 1, in which case the null distribution has a uniform distribution, $\hat{\pi}_0$ seems to be biased downward. Shouldn't the histogram reflect the uniform distribution better than other skewed distributions (when the proportion is less than 1), thus a better estimation of $\pi_0$?

5. Though it is claimed that the last histogram height method has the minimum MSE, Figure 6 only shows the bias of the estimators. It would be better if there's another figure showing the variance and/or MSE for all estimators.

6. Please add axis labels and interpretable legends to every figure.

7. The default method when there are ties in the ranks is important. The authors set the default to "random", which has the downside that the results may differ on the repeated running of the same code. Can the authors discuss this choice? In addition, we think if there are ties, the functions should provide a warning or message about the presence of ties and how they are being handled.

**Is the rationale for developing the new software tool clearly explained?**
Yes

**Is the description of the software tool technically sound?**
Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**
Partly

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
Partly

**Are the conclusions about the tool and its performance adequately supported by the**

**findings presented in the article?**

Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Non-parametric efficient estimation

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 06 Oct 2021

**Megan Murray**, Vanderbilt University Medical Center, Nashville, USA

Thank you very much for your time and suggestions. We have incorporated those suggestions and revised accordingly, which resulted in an improved paper. Some key issues were noted, and we clarify those below.

1. This is a subtle but important issue. The false discovery rates and adjusted p-values are all computed using the provided 'pvalues' vector. The z-values are used only for graphing in the z-value space and for computing the lower bound on the FDR. Remember that the z-values alone do not uniquely determine the p-value; it matters if the p-values are two-sided, one-sided in either direction, or adjusted for continuity or multiple looks. Accordingly, we found it simpler to require the p-values instead of hard-coding certain assumptions. We added a warning message for when the 'pvalues' vector is missing. It reads: "Error: The 'pvalues' vector is required for all FDR computations."

2. This is already included in the function. The lower bound is applied in all cases where the user has only supplied a single p-value. We have changed the class name of the output for this situation to improve user clarity.

3. Perhaps. However, we call the graphics package function "hist()", which automatically computes and stores every bin height. So, changing this would not improve the run time. However, the all the bin heights do not have to be stored for this method.

4. When the true null proportion is one, all of methods produce estimates that are biased downward (largely because the estimates are capped at 1). Accordingly, there is no unbiased estimate of the null proportion. Moreover here "better" is open for debate, as there is a bias-variance trade-off among the methods.

5. We have included the plots of the MSE in new Figure 7 so readers can assess for themselves the relative improvements. In our opinion, 'Last Histogram Height' and 'Storey' methods were the two methods that performed best over the widest range of cases.

6. We have submitted an updated version with a few axis labels and interpretable legends added. For Figures 1,2,4,8, and 9 we do not label the y-axis label to avoid

confusion between the adjusted p-values and raw p-values (the y-axis space is the unit interval).

7. Thank for your input here. We added the suggested warning to alert the user when ties are present. We chose the default "random" because in most cases the order of the p-values vector is not contextually meaningful. To address the reproducibility issue, the user can set a seed before running the function and reproduce the exact results.

*Competing Interests:* No competing interests were disclosed.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

**F1000Research**