A frequentist perspective on the local false discovery rate

Daniel Xiang[1], Jake A. Soloff[1], and William Fithian[2]

[1] *University of Chicago*
[2] *University of California, Berkeley*

November 14, 2022

Abstract

The two-groups model is a Bayesian multiple testing framework that models the truth status of hypotheses as random. Within this model, the local false discovery rate (lfdr, Efron et al., 2001) plays a fundamental role in testing hypotheses and in interpreting the resulting discoveries. We propose an alternative, frequentist definition of the lfdr based on the intuitive idea that lfdr is the local frequency of nulls in a small region of the sample space, which can be made precise by taking the limit of the marginal false discovery rate (mFDR) in a neighborhood shrinking to a point. Multiple testing procedures often summarize the rate at which they make false claims by reporting their FDR (Benjamini and Hochberg, 1995). Such procedures aim to control the average quality of their rejections. We instead propose to control the recognizable false discovery rate (RFDR), the quality of the least promising rejection.

1 Introduction

Suppose that we are testing some scientific hypothesis, and we observe a t-statistic equal to 3.5. How confident can we be in rejecting the corresponding null hypothesis? One way to quantify our confidence is to calculate a p-value, say $p = 0.001$ if there are 50 degrees of freedom. A common lay mistake is to interpret this 0.001 as the probability that the null is true in light of the data, but in fact if we wanted to calculate that quantity we would need two other pieces of information: the prior probability that the null was true, and the distribution the test statistic would take under the alternative. In many or most situations, different observers may disagree about these quantities; moreover, even to speculate about them is to first stipulate that the truth or falsehood of the null hypothesis is a random variable, which many scientists find difficult to swallow (see e.g. Goodman (1999)).

If we were to observe many comparable experiments — say, if we came upon a large data set that included a large number of studies from the same sub-area of the psychology literature, published around the same time — then we might at least be able to eliminate most of the subjectivity in the previous setting, by *estimating* a prior distribution for effect sizes in

experiments in the same category using empirical Bayes or hierarchical Bayes methods. If we were then willing to assume that the effect sizes for studies in this category are independent and identically distributed, or at least exchangeable, then we could arrive at a fairly satisfying answer to the question posed above, which might satisfy a wide range of observers. However, this would only make sense if we really were willing to accept not only the exchangeability assumption (which would almost certainly recede in its plausibility if we actually read the papers in question to discover what the different experiments were) but also the randomness of the original hypothesis' truth value.

These questions of relevance will almost inevitably prevent us from being fully satisfied that the results of a simple Bayesian analysis accurately reflect our (or anyone else's) true subjective posterior attitude. Nevertheless, methods like this are very appealing from a pragmatic perspective. If we choose to restrain ourselves from pulling all of these Bayesian threads, and instead stop at a simple empirical Bayes analysis, what interpretation does the result have? This paper aims to give a satisfying frequentist answer to the question of what quantity we have estimated, *without* assuming that the parameters themselves are exchangeable, or even random.

Rather than attempt to calculate a subjective probability, which would only be valid if these highly restrictive assumptions actually described our subjective state of mind, we instead ask the question: on average, of all the studies in this literature with t-statistics close to 3.5, what fraction had true null hypotheses? This question makes sense without requiring that the effects be exchangeable and can be answered consistently across sub-categories of the literature. Moreover, if scientists in the field agreed to report this number, say 10%, in addition to their t-statistic, then approximately 10% of the ones who reported this number would have made false discoveries. An analogy can be made to *calibration* in forecasting, which we discuss in Section 4.

In the next section, we analyze a dataset of "nudge" experiments from the behavioral psychology literature, in order to motivate and provide an instance of our answer to the question at the start: How confident are we in rejecting the hypothesis corresponding to a p-value near 0.001?

1.1 Example: Mertens et al. (2022a) nudge meta-analysis

Thaler and Sunstein (2009) describe "nudging" as an attempt to influence people's behavior in socially desirable ways without necessarily restricting their options. In their recent meta-analysis, Mertens et al. (2022a) collected data from 447 nudges in the behavioral psychology literature, to assess their overall effectiveness. For example, one of the nudges in the dataset set a healthy default option on a breakfast menu and measuring how much more often this option was selected. Another nudge was designed to promote organ donor registration in the U.K. by adding extra text to a prompt that was shown to drivers after renewing their licenses online. For instance, one prompt asked a question about reciprocity,

> If you needed an organ transplant, would you have one? If so please help others. (rn)

Another nudge in the same organ donation study encouraged drivers to sign up by suggesting that doing so was a social norm,

> Every day thousands of people who see this page decide to register.

People who saw the latter nudge were significantly less likely to join the organ donor registry, with a t-statistic larger than 13!

The goal of the meta-analysis by Mertens et al. (2022a) was to analyze these nudges in aggregate along with hundreds of others to determine whether or not nudges were effective; the formulation of this question and the authors' conclusion was the subject of some debate, see e.g. Maier et al. (2022), Mertens et al. (2022b), Szaszi et al. (2022).

Given a p-value for each nudge study, we might think to report the rate of false discoveries (FDR, Benjamini and Hochberg, 1995) among the studies deemed significant at the 5% level, of which there were 261 out of the total 447 nudge studies in the dataset. The p-values below the significance threshold have been re-normalized by dividing by the threshold, and plotted in the left panel of Figure 1. This pre-processing step is a way to work around the publication bias present in scientific journals; although the fraction of nudges which either backfire or have no effect may be under-represented among published studies in the aggregated dataset, the nulls falling in the significance region are less prone to censorship (Hung and Fithian, 2020).

The Storey estimate (Storey, 2002) of the proportion of nulls among these p-values is $\hat{\pi}_0 = 0.28$. In other words, nearly one in three reported significant results is a false discovery. One way to mitigate this high rate of false discoveries is to first specify our FDR tolerance level, say 0.1, and then run a multiple testing procedure which guarantees a 10% rate of false positives. Running the Storey-adjusted BH procedure (Benjamini and Hochberg, 1995) at level $q = 0.1$ yields a rejection threshold $\tau_q^{\text{BH}} = 0.27$, as shown in the right panel of Figure 1.

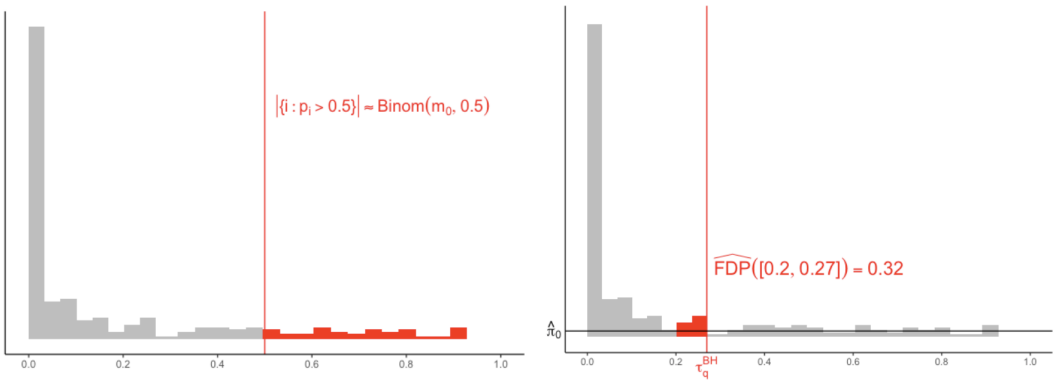

Figure 1: p-values below the significance threshold are re-normalized by scaling by 40 (the reciprocal of the 2.5% one-sided significance threshold) and plotted in the left panel. On the right, the Storey estimate is plotted as a horizontal line, and the FDP towards the edge of the BH set is estimated at 32%.

As the right panel of Figure 1 shows, the proportion below the horizontal line at $\hat{\pi}_0$ can exceed 10% for certain subsets of rejections. In Figure 2, the order statistics are used to

estimate the FDP on a subset of rejections, and there is enough data to see that it increases well above the specified 10% level within the second half of the rejection set. The BH set is overly liberal in its last few rejections, which may be of arbitrarily low quality as judged on the basis of the order statistics.

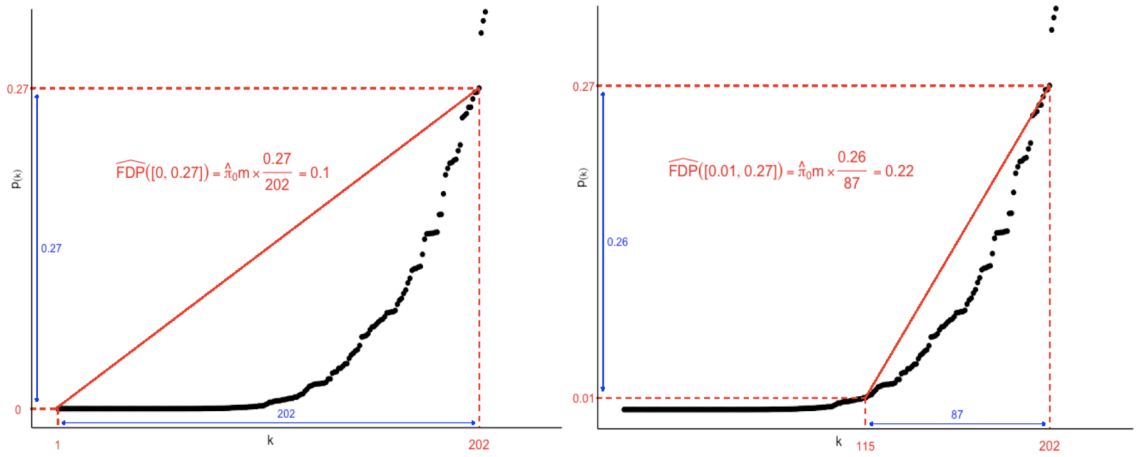

Figure 2: The order statistics below the BH rejection threshold are plotted against their rank, and the FDP over the entire set of rejections is estimated at 0.1 (left), whereas within the second half of the rejections the estimate is 0.22 (right). The FDP estimate over a subset of the rejection region is obtained via $\frac{V[a,b]}{R[a,b]} \approx \frac{m\hat{\pi}_0 \times (b-a)}{R[a,b]}$, where $[a, b] \subset [0, \tau_q^{\text{BH}}]$, $V[a, b]$ is the number of nulls in $[a, b]$, and $R[a, b]$ is the number of p-values in $[a, b]$. The estimate is proportional to the slope of the secant over the subset, and increases towards the edge of the rejection set.

Continuing the thought experiment further, we could imagine shrinking a subset of rejections down to a single one, and asking about whether a particular nudge was effective or not. At this point, if we use a Bayes model on the hypotheses,

$$H_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ nudge has the desired effect} \\ 0 & \text{if the } i^{\text{th}} \text{ nudge has no effect or backfires.} \end{cases}$$

we can answer the question by estimating the posterior probability that the hypothesis is null, conditional on the observation.

In the Bayes two-groups model, the truth statuses of hypotheses are modeled as Bernoulli random variables. For $i = 1, \ldots, m$,

$$H_i \overset{\text{iid}}{\sim} \text{Bern}(1 - \pi_0), \tag{1}$$

$$p_i \mid H_i \sim \begin{cases} f_0, & \text{if } H_i = 0 \\ f_1 & \text{if } H_i = 1, \end{cases}$$

where $\pi_0 \in [0, 1]$ is the probability that a hypothesis is null, f_0 and f_1 are the null and alternative densities, and $p_1, \ldots, p_m \in \mathbb{R}$ are p-values. In this setting, the *local false discovery*

rate (lfdr, Efron et al., 2001) is defined,

$$\text{lfdr}(t) = \mathbb{P}(H_i = 0 \mid p_i = t) = \frac{\pi_0 f_0(t)}{f(t)}, \quad f \coloneqq \pi_0 f_0 + (1 - \pi_0) f_1 \tag{2}$$

and models our uncertainty in a hypothesis conditional on the observation. As a posterior probability, the lfdr appears to rely crucially on the Bayesian assumption (1). However, we were able to recognize a bad subset of the rejections by estimating the false discovery proportion in local regions of the sample space, without making any assumptions about how the hypotheses were generated. From a frequentist point of view, it still makes sense to ask which of the nudges are individually promising. Without the random effects assumption, the lfdr can still be non-trivially defined, but its interpretation is no longer epistemic in nature. As we have seen with the nudge data, a local false discovery rate can still be estimated, but what does this estimate mean in the absence of the assumption (1)?

1.2 Outline of the paper

In Section 2 we propose a definition of the lfdr for the fixed effects model. In Section 2.1 we outline a connection between the definition (4) and an optimal procedure for minimizing a compound risk. Closely related is the notion of a procedure's recognizable FDR, defined in Section 2.3. In Section 3 we continue our analysis of the nudge data, and propose a new range of thresholds for assessing statistical significance of studies published in the nudge literature. We also demonstrate an extension of SL procedure for analyzing z-scores coming from an HIV microarray dataset exhibiting asymmetry. Section 4 concludes the paper with a discussion, and Section A contains all proofs not provided in the exposition.

2 A frequentist local false discovery rate

Let $H_1, \ldots, H_m \in \{0, 1\}$ be fixed, and suppose each p-value is independently distributed according to a probability measure on $[0, 1]$,

$$p_i \overset{\text{ind}}{\sim} P^{(i)} \quad \text{for } i = 1, \ldots, m. \tag{3}$$

The local false discovery rate (lfdr) is defined as the probability, conditional on *some* p-value having been realized at t, that the hypothesis corresponding to *that* p-value is null,

$$\text{lfdr}(t) \coloneqq \mathbb{P}(H_J = 0 \mid p_J = t \text{ for some } J \in [m]). \tag{4}$$

Suppose for simplicity that each $P^{(i)}$ has a Lebesgue continuous density $f^{(i)}$. Then by independence, on the event that J exists, it is essentially unique. No more than one hypothesis H_J corresponds to an observation at t, and $\text{lfdr}(t)$ models our uncertainty about whether or not this hypothesis is null. Here the randomness in the index J, representing *which* p-value is realized at t, substitutes for the random truth value of each H_j.

At first glance, one might be concerned about conditioning on the event that a continuous random variable is exactly equal to a point, since this is a zero probability occurrence. The

conditional probability can be understood as a limit of posterior probabilities $\mathbb{P}(H_J = 0 \mid p_J \in N_\varepsilon(t)$ for some $J \in [m])$ for a collection of neighborhoods $N_\varepsilon(t)$ shrinking to $\{t\}$ as $\varepsilon \to 0$. Since each p_i has a continuous density on $[0,1]$, the manner in which the neighborhood $N_\varepsilon(t)$ shrinks to $\{t\}$ is arbitrary, always yielding the same limit. When there are multiple p-values in $N_\varepsilon(t)$, the selected p_J in (4) can be understood as a uniform draw from among them. In the $\varepsilon \to 0$ limit, the posterior probability is proportional to the ratio between the average null density and the overall average density of the p-values.

Theorem 2.1. *Fix $H_1, \ldots, H_m \in \{0,1\}$ and suppose $p_1, \ldots, p_m$ are generated independently from (3). Then if each $P^{(i)}$ is a continuous distribution with density $f^{(i)}$, we have*

$$\mathrm{lfdr}(t) = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)}, \tag{5}$$

where $\bar{f}_0 = \frac{1}{m\bar{\pi}_0} \sum_{i:H_i=0} f^{(i)}$ and $\bar{f} = \frac{1}{m}\sum_{i=1}^m f^{(i)}$ are the null average and overall average densities, and $\bar{\pi}_0 = \frac{1}{m}\sum_{i=1}^m (1 - H_i)$ is the proportion of nulls. If $(H_i, p_i)_{i=1}^m$ are generated independently from the two-groups model (1), then

$$\mathrm{lfdr}(t) = \frac{\pi_0 f_0(t)}{f(t)},$$

where $f = \pi_0 f_0 + (1 - \pi_0) f_1$ is the marginal density of each observation.

Definition (4) recovers the Bayes lfdr within the two-groups model. It is non-trivial in the frequentist model because by conditioning on t being in the *set* of observations, and not on the observations themselves, there is still uncertainty that can be measured on the basis of the order statistics. Having conditioned only on the set of observations, we can ask about hypotheses corresponding to particular points in the set. For example, the hypotheses corresponding to the smallest few p-values are a random subset of hypotheses, even if the effects are fixed.

The lfdr answers the question of how confident we should be in rejecting a hypothesis corresponding to a particular p-value. Sun and Cai (2007) show that lfdr is the right quantity to look at for deciding whether or not to reject a hypothesis in the Bayes model. The conclusion continues to hold in the frequentist model, where optimality is defined with respect to a weighted combination of type 1 and 2 errors. We formalize this claim in the next section.

2.1 A compound decisions perspective

Consider data $p = (p_1, \ldots, p_m)$ drawn independently from model (3) with fixed effects $H = (H_1, \ldots, H_m) \in \{0,1\}^m$. In this setting, a multiple testing procedure can be represented by a set of decisions $\delta(p) = (\delta_1(p), \ldots, \delta_m(p)) \in \{0,1\}^m$ with average loss,

$$L(H, \delta) = \frac{1}{m}\sum_{i=1}^m \ell(H_i, \delta_i).$$

for some non-negative function $\ell : \{0,1\}^2 \to \mathbb{R}_+$. If the decision rule is separable, i.e. $\delta_i(p) = \mathfrak{d}(p_i)$ for some function $\mathfrak{d} : [0,1] \to \{0,1\}$, then the expectation of the average loss coincides with the Bayes risk for a single pair $(H_I, p_I) \in \{0,1\} \times \mathbb{R}$,

$$\mathbb{E}L(H, \delta) = \mathbb{E}\ell(H_I, \mathfrak{d}(p_I)), \tag{6}$$

where (H_I, p_I) is drawn from a two-groups model:

$$\begin{aligned} H_I &\sim \text{Bern}(1 - \bar{\pi}_0), \\ p_I \mid H_I = h &\sim \begin{cases} \bar{f}_0 & \text{if } h = 0 \\ \bar{f}_1 & \text{if } h = 1 \end{cases} \end{aligned} \tag{7}$$

Here, $\bar{f}_0 \coloneqq \frac{1}{m_0} \sum_{i:H_i=0} f^{(i)}$ and $\bar{f}_1 \coloneqq \frac{1}{m_1} \sum_{i:H_i=1} f^{(i)}$ denote the average null and alternative densities respectively. We call (7) the *oracle two-groups model* to emphasize that it depends on unknown parameters such as $\bar{f}_1$ and m_0, and that it holds in the frequentist setting where the truth status of each hypothesis is fixed.

To summarize, this reformulation of the compound risk shows that minimizing the average case loss is equivalent to achieving the univariate Bayes risk in the oracle two-groups model. In particular, the smallest risk among all separable rules is attained by the Bayes rule within this model, characterized by the local false discovery rate,

$$\mathbb{P}(H_I = 0 \mid p_I = t) = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)}, \tag{8}$$

where $\bar{f} \coloneqq \bar{\pi}_0 \bar{f}_0 + (1 - \bar{\pi}_0) \bar{f}_1$ is the average density of the observations. Yekutieli and Weinstein (2019) noted that the procedure

$$\mathfrak{d}^*(p_i) = 1_{\{\text{lfdr}(p_i) \le \alpha\}}, \quad \alpha \in (0,1), \tag{9}$$

optimizes a trade-off between marginal false discovery/non-discovery rates, which is equivalent to minimizing the weighted misclassification risk,

$$\mathbb{E}L(H, \delta) = \mathbb{E}\ell(H_I, \mathfrak{d}(p_I)) = \mathbb{P}(H_I = 1, \mathfrak{d}(p_I) = 0) + \lambda \cdot \mathbb{P}(H_I = 0, \mathfrak{d}(p_I) = 1),$$

for a parameter $\lambda = \alpha^{-1} - 1 > 0$ specifying the cost of making a type 1 error relative to a type 2 error. According to the first part of Theorem 2.1, formula (8) represents our uncertainty in a hypothesis corresponding to a p-value realized at t. Instead of describing our beliefs in a fixed effects model, lfdr gives a way to decide whether or not to reject a hypothesis. This can be done without making claims about the probability that a fixed hypothesis is null, or even assuming in the first place that the truth status of each hypothesis is random. The lfdr can be more generally understood as the limiting rate of nulls in a shrinking region of the sample space.

2.2 Local frequency of nulls

Although the reduction to an oracle two-groups model is useful for seeing why the lfdr is a fundamental quantity for fixed effects testing, we don't actually need to reference any kind of Bayes two-groups model to make sense of the lfdr. It can be understood on its own as the local frequency of nulls near t. Define the interval FDP

$$\mathrm{FDP}([s,t]) = \frac{\#\{i : H_i = 0, p_i \in [s,t]\}}{1 \vee \#\{i : p_i \in [s,t]\}}$$

and interval pFDR

$$\mathrm{pFDR}([s,t]) = \mathbb{E}\left(\mathrm{FDP}([s,t]) \mid p_i \in [s,t] \text{ for some } i\right).$$

As the interval shrinks to a point, its pFDR tends to the local false discovery rate.

Theorem 2.2. *Suppose the p-values are independently distributed according to (3). If each $P^{(i)}$ has a continuous density $f^{(i)}$, then*

$$\lim_{\varepsilon \to 0} \mathrm{pFDR}([t-\varepsilon, t]) = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)}, \qquad t \in (0,1). \tag{10}$$

If some $P^{(i)}$ has an atom at t, then

$$\lim_{\varepsilon \to 0} \mathrm{mFDR}([t-\varepsilon, t]) = \frac{\bar{\pi}_0 \bar{P}_0(\{t\})}{\bar{P}(\{t\})},$$

where $\bar{P}_0 \coloneqq \frac{1}{m_0} \sum_{i:H_i=0} P^{(i)}$ and $\bar{P} \coloneqq \frac{1}{m} \sum_{i=1}^m P^{(i)}$ are the average null and overall average probability measures.

The limiting frequency of nulls near a point in the sample space is a function of unobserved quantities, such as the null proportion $\bar{\pi}_0$ and average density $\bar{f}$. When the p-values are drawn iid from a Bayes two-groups model, the law of large numbers implies convergence of the empirical distribution function F_m to the marginal cdf,

$$F_m(t) \coloneqq \frac{1}{m} \sum_{i=1}^m 1_{\{p_i \le t\}} \to F(t) \text{ for every } t \in [0,1] \text{ as } m \to \infty.$$

To estimate the lfdr, Strimmer (2008) proposed to use the Grenander estimator, which estimates the marginal density using slopes of the least concave majorant of the empirical distribution function (Grenander, 1956). By contrast, within the frequentist model where the p-values are not identically distributed, F_m is unbiased for their average distribution function, and the Grenander estimator targets the average density $\bar{f}$. This estimate was used by Soloff et al. (2022) to assess the rejections made at the margin of the BH rejection set, which were shown to be of low quality in a simulation setting from the original BH paper.

2.3 Quality control at the margin

Bounding $\bar{\pi}_0 \leq 1$ and plugging in the Grenander estimate for $\bar{f}$ gives an estimate of the lfdr, which can be used to define the Support Line (SL) procedure (see Section 3.2 of Soloff et al. (2022)). This procedure controls the quality of its least promising discoveries by rejecting the hypotheses corresponding to the R smallest p-values, where

$$R \coloneqq \underset{k=0,\ldots,m}{\operatorname{argmax}} \left\{ \frac{\alpha k}{m} - p_{(k)} \right\}, \quad p_{(0)} \coloneqq 0. \tag{11}$$

In the Bayes model, the SL procedure controls its maximum lfdr over its rejection set in expectation (Theorem 1, Soloff et al. (2022)). SL also satisfies an exact bound on its rate of false discoveries under fairly mild assumptions in the frequentist setting. Namely, when nulls are independent and uniform, and the alternative densities are non-increasing, SL controls its rate of false discoveries uniformly over subsets of the rejections. In contrast, the BH procedure can have a high rate of false discoveries in the latter half of the rejections (see e.g. Figure 2).

Intuitively, the recognizable false discovery rate (RFDR) is the highest rate of false discoveries on a subset of rejections that can be selected by looking only at the order statistics. More precisely, let $P_m = \frac{1}{m}\sum_{i=1}^m \delta_{p_i}$ denote the empirical distribution of the p-values, $\mathcal{R} = \{i \in [m] : H_{(i)} \text{ rejected}\}$ denote the ranks of the p-values whose hypotheses we reject, and $\mathcal{H}_0 \coloneqq \{i \in [m] : H_{(i)} = 0\}$ are the ranks of the null p-values. Then the RFDR of the rejection set $\mathcal{R}$ is defined

$$\mathrm{RFDR}(\mathcal{R}) \coloneqq \mathbb{E}\left[\sup_{S \subset \mathcal{R}} \mathbb{E}\left(\frac{|\mathcal{H}_0 \cap S|}{|S|} \mid P_m\right)\right], \tag{12}$$

where the supremum is taken over all non-empty subsets $S \subset \mathcal{R}$ of ranks which are measurable with respect to the order statistics. This rules out, for instance, the singleton set $S = \{i : H_{(i)} = H_1\}$, since this index cannot be selected by looking at only the order statistics. An implication of this restricted supremum is that it doesn't reduce to the indicator of the event that some null was rejected, and so the RFDR avoids a trivial reduction to the family-wise error rate (FWER). Finner and Roters (2001) noted a fundamental difference between FDR and FWER control: any subset of a FWER-controlling rejection set also controls FWER at the same level, whereas the same cannot be said about a procedure's FDR (see e.g. Figure 2). The RFDR can be viewed as a corrective to the FDR, still yielding a more liberal notion of type 1 error than the FWER, but self-consistent in the sense that subsets of an RFDR-controlling rejection set inherit the RFDR guarantee of the overall rejection set.

The supremum in (12) is achieved by the singleton set containing the rank of the least promising rejection that can be recognized based on the order statistics, and coincides with the max-lfdr of Soloff et al., 2022 when the data are drawn iid from the Bayes model (1).

Lemma 2.1. *If p_i is drawn independently from (3) and $H_i \in \{0,1\}$ is fixed for each $i = 1,\ldots,m$, then*

$$\mathrm{RFDR}(\mathcal{R}) = \mathbb{E}\left[\sup_{J \in \mathcal{R}} \mathbb{P}\left(H_{(J)} = 0 \mid P_m\right)\right],$$

where the supremum is taken over all ranks $J \in \mathcal{R}$ which are measurable with respect to the order statistics $p_{(1)} \leq \cdots \leq p_{(m)}$. If $\{(H_i, p_i)\}_{i=1}^m$ are generated iid from the Bayes model (1), then

$$\mathrm{RFDR}(\mathcal{R}) = \mathbb{E}\left[\max_{i \in \mathcal{R}} \mathrm{lfdr}(p_{(i)})\right].$$

According to the second part of Lemma 2.1, SL controls its RFDR in a Bayes two-groups model with a decreasing alternative density, as a direct consequence of Theorem 1 in Soloff et al. (2022). In the fixed hypothesis setting, SL satisfies a bound on its RFDR closely mirroring the BH guarantee when a monotonicity assumption holds:

$$f^{(i)} \text{ is non-increasing for each } i. \tag{MA}$$

Theorem 2.3. *Fix $H_1, \ldots, H_m \in \{0,1\}$. If $p_1, \ldots, p_m$ are independent, $H_i = 0$ implies $f^{(i)} = 1_{[0,1]}$, and $H_i = 1$ implies $f^{(i)}$ is non-increasing, then*[1]

$$\mathrm{RFDR}(\mathcal{R}_\alpha) = \mathbb{P}(H_{(R)} = 0) = \bar{\pi}_0 \alpha,$$

where $H_{(1)}, \ldots, H_{(m)}$ are the hypotheses corresponding to $p_{(1)}, \ldots, p_{(m)}$ and $\mathcal{R}_\alpha := \{1, \ldots, R\}$ is defined by (11).

Proof. The first equality follows from Lemmas A.3 and 2.1, since R is a function of the order statistics. For the second equality, note that since the null p-values are exchangeable,

$$\mathbb{P}(H_{(R)} = 0) = m\bar{\pi}_0 \mathbb{P}(p_{(R)} = p_m),$$

where we have assumed without loss of generality[2] that $H_m = 0$. By Lemma 2 of Soloff (2020), the probability that a particular null achieves the optimum in (11) is equal[3] to $\frac{\alpha}{m}$. □

Under some regularity conditions on the average density of the test statistics, the lfdr at the rejection threshold $\hat{\tau}_\alpha := p_{(R)}$ is controlled with high probability, a result we state below and prove in Appendix A[4].

Theorem 2.4. *Suppose the p-values are independent and uniform under the null, and that $\bar{f}$ has a unique solution to the equation $\bar{f}(\tau_\alpha) = \alpha^{-1}$. If $\bar{f}$ is differentiable and decreasing over $(0, \tau_\alpha + \alpha)$, and for some constants $\delta, J > 0$ we have $J \leq |\bar{f}'(t)| \leq J^{-1}$ for all t with $|t - \tau_\alpha| \leq \varepsilon$, where $\varepsilon := \left(\frac{48}{\alpha J^2}\right)^{1/3} m^{-1/3} \log(2m/\delta)$, then with probability $\geq 1 - \delta$,*

$$\mathrm{lfdr}(\hat{\tau}_\alpha) \leq \bar{\pi}_0 \alpha + C m^{-1/3} \log(m/\delta),$$

for a constant $C > 0$ depending on α, J and δ.

[1] Here, $H_{(0)} := 1$ is used to indicate the event where no rejections are made, i.e. $R = 0$.

[2] If $\bar{\pi}_0 = 0$ then the result holds trivially.

[3] An alternative proof of the claim that $\mathbb{P}(p_{(R)} = p_m) = \frac{\alpha}{m}$ can be found in Appendix A.

[4] In Appendix B, we check that the requirement that $\bar{f}$ be decreasing over the entire interval $(0,1)$ need not hold for the threshold $\hat{\tau}_\alpha$ to approach the population threshold at the $m^{-1/3}$ rate.

3 Applications

We illustrate some of the key ideas introduced in this paper on two real datasets. We first continue our study of meta-analysis data from Section 1.1 on various types of nudge experiments, and then analyze an HIV microarray dataset from Efron (2012).

3.1 Nudge data

In one of the studies from the Mertens et al. (2022a) dataset, parents were split into two groups and shown a menu from which each parent would select one of two breakfast options for their child. The first group was shown a menu with the healthy option shown as the default, in the sense that it was displayed in large font in the center of the menu. A less healthy breakfast option was only shown in a footnote, illustrated in Figure 3. The second group saw the same menu but with the placement of the two options flipped. The researchers

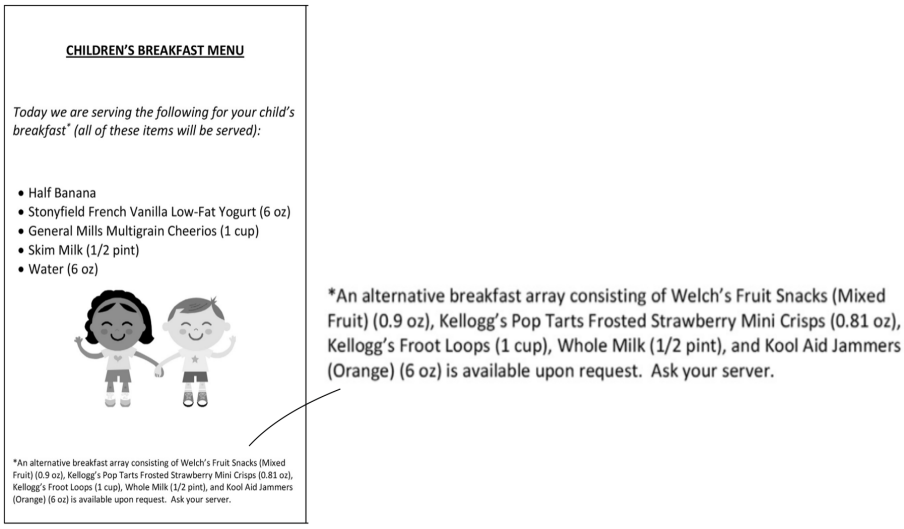

Figure 3: A breakfast menu from the Loeb et al. (2017) study with the healthy option set as default.

then recorded the proportion of parents within each group that selected the healthy option, and found that the first group was 79% more likely to choose the healthy option, with a large and positive t-statistic ($t = 6.2$, Loeb et al. (2017)), indicating a nudge effect in the direction predicted by the researchers.

However, not all nudges work as intended. As described in Section 1.1, another study in the Mertens et al. (2022a) meta-analysis contained data collected by researchers from the NHS Organ Donor Registry in the U.K. who hoped to devise nudges that could increase the number of organ donors. Drivers were split into groups and, after having their license renewed online, were prompted to become organ donors. One group was shown a panel which simply offered the option to join, while another group was shown the same panel, but with an additional image of smiling organ donors and extra text suggesting that organ donation is a social norm. These two prompts are displayed in Figure 4.

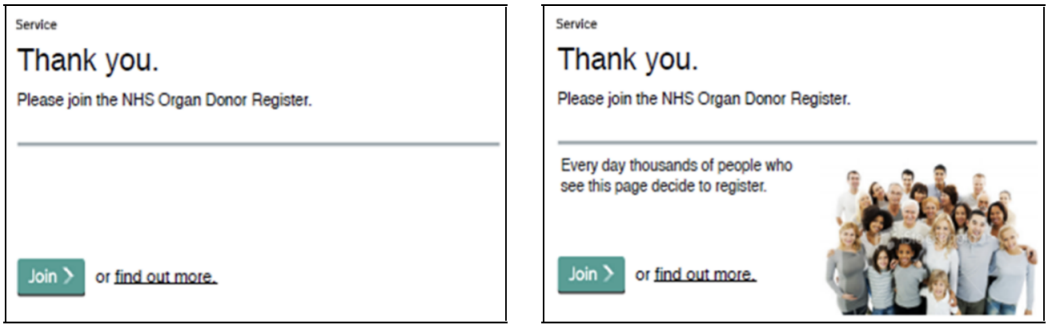

Figure 4: Two prompts from the study "Applying Behavioural Insights to Organ Donation".

The researchers measured the proportion in each group that joined the organ donor registry, and found that drivers in the group that saw the second panel in Figure 4 were slightly less likely to join, with an estimated decrease of 5% in the sign-up rate relative to the control group, and a corresponding t-statistic of -1.7. With a mildly negative t-statistic, this nudge might not have backfired, but there's certainly no evidence to suggest it promoted organ donor registration. By contrast, the nudge in the same organ donation study that used the reciprocity nudge (rn) produced a much stronger effect. The reciprocity nudge increased the sign-up rate by 35% relative to the control group, with an associated t-statistic of 13.1. The number of samples used to estimate the effect of the organ donation nudges was over 270,000.

The organ donation nudges and the breakfast nudge are just a few examples that demonstrate the heterogeneity in concept and effectiveness of the experiments summarized in the Mertens et al. (2022a) dataset.

In classical meta-analysis, the studies to be aggregated typically contain estimates of a common effect. For instance, Peto (1996) was interested in estimating the effect of the drug tamoxifen on survival rates among breast cancer patients using data from trials conducted in Europe and North America for different lengths of the treatment, ranging from two to five to ten years. In this setting, the effect of the drug tamoxifen is assumed to be constant across locations, and it is reasonable to aggregate the evidence and try to determine whether tamoxifen is effective, and to quantify the overall effect of each treatment. These estimates might then be used to better inform prescriptions for the duration of tamoxifen treatment on future breast cancer patients.

Mertens et al. (2022a) compiled data from 447 nudge experiments and estimated a positive "overall nudge effect", which they deemed statistically significant. Maier et al. (2022) responded with an article claiming there was "no evidence for nudging after adjusting for publication bias". In the examples presented above, evidence for the effectiveness of particular nudges ranged widely, and there is little doubt about the effectiveness of certain nudges. For instance, the reciprocity nudge in the organ donation study had a t-statistic of 13.1, whereas the social norm nudge was insignificant. The breakfast nudge was moderately significant, with a t-statistic near 6, but for other studies in the nudge dataset with t-statistics around

3 or 4, it is much less clear what to think.

Estimation of the overall effect is still a well-defined statistical task, but the estimand is not immediately meaningful in the nudge setting. What does it mean to aggregate a breakfast nudge with an organ donation nudge, and what use is there for an estimate of their average effect? Some of the nudges worked, some of them didn't and some of them may well have backfired; scientifically interesting departures from the global null can be consistent with a negligible overall effect (Szaszi et al., 2022). Instead of estimating the average effect of all nudges being studied in this area of the psychology literature, we might rather ask about the rate of false positives among findings deemed significant (Section 1.1), and about the threshold below which a p-value indicates strong evidence that a particular nudge was effective.

We estimate new default thresholds for results in this area of the behavioral psychology literature to be deemed statistically significant, roughly forty times smaller than the standard 0.05 cut-off. We argue that by using the smaller threshold, nudge researchers will control the rate at which they collectively make false discoveries, so that on average only 10% of the studies that just barely reach the new threshold of statistical evidence are false positives. Alternatively, to target a 5% false positive rate among the field's least promising discoveries, the standard p-value cut-off should be reduced by a factor of about a hundred. For a 20% false discovery rate at the margin, we find that the cut-off should be reduced by a factor of about twelve, which roughly agrees with a more general proposal to reduce the cut-off from 0.05 to 0.005 (see e.g. Greenwald et al. (1996) and Benjamin et al. (2018)). The suggested new cut-offs are summarized in Table 1.

lfdr cut-off	0.05	0.10	0.20	0.30
one sided p-value cut-off	2.44×10^{-4}	6.25×10^{-4}	2.04×10^{-3}	3.70×10^{-3}

Table 1: Suggested significance cut-offs for one-sided p-values in the nudge literature. See also the plot in Figure 6.

To estimate these significance thresholds for the nudge literature, we first estimate the base rate of nulls being published in this area of the literature at the 5% two-sided significance level. As explained in Section 1.1, restricting our attention to one-sided p-values below 2.5% is a way to work around the issue of publication bias, since these nulls are less likely to be censored by journals that adhere to the standard significance cut-off. If the nulls are super-uniform, then those falling below the significance cut-off are conditionally super-uniform$(0, 0.025)$, and thus multiplying each p-value by 40 restores validity of the p-values. After adjustment, the nulls are at least as likely to fall to the right of $1/2$ as they are to fall to the left, which implies a conservative approximation of the rate of false discoveries among significant studies, $\bar{\pi}_0 \approx 0.28$ (see the left panel of Figure 1). As we have seen in Section 1.1, incorporating this Storey-type estimate of $\bar{\pi}_0$ into the BH procedure at level 10% yields 202 rejections.

The Storey-modified SL procedure is proven to maintain its max-lfdr control within the Bayes model (Theorem 4, Soloff et al. (2022)), and a similar guarantee can be shown on its

RFDR within the fixed effects model under the conditions of Theorem 2.3. Incorporating the estimate of $\bar{\pi}_0$, the SL procedure run at the same 10% level makes about two thirds as many rejections as BH, 131 in total, pictured in the left panel of Figure 5. By reducing the number of rejections compared to the BH set, we ensure FDR control at the 10% level uniformly over all post-selected subsets identifiable from the un-ordered set of test statistics. Equivalently, running SL at level $\alpha = 0.1$ guarantees that the least promising effect among those deemed significant by this procedure is null with probability below 10%.

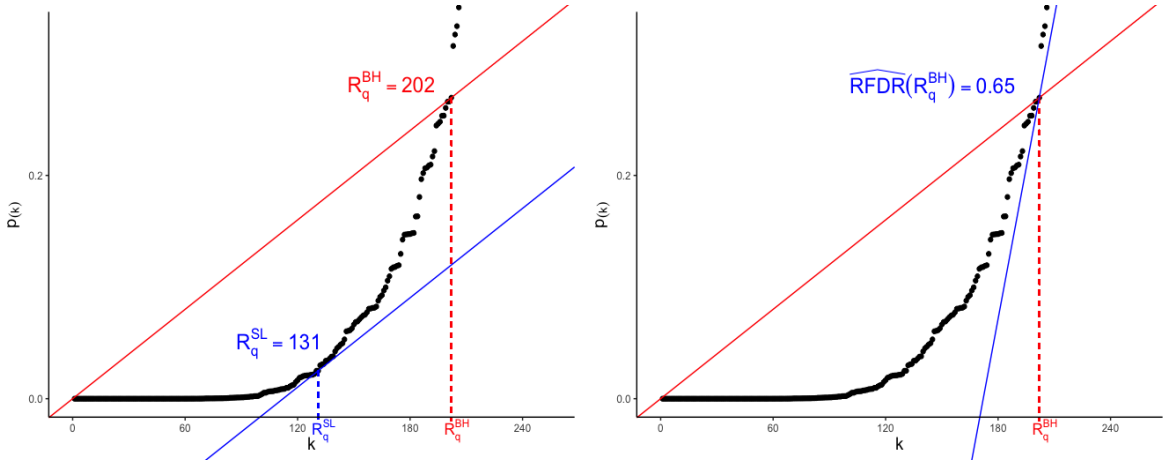

Figure 5: The left plot illustrates the Storey adjusted BH and SL procedures run at level 10% on the one sided p-values from the Mertens et al. (2022a) dataset. The right plot shows the estimate of the RFDR for the Storey-BH procedure run at level 10%, which is proportional to the slope of a supporting line at the largest p-value in the rejection region.

Conversely, given some threshold t, we can estimate the recognizable FDR of $\{i : p_{(i)} \leq t\}$ by asking about the level at which we would run SL to obtain this rejection set. Under the monotonicity assumption (MA), the least promising rejection is reducible to the one with the largest p-value below the threshold, and the recognizable FDR in this case is equivalent to the probability that the corresponding hypothesis is null. The level at which SL barely rejects the hypothesis corresponding to a particular order statistic is proportional to the slope of a line that supports the plot at that order statistic. Equivalently, the level is given by the estimate of $\bar{\pi}_0$ divided by the Grenander estimator evaluated at the largest p-value in the rejection region. This estimate for the FDR at the margin is 0.65 for the Storey-BH procedure targeting a 10% overall rate of false positives, illustrated in the right panel Figure 5.

One strategy to mitigate the high rate of false discoveries at the edge of the BH set is to run the procedure at a range of levels, lowering the tuning parameter α until the estimate of the RFDR falls below the targeted level. However, if the goal is to control the quality of individual rejections at a specified tolerance α, then we recommend directly running the SL procedure at level α, which has the same computational cost as the BH procedure.

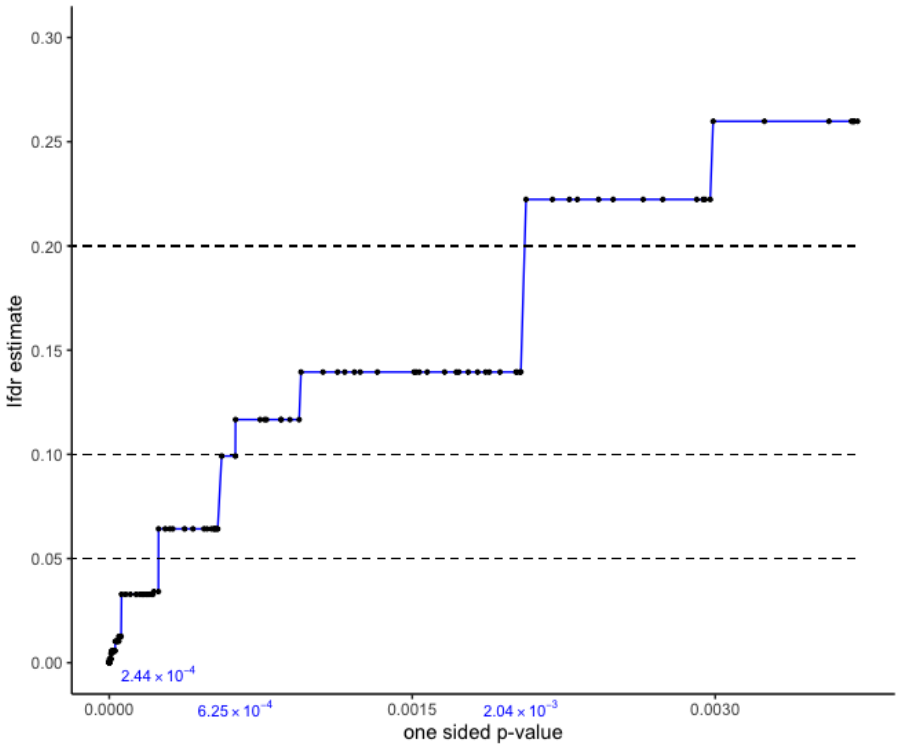

Figure 6: A plot is shown of the estimated local false discovery rate for the nudge data. Dashed horizontal lines indicate a cut-off for the estimated lfdr, and the corresponding one sided p-value cut-offs are indicated in blue font (see also Table 1). Black dots indicate the observed (un-adjusted) p-values and their estimated lfdr values.

3.2 HIV data

The HIV dataset contains $m = 7680$ entries, each representing a difference between the average gene expression level in two groups of patients. The first group consists of 4 HIV-negative individuals, while the second group consists of 4 HIV-positive individuals. The goal of the study is to identify a small subset of the genes that are potentially relevant for understanding HIV. Each difference has been standardized, and after some pre-processing they are modeled as a normal random variables with unit variance and unknown means,

$$Z_i \sim N(\theta_i, 1) \quad \text{independently for } i = 1, \ldots, 7680.$$

Formulated as a large scale testing problem, the objective is to identify which of the θ_i's are non-zero while guarding against making too many false discoveries. The hypotheses are formulated as

$$H_i = \begin{cases} 0 & \text{if } \theta_i = 0 \\ 1 & \text{if } \theta_i \neq 0. \end{cases} \qquad i = 1, \ldots, 7680$$

If the i^{th} hypothesis is null, $Z_i \sim N(0, 1)$, and the two sided p-value, defined $p_i := 2(1 - \Phi(|Z_i|))$, is uniformly distributed on the unit interval. Running the SL procedure (11) on these two

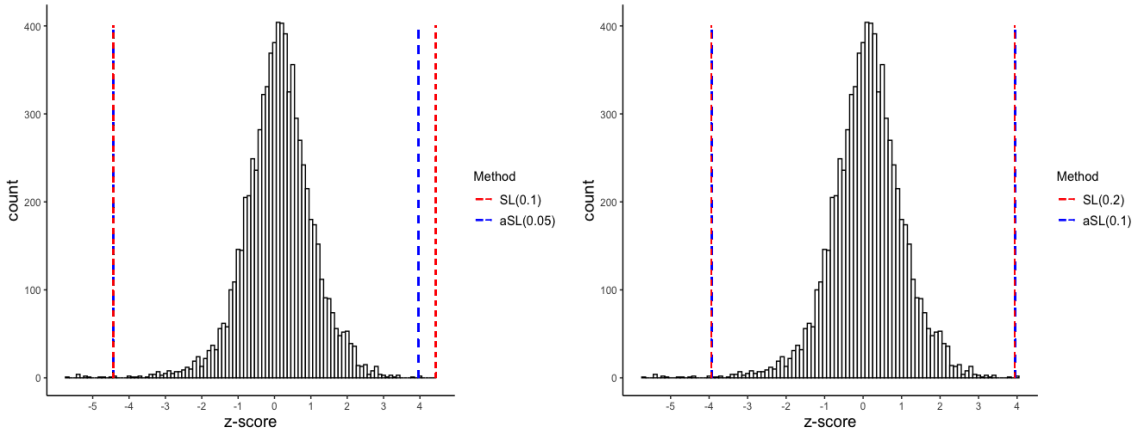

Figure 7: Dotted lines represent the rejection boundary for the SL procedures run on the HIV data. The asymmetric SL procedure (blue) is run at level $\alpha/2$ compared to the SL procedure (red), which is run at level $\alpha = 0.1$ (left) and $\alpha = 0.2$ (right).

sided p-values yields a rejection threshold $\hat{\tau}_\alpha$, which can be inverted to obtain the threshold on the scale of the original z-statistics,

$$\hat{z}_\alpha = \Phi^{-1}(1 - \hat{\tau}_\alpha/2).$$

For the HIV data, the threshold for running SL at level $\alpha = 0.2$ on the two sided p-values is $\hat{z}_\alpha = 3.94$. Note that this implies a symmetric threshold on the left -3.94 for the significantly negative test statistics.

The histogram of the z-scores is skewed, but the SL procedure run on the two sided p-values always gives a symmetric rejection region. To calibrate a separate threshold on each side of the origin, we can compute one-sided p-values and run the SL procedure separately on the left-tail p-values, $p_i = \Phi(Z_i)$, and the right-tail p-values, defined similarly. Equivalently, the rejection set $\mathcal{R}$ will be a union of two sets $\mathcal{R} = \mathcal{R}_- \cup \mathcal{R}_+$, where $\mathcal{R}_- := \{i \in [m] : i \leq R_-\}$ and $\mathcal{R}_+ := \{i \in [m] : i \geq R_+\}$ are defined by

$$R_- := \operatorname*{argmax}_{k=0,\ldots,m} \left\{ \frac{\alpha k}{m} - \Phi(Z_{(k)}) \right\}, \quad Z_{(0)} := -\infty. \tag{13}$$

$$R_+ := \operatorname*{argmin}_{k=1,\ldots,m+1} \left\{ \frac{\alpha k}{m} - \Phi(Z_{(k)}) \right\}, \quad Z_{(m+1)} := +\infty. \tag{14}$$

This yields a potentially asymmetric rejection region $[-\infty, Z_{(R_-)}] \cup [Z_{(R_+)}, +\infty]$, and it follows directly from Theorem 2.3 that the rate of false discoveries at the edges of the rejection set are controlled under mild assumptions[5],

$$\mathbb{P}(H_{(R_-)} = 0) = \mathbb{P}(H_{(R_+)} = 0) = \bar{\pi}_0 \alpha,$$

[5]Here, the convention $H_{(0)} := 1$ (resp. $H_{(m+1)} := 1$) is used to indicate the event where no rejections are made on the left (resp. right).

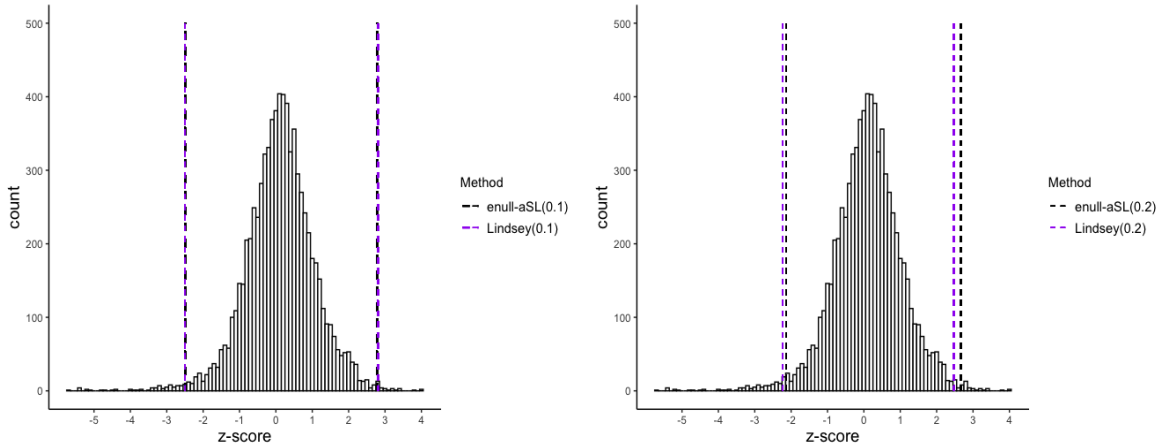

Figure 8: Dotted lines represent the rejection boundaries for the lfdr thresholding procedures run on the HIV data. The asymmetric SL procedure (black) and Lindsey's method (purple) are run at level $\alpha = 0.1$ (left) and $\alpha = 0.2$ (right).

since $\Phi(Z_i) \sim \text{Uniform}(0,1)$ when $H_i = 0$. Under the further assumption that $H_i = 1$ implies $Z_i \sim f_1$, a density for which the lfdr is unimodal about zero, it can be shown that

$$\text{RFDR}(\mathcal{R}_-) = \text{RFDR}(\mathcal{R}_+) = \bar{\pi}_0 \alpha.$$

Since the supremum over $\mathcal{R}$ in the definition of RFDR is bounded by the sum of the suprema taken separately, the above equalities imply

$$\text{RFDR}(\mathcal{R}) = \mathbb{E}\left[\sup_{J \in \mathcal{R}} \mathbb{P}(H_{(J)} = 0 \mid P_m)\right] \leq \text{RFDR}(\mathcal{R}_-) + \text{RFDR}(\mathcal{R}_+) = 2\bar{\pi}_0 \alpha. \tag{15}$$

We run both versions of the SL procedure on the HIV data and plot the results in Figure 7. The red dashed lines are equidistant from the origin, and represent the SL procedure run on the two sided p-values. The blue dashed lines designate the rejection thresholds of the asymmetric SL procedure defined by (13) and (14), which is run at half the level $\alpha \in \{0.1, 0.2\}$ at which the SL procedure was run. Compared to the SL procedure run at level $\alpha = 0.1$ on the two sided p-values, the asymmetric SL procedure, run at level $\alpha = 0.05$, yields a slightly more liberal rejection threshold on the right, as depicted in the left panel of Figure 7. When $\alpha = 0.2$ (and 0.1 for the asymmetric procedure), both procedures result in the same set of rejections.

The analysis presented above hinges on the assumption that under the null, the z-statistics are distributed $N(0,1)$. However, the histogram of the data isn't centered at zero, suggesting a violation of the assumptions that led to the theoretical (standard normal) null distribution. Efron (2012) proposed to estimate the mean and variance of the null distribution based on the central bulk of values in the histogram, using maximum likelihood estimates. For the HIV data, this leads to the empirical null,

$$\hat{f}_0 \sim N(0.12, 0.75^2), \quad \hat{\pi}_0 = 0.93. \tag{16}$$

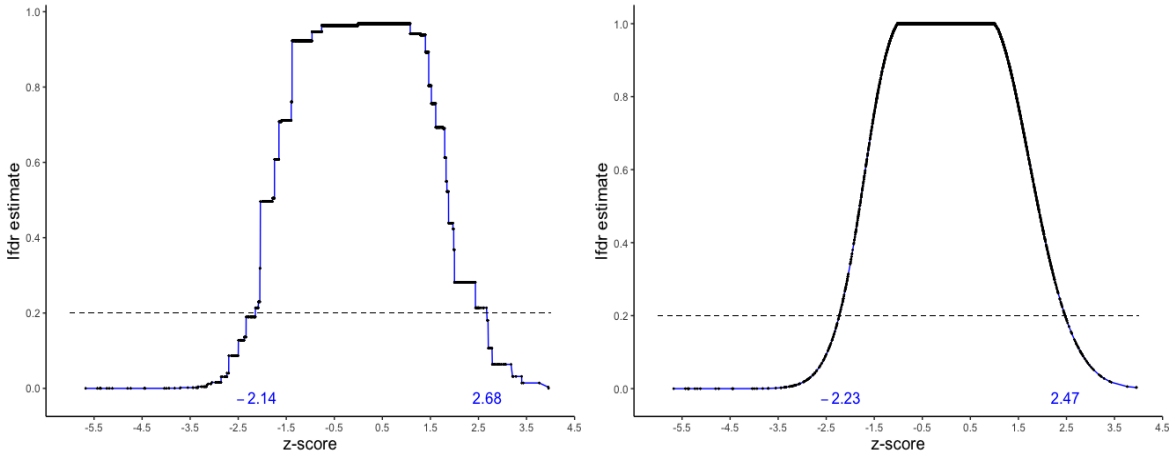

Figure 9: The Storey-adjusted Grenander estimate of lfdr for the HIV data is plotted on the left, and Lindsey's method on right. Black dots indicate the observations. The $\alpha = 0.2$ cut-offs on the z-score scale are written in blue, and are also plotted with vertical dashed lines on the histogram with the data on the right panel of Figure 8.

Replacing the theoretical null cdf Φ with the estimated null cdf $\hat{F}_0(z) := \Phi\left(\frac{z-0.12}{0.75}\right)$ in expressions (13) and (14), and running this procedure at level $\frac{1}{2} \cdot \frac{\alpha}{\hat{\pi}_0}$ controls the RFDR below α (see expression (15)). Efron (2012) also proposed to estimate the marginal density of the z-statistics using a method using Poisson regression due to Lindsey (Lindsey (1974a), Lindsey (1974b)). Together with the empirical null (16), this yields an estimate for the lfdr, which we have plotted in Figure 9. Thresholding this estimate at level $\alpha \in \{0.1, 0.2\}$ gives a rejection region, displayed in Figure 8 along with the rejection region of the empirical-null-modified asymmetric SL procedure run at the same level.

4 Discussion

We return to the question posed at the beginning of this paper: on average, of all the studies in this literature with p-values close to 0.001, for what fraction was the null hypothesis actually true? The answer is given by the local false discovery rate evaluated at $t = 0.001$. When $m = 1$, i.e. there is one fixed hypothesis to test, the formula (5) for the lfdr at any $t \in [0, 1]$ reduces to either 0 or 1, reflecting the difficulty in making posterior inferences based on a single observation. Not much can be done in this case without specifying a subjective prior probability on the truth status of the hypothesis. As Theorem 2.2 demonstrates, when there are many hypotheses to test, a frequentist interpretation can be given: lfdr is the local frequency of nulls in a small neighborhood of the sample space,

$$\text{lfdr}(t) = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)} \approx \frac{\#\{i : H_i = 0, p_i \approx t\}}{\#\{i : p_i \approx t\}}. \tag{17}$$

4.1 Calibration

Viewing each p-value as a summary of evidence for the presence of an effect, the lfdr converts this number to the correct scale for calibration in the sense of (17). We have argued in Section 3 for the nudge data that by reporting statistical significance on the scale of the lfdr (as opposed to the p-value) in this area of the psychology literature, researchers claiming a certain false positive probability will collectively achieve the rate they claim. For instance, a researcher who publishes a result with one sided p-value near 0.002 implicitly reports a corresponding lfdr estimate of $\text{lfdr}(0.002) \approx 20\%$, according to our estimates based on the meta-analysis data collected by Mertens et al. (2022a). We estimate that about 20% of researchers in this position will be making false discoveries. For studies with p-values ten times as small as this, we estimate fewer than 5% of them to be false discoveries (see Table 1).

The p-value is often misinterpreted as a posterior probability that the tested hypothesis is null (see e.g. Goodman (1999)). However, the numeric value of the observed p_i can differ substantially from the proportion of p-values near p_i corresponding to true null hypotheses, which we have argued is close to the local false discovery rate at p_i when there are many independent tests (Theorem 2.2). For a numerical illustration of this difference, see Section 2 of Sellke et al. (2001). In their paper, Sellke et al. (2001) also proposed a method for calibrating a p-value, which is to compute

$$B(p_i) = -ep_i \log(p_i),$$

for $p_i < 1/e \approx 0.368$, and interpret this number as a lower bound on the Bayes factor for $H_i = 0$ against $H_i = 1$. Multiplying this value by the prior odds gives a lower estimate on the posterior odds,

$$\frac{\text{lfdr}(p_i)}{1 - \text{lfdr}(p_i)} \geq B(p_i) \times \frac{\bar{\pi}_0}{1 - \bar{\pi}_0},$$

which can be converted to a lower estimate on $\text{lfdr}(p_i)$. Plugging in our estimate $\bar{\pi}_0 \approx 0.28$ from Section 3.1 gives a lower bound on the local false discovery rate at various p-value cut-offs. Four of these estimates are displayed in Table 2, corresponding to our suggested p-value cut-offs for the nudge literature.

one sided p-value cut-off	2.44×10^{-4}	6.25×10^{-4}	2.04×10^{-3}	3.70×10^{-3}
lfdr cut-off	0.05	0.10	0.20	0.30
Sellke et al. (2001) lower bound	0.045	0.09	0.17	0.23

Table 2: Suggested significance cut-offs for the nudge literature with corresponding lfdr estimates.

4.2 Towards RFDR-controlling procedures

We have also argued against assigning the same FDR summary statistic to two rejection sets where one has recognizably worse rejections than the other. For fixed configurations

of hypotheses, the RFDR of the SL procedure is controlled under weak assumptions on the densities of the p-values. While the proof of an exact finite-sample bound on the RFDR for the SL procedure is not as extensible as the martingale and leave-one-out proofs of FDR control for BH, our view is to prioritize controlling a multiple testing procedure's least promising rejection, perhaps only approximately, over designing procedures that achieve exact finite-sample control over the average quality of their rejections.

Exact guarantees like the ones in Theorem 2.3 and in Theorem 1 of Benjamini and Hochberg (1995) are useful to have in finite samples, where we don't know if we've done a good job estimating the signal distribution based on the few non-nulls distinguishable from the bulk of the data. However, the prioritization of exact FDR control guarantees has led attention away from the wider implications of the BH idea, namely that average case guarantees let bad bets in through the cracks. SL also satisfies an exact bound, but its true merit comes from the fact that it is consistent with what an oracle would do. An oracle would not choose to control FDR; they would minimize a combination of type 1 and 2 errors, and lfdr is the right tool for converting p-values to the appropriate scale for making this trade-off.

The RFDR criterion proposed in this paper characterizes the size of a multiple testing procedure in a coherent way, and while it treats the p-values exchangeably by conditioning only on their order statistics, there is room to extend the ideas in this paper to problems where additional distinguishing information is observed, or where the primary goal is to select a subset of variables for prediction of a response. Unfortunately, even in the simple setting discussed in this paper, the proof of Theorem 2.3 does not appear to generalize beyond the stated conditions, such as to discretely uniform nulls, or to p-values with dependence. Proving exact control for a procedure's RFDR may seem challenging in more structured settings, but it is still worthwhile to consider modifying multiple testing procedures with finite-sample FDR bounds (Barber and Candès (2015), Lei and Fithian (2018)) to instead target the RFDR, and we intend to further investigate this important direction in future work.

Acknowledgements

We thank Bradley Efron and Peter McCullagh for their insightful comments. J. A. Soloff was supported by NSF Grant DMS-2023505 and by the Office of Naval Research under the Vannevar Bush Fellowship. William Fithian was supported by the NSF DMS-1916220 and a Hellman Fellowship from Berkeley.

References

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs, *The Annals of Statistics* **43**(5): 2055–2085.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. et al. (2018). Redefine statistical significance, *Nature human behaviour* **2**(1): 6–10.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* **57**(1): 289–300.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1, Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American statistical association* **96**(456): 1151–1160.

Finner, H. and Roters, M. (2001). On the false discovery rate and expected type i errors, *Biometrical Journal* **43**(8): 985–1005.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy, *Annals of internal medicine* **130**(12): 995–1004.

Greenwald, A., Gonzalez, R., Harris, R. J. and Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated?, *Psychophysiology* **33**(2): 175–183.

Grenander, U. (1956). On the theory of mortality measurement: Part II, *Scandinavian Actuarial Journal* **1956**(2): 125–153.

Hoeffding, W. (1956). On the distribution of the number of successes in independent trials, *The Annals of Mathematical Statistics* pp. 713–721.

Hung, K. and Fithian, W. (2020). Statistical methods for replicability assessment, *The Annals of Applied Statistics* **14**(3): 1063–1087.

Lei, L. and Fithian, W. (2018). AdaPT: An interactive procedure for multiple testing with side information, *Journal of the Royal statistical society: series B (Statistical Methodology)* **80**: 649–679.

Lindsey, J. (1974a). Comparison of probability distributions, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(1): 38–47.

Lindsey, J. (1974b). Construction and comparison of statistical models, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(3): 418–425.

Loeb, K. L., Radnitz, C., Keller, K., Schwartz, M. B., Marcus, S., Pierson, R. N., Shannon, M. and DeLaurentis, D. (2017). The application of defaults to optimize parents' health-based choices for children, *Appetite* **113**: 368–375.

Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J. and Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias, *Proceedings of the National Academy of Sciences* **119**(31): e2200300119.

Mertens, S., Herberz, M., Hahnel, U. J. and Brosch, T. (2022a). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains, *Proceedings of the National Academy of Sciences* **119**(1): e2107346118.

Mertens, S., Herberz, M., Hahnel, U. J. and Brosch, T. (2022b). Reply to maier et al., szaszi et al., and bakdash and marusich: The present and future of choice architecture research, *Proceedings of the National Academy of Sciences* **119**(31): e2202928119.

Peto, R. (1996). Five years of tamoxifen–or more?, *Journal of the National Cancer Institute* **88**(24): 1791–1793.

Sellke, T., Bayarri, M. and Berger, J. O. (2001). Calibration of p-values for testing precise null hypotheses, *The American Statistician* **55**(1): 62–71.

Soloff, J. A. (2020). ebpy: Nonparametric empirical Bayes in Python, https://github.com/jake-soloff/ebpy.

Soloff, J. A., Xiang, D. and Fithian, W. (2022). The edge of discovery: Controlling the local false discovery rate at the margin.

Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(3): 479–498.

Strimmer, K. (2008). A unified approach to false discovery rate estimation, *BMC bioinformatics* **9**(1): 1–14.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control, *Journal of the American Statistical Association* **102**(479): 901–912.

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S. and Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions, *Proceedings of the National Academy of Sciences* **119**(31): e2200732119.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*, Penguin.

Yekutieli, D. and Weinstein, A. (2019). Hierarchical bayes modeling for large-scale inference, *arXiv preprint arXiv:1908.08444* .

A Proofs

Proof of Theorem 2.1. For the first part, the lfdr is a limit of posterior probabilities

$$\begin{aligned}\text{lfdr}(t) &= \lim_{\varepsilon\to 0} \mathbb{P}(H_J = 0 \mid |p_J - t| \le \varepsilon \text{ for some } J \in [m]) \\ &= \lim_{\varepsilon\to 0} \frac{\mathbb{P}(H_J = 0, |p_J - t| \le \varepsilon \text{ for some } J \in [m])}{\mathbb{P}(|p_J - t| \le \varepsilon \text{ for some } J \in [m])}. \end{aligned} \tag{18}$$

The event in the numerator can be written as a union,

$$\{H_J = 0, |p_J - t| \le \varepsilon \text{ for some } J \in [m]\} = \bigcup_{j:H_j=0} \{|p_j - t| \le \varepsilon\}.$$

By independence,

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{j:H_j=0} \{|p_j - t| \le \varepsilon\}\Big) &= 1 - \mathbb{P}\Big(\bigcap_{j:H_j=0} \{|p_j - t| > \varepsilon\}\Big) \\ &= 1 - \prod_{j:H_j=0} \mathbb{P}\big(|p_j - t| > \varepsilon\big) \\ &= 1 - \prod_{j:H_j=0} \Big(1 - \Big(F^{(j)}(t+\varepsilon) - F^{(j)}(t-\varepsilon)\Big)\Big) \\ &= 1 - \prod_{j:H_j=0} \Big(1 - 2\varepsilon f^{(j)}(\xi_j)\Big),\end{aligned}$$

for some $\xi_1, \dots, \xi_m \in [t-\varepsilon, t+\varepsilon]$ by the mean value theorem. Now since $\varepsilon > 0$ can be arbitrarily small and each $f^{(j)}$ is continuous,

$$\prod_{j:H_j=0} \Big(1 - 2\varepsilon f^{(j)}(\xi_j)\Big) \sim \exp\big(-2m_0\varepsilon \bar{f}_0(t)\big) \quad \text{as } \varepsilon \to 0,$$

where $f_0(t) = \frac{1}{m_0}\sum_{j:H_j=0} f^{(j)}(t)$ is the average null density. The above implies

$$\mathbb{P}\Big(\bigcup_{j:H_j=0} \{|p_j - t| \le \varepsilon\}\Big) \sim 2m_0\varepsilon \bar{f}_0(t) \quad \text{as } \varepsilon \to 0.$$

An identical argument will show

$$\mathbb{P}\Big(\bigcup_{j=1}^{m} \{|p_j - t| \le \varepsilon\}\Big) \sim 2m\varepsilon \bar{f}(t) \quad \text{as } \varepsilon \to 0,$$

where $f(t) = \frac{1}{m}\sum_{j=1}^{m} f^{(j)}(t)$ is the average density. Dividing these two expressions gives

$$\lim_{\varepsilon\to 0} \frac{\mathbb{P}(H_J = 0, |p_J - t| \le \varepsilon \text{ for some } J \in [m])}{\mathbb{P}(|p_J - t| \le \varepsilon \text{ for some } J \in [m])} = \lim_{\varepsilon\to 0} \frac{2m_0\varepsilon \bar{f}_0(t)}{2m\varepsilon f(t)} = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)}.$$

Now suppose that the test statistics are generated from the independently from the two-groups model (1). The lfdr is a limit of posterior probabilities,

$$\text{lfdr}(t) = \lim_{\varepsilon \to 0} \frac{\mathbb{P}\left(\bigcup_{j=1}^m \{H_j = 0, |p_j - t| \le \varepsilon\}\right)}{\mathbb{P}\left(\bigcup_{j=1}^m \{|p_j - t| \le \varepsilon\}\right)}. \tag{19}$$

Since the pairs (H_j, p_j) are independent across $j = 1, \dots, m$, the numerator is equal to

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{j=1}^m \{H_j = 0, |p_j - t| \le \varepsilon\}\right) &= 1 - \prod_{j=1}^m \left(1 - \mathbb{P}(H_j = 0, |p_j - t| \le \varepsilon)\right) \\
&= 1 - \prod_{i=1}^m \left(1 - \pi_0(F_0(t+\varepsilon) - F_0(t-\varepsilon))\right) \\
&= 1 - \left(1 - \pi_0(F_0(t+\varepsilon) - F_0(t-\varepsilon))\right)^m \\
&= 1 - \left(1 - \pi_0 \cdot 2\varepsilon f_0(\xi)\right))^m
\end{aligned}$$

for some $\xi \in (t - \varepsilon, t + \varepsilon)$ by the mean value theorem. As $\varepsilon \to 0$,

$$\left(1 - \pi_0 \cdot 2\varepsilon f_0(\xi)\right))^m \sim \exp\left(-m\pi_0 \cdot 2\varepsilon f_0(\xi)\right)) \sim 1 - 2m\pi_0 \varepsilon f_0(t)$$

since f_0 is continuous and $\xi \to t$ as $\varepsilon \to 0$. It follows that

$$\mathbb{P}\left(\bigcup_{j=1}^m \{H_j = 0, |p_j - t| \le \varepsilon\}\right) \sim 2m\pi_0 \varepsilon f_0(t) \quad \text{as } \varepsilon \to 0.$$

An identical argument shows

$$\mathbb{P}\left(\bigcup_{j=1}^m \{|p_j - t| \le \varepsilon\}\right) \sim 2m\varepsilon f(t),$$

which implies that the ratio (19) tends to $\frac{\pi_0 f_0(t)}{f(t)}$, as desired.

□

Proof of Lemma 2.1. Since S is a function of F_m,

$$\begin{aligned}
\text{RFDR}(\mathcal{R}) &= \mathbb{E}\left[\sup_{S \subset \mathcal{R}} \frac{1}{|S|} \sum_{i \in S} \mathbb{P}\left(H_{(i)} = 0 \mid F_m\right)\right] \\
&\le \mathbb{E}\left[\sup_{S \subset \mathcal{R}} \frac{1}{|S|} \cdot |S| \cdot \sup_{J \in \mathcal{R}} \mathbb{P}\left(H_{(J)} = 0 \mid F_m\right)\right] = \mathbb{E}\left[\sup_{J \in \mathcal{R}} \mathbb{P}\left(H_{(J)} = 0 \mid F_m\right)\right].
\end{aligned}$$

Conversely,

$$\text{RFDR}(\mathcal{R}) \ge \mathbb{E}\left[\sup_{S \subset \mathcal{R}: |S| = 1} \mathbb{E}\left(\frac{|\mathcal{H}_0 \cap S|}{|S|} \mid F_m\right)\right] = \mathbb{E}\left[\sup_{J \in \mathcal{R}} \mathbb{P}\left(H_{(J)} = 0 \mid F_m\right)\right].$$

The right hand side of the above is equal to the max-lfdr criterion when the data are drawn iid from the two-groups model (1),

$$\begin{aligned}\mathbb{P}(H_{(J)} = 0 \mid F_m) &= m \cdot \mathbb{P}(p_{(J)} = p_1, H_1 = 0 \mid F_m) \\ &= m \cdot \mathbb{P}(H_1 = 0 \mid p_{(J)} = p_1, F_m) \cdot \mathbb{P}(p_{(J)} = p_1 \mid F_m) \\ &= m \cdot \mathrm{lfdr}(p_{(J)}) \cdot \frac{1}{m},\end{aligned}$$

where the first equality is by exchangeability, the second is by Bayes rule, and the third follows from independence between H_1 and F_m given the value of p_1.

□

Proof of Theorem 2.2. The conditional expectation is equal to

$$\mathrm{pFDR}([t-\varepsilon, t]) = \frac{\mathbb{E}(\mathrm{FDP}([t-\varepsilon, t]) \cdot 1_{\{p_i \in [t-\varepsilon,t] \text{ for some } i\}})}{\mathbb{P}(p_i \in [t-\varepsilon, t] \text{ for some } i)}$$

Observe that

$$\begin{aligned}\mathrm{FDP}([t-\varepsilon, t]) \cdot 1_{\{p_i \in [t-\varepsilon,t] \text{ for some } i\}} &\le 1_{\{p_i \in [t-\varepsilon,t] \text{ and } H_i = 0 \text{ for some } i\}} \\ \mathrm{FDP}([t-\varepsilon, t]) \cdot 1_{\{p_i \in [t-\varepsilon,t] \text{ for some } i\}} &\ge 1_{\{p_i \in [t-\varepsilon,t] \text{ and } H_i = 0 \text{ for exactly one } i\}}.\end{aligned}$$

It follows from continuity of the densities that the numerator and denominator are

$$\begin{aligned}\mathbb{E}(\mathrm{FDP}([t-\varepsilon, t]) \cdot 1_{\{p_i \in [t-\varepsilon,t] \text{ for some } i\}}) &\sim \varepsilon m_0 \bar{f}_0(t) \\ \mathbb{P}(p_i \in [t-\varepsilon, t] \text{ for some } i) &\sim \varepsilon m \bar{f}(t),\end{aligned}$$

so their ratio tends to $\mathrm{lfdr}(t)$ as $\varepsilon \to 0$. The same argument shows that the mFDR tends to the lfdr as $\varepsilon \to 0$ when all p-values are continuously distributed. When there is an atom at t, the mFDR tends to

$$\mathrm{mFDR}([t-\varepsilon, t]) = \frac{\mathbb{E}\left(\#\{i : H_i = 0, p_i \in [t-\varepsilon, t]\}\right)}{\mathbb{E}\left(\#\{i : p_i \in [t-\varepsilon, t]\}\right)} \to \frac{m_0 \bar{P}_0(\{t\})}{m \bar{P}(\{t\})} \quad \text{as } \varepsilon \to 0.$$

□

Proof of Theorem 2.3. Suppose without loss of generality that $H_m = 0$. Then since the nulls are exchangeable, this probability is

$$\mathbb{P}(H_{(R)} = 0) = m \bar{\pi}_0 \mathbb{P}(p_{(R)} = p_m).$$

Let $q_{(1)} \le \cdots \le q_{(m-1)}$ denote the order statistics of $p_1, \ldots, p_{m-1}$, and note that p_m achieves the maximum in (11) as the $(k+1)^{\text{th}}$ order statistic if $q_{(k)} < p_m < q_{(k+1)}$ and

$$\frac{\alpha(k+1)}{m} - p_m > \left[\max_{j=k+1,\ldots,m-1} \left\{\frac{\alpha(j+1)}{m} - q_{(j)}\right\}\right] \vee \left[\max_{j=0,\ldots,k} \left\{\frac{\alpha j}{m} - q_{(j)}\right\}\right].$$

Rearranging the above inequalities gives the range in which p_m achieves the maximum as the $(k+1)^{\text{th}}$ order statistic,

$$q_{(k)} < p_m < \frac{\alpha k}{m} - \left[\max_{j=k+1,\dots,m-1}\left\{\frac{\alpha j}{m} - q_{(j)}\right\}\right] \vee \left[\max_{j=0,\dots,k}\left\{\frac{\alpha j}{m} - q_{(j)}\right\} - \frac{\alpha}{m}\right].$$

This range is non-empty when $\Delta_k := \frac{\alpha k}{m} - q_{(k)}$ exceeds each of $\Delta_{k+1}, \dots, \Delta_{m-1}$ as well as $\max_{j=0,\dots,m-1} \Delta_j - \frac{\alpha}{m}$, and has length $\Delta_k - (\max_{j=k+1,\dots,m-1} \Delta_j) \vee (\max_{j=0,\dots,k} \Delta_j - \frac{\alpha}{m})$. The sum of lengths of the non-empty ranges is telescoping and equal to $\frac{\alpha}{m}$, as illustrated in Figure 10.

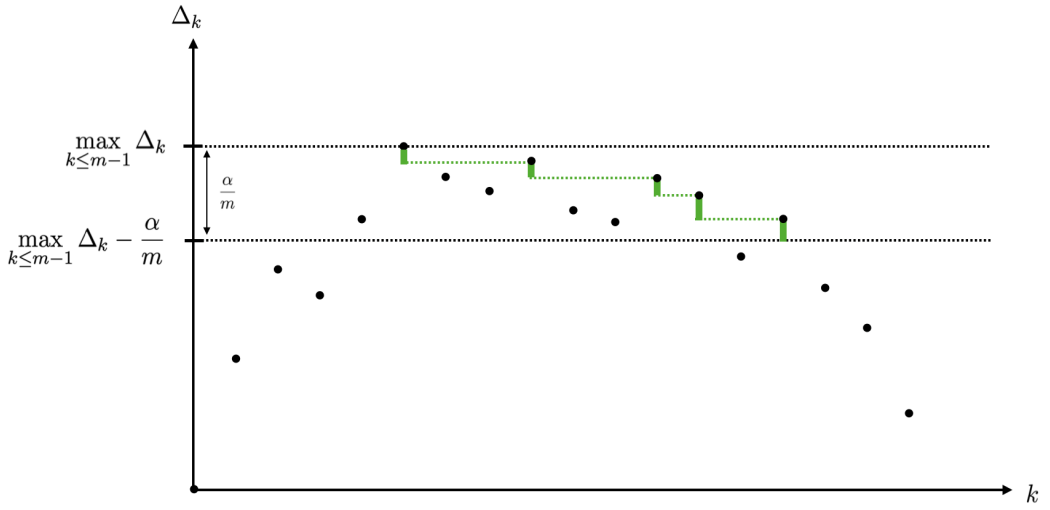

Figure 10: The length of each range in which p_m achieves the maximum in (11) is indicated by the green highlight, and the sum of these lengths is $\frac{\alpha}{m}$.

□

Proof of Theorem 2.4. By Lemma A.5 that for m large enough, we have with probability $\geq 1 - \delta$ that

$$\hat{\tau}_\alpha - \tau^*_\alpha \leq \varepsilon := C' m^{-1/3} \log(2m/\delta), \tag{20}$$

for some constant $C' > 0$ depending on α, J and δ. Since f is decreasing on $(0, \tau^*_\alpha + \alpha)$ and has derivative greater than $-J^{-1}$ over the interval $(\tau^*_\alpha, \tau^*_\alpha + \varepsilon)$, the above inequality implies

$$f(\hat{\tau}_\alpha) \geq f(\tau^*_\alpha) - J^{-1}\varepsilon = \alpha^{-1} - J^{-1}\varepsilon.$$

It follows that

$$\mathrm{lfdr}(\hat{\tau}_\alpha) = \frac{\pi_0}{f(\hat{\tau}_\alpha)} \leq \pi_0 \alpha + C m^{-1/3} \log(m/\delta),$$

for another constant $C > 0$ depending on α, J and δ.

□

Theorem A.1. *Let $H_1, \dots, H_m \in \{0,1\}$ be fixed and suppose $Z_1, \dots, Z_m$ are independent where $H_i = 0$ implies $Z_i \sim f_0$, a continuous and symmetric density with $F_0^{-1}(1/2) = 0$. Let $H_{(0)} \coloneqq 1$ and $H_{(m+1)} \coloneqq 1$. Then*

$$\mathbb{P}(H_{(R_-)} = 0) = \mathbb{P}(H_{(R_+)} = 0) = \bar{\pi}_0 \alpha. \tag{21}$$

Further, if $H_i = 1$ implies $Z_i \sim f_1$, a continuous density for which lfdr *is unimodal about zero and $\alpha \in (0, 1/2)$, then*

$$\mathrm{RFDR}(\mathcal{R}_-) = \mathrm{RFDR}(\mathcal{R}_+) = \bar{\pi}_0 \alpha \quad \textit{and} \quad \mathrm{RFDR}(\mathcal{R}) \leq 2\bar{\pi}_0 \alpha. \tag{22}$$

Proof. For (21), define the p-values $p_i \coloneqq F_0(Z_i)$ so that $p_i \sim \mathrm{Uniform}(0,1)$ when $H_i = 0$. Then since the nulls are exchangeable,

$$\mathbb{P}(H_{(R_-)} = 0) = m\bar{\pi}_0 \mathbb{P}(p_{(R_-)} = p_m),$$

where we have assumed $H_m = 0$ without loss of generality[6]. By Lemma 2 of Soloff et al. (2022), the right hand side above is

$$m\bar{\pi}_0 \mathbb{P}(p_{(R_-)} = p_m) = m\bar{\pi}_0 \mathbb{E}(\mathbb{P}(p_{(R_-)} = p_m \mid p_1, \dots, p_{m-1})) = \bar{\pi}_0 \alpha.$$

Lemma A.1 (Lemma 2 of Soloff et al., 2022)**.** *Fix $p_1, \dots, p_{m-1} \in [0,1]$, and suppose $p_m \sim$ Uniform$(0,1)$. Then $\mathbb{P}(p_{(R_-)} = p_m) = \frac{\alpha}{m}$.*

Since multiplication by -1 reverses the ranking of the order statistics, $H_{(R_+)}$ is the hypothesis corresponding to the $m - \widetilde{R}_-$ largest p-value, where $\widetilde{R}_-$ is the index (13) computed on $-Z_1, \dots, -Z_m$. Since F_0 is symmetric, $H_i = 0$ implies $-Z_i \sim F_0$, and it follows from what we have just shown that

$$\mathbb{P}(H_{(R_+)} = 0) = \mathbb{P}(H_{(\widetilde{R}_-)} = 0) = \bar{\pi}_0 \alpha,$$

completing the proof of (21).

For the next claim, note that we always have $F_0(Z_{(R_-)}) \leq \frac{\alpha}{2}$, because otherwise the objective in (13) would be smaller at $k = R_-$ than at $k = 0$. Since the median of the null distribution is 0 and $\alpha < 1$, this implies that $Z_{(R_-)}$ always falls to the left of zero. We claim that for any $J \in \mathcal{R}_-$,

$$\mathbb{P}(H_{(J)} = 0 \mid F_m) \leq \mathbb{P}(H_{(R_-)} = 0 \mid F_m) \tag{23}$$

where F_m denotes the ecdf of $\{Z_1, \dots, Z_m\}$. By exchangeability of the nulls,

$$\mathbb{P}(H_{(J)} = 0 \mid F_m) = \sum_{i : H_i = 0} \mathbb{P}(Z_{(J)} = Z_i) = m\pi_0 \mathbb{P}(Z_{(J)} = Z_m \mid F_m),$$

where we have assumed $H_m = 0$ without loss of generality. By Lemma A.2, $J \leq R_-$ implies that $\mathbb{P}(Z_{(J)} = Z_m \mid F_m) \leq \mathbb{P}(Z_{(R_-)} = Z_m \mid F_m)$ since the unimodal assumption implies

[6] if $\pi_0 = 0$, then all sides of equation (21) are 0

f_0/f_1 is increasing on $(-\infty, 0]$. Plugging this bound in the above and reversing the chain of equalities above gives (23), which implies

$$\mathrm{RFDR}(\mathcal{R}_-) = \mathbb{E}\left[\sup_{J\in\mathcal{R}_-} \mathbb{P}(H_{(J)} = 0 \mid F_m)\right] = \mathbb{E}\left[\mathbb{P}(H_{(R_-)} = 0 \mid F_m)\right] = \mathbb{P}(H_{(R_-)} = 0) = \frac{\pi_0\alpha}{2}.$$

An identical argument shows that any $J \geq R_+$ has

$$\mathbb{P}(H_{(J)} = 0 \mid F_m) \leq \mathbb{P}(H_{(R_+)} = 0 \mid F_m), \tag{24}$$

which implies the analogous result for $\mathcal{R}_+$.

Finally, since the maximum is bounded by the sum, (23) and (24) imply

$$\mathrm{RFDR}(\mathcal{R}) = \mathbb{E}\left[\sup_{J\in\mathcal{R}} \mathbb{P}(H_{(J)} = 0 \mid F_m)\right] \leq \mathbb{E}\left[\mathbb{P}(H_{(R_-)} = 0 \mid F_m) + \mathbb{P}(H_{(R_+)} = 0 \mid F_m)\right].$$

By the tower property, the right hand side is equal to $\mathbb{P}(H_{(R_-)} = 0)+\mathbb{P}(H_{(R_+)} = 0) = \pi_0\alpha$. □

A.1 Technical Lemmas

Lemma A.2. *Suppose $Z_i \sim f^{(i)}$ independently for $i = 1, \ldots, m$ where $H_i = 0$ implies $f^{(i)} = f_0$ and $H_i = 1$ implies $f^{(i)} = f_1$, for two densities f_0, f_1 whose ratio f_0/f_1 is increasing on $(-\infty, 0]$ and decreasing on $[0, \infty)$, and $F_0^{-1}(1/2) = 0$. If $H_m = 0$, then for any $J \leq R_-$ defined by (13),*

$$\mathbb{P}(Z_{(J)} = Z_m \mid F_m) \leq \mathbb{P}(Z_{(R_-)} = Z_m \mid F_m).$$

Similarly, if $J \geq R_+$ defined in (14),

$$\mathbb{P}(Z_{(J)} = Z_m \mid F_m) \leq \mathbb{P}(Z_{(R_+)} = Z_m \mid F_m).$$

Proof. We show the first inequality since the second follows by an identical argument. For $j \leq m$, let $[j] \in \{1, \ldots, m\}$ denote the index for which $Z_{[j]} = Z_{(j)}$. By the tower property, it suffices to show

$$\mathbb{P}(Z_{(J)} = Z_m \mid \mathcal{G}_-) \leq \mathbb{P}(Z_{(R_-)} = Z_m \mid \mathcal{G}_-), \tag{25}$$

where $\mathcal{G}_-$ is the sigma field containing knowledge of the order statistics and all indices besides $[J], [R_-]$,

$$\mathcal{G}_- := \sigma(F_m, \{[1], \ldots, [m]\}\backslash\{[J], [R_-]\}).$$

If $m \notin \{[J], [R_-]\}$, then both sides of (25) are zero and the inequality is trivially satisfied. Otherwise, $\{[J], [R_-]\} = \{m, k\}$ for some $1 \leq k < m$ and we have

$$\mathbb{P}(Z_{(J)} = Z_m \mid \mathcal{G}_-) \propto f^{(m)}(Z_{(J)}) f^{(k)}(Z_{(R_-)}) \prod_{i\notin\{m,k\}} f^{(i)}(Z_i)$$

$$\mathbb{P}(Z_{(R_-)} = Z_m \mid \mathcal{G}_-) \propto f^{(m)}(Z_{(R_-)}) f^{(k)}(Z_{(J)}) \prod_{i\notin\{m,k\}} f^{(i)}(Z_i).$$

Since both probabilities have the same normalizing constant, and $H_m = 0 \Rightarrow f^{(m)} = f_0$, it follows that the ratio of posterior probabilities is

$$\frac{\mathbb{P}(Z_{(J)} = Z_m \mid \mathcal{G}_-)}{\mathbb{P}(Z_{(R_-)} = Z_m \mid \mathcal{G}_-)} = \begin{cases} 1 & \text{if } H_k = 0 \\ \frac{f_0}{f_1}(Z_{(J)}) \cdot \frac{f_1}{f_0}(Z_{(R_-)}) & \text{if } H_k = 1. \end{cases}$$

Since f_0/f_1 is increasing on $(-\infty, 0]$, and $Z_{(J)} \le Z_{(R_-)} \le 0$ almost surely, we have shown

$$\frac{\mathbb{P}(Z_{(J)} = Z_m \mid \mathcal{G}_-)}{\mathbb{P}(Z_{(R_l)} = Z_m \mid \mathcal{G}_-)} \le 1.$$

Multiplying both sides by $\mathbb{P}(Z_{(R_-)} = Z_m \mid \mathcal{G}_-)$, the proof is complete by the tower property. □

Lemma A.3. *Suppose $p_i \sim f^{(i)}$ independently for $i = 1, \dots, m$, and each density $f^{(i)}$ is non-increasing. Let R_i denote the rank of p_i among $p_1, \dots, p_m$, where $R_i = 1$ when $p_i = \min_{j=1,\dots,m} p_j$. Then for any $p_i \sim$ Uniform$(0,1)$, the conditional pmf of R_i is non-decreasing,*

$$\mathbb{P}(R_i = j \mid F_m) \le \mathbb{P}(R_i = j+1 \mid F_m), \quad j = 1, \dots, m-1$$

where F_m is the ecdf of $p_1, \dots, p_m$.

Proof. For $j \le m$, let $[j] \in \{1, \dots, m\}$ denote the index for which $R_{[j]} = j$, or equivalently, $p_{[j]} = p_{(j)}$. By the tower property, it suffices to show

$$\mathbb{P}(R_i = j \mid \mathcal{G}_j) \le \mathbb{P}(R_i = j+1 \mid \mathcal{G}_j), \tag{26}$$

where $\mathcal{G}_j$ is the sigma field containing the order statistics and all anti-ranks besides $[j], [j+1]$,

$$\mathcal{G}_j := \sigma(F_m, \{[1], \dots, [m]\} \backslash \{[j], [j+1]\}).$$

If $i \notin \{[j], [j+1]\}$, then both sides of (26) are equal to zero and the inequality is satisfied. Now suppose $\{[j], [j+1]\} = \{i, k\}$ for some $1 \le k \le m$. Then

$$\mathbb{P}(R_i = j \mid \mathcal{G}_j) \propto f^{(i)}(p_{(j)}) f^{(k)}(p_{(j+1)}) \prod_{\ell \notin \{i,k\}} f^{(\ell)}(p_\ell)$$

$$\mathbb{P}(R_i = j+1 \mid \mathcal{G}_j) \propto f^{(i)}(p_{(j+1)}) f^{(k)}(p_{(j)}) \prod_{\ell \notin \{i,k\}} f^{(\ell)}(p_\ell),$$

where the constant of proportionality is the same in both cases,

$$C := \frac{1}{\left(f^{(i)}(p_{(j)}) f^{(k)}(p_{(j+1)}) + f^{(i)}(p_{(j+1)}) f^{(k)}(p_{(j)})\right) \prod_{\ell \notin \{i,k\}} f^{(\ell)}(p_\ell)}.$$

Since $f^{(i)} = 1_{[0,1]}$, the ratio of conditional probabilities is

$$\frac{\mathbb{P}(R_i = j \mid \mathcal{G}_j)}{\mathbb{P}(R_i = j+1 \mid \mathcal{G}_j)} = \frac{f^{(k)}(p_{(j+1)})}{f^{(k)}(p_{(j)})} \le 1,$$

since $f^{(k)}$ is non-increasing. Multiplying both sides of the above inequality by $\mathbb{P}(R_i = j+1 \mid \mathcal{G}_j)$ and taking expectation with respect to the conditional distribution given F_m completes the proof. □

Lemma A.4. *Let $\tau_\alpha^* < 1/2$ be a solution to $f(\tau_\alpha^*) = \alpha^{-1}$ and $\hat{\tau}_\alpha$ is the rejection threshold of the SL(α) procedure (11). If $\hat{\tau}_\alpha > \tau_\alpha^* + \varepsilon$, then there exists an index $k \geq 1$ for which*

$$p_{(i^*+k)} \leq \tau_\alpha^* + \frac{\alpha k}{m} \quad \text{and} \quad k > \frac{m\varepsilon}{\alpha},$$

where $i^* := \max\{i : p_{(i)} \leq \tau_\alpha^*\}$ and $i^* = 0$ if no such i exists.

Proof. Let $\hat{k}$ be the index for which $\hat{\tau}_\alpha = p_{(i^*+\hat{k})}$. The first inequality can be written

$$\frac{i^* + \hat{k}}{m} - \frac{i^*}{m} - \alpha^{-1}(p_{(i^*+\hat{k})} - \tau_\alpha^*) \geq 0,$$

which holds because $F_m(t) - \alpha^{-1}F_0(t)$ is maximized at $t = p_{(i^*+\hat{k})}$. Since $\hat{\tau}_\alpha > \tau_\alpha^* + \varepsilon$, the above inequality implies $\hat{k} > \frac{m\varepsilon}{\alpha}$.

□

Lemma A.5. *Let τ_α^* and $\hat{\tau}_\alpha$ be defined as in Lemma A.4, let $\delta > 0$ and suppose $\bar{f}$ is decreasing on $[\tau_\alpha^*, \tau_\alpha^* + \alpha]$ and that there exists some $J > 0$ for which $J \leq |\bar{f}'(t)| \leq J^{-1}$ for all t with $|t - \tau_\alpha^*| \leq \varepsilon$, where $\varepsilon := \left(\frac{24}{\alpha L^2}\right)^{1/3} m^{-1/3} \log(2m/\delta)$. Then*[7]

$$\mathbb{P}(\hat{\tau}_\alpha > \tau_\alpha^* + \varepsilon) \leq \delta,$$

for any $m \geq C(\alpha, J, \delta)$, a constant depending only on α, J and δ.

Proof of Lemma A.5. Applying Lemma A.4 with ε defined as above, we have

$$\begin{aligned}\mathbb{P}(\hat{\tau}_\alpha > \tau_\alpha^* + \varepsilon) &\leq \sum_{k > \frac{m\varepsilon}{\alpha}} \mathbb{P}\left(p_{(i^*+k)} \leq \tau_\alpha^* + \frac{\alpha k}{m}\right) \\ &= \sum_{\frac{m\varepsilon}{\alpha} < k \leq \frac{m\varepsilon \log m}{\alpha}} \mathbb{P}\left(N_k \geq k\right) + \sum_{k > \frac{m\varepsilon \log m}{\alpha}} \mathbb{P}\left(N_k \geq k\right), \end{aligned} \tag{27}$$

where i^* is defined in Lemma A.4, and N_k is the number of p-values between τ_α^* and $\tau_\alpha^* + \frac{\alpha k}{m}$, distributed Generalized-Binomial with sample size m and average success probability $\bar{F}(\tau_\alpha^* + \alpha k/m) - \bar{F}(\tau_\alpha^*)$,

$$N_k = \sum_{j=1}^{m} 1_{\{p_j \in (\tau_\alpha^*, \tau_\alpha^* + \alpha k/m)\}} \Rightarrow \mathbb{E}N_k = m(\bar{F}(\tau_\alpha^* + \alpha k/m) - \bar{F}(\tau_\alpha^*)),$$

where $\bar{F} := \frac{1}{m}\sum_{i=1}^{m} F^{(i)}$ is the average cdf of the p-values. Note that since $\bar{F}' = \bar{f}$, we have by the mean value theorem that

$$\mathbb{E}N_k = m(\bar{F}(\tau_\alpha^* + \alpha k/m) - \bar{F}(\tau_\alpha^*)) = m\bar{f}(\xi) \cdot \frac{\alpha k}{m},$$

[7] In Appendix B, we show a simulation in which the average density is decreasing $(0, \tau_\alpha + \alpha)$ (but not on the entire unit interval) where the estimate $\hat{\tau}_\alpha$ tends to τ_α^* at the $m^{-1/3}$ rate. Namely, the assumption that $\bar{f}$ be decreasing over the entire unit interval is not necessary for the threshold based on the Grenander estimator to be accurate in large samples.

for some $\xi \in (\tau_\alpha^*, \tau_\alpha^* + \alpha k/m)$. By the monotonicity assumption, $\bar{f}(\xi) \le \bar{f}(\tau_\alpha^*) = \alpha^{-1}$ implies we have $\mathbb{E}N_k \le k$. Consider the corresponding Binomial random variable, $\widetilde{N}_k \sim \text{Binomial}(m, \bar{F}(\tau_\alpha^* + \alpha k/m) - \bar{F}(\tau_\alpha^*))$. Since $\mathbb{E}\widetilde{N}_k = \mathbb{E}N_k \le k$, it follows from Theorem 5 in Hoeffding (1956) that

$$\mathbb{P}\left(N_k \ge k\right) \le \mathbb{P}\left(\widetilde{N}_k \ge k\right) = \mathbb{P}\left(\widetilde{N}_k \ge \mathbb{E}\widetilde{N}_k \cdot \frac{k}{\mathbb{E}\widetilde{N}_k}\right).$$

To bound the probability on the right hand side, we use the following estimates on the expectation $\mathbb{E}\widetilde{N}_k = m(\bar{F}(\tau_\alpha^* + \alpha k/m) - \bar{F}(\tau_\alpha^*))$,

$$\mathbb{E}\widetilde{N}_k \le k + \frac{Jm\varepsilon^2}{2} - J\alpha k\varepsilon \tag{28}$$

$$\mathbb{E}\widetilde{N}_k \ge \frac{m\varepsilon}{2\alpha}. \tag{29}$$

Before proving inequalities (28) and (29), we show how they can be used to complete the proof. When $\frac{m\varepsilon}{\alpha} < k \le \frac{m\varepsilon \log m}{\alpha}$, the upper bound (28) gives $\mathbb{E}\widetilde{N}_k \le k + \frac{Jm\varepsilon^2}{2} - J\alpha\varepsilon \cdot \frac{m\varepsilon}{\alpha}$, which implies

$$\mathbb{P}\left(\widetilde{N}_k \ge \mathbb{E}\widetilde{N}_k \cdot \frac{k}{\mathbb{E}\widetilde{N}_k}\right) \le \mathbb{P}\left(\widetilde{N}_k \ge \mathbb{E}\widetilde{N}_k \cdot \frac{k}{k - \frac{Jm\varepsilon^2}{2}}\right).$$

Now since $\frac{1}{1-x} \ge 1 + x$, the rhs of the above is

$$\le \mathbb{P}\left(\widetilde{N}_k \ge \mathbb{E}\widetilde{N}_k \cdot \left(1 + \frac{Jm\varepsilon^2}{2k}\right)\right) \le \exp\left(-\frac{1}{3} \cdot \mathbb{E}\widetilde{N}_k \cdot \left(\frac{Jm\varepsilon^2}{2k}\right)^2\right),$$

where the last inequality follows from a Binomial tail bound, recorded in Lemma A.7. Now using $k \le \frac{m\varepsilon \log m}{\alpha}$ and applying the lower bound (29), we obtain

$$\le \exp\left(-\frac{1}{3} \cdot \frac{m\varepsilon}{2\alpha} \cdot \left(\frac{J\alpha\varepsilon}{2\log m}\right)^2\right).$$

Simplifying, we have shown that when $\frac{m\varepsilon}{\alpha} < k \le \frac{m\varepsilon \log m}{\alpha}$,

$$\mathbb{P}\left(\widetilde{N}_k \ge k\right) \le \exp\left(-\frac{\alpha J^2 m\varepsilon^3}{24\log^2 m}\right).$$

Plugging in the formula for ε, the above implies that the first piece of (27) is bounded,

$$\sum_{\frac{m\varepsilon}{\alpha} < k \le \frac{m\varepsilon \log m}{\alpha}} \mathbb{P}(N_k \ge k) \le m \exp\left(-\log(2m/\delta)\right) = \delta/2. \tag{30}$$

When $k > \frac{m\varepsilon \log m}{\alpha}$, the upper bound (28) gives

$$\mathbb{E}\widetilde{N}_k \le k + \frac{Jm\varepsilon^2}{2} - J\alpha k\varepsilon = k\left(1 + \frac{Jm\varepsilon^2}{2k} - J\alpha\varepsilon\right) \le k\left(1 - \frac{J\alpha\varepsilon}{2}\right),$$

for m large enough, since $\frac{Jm\varepsilon^2}{2k} \le O\left(\frac{\varepsilon}{\log m}\right)$. Again using $\frac{1}{1-x} \ge 1+x$, this upper bound on $\mathbb{E}\widetilde{N}_k$ implies

$$\begin{aligned}
\mathbb{P}\left(\widetilde{N}_k \ge \mathbb{E}\widetilde{N}_k \cdot \frac{k}{\mathbb{E}\widetilde{N}_k}\right) &\le \mathbb{P}\left(\widetilde{N}_k \ge \mathbb{E}\widetilde{N}_k \cdot \frac{k}{k\left(1-\frac{J\alpha\varepsilon}{2}\right)}\right) \\
&\le \mathbb{P}\left(\widetilde{N}_k \ge \mathbb{E}\widetilde{N}_k \cdot \left(1+\frac{J\alpha\varepsilon}{2}\right)\right) \\
&\le \exp\left(-\frac{1}{3}\cdot \mathbb{E}\widetilde{N}_k \cdot \left(\frac{J\alpha\varepsilon}{2}\right)^2\right) && \text{(by Lemma A.7)} \\
&\le \exp\left(-\frac{1}{3}\cdot \frac{m\varepsilon}{2\alpha} \cdot \left(\frac{J\alpha\varepsilon}{2}\right)^2\right) && \text{(by (29))} \\
&= \exp\left(-\frac{\alpha J^2 m\varepsilon^3}{24}\right).
\end{aligned}$$

Since $\delta \le 1$, the above implies that the second piece of (27) is bounded,

$$\sum_{k > \frac{m\varepsilon \log m}{\alpha}} \mathbb{P}(N_k \ge k) \le m \exp\left(-\log^3(2m/\delta)\right) \le \delta/2.$$

Together with (30), we have shown

$$\mathbb{P}(\hat{\tau}_\alpha > \tau^*_\alpha + \varepsilon) \le \sum_{k > \frac{m\varepsilon}{\alpha}} \mathbb{P}(N_k \ge k) \le \delta.$$

It remains to verify (28) and (29). To show (28), note that for any $t \in [\tau^*_\alpha, \tau^*_\alpha + \varepsilon]$, the mean value theorem gives

$$\bar{f}(t) - \bar{f}(\tau^*_\alpha) \le -J(t - \tau^*_\alpha)$$

since $\bar{f}' \le -J$ on $[\tau^*_\alpha, \tau^*_\alpha + \varepsilon]$. Since $\bar{f}$ is decreasing on $[\tau^*_\alpha, \tau^*_\alpha + \alpha]$, this implies

$$\bar{f}(t) \le \begin{cases} \bar{f}(\tau^*_\alpha) - J(t - \tau^*_\alpha) & \tau^*_\alpha \le t \le \tau^*_\alpha + \varepsilon \\ \bar{f}(\tau^*_\alpha) - J\varepsilon & \tau^*_\alpha + \varepsilon < t \le \tau^*_\alpha + \alpha. \end{cases}$$

Thus the expectation can be bounded,

$$\begin{aligned}
\mathbb{E}\widetilde{N}_k &= m \int_{\tau^*_\alpha}^{\tau^*_\alpha + \frac{\alpha k}{m}} \bar{f}(t)\mathrm{d}t \\
&= m \int_{\tau^*_\alpha}^{\tau^*_\alpha + \varepsilon} \bar{f}(t)\mathrm{d}t + m \int_{\tau^*_\alpha + \varepsilon}^{\tau^*_\alpha + \frac{\alpha k}{m}} \bar{f}(t)\mathrm{d}t \\
&\le m \int_{\tau^*_\alpha}^{\tau^*_\alpha + \varepsilon} \bar{f}(\tau^*_\alpha) - J(t - \tau^*_\alpha)\mathrm{d}t + m \int_{\tau^*_\alpha + \varepsilon}^{\tau^*_\alpha + \frac{\alpha k}{m}} \bar{f}(\tau^*_\alpha) - J\varepsilon \mathrm{d}t \\
&= m\left[\bar{f}(\tau^*_\alpha)\cdot \frac{\alpha k}{m} - \frac{J(t-\tau^*_\alpha)^2}{2}\Big|_{\tau^*_\alpha}^{\tau^*_\alpha + \varepsilon} - J\varepsilon\left(\frac{\alpha k}{m} - \varepsilon\right)\right] \\
&= k - \frac{Jm\varepsilon^2}{2} - J\alpha k\varepsilon + Jm\varepsilon^2 = k + \frac{Jm\varepsilon^2}{2} - J\alpha k\varepsilon,
\end{aligned}$$

which shows (28). For (29), note that the mean value theorem and the condition $\bar{f}' \geq -J^{-1}$ on $[\tau_\alpha^*, \tau_\alpha^* + \varepsilon]$ imply that $\bar{f}(t) \geq \bar{f}(\tau_\alpha^*) - J^{-1}(t - \tau_\alpha^*)$ for any $t \in [\tau_\alpha^*, \tau_\alpha^* + \varepsilon]$. Thus we have

$$\begin{aligned}
\mathbb{E}\widetilde{N}_k &= m \int_{\tau_\alpha^*}^{\tau_\alpha^* + \frac{\alpha k}{m}} \bar{f}(t) \mathrm{d}t \\
&\geq m \int_{\tau_\alpha^*}^{\tau_\alpha^* + \varepsilon} \left(\bar{f}(\tau_\alpha^*) - J^{-1}(t - \tau_\alpha^*) \right) \mathrm{d}t \\
&= m\varepsilon \bar{f}(\tau_\alpha^*) - \frac{m\varepsilon^2}{2J} \\
&= \frac{m\varepsilon}{\alpha} - \frac{m\varepsilon^2}{2J} \geq \frac{m\varepsilon}{2\alpha},
\end{aligned}$$

since for m larger than some constant $C(\alpha, J, \delta) > 0$, we have $\frac{m}{\log^3(2m/\delta)} \geq 24\alpha^2/J^5$, which is equivalent to the last inequality above.

□

A high probability lower bound can be shown under an extended monotonicity constraint of f over the interval $(0, \tau_\alpha^*)$, as described in the next lemma.

Lemma A.6. *Let $\delta > 0$. Suppose f is decreasing on the interval $(0, \tau_\alpha^*)$ and that there exists some $J > 0$ for which $|f'(t)| \geq J$ for all t with $|t - \tau_\alpha^*| \leq \varepsilon$, where $\varepsilon := \left(\frac{48}{\alpha J^2}\right)^{1/3} m^{-1/3} \log(2m/\delta)$. Then*

$$\mathbb{P}(\hat{\tau}_\alpha < \tau_\alpha^* - \varepsilon) \leq \delta,$$

for any $m \geq C(\alpha, J, \delta)$, a constant depending only on α, J and δ.

Proof. Define i^* as in Lemma A.4. If $\hat{\tau}_\alpha < \tau_\alpha^* - \varepsilon$, then there exists some $0 \leq k \leq i^*$ for which $\hat{\tau}_\alpha = p_{(i^* - k)}$ and thus

$$p_{(i^*-k)} - \frac{\alpha(i^* - k)}{m} \leq p_{(i^*)} - \frac{\alpha i^*}{m} \quad \text{and} \quad p_{(i^*-k)} < \tau_\alpha^* - \varepsilon.$$

Since $p_{(i^*)} \leq \tau_\alpha^*$, it follows that the probability can be bounded,

$$\begin{aligned}
\mathbb{P}(\hat{\tau}_\alpha < \tau_\alpha^* - \varepsilon) &\leq \mathbb{P}\left(\bigcup_{k=0}^{m} \left\{ p_{(i^*-k)} \leq \left(\tau_\alpha^* - \frac{\alpha k}{m} \right) \wedge (\tau_\alpha^* - \varepsilon) \right\} \cap \{ i^* \geq k \} \right) \\
&\leq \mathbb{P}\left(\bigcup_{0 \leq k \leq \frac{m\varepsilon}{\alpha}} \{ p_{(i^*-k)} \leq \tau_\alpha^* - \varepsilon \} \cap \{ i^* \geq k \} \right) \qquad (31) \\
&\quad + \mathbb{P}\left(\bigcup_{k > \frac{m\varepsilon}{\alpha}} \left\{ p_{(i^*-k)} \leq \tau_\alpha^* - \frac{\alpha k}{m} \right\} \cap \{ i^* \geq k \} \right). \qquad (32)
\end{aligned}$$

For (31), note that

$$p_{(i^*-k)} \leq \tau_\alpha^* - \varepsilon \Rightarrow N_\varepsilon := \sum_{j=1}^{m} 1_{\{p_j \in [\tau_\alpha^* - \varepsilon, \tau_\alpha^*]\}} \leq k,$$

since if at least $i^* - k$ of the p-values fall below $\tau_\alpha^* - \varepsilon$, and exactly i^* of the p-values are below τ_α^*, then at most k of the p-values fall in the interval $[\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$. Since the p-values are independent, we again have $N_\varepsilon \sim$ Generalized-Binomial with sample size m and average success probability $\bar{F}(\tau_\alpha^*) - \bar{F}(\tau_\alpha^* - \varepsilon)$. By the mean value theorem, for some $\xi \in [\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$, we have

$$\mathbb{E}N_\varepsilon = m(\bar{F}(\tau_\alpha^*) - \bar{F}(\tau_\alpha^* - \varepsilon)) = m\bar{f}(\xi)\varepsilon \geq m\bar{f}(\tau_\alpha^*)\varepsilon \geq k,$$

since $\bar{f}$ is decreasing on $(0, \tau_\alpha^*)$, $\bar{f}(\tau_\alpha^*) = \alpha^{-1}$, and $k \leq \frac{m\varepsilon}{\alpha}$. It follows from Theorem 5 in Hoeffding (1956) that

$$\mathbb{P}(p_{(i^*-k)} \leq \tau_\alpha^* - \varepsilon, i^* \geq k) \leq \mathbb{P}(N_\varepsilon \leq k) \leq \mathbb{P}(\widetilde{N}_\varepsilon \leq k), \tag{33}$$

where $\widetilde{N}_\varepsilon \sim \text{Binomial}(m, \bar{F}(\tau_\alpha^*) - \bar{F}(\tau_\alpha^* - \varepsilon))$. Further note that for any $t \in [\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$, the mean value theorem and the condition $\bar{f}' \leq -J$ on $[\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$ imply

$$\bar{f}(\tau_\alpha^*) - \bar{f}(t) = \bar{f}'(\xi)(\tau_\alpha^* - t) \leq -J(\tau_\alpha^* - t),$$

which further implies the following lower bound on the mean,

$$\begin{aligned}
\mathbb{E}\widetilde{N}_\varepsilon &= m \int_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} \bar{f}(t)\mathrm{d}t \\
&\geq m \int_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} \bar{f}(\tau_\alpha^*) + J(\tau_\alpha^* - t)\mathrm{d}t \\
&= m\bar{f}(\tau_\alpha^*)\varepsilon - \frac{mJ}{2}(\tau_\alpha^* - t)^2 \Big|_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} = \frac{m\varepsilon}{\alpha} + \frac{mJ\varepsilon^2}{2}.
\end{aligned} \tag{34}$$

It follows that (33) is bounded,

$$\begin{aligned}
\mathbb{P}(\widetilde{N}_\varepsilon \leq k) &= \mathbb{P}\left(\widetilde{N}_\varepsilon \leq \mathbb{E}\widetilde{N}_\varepsilon \cdot \frac{k}{\mathbb{E}\widetilde{N}_\varepsilon}\right) \\
&\leq \mathbb{P}\left(\widetilde{N}_\varepsilon \leq \mathbb{E}\widetilde{N}_\varepsilon \cdot \frac{k}{\frac{m\varepsilon}{\alpha}\left(1 + \frac{J\alpha\varepsilon}{2}\right)}\right) \\
&\leq \mathbb{P}\left(\widetilde{N}_\varepsilon \leq \mathbb{E}\widetilde{N}_\varepsilon \cdot \frac{1}{1 + \frac{J\alpha\varepsilon}{2}}\right). \qquad (k \leq \tfrac{m\varepsilon}{\alpha})
\end{aligned}$$

Now since $\frac{1}{1+x} \leq 1 - x/2$ for $x \in [0, 1]$, and since $\frac{J\alpha\varepsilon}{2} \leq 1$ for m larger than a constant, the above is bounded

$$\begin{aligned}
&\leq \mathbb{P}\left(\widetilde{N}_\varepsilon \leq \mathbb{E}\widetilde{N}_\varepsilon \left(1 - \frac{J\alpha\varepsilon}{4}\right)\right) \\
&\leq \exp\left(-\frac{1}{3} \cdot \mathbb{E}\widetilde{N}_\varepsilon \cdot \left(\frac{J\alpha\varepsilon}{4}\right)^2\right) \qquad \text{(Lemma A.7)} \\
&\leq \exp\left(-\frac{1}{3} \cdot \frac{m\varepsilon}{\alpha} \cdot \left(\frac{J\alpha\varepsilon}{4}\right)^2\right),
\end{aligned}$$

since (34) implies $\mathbb{E}\widetilde{N}_\varepsilon \geq \frac{m\varepsilon}{\alpha}$. Plugging the definition of ε, we have shown

$$\mathbb{P}(\widetilde{N}_\varepsilon \leq k) \leq \exp\left(-\frac{\alpha J^2 m\varepsilon^3}{48}\right) = \exp\left(-\log^3(2m/\delta)\right) \leq \frac{\delta}{2m},$$

so by the union bound, (31) is no larger than $\delta/2$.

For (32), similar to the first step in the analysis of (31), we have the implication

$$p_{(i^*-k)} \leq \tau_\alpha^* - \frac{\alpha k}{m} \Rightarrow N_k \coloneqq \sum_{j=1}^m 1_{\{p_j \in [\tau_\alpha^* - \frac{\alpha k}{m}, \tau_\alpha^*]\}} \leq k.$$

We have $N_k \sim$ Generalized-Binomial with sample size m and average success probability $\bar{F}\left(\tau_\alpha^*\right) - \bar{F}(\tau_\alpha^* - \frac{\alpha k}{m})$ because the p-values are independent. By the mean value theorem, for some $\xi \in [\tau_\alpha^* - \frac{\alpha k}{m}, \tau_\alpha^*]$, we have

$$\mathbb{E}N_k = m\bar{f}(\xi) \cdot \frac{\alpha k}{m} \geq k,$$

since $\bar{f}$ is decreasing on $(0, \tau_\alpha^*)$ and $\bar{f}(\tau_\alpha^*) = \alpha^{-1}$. It thus follows from Theorem 5 in Hoeffding (1956) that

$$\mathbb{P}\left(p_{(i^*-k)} \leq \tau_\alpha^* - \frac{\alpha k}{m}, i^* \geq k\right) \leq \mathbb{P}(N_k \leq k) \leq \mathbb{P}(\widetilde{N}_k \leq k),$$

where $\widetilde{N}_k \sim \text{Binomial}\left(m, \bar{F}(\tau_\alpha^*) - \bar{F}\left(\tau_\alpha^* - \frac{\alpha k}{m}\right)\right)$. For any $t \in [\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$, the mean value theorem gives

$$\bar{f}(\tau_\alpha^*) - \bar{f}(t) \leq -J(\tau_\alpha^* - t)$$

since $\bar{f}' \leq -J$ on $[\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$. Since $\bar{f}$ is decreasing on $(0, \tau_\alpha^*)$, this implies

$$\bar{f}(t) \geq \begin{cases} \bar{f}(\tau_\alpha^*) + J(\tau_\alpha^* - t) & \tau_\alpha^* - \varepsilon \leq t \leq \tau_\alpha^* \\ \bar{f}(\tau_\alpha^*) + J\varepsilon & t < \tau_\alpha^* - \varepsilon. \end{cases}$$

Thus $\mathbb{E}\widetilde{N}_k$ is bounded below,

$$\begin{aligned}
\mathbb{E}\widetilde{N}_k &= m \int_{\tau_\alpha^* - \frac{\alpha k}{m}}^{\tau_\alpha^*} \bar{f}(t)\mathrm{d}t \\
&= m \int_{\tau_\alpha^* - \frac{\alpha k}{m}}^{\tau_\alpha^* - \varepsilon} \bar{f}(t)\mathrm{d}t + m \int_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} \bar{f}(t)\mathrm{d}t \qquad (k > \tfrac{m\varepsilon}{\alpha}) \\
&\geq m \int_{\tau_\alpha^* - \frac{\alpha k}{m}}^{\tau_\alpha^* - \varepsilon} (\bar{f}(\tau_\alpha^*) + J\varepsilon)\mathrm{d}t + m \int_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} f(\tau_\alpha^*) + J(\tau_\alpha^* - t)\mathrm{d}t \\
&= m\bar{f}(\tau_\alpha^*) \cdot \frac{\alpha k}{m} + Jm\varepsilon\left(\frac{\alpha k}{m} - \varepsilon\right) - \frac{mJ}{2}(\tau_\alpha^* - t)^2\Big|_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} \\
&= k + J\alpha k\varepsilon - mJ\varepsilon^2 + \frac{mJ\varepsilon^2}{2}.
\end{aligned}$$

Simplifying, we have shown

$$\begin{aligned}\mathbb{E}\widetilde{N}_k &\geq k + J\alpha k\varepsilon - \frac{mJ\varepsilon^2}{2} \\ &> k + J\alpha k\varepsilon - \frac{J\alpha k\varepsilon}{2} && (m\varepsilon < \alpha k) \\ &= k\left(1 + \frac{J\alpha\varepsilon}{2}\right). && (35)\end{aligned}$$

Now since $\frac{1}{1+x} \leq 1 - x/2$ for $x \in [0,1]$, and since $\frac{J\alpha\varepsilon}{2} \leq 1$ for m larger than a constant, we have

$$\begin{aligned}\mathbb{P}(\widetilde{N}_k \leq k) &= \mathbb{P}\left(\widetilde{N}_k \leq \mathbb{E}\widetilde{N}_k \cdot \frac{k}{\mathbb{E}\widetilde{N}_k}\right) \\ &\leq \mathbb{P}\left(\widetilde{N}_k \leq \mathbb{E}\widetilde{N}_k \cdot \left(1 - \frac{J\alpha\varepsilon}{4}\right)\right) \\ &\leq \exp\left(-\frac{1}{3} \cdot \mathbb{E}\widetilde{N}_k \cdot \left(\frac{J\alpha\varepsilon}{4}\right)^2\right) \\ &\leq \exp\left(-\frac{1}{3} \cdot \frac{m\varepsilon}{\alpha} \cdot \frac{J^2\alpha^2\varepsilon^2}{16}\right),\end{aligned}$$

since (35) together with $k > \frac{m\varepsilon}{\alpha}$ imply $\mathbb{E}\widetilde{N}_k \geq \frac{m\varepsilon}{\alpha}$. Plugging in the definition of ε, we have shown

$$\mathbb{P}(\widetilde{N}_\varepsilon \leq k) \leq \exp\left(-\frac{\alpha J^2 m\varepsilon^3}{48}\right) = \exp\left(-\log^3(2m/\delta)\right) \leq \frac{\delta}{2m},$$

so by the union bound, (32) is no larger than $\delta/2$. Since we've now shown that both terms (31) and (32) are below $\delta/2$, the proof is complete.

□

Lemma A.7. *Let $X \sim$ Binomial(n,p). Then for any $0 < \delta < 1/2$, we have*

$$\mathbb{P}(X \geq np(1+\delta)) \leq \exp\left(-\frac{1}{3}np\delta^2\right).$$

Proof. By Markov's inequality, for any $t \geq 0$ we have

$$\mathbb{P}(X \geq np(1+\delta)) \leq \frac{\mathbb{E}e^{tX}}{e^{tnp(1+\delta)}} = \frac{(1-p+pe^t)^n}{e^{tnp(1+\delta)}} \leq \exp\left(np(e^t-1) - tnp(1+\delta)\right).$$

Letting $t = \log(1+\delta)$, we have

$$\mathbb{P}(X \geq np(1+\delta)) \leq e^{np(\delta - (1+\delta)\log(1+\delta))}.$$

Now since $(1+\delta)\log(1+\delta) \geq \delta + \frac{1}{3}\delta^2$ for any $\delta \in (0, 1/2)$, we obtain the result. □

B Simulation

Theorem 2.4 only requires the assumption that the average density $\bar{f}$ is decreasing over $(0, \tau_\alpha + \alpha)$, and the following simulation checks empirically that $\bar{f}$ need not be decreasing over the entire unit interval in order for $\hat{\tau}_\alpha$ to closely approximate the population threshold τ_α^*.

For each sample size $m \in \{10^k : k = 1, \ldots, 6\}$, we draw m p-values from the Beta(0.5,0.5) distribution, whose density tends to infinity at 0 and 1. We compute $\hat{\tau}_\alpha$ at level $\alpha = 0.1$ for $N = 10^4$ Monte Carlo trials to estimate the expected difference $\mathbb{E}|\hat{\tau}_\alpha - \tau_\alpha^*|$ for each m. In Figure 11, we plot the logged estimates of these differences against the log of the sample size. The slope is roughly $-1/3$, which matches the theoretical prediction of Theorem 2.4, as lfdr in this setting is a continuous function.

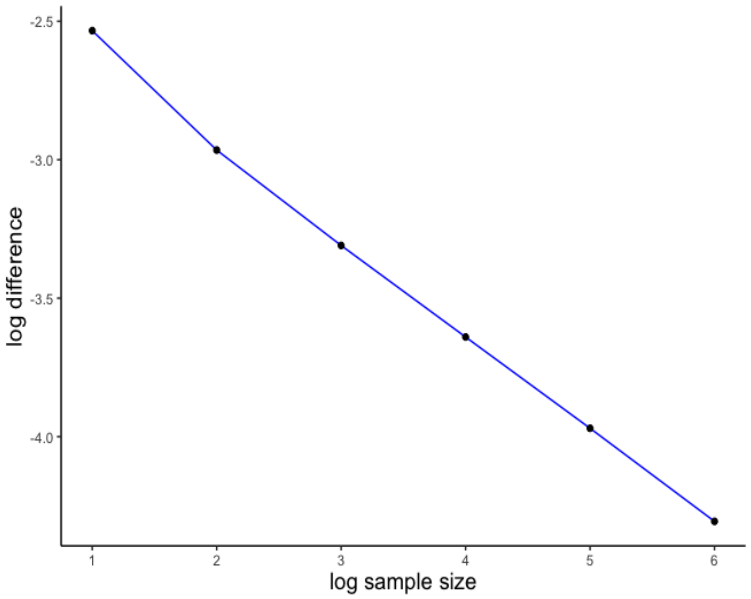

Figure 11: The logged absolute differences (vertical axis) are plotted against the log sample size (horizontal axis). The least squares estimate for the slope is $\hat{\beta}_1 \in -0.348 \pm 2(0.0084) = (-0.365, -0.331)$