ON FALSE DISCOVERY RATES FOR SECOND-GENERATION P-VALUES

By

Valerie Frances Welty

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Biostatistics

August 11, 2023

Nashville, Tennessee

Approved:

Amber J. Hackstadt, Ph.D.

Jeffrey D. Blume, Ph.D.

Thomas G. Stewart, Ph.D.

Melinda C. Aldrich, Ph.D., MPH

ACKNOWLEDGMENTS

I owe an abundance of gratitude to a great number of people for shaping this journey and supporting me along the way.

I would like to first and foremost thank my advisor, Dr. Jeffrey Blume. Your support has been multi-faceted and encompasses so many different aspects of my doctoral career. From recruitment, to early year coursework, to 1st year summer research diving into second-generation p-values (which clearly stuck!), to applied biostatistical advisement, and on to dissertation research. You have provided an endless amount of time, knowledge, opportunity, words of encouragement, insightful discussions, and advice in so many different areas. I am grateful that you always extended guidance and support, yet fostered an environment where I could develop my skills and confidence as an independent researcher. I could not have asked for a better mentor.

To my dissertation committee, Dr. Amber Hackstadt, Dr. Thomas Stewart, and Dr. Melinda Aldrich, I would like to extend my sincere thanks for your guidance, for helping me shape the direction of my research, and your thoughtful feedback which greatly improved this work. Additionally, I am extremely grateful for your time, flexibility, and encouragement and kindness throughout this process.

This endeavor would not have been possible without the support of the TREAT Lung Cancer Program, including Dr. Eric Grogan, Dr. Melinda Aldrich, Dr. Kim Sandler and Dr. Stephen Deppen. I am grateful for the opportunity to foster my skills as an applied Biostatistician with this knowledgeable, passionate, and multi-disciplinary team. Thank you all for the support, mentorship, opportunity, engaging research, and inspiration provided over these past years.

I would like to acknowledge the Biostatistics department and graduate program, including our caring graduate program directors, our infinitely helpful and expert program managers, and the brilliant professors who taught our courses and enthusiastically supported the students in many other ways. I am thankful for having been part of such a unique graduate program.

To my entire cohort, I'm extremely grateful for your support, encouragement, interesting methodological discussions, and being able to commiserate together in moments of overwhelm – which were plenty. It's hard to say how I would have fared were it not for this welcoming group. I'd like to also give my thanks to all the other students and fellows, both in Biostatistics and other biomedical and clinical departments across campus, with whom I have had the pleasure of meeting or working with, and who have greatly enriched my time as a graduate student.

I would like to thank my peers and mentors from Georgia Southern for further cultivating my interest in research. I am thankful to Dr. Patricia Humphrey, for your immense support and advisement throughout my honors thesis. I would be remiss not to extend a thank you to my friend Michael Ackerman, from whom I first learned about the excellence of the Vanderbilt Biostatistics program!

To my family, without your support, guidance and understanding, I would not be where I am today. I am grateful for your development of our independent learning, which ultimately provided me with the ability and confidence to pursue this path. Mom, I would like to thank you for sharing with me when you read the title "Biostatistician" at the bottom of a work email and thought this career could be of interest to me – it turns out you were right! Also, I would like to thank my Fowler family, for your immense encouragement and care over the years. I am lucky to have such a large group of cheerleaders on my side.

Last, I would like to say that I am forever indebted to my amazing husband, Gregory Fowler, who is my everything. There are not enough words to describe the ways in which you have supported me throughout this journey and everything leading up to it. You may not know how much you've done for me – you have been my partner, my comfort, my inspiration, my constant source of laughter and joy, and my best friend without fail. I want to thank you for your unending patience throughout the last seven years, and particularly the last year or two, and for taking on burdens so that I wouldn't have to shoulder them. Without a doubt, this would not have been possible without you.

TABLE OF CONTENTS

Page

LIST OF TABLES ... vii

LIST OF FIGURES ... viii

1 Introduction ... 1

2 A Statistical Primer on False Discovery Quantities ... 4

2.1 Introduction ... 4
2.1.1 Motivating examples ... 4
2.1.2 Setup and notation ... 5
2.2 Error control in multiple testing ... 7
2.2.1 Family wise error rate and adjusted p-values ... 8
2.2.2 False discovery proportion ... 9
2.2.3 False discovery rate ... 10
2.3 Estimating false discovery quantities ... 12
2.3.1 The positive false discovery rate ... 13
2.3.2 Bayesian interpretation of the pFDR ... 14
2.3.3 q-Value ... 14
2.3.4 Local FDR ... 15
2.3.5 Estimation approaches ... 15
2.4 Discussion ... 21
2.4.1 FDR testing and estimation ... 21
2.4.2 Advanced topics ... 23
2.4.3 Recommendations for routine use of FDRs ... 24
2.5 Appendix A: Remarks and supplemental content ... 25
2.5.1 Remarks ... 25

3 The Positive False Discovery Rate for Second-Generation p-Values ... 40

3.1 Introduction ... 40
3.2 Second-generation p-value background ... 42
3.2.1 Operational characteristics ... 43
3.3 False discovery rate quantity definitions ... 48
3.3.1 Example: simple null and simple alternative hypotheses ... 50
3.3.2 Composite hypotheses approaches ... 52
3.3.3 Common scenarios ... 54
3.3.4 Numerical illustrations and asymptotic behavior ... 56
3.4 Large-scale inference: empirical estimation ... 58
3.4.1 Empirical Bayes mixture estimates ... 58
3.4.2 False discovery rate estimates ... 61
3.5 Closing remarks ... 64
3.6 Appendix A: Remarks and supplemental content ... 66
3.6.1 Remarks ... 66

4 False Discovery Rate Quantities for Second-Generation p-Values: Behavior, FDR Control, and Impact of Trivial Effects ... 68

4.1 Introduction ... 68
4.2 The FDR and other relevant metrics for the SGPV ... 70
4.2.1 Direct calculations for fixed variance ... 71
4.2.2 Discussion: FDR control and pFDR minimization ... 73
4.3 Simulations for sample variance ... 74

4.4 Rejections of trivial effects for an interval null hypothesis 79
4.4.1 Results 80
4.4.2 Illustration: which type of tests rejected by SGPV vs. p-value approaches....................80
4.4.3 Hybrid procedures 83
4.5 Discussion 84
4.6 Appendix A: Remarks and supplemental content....................91

5 Conclusion....................107

REFERENCES 109

LIST OF TABLES

Table Page

2.1 Data table summarizing the results of a multiple testing procedure....7
2.2 Data table summarizing the results of a multiple testing analysis for the simple simulation for various rejection regions or procedures....8
3.1 Description of three important outcomes for second-generation p-value....43
3.2 Summary of design probabilities for SGPV inference....47
3.3 Summary of false discovery quantities for classical p-values and second-generation p-values....49
3.4 False discovery, non-discovery, and confirmation rate quantity results for classical p-value significance testing and second-generation p-value based cutoffs....57
3.5 True mixture probabilities and empirical estimates for various values of null proportion, n, and m....60
3.6 True false discovery quantities and estimated false discovery quantities with the empirical mixture probability estimate for the simple simulation for various values of null proportion, n, and m....62
4.1 Example scenarios for controlled FDR....73

LIST OF FIGURES

Figure Page

2.1 Histograms of p-values and z-values for the examples 6
2.2 Distributions of R, V, and Q for the simple simulation from 100,000 simulation replications 11
2.3 Estimated FDR quantities using the theoretical null and ECDF-based estimators 17
2.4 Connection between the Benjamini-Hochberg procedure and the pFDR and q-value 18
2.5 Estimated FDR quantities using the theoretical null and Lindsey's method estimators 20
2.6 Estimated FDR quantities using the empirical Null distribution with Lindsey's method mixture estimate 22
3.1 Operational characteristics of the second-generation p-value 46
3.2 Comparison of FDR and FCR for SGPV to FDR and FNR for the classical p-value 51
3.3 Simple FDR as a function of sample size for varying values of the proportion of tests which are null ..52
4.1 Relevant false discovery rate quantities for the second-generation p-value rejection rule 72
4.2 Simulation estimates of false discovery rate quantities for a setting with only null and non-trivial alternative effects, and with tests having a common variance 76
4.3 Simulation estimates of false discovery rate quantities for a setting with only null and non-trivial alternative effects, and with two possible variance values distributed randomly among tests 77
4.4 Simulation estimates of false discovery rate quantities for a setting with only null and non-trivial alternative effects, and with null tests having smaller variance and alternative tests having larger variance 78
4.5 Simulation estimates of false discovery rate quantities for a setting with null, trivial, and non-trivial alternative effects, and with tests having a common variance 81
4.6 Illustration of rejected intervals for each testing method in two example settings 82
4.7 Simulation estimates of false discovery rate quantities, including results from the hybrid procedures, for a setting with null, trivial, and non-trivial alternative effects, and with tests having a common variance 85
4.8 Simulation estimates of false discovery rate quantities, including results from the hybrid procedures, for a setting with null, trivial, and non-trivial alternative effects, and with null tests having smaller variance and alternative tests having larger variance 86

CHAPTER 1

Introduction

In statistical inference, when a single significance test is performed, we primarily concern ourselves with the type I error of the test. On the other hand, in the case of multiple testing, we are concerned with the accumulation of errors across tests. This was originally addressed with the concept of family wise error (FWER), the probability of one or more tests falsely rejecting (Tukey 1953; Dunn 1961).

An alternative proposal, in considering the rate of false rejections, was formalized in (Benjamini and Hochberg 1995) with their definition of the false discovery rate (FDR). However, a distinction must be made between the formal BH definition of the FDR, and another quantity: the probability that tests are null conditioned on their rejection. This latter Bayesian quantity is equivalent to the rate of false discoveries conditional on one or more rejections being observed (formally defined as the positive false discovery rate (pFDR) in (Storey 2003), which also established this equivalency), which is a condition not required for the BH definition of the FDR. The FDR is useful when examining the properties of a study design, and when strict control is desired. However, out of the two quantities, the pFDR answers the natural question that arises when faced with a set of rejected tests – that is, how many do we expect to correspond to the null hypothesis of no effect? Control of the FDR by a multiple testing algorithm does not necessarily guarantee a small pFDR.

A general class of methods have been proposed in the literature to estimate this quantity via a family of empirical Bayes approaches, made possible by the equivalency established by Storey (2003) between the pFDR and the posterior/Bayes FDR quantity (for example, those described in (Allison et al. 2002; Pounds and Morris 2003; Aubert et al. 2004; Liao et al. 2004; Pounds and Cheng 2004; Tang et al. 2007; Efron 2010b)).

In Chapter 2, we provide a thorough review of the topics outlined above. In particular, we present a unified framework used to introduce a variety of ideas presented in the literature under a common notation, aiming to highlight the relationships between them. We discuss and illustrate the distinction between the classically defined FDR and the positive false discovery rate. Importantly, the FDR measures overall frequentist properties of a rejection procedure, but the pFDR measures the probability of observed rejections to be misleading, which is the scientific question which most interests researchers in practice. Further, we review a selection of empirical methods for pFDR estimation on the basis of classical z or p-values, focusing on those presented in (Storey 2002; Efron 2010b). These methods are illustrated with a simulation example consisting of underdispersed z-values, as well as with a real-world large-scale inference example studying the association between 224,866 SNPs and prostate cancer, with accompanying code provided online. Chapter 2 concludes with our recommendations for routine use of false discovery quantities. We intend for this chapter to serve as an important resource for those seeking to gain a better understanding of false discovery rates and their implications in traditional statistical inference. The topics covered represent some of the most fundamental FDR concepts, with inclusion of and reference to a wide variety of ideas, however, are not necessarily fully comprehensive.

Next, we turn our focus to an alternative paradigm, with the incorporation of an interval null hypothesis in lieu of a standard point null hypothesis. This interval null is defined to consist of effects which are null or practically null, i.e., those which are not of scientific interest or have negligible impact in real practice. One such methodology that utilizes this paradigm is that of Blume et al. (2018), in the definition of the "second-generation p-value" as an alternative metric to the classical p-value. This approach formalizes the concept of practical significance, rather than only statistical significance, into the inference process. Blume et al. (2018) establish several improved properties of the SGPV and introduces a preliminary definition of the pFDR – as well as an analogous quantity, the false confirmation rate (FCR) – for second-generation p-values. The assumptions required to calculate these FDR quantities, however, are quite restrictive and ultimately do not directly account for the interval null, despite this being the key underlying concept. In Chapters 3 and 4, we undertake an examination of several different topics at the intersection of second-generation p-values and false discovery rates.

First, in Chapter 3, we address a specific question – how do we generally define the pFDR, i.e., the Bayes FDR, for second-generation p-values, accounting for interval null and alternative hypotheses, and how may we empirically estimate it in practice? We begin by providing a further examination of the (Blume et al. 2018) definition of the SGPV FDR and FCR using simple null and alternative hypothesis specifications. Next, building on this work, we define a general form of the Bayes FDR quantities. Our proposed solution includes specifying a weighting function for the parameters across the null or alternative spaces, hypothetically representing the true underlying distribution of effects, and marginalizing across the power curve of the SGPV to obtain the respective design probabilities conditional on each of the composite hypotheses. We provide several different approaches for specification of these weighting distributions. We find that regardless of the approach used, the SGPV Bayes FDR converges to zero as a function of the sample size underlying the tests. This represents an important generalization of the findings of (Blume et al. 2018). However, for finite sample sizes, the specifications of the weighting functions can result in FDR estimates which vary widely. Further, we note that the null weights do not impact the asymptotic behavior of the Bayes FCR for the second-generation p-value. However, the alternative weighting specification might – namely, in the instance where a point mass is placed at the minimum scientifically important effect, in which case the FCR converges to a lower bound greater than zero (but less than α).

We conclude with proposing an estimator of the Bayes FDR for second-generation p-values which utilizes empirical Bayes methods, similar to those described in Chapter 2, for part of the calculation. This results in a reduction in the number of specifications and assumptions required. The component that is empirically estimated, the mixture probability of a SGPV rejection, behaves well, particularly with large numbers of tests. However, the specification of the other components, such as the null design probability and null proportion, can still have a substantial impact on the resulting estimates. An upper bound on this estimator can be found without much trouble, although it may be an appreciable overestimate in some cases. Future work on empirical estimation of the design probabilities, and of the proportion of null tests, is desired to produce reliable Bayes FDR, i.e., pFDR, estimates for second-generation p-values for use in practice.

Next, in Chapter 4, we broaden our scope, and provide a holistic examination of second-generation p-value performance in large-scale inference. This chapter addresses several open questions related to the SGPV, including its comparison with p-value approaches that adjust for multiple comparisons, and an assessment of overall FDR control (referring back to the original BH definition of the false discovery rate). Several assumptions are relaxed, particularly those of fixed, known variance and of common variance among tests, which previously limited the applicability of the findings. We use extensive simulations to examine the impact of factors such as sample size, number of tests and variance on the overall FDR control, the pFDR, the probability of observing rejections, and power as defined by the rate of truly scientifically meaningful tests rejected.

First, if scientific relevance is ignored and we focus on the rejection of exactly null effects, we find that – unlike with the Benjamini-Hochberg approach – the FDR is not generally controlled by the SGPV for all sample sizes, although the main scientific quantity of interest (the pFDR) is often comparable or reduced. However, when trivial effects (i.e., those where the true effect is non-zero, but not large enough to be scientifically relevant) are considered, neither the BH or SGPV methods control the FDR for finite samples, and only the SGPV pFDR is asymptotically controlled. In general, for most settings, SGPV FDR control occurs only when the pFDR itself falls below the desired threshold, which is guaranteed for large enough sample sizes. We conclude by examining two hybrid methods, one of which we propose as an intersection between the SGPV and BH, and another which is a simple extension of a method described previously by Goodman et al. (2019). These hybrid methods provide a better overall balance between FDR control and pFDR minimization. However, they are not universal improvements over the SGPV alone in regard to the pFDR and power. As a whole, we show that standard p-value methods are not sufficient when scientific relevance is considered, and methods such as the second-generation p-value, or a hybrid approach, should be implemented instead. Ultimately, the best choice of methods will depend on factors such as study priorities and expected properties of the data.

Overall, this dissertation provides contributions to the field of false discovery rates in large-scale inference, with elucidation of established methodology, and with further development of theory and implementation of a newer metric, the second-generation p-value. Chapter 2 provides an accessible yet thorough introduction to the field of false discovery rates, which will prove useful for statisticians endeavoring to better understand this methodological area. Chapter 3 establishes a more general framework for the second-generation p-value Bayes FDR (i.e., pFDR). The illustration of small and large-sample properties of proposed methods provides an improved understanding of this metric, while a proposed empirical estimator illustrates the need for further work in this area, with the end goal of establishing a robust method to estimate the probability that findings are null or practically null – a key question in large-scale inference. The comprehensive overview of second-generation p-value behavior in large-scale inference, found in Chapter 4, will provide insight into which method to use in practice in various contexts, ultimately leading to improved, scientifically meaningful statistical inference.

CHAPTER 2

A Statistical Primer on False Discovery Quantities

Valerie F. Welty, Jeffrey D. Blume

2.1 Introduction

The problem of simultaneous multiple inferences has been studied for quite some time, with interest beginning around the 1950's. There is by now a rich and vast literature on multiple testing and multiple comparisons (see for example, Tukey 1953, 1991; Miller 1981; Hochberg and Tamhane 1987; Benjamini and Braun 2002). The problem of multiple inferences was originally studied in the context of relatively few tests (e.g., on the order of 10 or 20 tests). Now, however, it is commonplace to see multiple testing problems with thousands or even millions of simultaneous tests. For example, in genome-wide association studies (GWAS), genetic variants may be studied to assess their association with a particular outcome, such as prostate cancer. When the number of tests become large, the strategy of controlling the probability that at least one false positive occurs is debatable. In these contexts, it is often more desirable to allow a small number of false positives to occur in order to increase the probability of identifying true signals. This idea was formalized in a seminal 1995 paper by Benjamini and Hochberg. They proposed an algorithm for controlling the false discovery rate instead of the family-wise error rate (Benjamini and Hochberg 1995). An important aspect of Benjamini and Hochberg's seminal paper was that they proposed a workable definition of the "false discovery rate", building on the work and ideas of others at the time (Spjøtvoll 1972; Simes 1986; Berger and Sellke 1987; Sorić 1989).

The Benjamini-Hochberg procedure has become quite popular; as of 2005, it was 7th on the list of most-cited papers published since 1993 with 294 citations (Ryan and Woodall 2005). At the time of this writing, the original paper has amassed over 78,000 citations. However, advancements and refinements in false discovery methodology have not received quite as much attention. The goal of this chapter is to serve as a statistical primer on false discovery quantities and some popular advancements, particularly in the area of estimating relevant quantities. We introduce fundamental and advanced FDR concepts, present a unified notational structure for easy reading, share simulations and a real-world example to illustrate key concepts, and present some simple recommendations for using false discovery quantities in practice.

2.1.1 Motivating examples

A common context is a GWAS study where single-nucleotide polymorphism (SNP) data is collected in diseased and non-diseased individuals. For each SNP, researchers might fit a logistic regression model of disease status on the number of variant alleles (this approach is referred to as the additive model) and adjust for demographic covariates. The coefficient for the SNP is the coefficient of interest, and there is one for each SNP when utilizing the additive model. For a simple motivating simulation example, imagine that we have 1,000 SNPs with resulting regression coefficients $\theta_1, \dots, \theta_{1,000}$. We assume that 85% of these SNPs are truly not associated with the outcome (i.e., they are null) while the remaining 15% come from a common non-null distribution with a positive mean, and that all SNPs are independent (a tenuous assumption in genetic settings). The mixture distribution of the z-values and p-values, for one such realization of this simple scenario, are shown in Figure 2.1(a) along with the theoretical null distributions. In large-scale inference settings, the underlying structure of the data is generally not so simple, containing for example a complex correlation structure and a variety of non-null effect sizes. To illustrate, we used a real-world example to explore the association between prostate cancer and 224,866 single-nucleotide polymorphisms (SNPs) from chromosome six. The data was collected from 3,894 individuals by the International Consortium for Prostate Cancer Genetics (ICPCG) (Schaid and Chang 2005). The distribution of the z-values and p-values for this study are provided in Figure 2.1(b). To illustrate the importance of flexible FDR methods, we use a second simulation scenario that has an underdispersed null distribution, which is one type of departure from the theoretical null that can occur in genetics and imaging scenarios. We set the number of tests at 50,000, which is a more typical size of modern large-scale inference experiments. It is assumed that 90% of the tests are null and the alternative distribution is generated as a mixture distribution comprised of three effect sizes. The underdispersion can be seen in Figure 2.1(c) when contrasting to the red line, which is the theoretical null of $N(0,1)$ (for the z-values) or $U(0,1)$ (for the p-values). The departure from the theoretical null can be seen most clearly in the upper range of the observed p-value distribution. Further details on the simulations and the prostate cancer genetic SNP study are provided in the Appendix (Remark 2.A).

2.1.2 Setup and notation

We are interested in testing m null hypotheses, for some very large m, which we denote $\mathcal{H} = (H_0^1, \dots, H_0^m)$. For each hypothesis there is an associated test statistic and p-value. Some of these null hypotheses will be correct, say $m_0 \leq m$, while the rest, $m_1 = m - m_0$, are false. It is convenient to work with the proportion of truly null hypotheses, defined as $\pi_0 = m_0/m$, and the proportion of non-null hypotheses, defined as $\pi_1 = m_1/m$. Throughout the chapter, we will assume a point null hypothesis is being used. Note that in large-scale settings it is unlikely that all effects will be exactly equal to zero, and some methodologies are robust to this in the sense that they allow for a composite null hypothesis comprised of small non-relevant effects (Efron 2010b). See also in (Cabras 2010; Chi 2010; Sun and McLain 2012; Blume et al. 2018).

A helpful step in false discovery estimation is mapping the observed test statistics $T_1, \dots, T_m$, or their p-values, to z-values for the purpose of leveraging their normality properties. The p-values can be directly transformed to z-values by $Z_i = \Phi^{-1}(p_i)$, however this definition does not guarantee that the z-values have the same sign as the original test statistics. For example, if the p-values were calculated as upper-tail one-sided p-values of test statistics such as t-statistics, the resulting Z_i will have the reverse sign. To utilize this p-value, we would need to define the special transformation $Z_i = \Phi^{-1}(1 - p_i)$ to maintain the original sign. A specialized transformation is also required if using two-sided p-values. More discussion and examples are given in the Appendix (Remark 2.B) and in (Efron

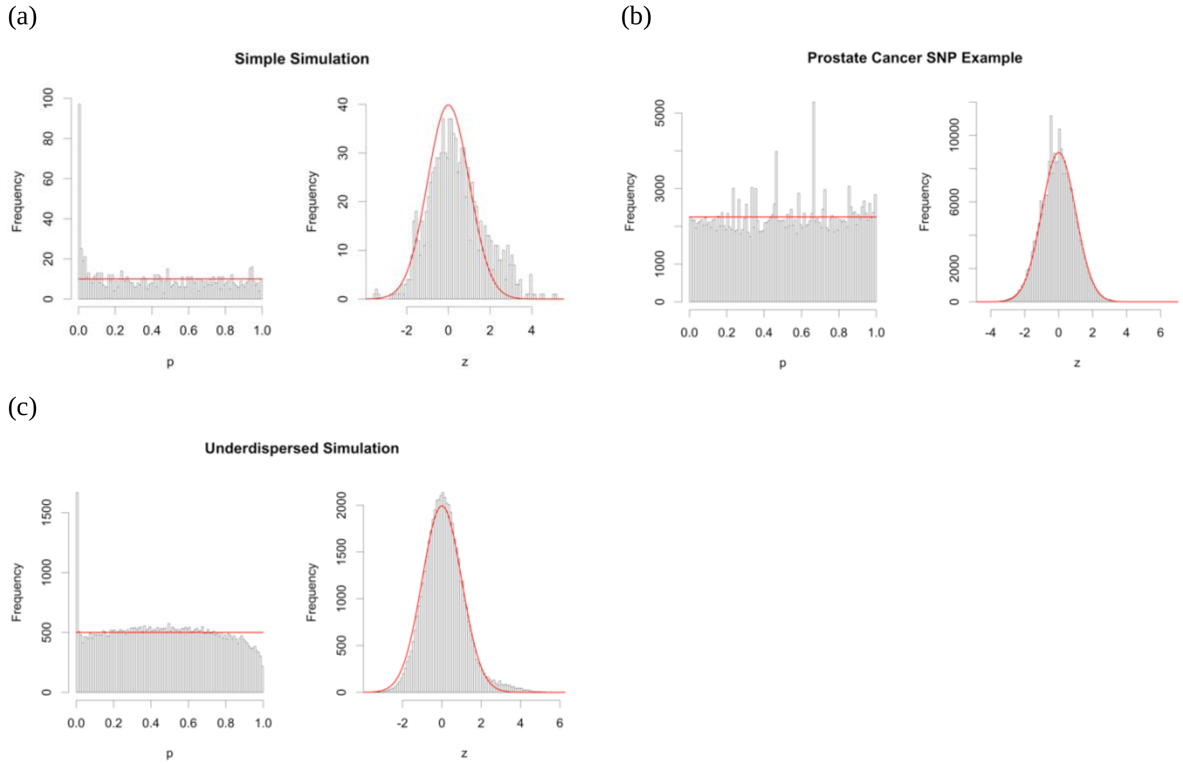


Figure 2.1 Histograms of p-values and z-values for the examples. The p-values are one-sided upper tail p-values. The red line is either the theoretical null distribution of $U(0,1)$ assuming all tests are null (for the p-values) or the theoretical null distribution of $N(0,1)$ (for the z-values). (a) simple simulation; (b) real-world SNP example; (c) underdispersed simulation.

2010b). Alternately, we can use a more straightforward transformation if working directly with the original test statistics: for test statistics $T_1, \dots, T_m$ with distribution function $T_i \sim F_0(T)$ under the null hypothesis, the z-value transformation is $Z_i = \Phi^{-1}\big(F_0(T_i)\big)$. For example, suppose $T_i = 2$ comes from a one-sided upper-tail t-test with 10 degrees of freedom, such that $p_i = 0.0367$. The corresponding z-score is then $Z_i = \Phi^{-1}\big(F_0(2)\big) = \Phi^{-1}(1 - p_i) = \Phi^{-1}(0.9633) = 1.7904$. For the remainder of the chapter, we will assume a transformation has yielded z-values $Z_1, \dots, Z_m$ such that $Z_i | H_0^i \sim N(0,1)$.

Table 2.1 displays the data table, which summarizes the outcomes of a large-scale inference procedure. In the set of m null hypotheses under consideration, previously defined as $\mathcal{H}$, the set of those that are rejected will be denoted by $\mathcal{R}$ and the size of that set is R. The set of hypotheses that were not rejected is denoted by $\mathcal{J}$ and has size $m - R$. It is also common to denote the number of rejected hypotheses that correspond to true null hypotheses by V, and the number of rejected hypothesis that correspond to false null hypotheses is denoted by U. The quantities m, m_0, and m_1 are fixed, although only m is known to the analyst. The total column is also observed with realized values r and $m - r$ respectively. The interior of the table, along with m_0 and m_1, are not observable in given set of data. The quantities V (false rejections) and U (correct rejections) are random variables; they change from table

Table 2.1 Data table summarizing the results of a multiple testing procedure. The rows correspond to the results of the analysis, i.e., reject or inconclusive (fail to reject), while the columns correspond to the truth or falsehood of the m null hypotheses. Note that only the last column, containing the totals, is known after the analysis; the other columns contain unknown quantities (shaded in grey), unless the data are simulated.

	Null ($\mathcal{T}$)	Alt. ($\mathcal{F}$)	Total
Reject ($\mathcal{R}$)	V	U	R
Inconclusive ($\mathcal{J}$)	$m_0 - V$	$m_1 - U$	$m - R$
Total	m_0	m_1	m

to table under the same data generating mechanism in the same way that R does. But for any given set of data, these random variables have fixed but unknown values, denoted by v and u, and these are important false discovery quantities.

An important aspect of a data table like this is that it is dependent on the definition of a rejected hypothesis. The set and number of rejections ($\mathcal{R}$ and R) vary depending on rejection criterion. For example, in routine cases, $\mathcal{R}$ depends on the pre-specified size of the rejection region, say $[1.64, \infty)$ for a one sided 5%-sized test. Tables 2.2(a)-(c) provide example data tables for 1,000 z-values generated from the simple simulation setting for three different rejection criteria (an unadjusted procedure and two multiple testing procedures, Bonferroni and Benjamini-Hochberg). These tables are augmented with the marginal row percentages. The proportion of rejected hypotheses that are null, V/R, is called the false discovery proportion (FDP) and is the basis for nearly all false discovery quantities.

2.2 Error control in multiple testing

For a single significance test, the p-value rejection threshold c is chosen to control the Type I Error rate at some pre-specified level α. A consequence of this strategy is that in large-scale testing, where m significance tests are performed, the group-wise Type I Error rate will greatly exceed α. However, the group-wise Type I Error rate can be controlled with classical adjustments like Bonferroni, Šidák, or Simes (Dunn 1961; Šidák 1967; Holm 1979; Simes 1986; Hochberg 1988). These solutions work by allowing the individual Type I Error rates to shrink to zero as the number of tests grows. For nearly 70 years, this has been the standard approach. But in the 1990's an alternative approach emerged, building on preceding ideas (Spjøtvoll 1972; Berger and Sellke 1987; Sorić 1989; Benjamini and Hochberg 1995). Instead of fixing the (pre-test) family-wise Type I Error rate at some level, we might instead focus on a different goal and try to bound the (post-test) probability that the observed rejection is mistaken. This latter approach is based on false discovery quantities. Note that all approaches attempt to control the total number of false rejections (the random variable V in our data table) in some way.

Table 2.2 Data table summarizing the results of a multiple testing analysis for the simple simulation for various rejection regions or procedures. The rows correspond to the results of the analysis (reject or inconclusive), while the columns correspond to the truth or falsehood of the m null hypotheses. Unknown values are shaded in grey. Also shown are the marginal row percentages, corresponding in one case to the realized false discovery proportion (FDP) value and in another to the realized false non-discovery proportion (FNP). (a) Unadjusted upper-tail z-value rejection region of $[1.64, \infty)$, (b) Bonferroni procedure, and (c) Benjamini-Hochberg procedure.

(a) Unadjusted

	Null ($\mathcal{T}$)	Alt. ($\mathcal{F}$)	Total	Row Pct. Null	Row Pct. Alt.
Reject ($\mathcal{R}$)	50	123	173	$0.289 = FDP$	0.711
Inconclusive ($\mathcal{I}$)	800	27	827	0.967	$0.033 = FNP$
Total	850	150	1,000	$0.85 = \pi_0$	$0.15 = \pi_1$

(b) Bonferroni

	Null ($\mathcal{T}$)	Alt. ($\mathcal{F}$)	Total	Row Pct. Null	Row Pct. Alt.
Reject ($\mathcal{R}$)	1	10	11	$0.091 = FDP$	0.909
Inconclusive ($\mathcal{I}$)	849	140	989	0.858	$0.142 = FNP$
Total	850	150	1,000	$0.85 = \pi_0$	$0.15 = \pi_1$

(c) Benjamini-Hochberg (BH)

	Null ($\mathcal{T}$)	Alt. ($\mathcal{F}$)	Total	Row Pct. Null	Row Pct. Alt.
Reject ($\mathcal{R}$)	5	55	60	$0.083 = FDP$	0.917
Inconclusive ($\mathcal{I}$)	845	95	940	0.899	$0.101 = FNP$
Total	850	150	1,000	$0.85 = \pi_0$	$0.15 = \pi_1$

2.2.1 Family wise error rate and adjusted p-values

Perhaps the most well-known multiple testing error metric is the family-wise error rate (FWER), which is the probability of making at least one false rejection, written as $P(V \geq 1) = 1 - (1 - \alpha)^m$ under the usual simplifying assumptions. It was originally referred to as the “experiment-wise error rate” or “error rate per-experiment” (Tukey 1953; Ryan 1959) and is also sometimes called the “group-wise error rate”. Here α is the “per-comparison error rate”. If we perform m significance tests, each having a per-comparison Type I Error rate of 5%, then the family-wise error rate grows large quickly as the number of tests grow. For example, performing 45 tests with individual Type I Error of $\alpha = 0.05$ will result in an overall FWER of 0.9. Large-scale biomedical data often has tens of thousands or hundreds of thousands of tests. In these settings, FWER adjustments incur a very large Type II Error rate penalty, and this makes the FWER approach much less desirable because the power is reduced dramatically (Brown and Russell 1997; Perneger 1998).

In general, we define a "multiple testing procedure" as a mapping from a list of p-values $(p_1, \dots, p_m)$, or z-values $(Z_1, \dots, Z_m)$, to a list of rejected and not rejected null hypotheses. This mapping can be done to meet any constraint, on any error metric, related to the group of experiments. The most popular multiple testing procedure for controlling the FWER at level α^* is the Bonferroni procedure, which sets the p-value rejection threshold at α^*/m. The Bonferroni threshold for controlling the FWER at 5% in the SNP example is $0.05/224{,}866 = 2.2 \times 10^{-7}$ and in the simple simulation is $0.05/1{,}000 = 5 \times 10^{-5}$. Because Bonferroni uses a p-value threshold that is much more severe than the original unadjusted procedure, much fewer hypotheses are rejected. In the SNP example we reject the null hypothesis for only 16 of the tests compared to $10{,}637$ rejected tests with the unadjusted procedure. The Bonferroni procedure is an approximation to the Šidák procedure, both of which are examples of "fixed bounds", that is the bounds are comprised of known quantities or quantities specified a priori (e.g., α^* and m). Other FWER-controlling procedures are dependent upon the observed data (via p_i, z_i, or even the original data) and are known as "adaptive procedures". Holm's procedure for controlling the FWER is an example of an adaptive procedure. Adaptive FWER procedures are sometimes classified according to the order in which tests are compared, e.g., starting with the smallest p-value and working backwards ("step-down") or starting with the largest p-value and working forwards ("step-up").

A common misconception is that a procedure designed to control the FWER at 5% will have a FWER equal to 5%. Controlling the family-wise error rate means that it has an upper bound, not that it is exactly equal to the pre-specified rate. Several factors (such as correlation among the tests and the proportion of hypotheses that are truly null) can result in a much lower FWER. This can be seen in our underdispersed simulation example where the Bonferroni procedure results in a FWER of 0.3%, substantially smaller than the apparent 5% level.

A helpful concept is that of an adjusted p-value. It can allow for a much simpler rejection scheme when the adjusted p-value is well defined. For example, we reject hypothesis H_i if the adjusted p-value $\tilde{p}_i$ is less than 0.05. In the case of the Bonferroni procedure, the adjusted p-value is $\tilde{p}_i^{\mathrm{Bonf}} = \min(p_i \cdot m, 1)$. Multiple testing algorithms can be quite complex, and the derivation of the adjusted p-value is not often straightforward. Shaffer (1995) provides the following general definition: "Given any test procedure, the adjusted p-value corresponding to the test of a single hypothesis H_i can be defined as the level of the entire test procedure at which H_i would just be rejected, given the values of all test statistics involved." We will use this idea for false discovery quantities as well.

2.2.2 False discovery proportion

In large-scale settings, where false positive findings are virtually guaranteed and inference is typically more exploratory, it makes more sense to try and control the rate of false positives. That is, rather than using a (controversial) statistical adjustment to try and prevent all false rejections, the investigator should simply estimate how **many** false rejections V might have occurred in the set of m significance tests or the proportion of false rejections V/R. The "False Discovery Proportion" (FDP) is a good starting point for false discovery concepts and is loosely defined as the proportion of false discoveries that were observed. Formally, the FDP is

$$FDP \coloneqq Q = \frac{V}{\max(R,1)} = \begin{cases} V/R & \text{when } R > 0 \\ 0 & \text{when } R = 0 \end{cases}. \tag{2.1}$$

This composite definition is necessary because when no rejections occur ($R = 0$), there can be no false rejections, so V/R is technically undefined. The "False Non-Discovery Proportion" (FNP) is an analogous quantity for the rate of missed discoveries defined as $(m_1 - U)/\max[(m - R), 1]$. Both quantities assess the degree to which an observed rejection or non-rejection is reliable. Ideally, both would be reported in large-scale experiments where estimation of these false discovery quantities is possible.

To fix ideas, we simulated large-scale data and plotted the joint distribution of R and V (Figure 2.2(a)) and the distribution of the false discovery proportion Q (Figure 2.2(b)) for the fixed z-value rejection region of $[1.64, \infty)$. The simulated false discovery proportion ranged from 0.134 to 0.386 and was $\bar{Q} = 0.267$ on average. The observed false discovery proportion is a random variable, however properties of Q such as its expected value are not. The average FDP is called the False Discovery Rate, and an array of methods for controlling it exist. This is described in detail in the next section.

2.2.3 False discovery rate

Benjamini and Hochberg (1995) showed how to control the expected value of the false discovery proportion, which they named the False Discovery Rate (FDR). The idea was inspired by Sorić (1989), and the proposed procedure is a simple modification of (Simes 1986). Benjamini and Hochberg's proposal was considered revolutionary because it abandoned control of the FWER, in favor of FDR control, in order to increase power in multiple comparisons settings. The FDR, typically denoted by Q_e, can be expressed as

$$\text{FDR} \coloneqq Q_e = E[Q] = E\left[\frac{V}{R}\,\middle|\, R > 0\right] \cdot \Pr(R > 0). \tag{2.2}$$

We see that Q_e is equal to the scaled expected proportion of false discoveries when at least one rejection is observed. The scaling factor is the probability of at least one rejection, which quickly approaches one in medium to large data sets. Benjamini and Hochberg showed that controlling the false discovery rate Q_e is a more powerful procedure than controlling the FWER. Because Q is a random variable bounded by 0 and 1, the FDR is always less than or equal to the FWER ($E[Q] \leq \Pr(Q > 0) = \Pr(V \geq 1)$). They also noted that when all hypotheses are truly null, the FWER and FDR are equal ($E[Q] = \Pr(Q > 0) = \Pr(V \geq 1)$). Remark 2.D describes how to compute Q_e for procedures where the distribution of R and $V|R$ are known; otherwise, simulation of Q_e is straightforward.

Let the p-value order statistics be $p_{(1)} \leq \cdots \leq p_{(i)} \leq \cdots \leq p_{(m)}$. The Benjamini and Hochberg (BH) algorithm to control Q_e at level q is:

1) Define i_{max} to be the largest index for which $p_{(i)} \leq q \cdot i/m$, and
2) Reject all null hypothesis H_0^i corresponding to $p_{(i)}$ if $i \leq i_{max}$.

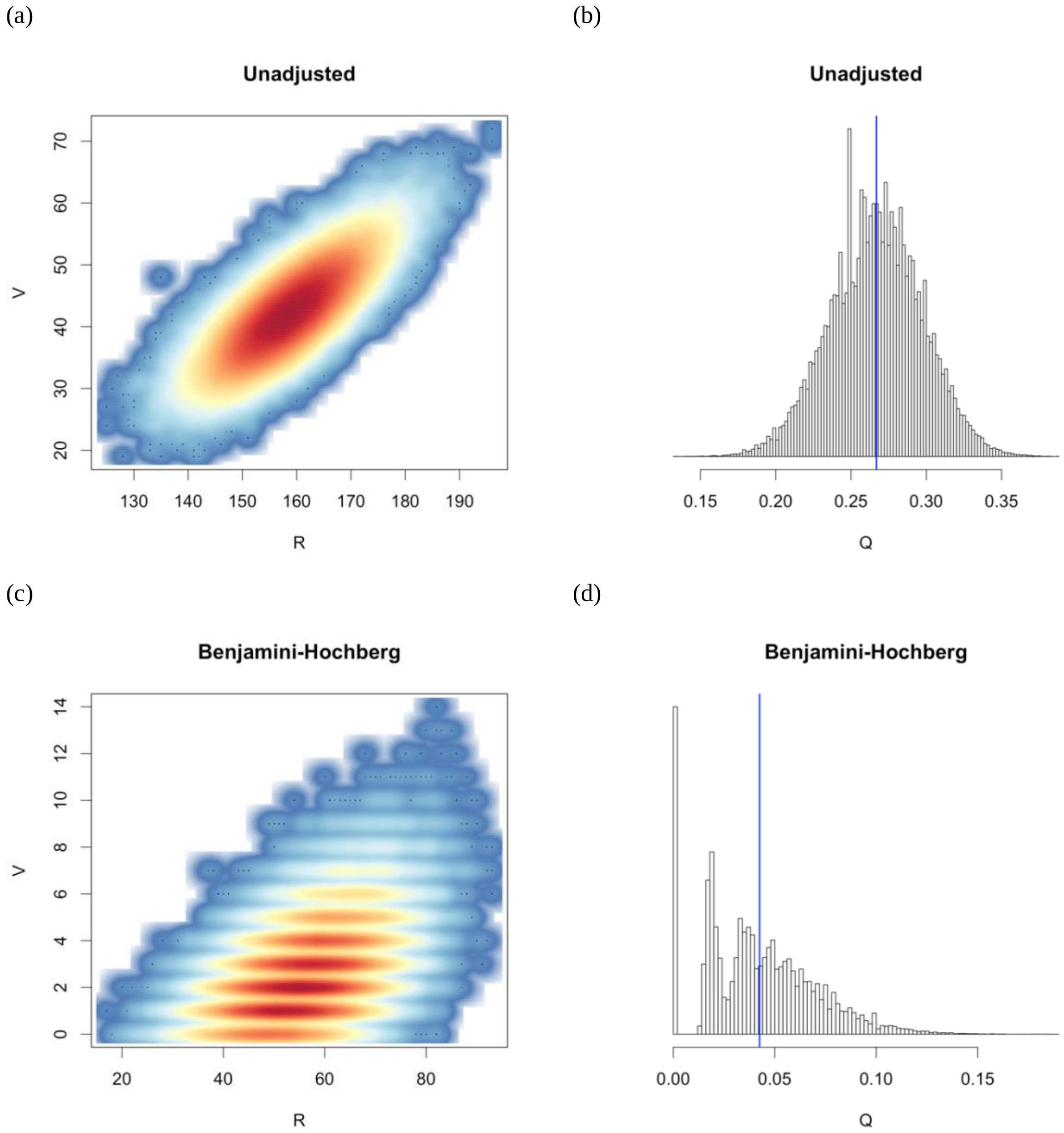


Figure 2.2 Distributions of R, V, and Q for the simple simulation from 100,000 simulation replications. $\bar{Q}$ is the mean across the simulated values of Q. (a) Joint distribution of R and V for the unadjusted z-value rejection region $[1.64, \infty)$, as a heat map. (b) Distribution of Q for the unadjusted rejection region. The vertical blue line is at $\bar{Q} = 0.267$. (c) Joint distribution of R and V for the BH procedure, as a heat map. (d) Distribution of Q for the BH procedure. The vertical blue line is at $\bar{Q} = 0.0425$.

Note that it is common to use q, instead of α or α^*, to denote the level of FDR control because the quantities are fundamentally different. We can see from Tables 2.2(b)-(c) that in the simple simulation illustration data, the BH procedure rejects more hypotheses than the strict Bonferroni procedure, without an increase in the FDP. An FDR-controlled procedure, however, controls only the average FDP and not the FDP itself. In this example, the BH procedure results in an FDP of $Q = 0.083$, despite that fact that the average Q is less than or equal to $q = 0.05$. Using the same repeated simulation data from Figures 2.2(a)-(b), the distributions of R, V, and Q for the BH procedure are provided in Figures 2.2(c)-(d). Note that V and R are discrete variables and this is apparent in Figure 2.2(c). Figure 2.2(d) shows how the BH procedure is controlling the false discovery proportion on **average**. The distribution is skewed towards 0, and the spike at $Q = 0$ is large enough to hold the average false discovery proportion (the FDR) below 0.05.

The per-comparison and family-wise error rates are related to the pre-test probability that an error will be made, while the FDR is related to the post-test probability that the observed test result is mistaken (See (Blume 2008, 2011) for further discussion; see also Section 2.3.2). The FDR is in fact controlled at level $\pi_0 q$, such that $Q_e \leq \pi_0 q \leq q$ (under certain conditions for dependency (Benjamini and Yekutieli 2001)). For example, if 85% of the hypotheses are null, then the FDR is controlled at 0.0425 when $q = 0.05$. Correlation among the tests can reduce the true Q_e further below the control level, and in the unusual case of negative dependence between tests, FDR control is not guaranteed with the standard BH procedure (Heesen and Janssen 2015). A modified procedure has been proposed which guarantees control under general dependence but can be quite conservative (Benjamini and Yekutieli 2001). It can be helpful to re-write the BH procedure as an adjusted p-value procedure, namely reject H_0^i when $\tilde{p}_{(i)}^{BH} \leq q$ where $\tilde{p}_{(i)}^{BH} = \min_{k \geq i}\left(p_{(k)}/(k/m)\right)$. The adjusted p-values for the Bonferroni and Benjamini Hochberg procedures are illustrated in Supplemental Figure 2.1. Additional discussion on FDR topics not covered in detail here can be found in the Appendix (Remarks 2.F – 2.I), such as the marginal false discovery rate $E[V]/E[R]$, conditional false discovery rate $E[V]/r$, and extensions such as k-FDR.

2.3 Estimating false discovery quantities

Care must be taken for the proper interpretation of false discovery rate quantities. A common interpretation of results from a false discovery rate procedure is:

> *"We used the Benjamini-Hochberg procedure to control the false discovery rate in 1,000 candidate tests. This resulted in 60 significant findings and we expect 5% of these findings to be false positives."*

While this interpretation seems natural, it is not necessarily correct. The phrase "we expect 5% of these findings to be false positives" refers to the observed quantity v/r. This quantity can be seen as a realization of the random variable V/R given that $R > 0$. However, the expected value of this conditional random variable, $E[V/R\,|R > 0]$, is not the FDR because the scaling factor $P(R > 0)$ is missing. Some have also argued that the quoted statement above refers to a realization v/r conditional on $R = r$ (Tsai et al. 2003; Pounds and Cheng 2004), but we remain unconvinced. The BH procedure controls only $E[V/R\,|R > 0] \cdot P(R > 0) = E[Q]$ which is taken over all values of R including $R = 0$; it does not control $E[V/R\,|R > 0]$ or $E[V/R\,|R = r]$. An extreme example is that the FDR can be 5% with a 50% chance of rejecting any hypotheses so that $\Pr(R > 0) = 0.5$ and $E[V/R|R > 0] = 0.1$. In this case the mistaken interpretation of the statement is very misleading (the correct rate is 10%, not 5%). Even when $P(R > 0)$ is near 1, one cannot interpret an observed realization as the expectation, just as we cannot interpret the p-value as the Type I error.

Storey (2003) defined $E[V/R|R > 0]$ to be the positive False Discovery Rate (pFDR) which in fact has a nice probabilistic interpretation. In effect, it answers the question "*How reliable are the findings?*" This is a different question from asking "*How reliable is my study design?*" for which we use Type I and Type II error rates to answer that question (Blume 2008, 2011). Notice that the FDR is a scaled version of the pFDR, and that the FDR is a lower bound for the pFDR. In large-scale settings, they are frequently identical because typically $P(R > 0) \approx 1$. However, we keep focus on the pFDR because formally is the more relevant quantity. What the above means is that: (1) we should estimate $E[V/R|R > 0]$ because that is more relevant after an experiment is conducted and we wish to assess the reliability of findings, and (2) for experiments with a high probability of discovery, $E[V/R|R > 0]$ will be close to the FDR (and the conditional FDR), approaching from above.

2.3.1 The positive false discovery rate

Although formally named in (Storey 2003), the positive false discovery rate was discussed in (Benjamini and Hochberg 1995) but dismissed because it is strictly not controllable. By definition, pFDR = 1 when $\pi_0 = 1$ (all hypotheses are null). Storey reasons that "when $m_0 = m$, one would want the false discovery rate to be 1" and when no tests are significant, no one is interested in the false discovery rate anyway. Therefore, if we are interested in estimating false discovery quantities, this issue is not relevant and the pFDR is a natural statistic to focus on.

The pFDR, indeed all false discovery quantities, correspond to a particular rejection region. If we define Γ to be the rejection set for the z-value space such that we reject H_0^i if $z_i \in \Gamma$, then the quantity R is shorthand for the more explicit $R(\Gamma)$ which is the number of rejections R that result from rejecting H_0^i when $Z_i \in \Gamma$. Similarly, for V and $V(\Gamma)$. The more explicit definition for pFDR is

$$pFDR(\Gamma) \coloneqq E\left[\frac{V(\Gamma)}{R(\Gamma)}\,\middle|\, R(\Gamma) > 0\right], \tag{2.3}$$

with $pFDR(\Gamma) = 0$ when $R = 0$. We will also use the shorthand notation $pFDR(\Gamma) \coloneqq E[V/R\,|R > 0; \Gamma]$. The rejection region Γ need not be fixed ahead of time to yield proper estimates of the pFDR (Storey et al. 2004). For simplicity, we will consider only rejection regions of the one-sided tail-area form $\Gamma = [z, \infty)$ for definitions and discussion, although all results can be easily extended to two-sided rejection regions and some illustrations will show the results based on two-sided rejection regions. We will also make use of the further shorthand $pFDR(z) \coloneqq pFDR\big(\Gamma = [z, \infty)\big)$. The pFDR has a natural Bayesian interpretation and various techniques for estimating the pFDR make use of this connection.

2.3.2 Bayesian interpretation of the pFDR

The "Bayes false discovery rate" has been introduced as a probabilistic form of the positive false discovery rate (Efron, Storey, et al. 2001; Efron, Tibshirani, et al. 2001; Storey 2001a, 2001b, 2002, 2003). The key insight is to regard the truth of each null hypothesis as a random variable, with $H_i \sim Bernoulli(1-\pi_0)$ where $H_i = 0$ indicates that the null hypothesis is true. Here π_0 is the prior probability that the null hypothesis is true, i.e., $\pi_0 = \Pr\,(H = 0)$, rather than the observed proportion of hypotheses that are null. The Bayes FDR is defined as

$$\phi(\Gamma) \coloneqq \Pr(H = 0 | Z \in \Gamma). \tag{2.4}$$

Additionally, the distribution of the z-values is a mixture, $F(z) = \pi_0 \cdot F_0(z) + (1-\pi_0) \cdot F_1(z)$. Here $F_0(z)$ is the distribution of null z-values and $F_1(z)$ is the distribution of non-null z-values (with corresponding probability density functions $f_0(z)$, $f_1(z)$, and $f(z) = \pi_0 \cdot f_0(z) + (1-\pi_0) \cdot f_1(z)$). Efron (2008) discusses this so-called "two-groups model" and its properties. Under the assumptions outlined above, Storey (2003) showed that for m identical hypothesis tests with z-values $Z_1, \dots, Z_m$, significance region Γ, and for (Z_i, H_i) independent and identically distributed random variables,

$$E\left[\frac{V}{R} | \, R > 0; \Gamma\right] = Pr(H = 0 | Z \in \Gamma). \tag{2.5}$$

That is, Storey showed that $pFDR(\Gamma) = \phi(\Gamma)$.

We can apply Bayes' rule to re-express the Bayes FDR for a general rejection region Γ that can be any subset of the real line as

$$\phi(\Gamma) \coloneqq \Pr(H = 0 | Z \in \Gamma) = \frac{\Pr(H = 0) \cdot \Pr(Z \in \Gamma | H = 0)}{\Pr(Z \in \Gamma)} = \frac{\pi_0 \cdot F_0(\Gamma)}{F(\Gamma)}, \tag{2.6}$$

where the notation of the distribution function F and F_0 of a region Γ is used to mean $F(\Gamma) = \Pr\,(Z \in \Gamma) = \int_\Gamma f(Z) dZ$ and $F_0(\Gamma) = \Pr(Z \in \Gamma | H = 0) = \int_\Gamma f_0(Z) dZ$, respectively. In this notation, the (proper) cumulative distribution function $F(c)$ is $F\big((-\infty, c]\big) = \Pr(Z \leq c)$. As we will see later, the form in Equation (2.6) is helpful for estimating the Bayes FDR. It is called the "global FDR" because the conditioning event is $Z \in \Gamma$, whether the test statistic is in the rejection region. As a result, both $F_0(\Gamma)$ and $F(\Gamma)$ are tail area probabilities for typical cases such as $\Gamma = [c, \infty)$. Conditioning on just the observed test statistic, $Z = z_i$, leads to a "local FDR" that we introduce later. The global FDR quantity is very intuitive; it is the scaled ratio of the probability of a null test statistic rejecting to the probability of any test statistic rejecting. We can get a simple estimate of this quantity by using the plug-in principle with assumptions dictating $F_0(\Gamma)$, an empirical estimate of $F(\Gamma)$, and setting $\pi_0 = 1$ to be conservative. Section 2.3.5 describes this and other estimation approaches.

2.3.3 q-Value

The q-value was introduced to mimic the p-value in posterior space, by taking into account "more extreme" results. The q-value is defined as

$$q(z_i) \coloneqq \min_{c \leq z_i} pFDR\big([c,\infty)\big) = \min_{c \leq z_i} Pr\big(H = 0 \big| Z \in [c,\infty)\big) \tag{2.7}$$

for $z_i \in (z_1, \dots, z_m)$. For our purposes, we assume nested one-sided upper tail-area rejection regions which yields this simpler definition; a more general definition of the q-value is given in (Storey 2003). In this context, q-values are pFDR estimates that are modified to ensure monotonicity in z-value space. This is illustrated in Supplemental Figure 2.2. In practice, using a plug-in estimator, the monotonicity is forced directly upon the estimated pFDR by defining the estimated q-value $\hat{q}(z_i) \coloneqq \min_{c \leq z_i} \widehat{pFDR}\big([c,\infty)\big)$. Storey and Tibshirani (2003) explain that the estimated q-value can be used as "a measure of each feature's significance, automatically taking into account the fact that thousands are simultaneously being tested". Further, they argue that if features with estimated q-values less than a level q are taken to be significant, then $q \times 100\%$ of them are expected to be false discoveries. However, this remains to be proven for all classes of q-value estimates. In some situations, using such a q-value procedure would result in a more powerful FDR-controlled procedure than if the pFDR was used. There are also a variety of situations where the q-value does not add anything beyond the pFDR. We find that it is better to stick with the original pFDR estimate which is a more "natural" quantity that is easier to use and interpret.

2.3.4 Local FDR

The global FDR conditions on the rejection of a null hypothesis. We might also consider conditioning on the observed significance level of the test and the resulting FDR is called a local FDR. The global FDR assumes the results are only parsed as "rejected" or "not rejected", while the local FDR assumes the actual significance level for each test was communicated. Both FDRs are useful in practice as results are often communicated in different ways. Efron, Tibshirani, et al. (2001) define the "local false discovery rate" as

$$\phi_l(z_i) \coloneqq \Pr(H = 0 | Z = z_i) = \frac{\pi_0 \cdot f_0(z_i)}{f(z_i)}. \tag{2.8}$$

Note that this is a ratio of densities ($f(z)$ and $f_0(z)$) rather than a ratio of distribution functions. Local inference is possible in large-scale settings via empirical Bayes methods, but tends to be unstable in smaller samples because the density estimates are noisy. Technically, the local FDR is a local version of the Bayes FDR (a pFDR quantity), not the classic FDR $E[Q]$. In large samples however there is virtually no distinction and the literature follows this convention. More on the local FDR can be found in (Efron, Tibshirani, et al. 2001; Efron and Tibshirani 2002; Efron 2010b).

2.3.5 Estimation approaches

The Bayes false discovery rate $\phi(\Gamma)$, from Equation (2.6), can be computed without difficulty when $F_0(z)$ and $F_1(z)$ are known. π_0 can be set to 1 to provide a conservative upper bound for the FDR. While $F_0(z)$ often takes an assumed form under the null, $F_1(z)$ is typically left unspecified. We can skip estimation of $F_1(z)$ because we only need to know the mixture distribution $F(z)$, which can be estimated directly from the observed data (e.g., a histogram of the z-values or smooth version of the same). Typically, we might assume that $F_0(z) = N(0,1)$, but there are circumstances where this does not necessarily hold, e.g., with correlation among the z-values (Efron 2008). So, it is important to have methods that relax this assumption in our toolbox.

A wide variety of methods have been proposed for estimating the false discovery components. Here we focus our attention on methods described in (Efron 2010b) that have the benefits of working in the unconstrained z-value space, estimating the mixture distribution directly and flexibly, and are simple and straightforward to implement. Other approaches are worth exploring. Many rely on parametric estimation of the alternative distribution F_1 or on non-parametric estimation of the alternative or mixture distribution (Allison et al. 2002; Pounds and Morris 2003; Aubert et al. 2004; Liao et al. 2004; Pounds and Cheng 2004; Tang et al. 2007). Some discussion of these approaches is in (Tang et al. 2007). These methods can perform well under certain conditions, although some are less robust or more complex than the approaches described here.

2.3.5.1 The ECDF estimate

A simple non-parametric estimate of the cumulative distribution function comes from the ECDF of the observed z-values, $\hat{F}(c) = \#\{j: z_j \leq c\}/m$ for $j \in (1, \dots, m)$ (where $\#\{A\}$ denotes the number in set A). For a general region Γ, we define the ECDF-type estimator $\hat{F}(\Gamma) = \#\{j: z_j \in \Gamma\}/m$ for $j \in (1, \dots, m)$. For upper-tail rejection regions with $\pi_0 = 1$ assumed, we get an estimated global FDR of

$$\hat{\phi}\big([z_i, \infty)\big) = \frac{\pi_0 \cdot F_0\big([z_i, \infty)\big)}{\hat{F}\big([z_i, \infty)\big)} = \frac{1 \cdot [1 - \Phi(z_i)]}{i/m}, \tag{2.9}$$

where $\Phi(z)$ is the standard Normal distribution and the denominator is simplified under the assumption that the $(z_1, \dots, z_m)$ are the **reverse** order statistics for the z-values $z_1 \geq z_2 \geq \cdots \geq z_m$ such that $\#\{j: z_j \geq z_i\} = i$. Some estimators exceed 1 because of the plug-in estimation strategy and the conservative bounding of π_0. When $\hat{\phi}(z_i) > 1$ it is standard practice to truncate the estimate at 1. Technical notes are provided in the Appendix (Remark 2.O). This estimator is easy to use and apply. Figure 2.3(a) shows the true Bayes pFDR of the underdispersed simulation along with the estimate from Equation (2.9). The FDR estimate does not match the true FDR curve well (even when the true value for π_0 of 0.9 is used for Equation (2.9)) because the assumed theoretical null does not hold. The estimator applied to the SNP example (Figure 2.3(b)) is a good illustration of how non-smooth the ECDF estimator can be, and how the q-values could be helpful in such a situation. However, the q-values are less useful in Figure 2.3(a), where the roughly bimodal nature of the curve is due to erroneous assumptions about the null.

An important connection to note is that the BH adjusted p-value is equal to the estimated q-value when using Equation (2.9). That is, $\hat{q}(z_i) = \min_{z \leq z_i} \hat{\phi}(z) = \tilde{p}_i^{BH}$ for $\hat{\phi}(z)$ estimated by Equation (2.9). This derivation is given in the Appendix (Remark 2.P). The quantities are illustrated in Figure 2.4 for the simple simulation, which also helps to illustrate the distinction between the pFDR and q-value. This connection sometimes encourages BH adjusted p-values to be mistaken for FDRs. However, q-values are actually thresholded FDRs, and this equality only holds for BH derived q-values. In this setting rejecting features with q-values less than or equal to pre-specified

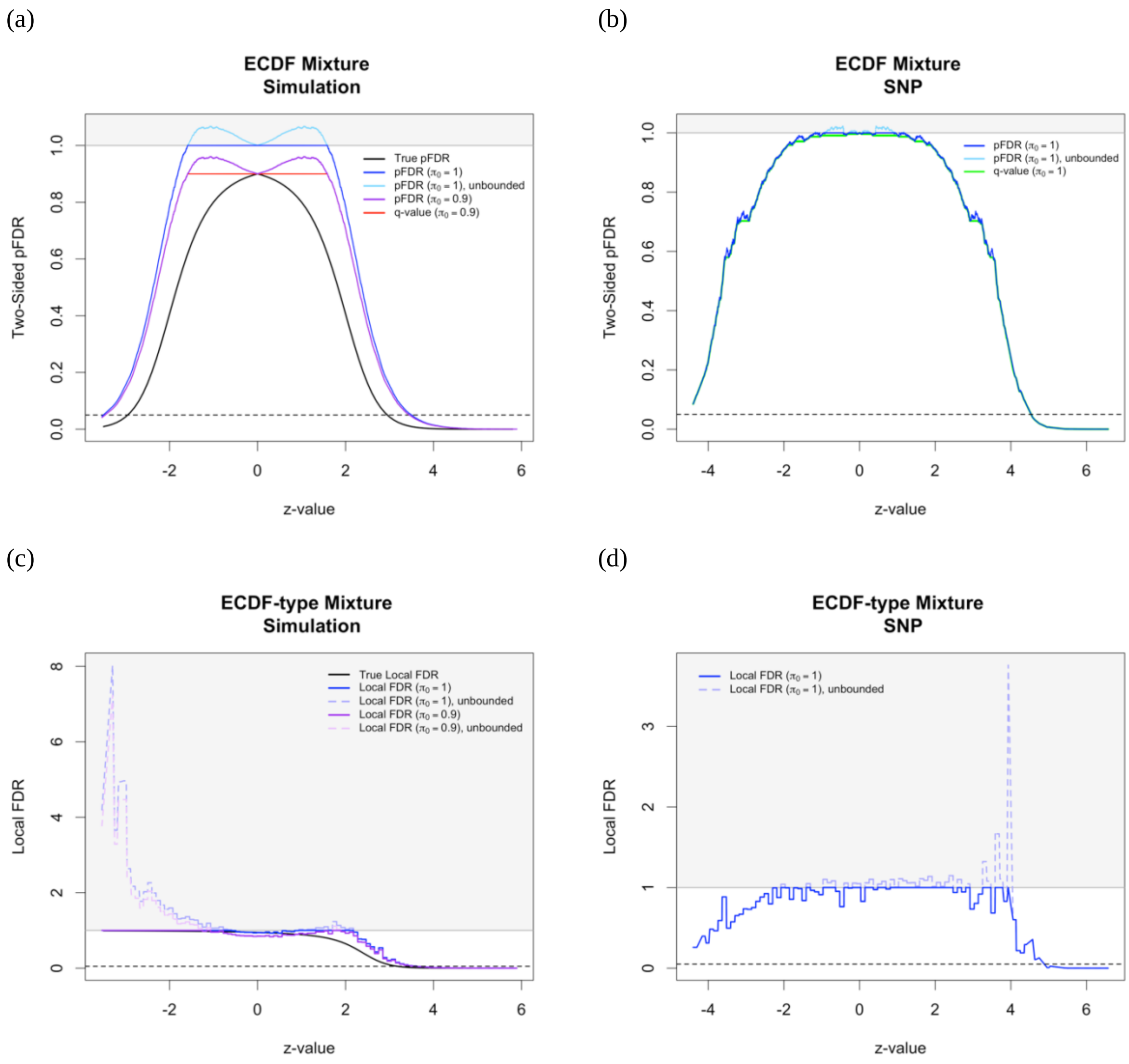


Figure 2.3 Estimated FDR quantities using the theoretical null and ECDF-based estimators (Equation (2.9) and corresponding local version). The two-sided pFDR is shown instead of the upper tail pFDR. When applicable, the original unbounded estimates are shown and the $FDR > 1$ region is shaded in grey. For the underdispersed simulation, the estimates using the true $\pi_0 = 0.9$ in place of $\pi_0 = 1$ are also provided. The horizontal black dashed line is at 0.05. (a) True and estimated pFDR (and q-values) for the underdispersed simulation, (b) Estimated pFDR (and q-values) for the SNP example, (c) True and estimated local FDR for the underdispersed simulation, (d) Estimated local FDR for the SNP example.

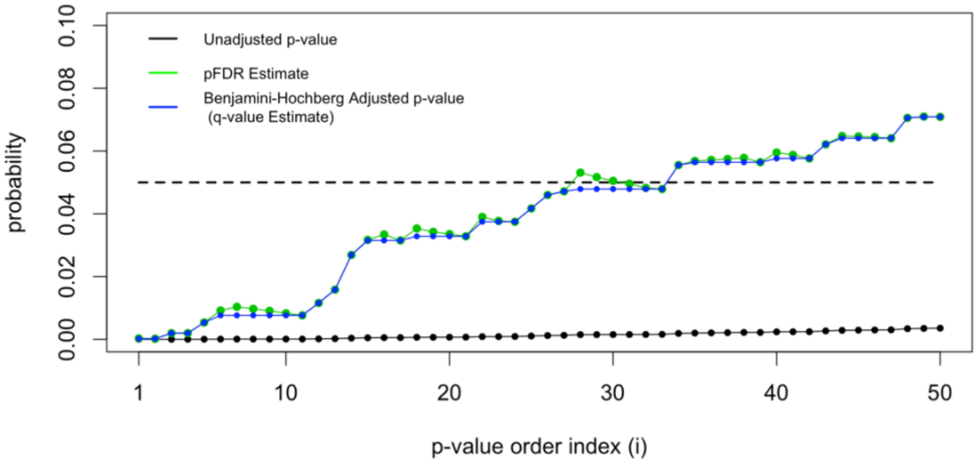


Figure 2.4 Connection between the Benjamini-Hochberg procedure and the pFDR and q-value. The pFDR estimated as described in Section 2.3.5.1 and the Benjamini-Hochberg adjusted p-value, which is equivalent to the q-value calculated from the pFDR using ECDF mixture estimates, for the simple simulation. The unadjusted p-values are also shown, along with an example 0.05 cutoff as the horizontal dashed black line.

level q would result in an FDR-controlled procedure because it is equivalent to the BH procedure. However, when q-values are derived by other means, FDR control does not necessarily follow (it must be evaluated on a case-by-case basis).

Estimation of the local FDR requires an estimate of the mixture density rather than the mixture distribution function, and a similar non-parametric approximation can be obtained by discretizing the z-values into bins (Efron 2010b); the details of this estimate are provided in the Appendix (Remark 2.Q). The resulting estimates are shown in Figures 2.3(c)-(d). These estimates are highly locally variable, particularly in areas where there is little information (i.e., few z-values) such as in the left tail of the distribution in the simulation or the right tail of the distribution in the SNP example. However, in our simulation, we see that they still tend to follow the true local FDR curve. Because of the instability and resulting local bias, this estimator is not recommended for use in practice, but is helpful for comparison purposes.

2.3.5.2 Lindsey's method

Smoothing the empirical mixture z-value distribution can help reduce the volatility of ECDF-based methods. Here we use Lindsey's method because it is flexible and relatively straightforward. Briefly, Lindsey's method uses the observed histogram of z-values to model the height of each bin, y_k, as a flexible function of the bin center point x_k (for example, as a J-order polynomial or spline function) using a Poisson count model (Lindsey 1974; Efron and Tibshirani 1996). The estimated mixture density is calculated by $\hat{f}(z) = \hat{y}(z)/(2\delta \cdot m)$ where 2δ is the histogram bin width and $\hat{y}(z)$ is the Poisson-model predicted count at z. Further details are given in the Appendix (Remark 2.R) and the fitted distributions corresponding to various values of J and δ are shown in Supplemental Figure 2.6.

This approach typically provides a nice smooth fit to the z-value distribution, although sometimes care must be taken with choosing the flexibility parameter J.

Using this approach, the estimate of the pFDR for the upper tail rejection region is

$$\hat{\phi}\big([z_i,\infty)\big) = \frac{\pi_0 \cdot F_0\big([z_i,\infty)\big)}{\int_{z_i}^{\infty} \hat{f}(z)dz} = \frac{1 \cdot [1 - \Phi(z_i)]}{\int_{z_i}^{\infty} \hat{f}(z)dz}, \tag{2.10}$$

assuming that $\pi_0 = 1$ and the theoretical null holds. The corresponding local FDR estimate is $\hat{\phi}_l(z_i) = \pi_0 \cdot f_0(z_i)/\hat{f}(z_i) = 1 \cdot f_0(z_i)/\hat{f}(z_i)$. Figures 2.5(a)-(d) show these pFDR and local FDR estimates for the underdispersed simulation and SNP examples, overlaid with the previous ECDF-based estimates from Equation (2.9). Lindsey's method yields a nice smooth pFDR estimate that is similar to the Equation (2.9) estimate but does not suffer from high local variability. However, smoothing does not address the irregular shape and bias of the pFDR curve in the simulation (Figure 2.5(a)), because this is partly due to the violation of the theoretical null assumption. For this, we will turn to estimating the null distribution in Section 2.3.5.4. As before, q-values are not particularly useful in this simulation, and in the SNP example (Figure 2.5(b)), q-values do not add anything beyond the pFDR estimates because the smoothed mixture results in a smooth and unimodal pFDR estimate (after bounding at 1).

2.3.5.3 Estimating the null proportion

Rather than assuming the conservative bound of 1 for π_0, many methods have been proposed for estimating π_0 (or m_0). While these might provide less biased FDR estimates, they increase the estimation variance. Benjamini and Hochberg originally formalized a graphical approach proposed in (Schweder and Spjøtvoll 1982) for estimating m_0, resulting in more powerful procedures for FWER control (Hochberg and Benjamini 1990) and FDR control (Benjamini and Hochberg 2000). Since then, many other methods have been described, including in (Storey and Tibshirani 2003; Efron 2010b; Murray and Blume 2021). Illustration of these selected estimation approaches are given in the Appendix (Remark 2.T). Some other methods proposed include, but are not limited to, (Pounds and Morris 2003; Broberg 2004; Langaas et al. 2005; Nettleton et al. 2005; Pawitan et al. 2005; Meinshausen 2006; Jiang and Doerge 2008). Review of some methods can be found in (Broberg 2005; Kang 2020). Any of these $\hat{\pi}_0$ estimates may be substituted for π_0 in Equations (2.9) or (2.10), or their local FDR counterparts. In Section 2.3.5.4, we describe an estimation process for the null distribution, which conveniently incorporates π_0 estimation.

2.3.5.4 The empirical null

Under the null, the assumption is that observed z-values are distributed as $N(0,1)$. However, this assumption may be violated. For example, correlation among tests (e.g., among SNPs or imaging voxels) under the null would not lead to a $N(0,1)$ mixture (it would be under- or over-dispersed). Efron proposes a direct estimation approach, assuming the null distribution has the form $f_0(z) = N(\delta_0, \sigma_0)$. Two methods to jointly estimate the set of parameters

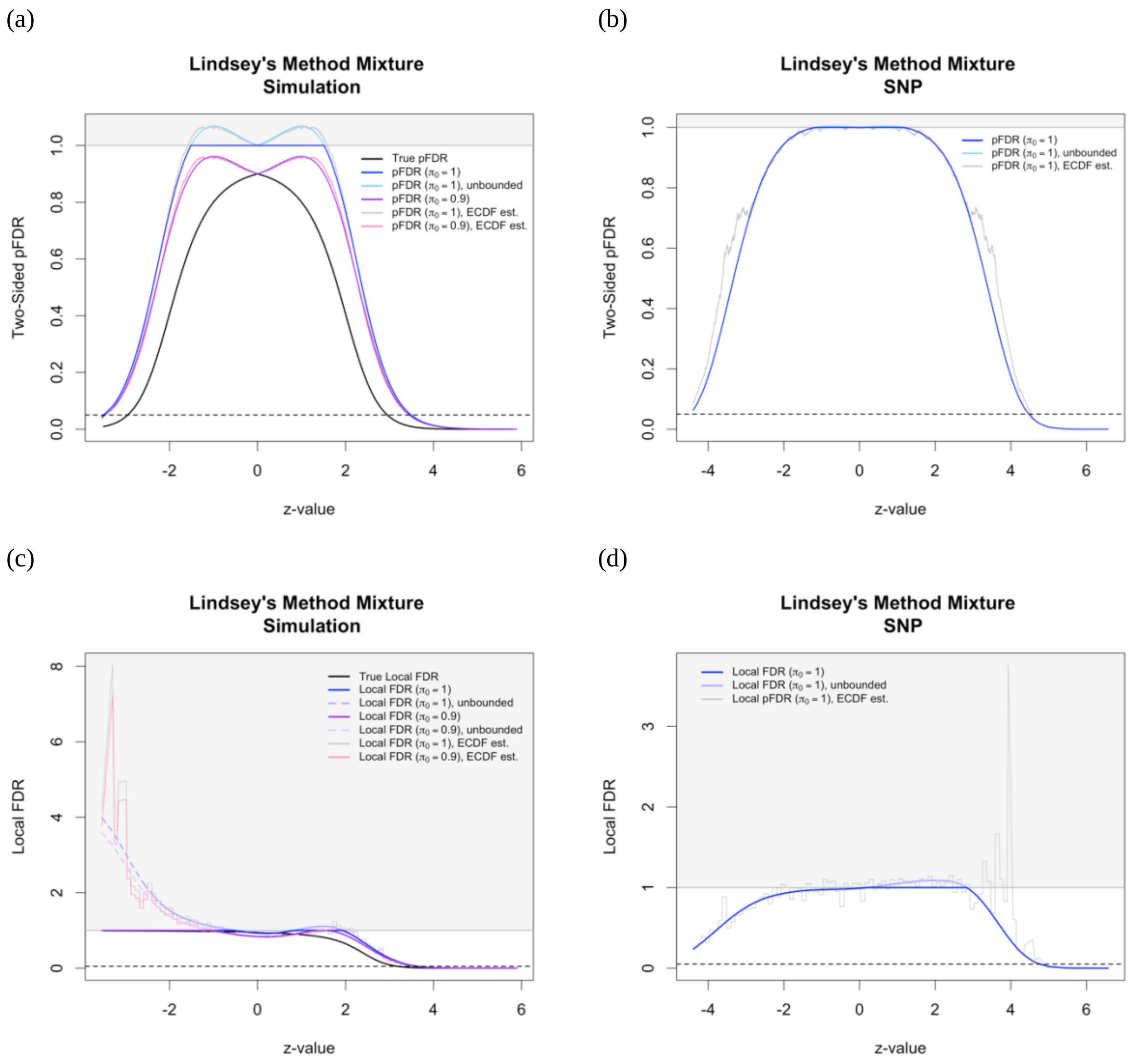


Figure 2.5 Estimated FDR quantities using the theoretical null and Lindsey's method estimators (Equation (2.10) and corresponding local version). The previous ECDF-based estimates are also given for comparison. The two-sided pFDR is shown instead of the upper tail pFDR. When applicable, the original unbounded estimates are shown and the $FDR > 1$ region is shaded in grey. For the underdispersed simulation, the estimates using the true $\pi_0 = 0.9$ in place of $\pi_0 = 1$ are also provided. The horizontal black dashed line is at 0.05. (a) True and estimated pFDR for the underdispersed simulation, (b) Estimated pFDR for the SNP example, (c) True and estimated local FDR for the underdispersed simulation, (d) Estimated local FDR for the SNP example.

$(\delta_0, \sigma_0, \pi_0)$ have been described, the central matching (CM) method and the maximum likelihood (ML) method (Efron 2004, 2007b, 2010a). These methods are outlined in the Appendix (Remark 2.U). Efron concludes that the maximum likelihood method tends to produce estimates that are less variable but more prone to bias than those from the central matching approach. In our underdispersed simulation, the CM estimates are $\hat{f}_0(z) = N(0.04, 0.87^2)$ with $\hat{\pi}_0 = 0.93$ and the ML estimates are $\hat{f}_0(z) = N(0.03, 0.9^2)$ with $\hat{\pi}_0 = 0.96$. Recall that the true null is $N(0, 0.9)$ and $\pi_0 = 0.9$ in this simulation. The resulting estimates from the SNP example are $\hat{f}_0(z) = N(-0.04, 0.98^2)$ with $\hat{\pi}_0 = 0.99$ for the CM approach and $\hat{f}_0(z) = N(-0.015, 0.91^2)$ with $\hat{\pi}_0 = 0.93$ for the ML approach. A visual illustration of the different fits is given in the Appendix (Remark 2.V) to help fix ideas.

Using an empirical null distribution gives the general empirical Bayes estimator for the upper tail pFDR,

$$\hat{\phi}\big([z_i, \infty)\big) = \frac{\hat{\pi}_0 \cdot \hat{F}_0\big([z_i, \infty)\big)}{\int_{z_i}^{\infty} \hat{f}(z)dz} = \frac{\hat{\pi}_0 \cdot \big[1 - \hat{F}_0(z_i)\big]}{\int_{z_i}^{\infty} \hat{f}(z)dz}, \tag{2.11}$$

where $\hat{F}_0(z)$ is $N\big(\hat{\delta}_0, \hat{\sigma}_0\big)$ with $\big(\hat{\delta}_0, \hat{\sigma}_0, \hat{\pi}_0\big)$ estimated from either the central matching or maximum likelihood method. Here the mixture density $\hat{f}(z)$ is estimated via Lindsey's method, however any approach may be used in combination with an empirical null estimate. The corresponding local FDR estimator is $\hat{\phi}_l(z_i) = \hat{\pi}_0 \cdot \hat{f}_0(z_i)/\hat{f}(z_i)$. The resulting pFDR and local FDR estimates for the underdispersed simulation and SNP examples are provided in Figures 2.6(a)-(d) for both empirical null approaches. While the null distribution estimates are not exact matches to the true null distribution in the simulation, the empirical null does capture the underdispersion and helps correct much of the bias in the pFDR curve induced by using the $N(0,1)$ null. We see that in the right tail of the local FDR curve in Figure 2.6(c), the estimates match the true FDR closely, but are unstable in the left tail of the curve. Because the local FDR estimate occurs at a single point (z-value) rather than averaging over a tail area, it is much more sensitive to the estimators that it is comprised of, particularly when a flexible non-parametric mixture estimator such as Lindsey's method is used. In the case of the SNP example, we do not know the true pFDR curve and so it is not known which of the empirical null or theoretical null estimates are a better estimate. However, Figure 2.6(d) illustrates that the local FDR estimate does not always suffer from instability issues. While the empirical null can be important in obtaining correct FDR estimates, we restate that the estimated null parameters can be prone to bias and thus care must be taken when considering this approach, particularly when the number of tests m is smaller.

2.4 Discussion

2.4.1 FDR testing and estimation

There are a number of benefits and drawbacks of FDR controlling procedures and of FDR estimation approaches. When all underlying necessary assumptions are met, we can achieve strict control of the false discovery rate, which may be important in some circumstances. Control of the false discovery rate allows for more potential discoveries than control of the family wise error rate, and control at the 5% level, for example, continues to provide a reasonable rate of errors in many large-scale inference contexts. A wide array of methods for FDR control have been proposed with well-established properties, and some of which are more robust to violations of assumptions such as

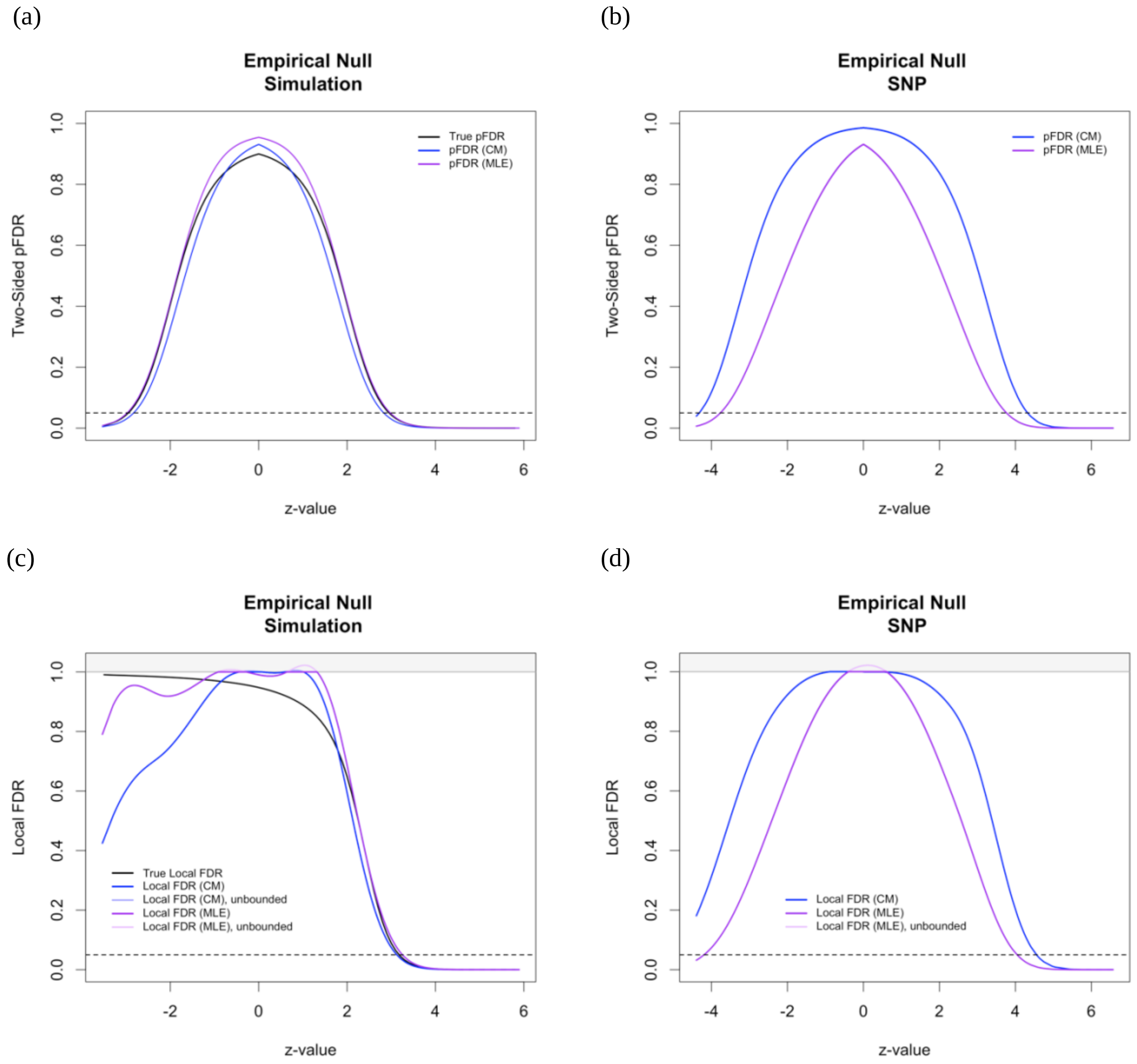


Figure 2.6 Estimated FDR quantities using the empirical Null distribution with Lindsey's method mixture estimate (Equation (2.11) and corresponding local version). The two-sided pFDR is shown instead of the upper tail pFDR. Both the maximum likelihood and central matching empirical null methods are shown. When applicable, the original unbounded estimates are shown and the $FDR > 1$ region is shaded in grey. The horizontal black dashed line is at 0.05. (a) True and estimated pFDR for the underdispersed simulation, (b) Estimated pFDR for the SNP example, (c) True and estimated local FDR for the underdispersed simulation, (d) Estimated local FDR for the SNP example.

independence among the tests. However, the prevailing drawback of FDR controlling algorithms is that they only focus on the overall FDR, i.e., the expected value of the random variable Q which also includes no rejections as no false rejections. For example, if we employ a procedure which controls the FDR at 5%, such as Benjamini-Hochberg, and this results in 100 rejected tests, can we say that we expect 5 or less of them to be false rejections? We have discussed that this is **not** the case; this is measured by the **positive false discovery rate**, and FDR control provides no information about what the pFDR is.

Unfortunately, it is not possible to develop a rejection procedure which controls the pFDR under all circumstances. However, the most important scientific question we would like the answer to is that which the pFDR answers. An analogy can be made for FDR control versus estimation with basic principles: a significance test results in a yes or no decision (reject or not reject), and has well-defined frequency properties, but we would then like to know the strength of evidence or the effect size. Similarly, control of the FDR with an appropriate procedure gives us a clear answer of reject or not reject, but we would like to know more – namely, the propensity for observed results to be misleading, or how many to expect to be true discoveries. Therefore, we may employ a pFDR estimation method to approximate this answer. The probabilistic equivalent definition of the pFDR allows for an intuitive class of empirical Bayes methods for estimation. These methods may be flexible and more robust to deviations from some common assumptions. However, one drawback is that the focus remains still on the central tendency measure of the expected value of the false discovery proportion, and no attention is paid to the variance or minimum or maximum value. On average, a procedure may have a pFDR of 5%, but the distribution of V/R could still range from 0 to 1. This means that the proportion of false rejections in a particular observed set of rejections could still be very large, such as 90% or 100%, with some probability greater than 0, even if on average it is 5%. Extensions to the standard pFDR – which conditions on a global tail area rejection region – such as the local pFDR or FDR may be considered, although their estimation is more challenging, and the resulting estimates may be biased or highly variable and could result in misleading interpretations of findings. One important topic of note that was not covered is variance estimates for false discovery rate estimates, either global or local. Efron covers this in detail, describing both theoretical approximation and bootstrap-based variance estimators (Efron 2007a, 2010a, 2010b). Currently this issue is often ignored in the literature in much the same way that the variability of the p-value is ignored.

One drawback of standard false discovery rate approaches – for both FDR controlling and pFDR estimation – is that they rely on z-values or p-values, which confound effect size and standard error. See (Ploner et al. 2006; Stephens 2017) for alternative approaches. False discovery rates based on second-generation p-values (Blume et al. 2018, 2019) also emphasize effect size and scientific relevance. This is covered in detail in Chapters 3 and 4.

2.4.2 Advanced topics

In this chapter we introduced several approachable and effective methods for estimating false discovery rate quantities and provided examples of how these methods behave in some cases to illustrate concepts. An evaluation and direct comparison of some important pFDR/FDR estimation methods are available in (Korthauer et al. 2019). Further developments in the field include false discovery rate estimates based on multi-dimensional statistics, which can be used to incorporate information from any type of auxiliary covariate available in the data (e.g., Chong et al. 2015; Alishahi et al. 2016; Chen et al. 2019). Some other notable advances are adaptive procedures which use an auxiliary or "informative" covariate to update the rejection threshold or a component of the FDR estimate (e.g., Boca and Leek 2018; Lei and Fithian 2018; Zhang et al. 2019) and the development of FDR estimates for "online" testing where the total number of tests m is not yet known (such as Javanmard and Montanari 2018; Ramdas et al. 2018; Robertson and Wason 2018).

2.4.3 Recommendations for routine use of FDRs

The critical takeaways for understanding and using false discovery quantities on a routine basis are:

- Controlling the false discovery rate is less strict than controlling the family wise error rate and perfectly sufficient for maintaining scientific rigor in large-scale inference settings.
- The **positive** false discovery rate (pFDR) is the expected fraction of observed results that are false. This is typically what scientists wish to control once results are observed and is therefore the relevant quantity to gauge and control.
- The q-value parallels the p-value in definition and is less intuitive for describing the reliability of observed results. It makes more sense to simply report the tendency for the observed results to be misleading, which is the pFDR.
- The local FDR is an even more direct inferential quantity. However, it faces additional challenges in estimation.
- A variety of (empirical) Bayes methods have been proposed for estimating false discovery quantities. Methods that estimate the entire mixture distribution (e.g., Lindsey's method), rather than the alternative distribution alone, are easier to implement and more suited for broad daily use (in our opinion).
- Estimating the null distribution (e.g., assuming $N(\delta_0, \sigma_0)$ rather than $N(0,1)$) adds a layer of flexibility and robustness that can be critical in some contexts. Relaxing this null assumption should considered as a routine sensitivity analysis.
- Assuming $\pi_0 = 1$ while estimating the FDR is usually not overly conservative and can simplify the pFDR estimation process considerably. Estimating π_0 can however lead to less conservative pFDR estimates and a greater number of discoveries.

2.5 Appendix A: Remarks and supplemental content

2.5.1 Remarks

2.5.1.1 Remark 2.A: Details of simulations and examples

In the simple simulation, generate $m = 1{,}000$ random z-values where $\pi_0 = 0.85$ of them come from the null $N(0,1)$ distribution and $1 - \pi_0 = 0.15$ of them come from an alternate distribution of $N(2.4,1)$. In the underdispersed simulation, generate $m = 50{,}000$ random z-values where $\pi_0 = 0.9$ of them come from an underdispersed null distribution $N(0,0.9)$, and $1 - \pi_0 = 0.1$ of them come from underdispersed alternate distributions $N(\theta, 0.9)$, evenly divided between effect sizes of $\theta = 0.3$, 1 and 3.

The real example is set in the context of a single nucleotide polymorphism (SNP) genetic study. A SNP is a position on human DNA which has been identified to commonly have variation in the nucleotide sequence or allele. These variations may be associated with disease and therefore are typically studied in connection with health conditions. Any allele at a SNP position that is prevalent in less than 50% of the population is referred to as a variant allele. For each person in the sample, we measure the number of variant alleles, with possible values 0, 1, or 2, because humans have 2 copies of each chromosome, and therefore the variant allele can be present at the specified genome position in none, one, or both of the copies of the chromosome. In the prostate cancer SNP example, data on 3,894 individuals (2,511 cases with prostate cancer and 1,383 controls without prostate cancer) comes from the International Consortium for Prostate Cancer Genetics (ICPCG) (Schaid and Chang 2005). To assess potential association between a SNP and prostate cancer, a logistic regression model was used with number of variant alleles treated as a continuous predictor variable. Further background on the data and how it was processed is detailed in (Blume et al. 2019). Because some of the SNPs had a constant (or nearly constant) number of variant alleles across all subjects, some models were not identifiable and these were screened out. Ultimately, we have 224,866 z-statistics corresponding to SNPs from chromosome six.

2.5.1.2 Remark 2.B: Mapping of test statistics to z-values

Much of the work in FDR estimation can be applied to the original p-values (with a number of original papers presented only in that context) but transformation to the z-value space can be beneficial, for working in the unconstrained $N(0,1)$ distribution rather than the $Unif(0,1)$ distribution. If the sign of the original test statistics is not relevant (as in the case of a χ^2 statistic), we can simply use the calculated p-values and the transformation $Z_i = \Phi^{-1}(p_i)$ to transition to the z-value space. In this case, small p-values will be mapped to larger negative z-values (e.g., $p_i = 0.029$ to $z_i = -1.896$) and larger p-values will be mapped to z-values near 0 and (with p-values near 1 mapped to large positive z-values).

For original test statistics where the sign of the test statistic can be helpful to preserve, such as with t-statistics, more care must be taken with the transformation. As noted in the main text, using the original statistics $T_1, \dots, T_m$ with distribution function $T_i \sim F_0(T)$ under the null hypothesis and the transformation $Z_i = \Phi^{-1}\big(F_0(T_i)\big)$ will give us a direct z-value mapping. If using the p-values, specialized transformations are needed, aside from lower-tail one-sided p-values where $Z_i = \Phi^{-1}(p_i) = \Phi^{-1}\big(F_0(T_i)\big)$ gives z-values with the correct original sign of

T_i. For upper-tail one-sided p-values, we need to use $Z_i = \Phi^{-1}(1 - p_i)$ to obtain z-values with the same original sign as T_i. For two-sided p-values, we cannot perform a sign-preserving transformation from the p-values alone and would need to know at least the sign of the original test statistic. Therefore, there is no reason not to use the direct mapping $Z_i = \Phi^{-1}\big(F_0(T_i)\big)$. The only exception would be an unusual case encountered where the p-values are provided along with the test statistics (or at least their signs), but the type of test statistic is not known such that $F_0(T)$ is not known. Then, the mapping $Z_i = \Phi^{-1}(p_i/2) \times \text{sign}(T_i) = \Phi^{-1}(p_i/2) \times (I(T_i > 0) \cdot 2 - 1) = \Phi^{-1}(p_i/2) \times \big(-I(T_i < 0) + I(T_i > 0)\big)$ could be used.

2.5.1.3 Remark 2.C: Scientific validity of controlling FDR vs FWER

When we perform a large number of tests, making a handful of false discoveries, while not desirable, may not be a big concern especially when it comes with the benefit of increased power such that we can reject a larger number of hypotheses. For example, if we reject 100 tests, 5 of which are false discoveries, the observed FDP is $v/\max(r, 1) = 5/\max(100,1) = 0.05$. This does not seem like such as bad deal. Hypothetically, if a multiple comparisons procedure makes only 5 false discoveries in 100 every time the experiment is hypothetically repeated, i.e., the FDP is controlled at 5%, this could be a desired testing procedure. This is particularly the case in large-scale inference where the mass significance testing is often used as a screening procedure for findings. That is, if a set of genes is found to be associated with prostate cancer occurrence, these results won't be used immediately to inform diagnostic medical practice. Rather, they will typically be used to inform further research. Therefore, while wasting resources is of concern, the impact of false discoveries in large-scale inference on human health is not generally thought to be an issue.

2.5.1.4 Remark 2.D: Calculation of FWER and FDR

For the random variable

$$FDP = Q = \frac{V}{\max(R, 1)} = \begin{cases} V/R & \text{when } R > 0 \\ 0 & \text{when } R = 0' \end{cases}$$

the family-wise error rate is $FWER = \Pr(V \geq 1) = \Pr(V > 0) = \Pr(Q > 0)$, such that

$$\begin{aligned} FWER = Pr(Q > 0) &= 1 - \Pr(Q = 0) \\ &= 1 - (\Pr(Q = 0|R = 0) \cdot \Pr(R = 0) + \Pr(Q = 0|R > 0) \cdot \Pr(R > 0)) \\ &= 1 - (1 \cdot \Pr(R = 0) + \Pr(V/R = 0|R > 0) \cdot \Pr(R > 0)) \\ &= 1 - \Pr(R = 0) - \Pr(V = 0|R > 0) \cdot \Pr(R > 0) \\ &= 1 - \Pr(R = 0) - \Pr(R > 0) \cdot \sum\nolimits_{r=1}^{m} \Pr(V = 0|R = r, R > 0) \cdot \Pr(R = r|R > 0) \\ &= 1 - \Pr(R = 0) - \Pr(R > 0) \cdot \sum\nolimits_{r=1}^{m} \Pr(V = 0|R = r) \cdot \frac{\Pr(R > 0|R = r) \cdot \Pr\,(R = r)}{\Pr(R > 0)} \\ &= 1 - \Pr(R = 0) - \sum\nolimits_{r=1}^{m} \Pr(V = 0|R = r) \cdot \Pr(R = r). \end{aligned}$$

The false discovery rate denoted by Q_e is

$$
\begin{aligned}
FDR \coloneqq Q_e = E[Q] &= E[V/R \,| R > 0] \cdot \Pr(R > 0) \\
&= \Pr(R > 0) \cdot \sum\nolimits_{r=1}^{m} E[V/R \,| R = r, R > 0] \cdot \Pr(R = r | R > 0) \\
&= \Pr(R > 0) \cdot \sum\nolimits_{r=1}^{m} \frac{E[V | R = r]}{r} \cdot \frac{\Pr(R > 0 | R = r) \cdot \Pr(R = r)}{\Pr(R > 0)} \\
&= \sum\nolimits_{r=1}^{m} \frac{1}{r} \cdot E[V | R = r] \cdot \Pr(R = r).
\end{aligned}
$$

The number of false rejections V is a Binomial random variable $V \sim Binom(m_0, \alpha)$ when assuming the tests are independent with fixed identical type I error rate α and the number of true rejections U is a Binomial random variable $U \sim Binom(m_1, 1 - \beta)$ when assuming the tests are independent with fixed identical type II error rate β. Thus $R = V + U$ is the convolution of these two independent Binomial random variables. The conditional distribution of $V|R$ (i.e., $V|V + U$, the conditional distribution of two Binomial random variables conditioned on their sum) follows Fisher's noncentral hypergeometric distribution. We can use `convpow` in the R package `distr` to easily obtain the pdf for R, $\Pr(R = r)$. The R package `BiasedUrn` provides the pdf $\Pr(V = v | R = r)$ using dFNCHypergeo(x=v, n=r, m1=m_0, m2=m_1, odds=γ) and the expected value $E[V | R = r]$ using meanFNCHypergeo(m1=m_0, m2=m_1, n= r, odds= γ), where $\gamma = (\alpha/(1 - \alpha))/((1 - \beta)/\beta)$.

2.5.1.5 Remark 2.E: Data tables for rejection procedures of interest

Supplemental Table 2.1(a) displays the known elements of the data table from our SNP example using an unadjusted rejection region of $\Gamma = [1.64, \infty)$. There are 10,637 z-values that fall in the rejection region for the SNP example, but we are unable to fill in the first two columns of the data table, because we do not know the truth of each hypothesis under consideration. Supplemental Table 2.1(b) displays the known elements of the data table for the Bonferroni procedure, and Supplemental Table 2.1(c) displays the known elements of the data table for the Benjamini-Hochberg (BH) procedure. We can see that there are much fewer rejections in than in the unadjusted setting, but that the BH procedure has almost double the rejections than for the Bonferroni procedure.

2.5.1.6 Remark 2.F: Adjusted p-values

The adjusted p-values for the simple simulation are shown in Supplemental Figure 2.1. We can see that the BH procedure provides a balance between the uncontrolled unadjusted procedure and the Bonferroni procedure.

2.5.1.7 Remark 2.G: The marginal false discovery rate

When Benjamini and Hochberg proposed Q_e as a measure of interest to control in large-scale testing, they discussed other alternative quantities that could be considered. One of the main reasons that Benjamini and Hochberg settled on Q_e is that many of the other quantities cannot be controlled in the strong or weak sense. Strong control of a quantity means that it can be controlled at a certain level no matter how many of the hypotheses are truly null, while

Supplemental Table 2.1 Partially filled data table summarizing the results of a multiple testing analysis for the SNP example for various rejection regions or procedures. The rows correspond to the results of the analysis (reject or inconclusive), while the columns correspond to the truth or falsehood of the m null hypotheses. Unknown values are shaded in grey. (a) Unadjusted upper-tail z-value rejection region of $[1.64, \infty)$, (b) Bonferroni procedure, and (c) Benjamini-Hochberg procedure.

(a) Unadjusted

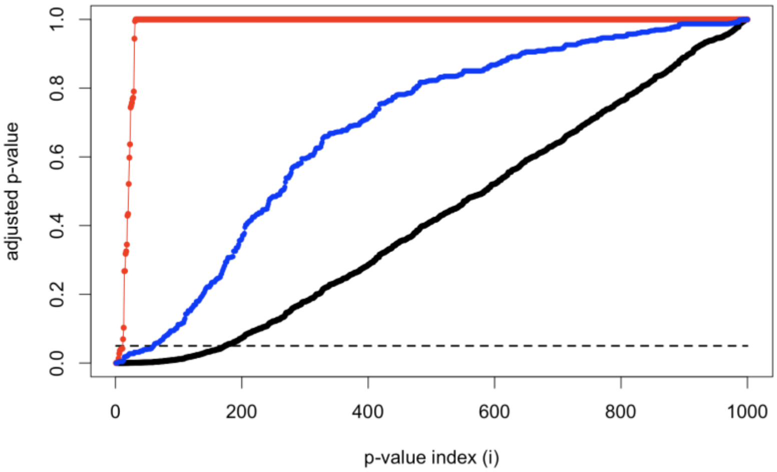
	Null $(\mathcal{T})$	Alt. $(\mathcal{F})$	Total
Reject $(\mathcal{R})$	?	?	10,637
Inconclusive $(\mathcal{I})$	?	?	214,229
Total	?	?	224,866

(b) Bonferroni

	Null $(\mathcal{T})$	Alt. $(\mathcal{F})$	Total
Reject $(\mathcal{R})$	?	?	16
Inconclusive $(\mathcal{I})$	?	?	224,850
Total	?	?	224,866

(c) Benjamini-Hochberg (BH)

	Null $(\mathcal{T})$	Alt. $(\mathcal{F})$	Total
Reject $(\mathcal{R})$	?	?	28
Inconclusive $(\mathcal{I})$	?	?	224,838
Total	?	?	224,866



Supplemental Figure 2.1 Adjusted p-values for the Benjamini-Hochberg procedure applied to the simple simulation. The black points are the unadjusted p-values, the red points are the Bonferroni adjusted p-values, and the blue points are the Benjamini-Hochberg adjusted p-values. The x-axis is the p-value index, with 1 indicating the smallest p-value in the data and 1,000 indicating the largest p-value in the data. The black dotted horizontal line is the classic 0.05 cutoff.

weak control for a quantity means that it can be controlled at a certain level only if all the hypotheses are truly null. The two major alternatives discussed by BH are $E[V]/E[R]$ and the pFDR (discussed in Section 2.3).

The quantity $E[V]/E[R]$ is often referred to in the literature as the marginal false discovery rate (mFDR). In many cases, the marginal FDR and the pFDR are approximately equal, and under a certain set of assumptions, they are exactly equivalent. However, there are drawbacks to the marginal FDR, particularly that it does not assess the joint behavior of V and R, which is of more theoretical interest. See (Storey 2003; Tsai et al. 2003; Benjamini 2010) for further discussion.

2.5.1.8 Remark 2.H: The conditional false discovery rate

Benjamini and Hochberg (1995) also discuss the conditional false discovery rate (cFDR), which is $cFDR = E(V/R\,|R = r) = E(V|R = r)/r$. This quantity is close to the quantity $E(V)/r$ discussed originally in (Sorić 1989) which was the inspiration for the FDR, but the conditional expectation is not equal to the marginal expectation $(E(V|R = r) \neq E(V))$. Instead, $E(V) = \sum_1^m E(V|R = r) \times \Pr(R = r)$. The cFDR is a more properly defined inferential quantity than $E(V)/r$ (sometimes called the empirical FDR), because it keeps the connection of the joint relationship between V and R. Tsai et al. (2003) and Pounds and Cheng (2004) (among others) have provided nice mathematical forms for and estimation of the cFDR and argue that it is the preferable quantity to the pFDR. In general, the cFDR and pFDR have a lot of properties in common. In fact, under the two-groups mixture model setup discussed in Section 2.3.2, the cFDR is equal to the pFDR and the mFDR. The argument that cFDR is better than the other quantities for summarizing observed results of an experiment because it conditions on the actual number of observed rejections does not necessarily provide us with any more precise inference. Rather, we would prefer to focus in on the local FDR (discussed in Section 2.3.4), which conditions on the value of the z-value or p-value, rather than conditioning on the less natural idea of other experiment repetitions which produce the same number of rejections.

2.5.1.9 Remark 2.I: k-family wise error rate and k-false false discovery rate

Alternatives to the traditional FWER and FDR which allow a larger level of false discoveries have been proposed. The family-wise error rate has been generalized to $kFWER = P(V \geq k)$, so that any number of false discoveries less than k does not "count" as an error (Dudoit et al. 2004). The false discovery rate has been generalized to $kFDR = E\big(V \cdot \mathbf{1}_{\{V \geq k\}}/R \cdot \mathbf{1}_{\{R>0\}} + 0 \cdot \mathbf{1}_{\{R=0\}}\big)$, where again we only start "counting" false discoveries once they have passed a certain threshold of k false discoveries (Sarkar 2007).

2.5.1.10 Remark 2.J: Exceedance control

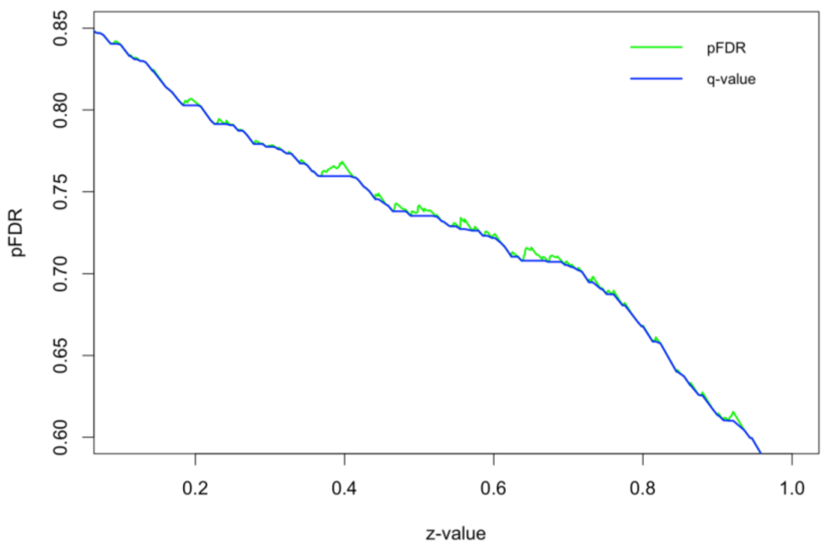
Another generalization, referred to as "exceedance control", bounds a different aspect of the FDP distribution. Instead of controlling the expectation of Q, the tail-area probability of Q is controlled. That is, $P(Q > \gamma)$ is bounded for a pre-specified value of $\gamma \in (0,1)$. More on this can be found in (Genovese and Wasserman 2002, 2004; Romano and Wolf 2007). Various methods proposed for such control have been described, see for example (Korn et al. 2004; van der Laan et al. 2004; Lehmann and Romano 2005; Genovese and Wasserman 2006; Döhler and Roquain 2020).

2.5.1.11 Remark 2.K: Rejection procedure and p-value ordering

In almost all accepted large-scale testing methods, the ranking of observed p-values is respected by the rejection rules. This means that, by convention, a hypothesis with a smaller p-value p_A may not fail to reject null hypothesis H_0^A if any hypothesis with a larger p-value p_B rejects null hypothesis H_0^B. False discovery rates usually follow this convention, although not in all cases (i.e., if one is willing to use another measure of evidence such as the Likelihood ratio, a posterior probability, or the second-generation p-value). This condition is sometimes imposed ad-hoc (such as with q-values, as we see in Section 2.3.3).

2.5.1.12 Remark 2.L: More illustration of q-value vs pFDR

An illustration of how the q-value operates, forcing the estimated pFDR to be monotone decreasing in the case of one-sided upper tail area rejection regions (or unimodal, in the case of two-sided rejection regions) is shown in Supplemental Figure 2.2. From this figure, we see that the pFDR and the q-value are often very close. However, these two quantities can be substantially different, as is seen in later sections of the chapter (e.g., Figure 2.3(a)).



Supplemental Figure 2.2 Illustration of q-values, showing an estimated pFDR and the corresponding estimated q-values.

2.5.1.13 Remark 2.M: Local FDR interpretation

In the rate-based form of the local FDR, $E[V/R\,|R > 0; \Gamma = \{z_i\}]$, the interpretation may be difficult because it is not that common for the results to contain more than 1 or 2 z-values that are exactly equal to z_i, even in large-scale inference. This issue may be dealt with by some expanded local region (e.g., $\Gamma = [z_i - \delta, z_i + \delta)$) for a carefully specified δ. Alternatively, in the probabilistic form of the local FDR, $\Pr(H = 1|Z = z_i)$, the interpretation is more straightforward.

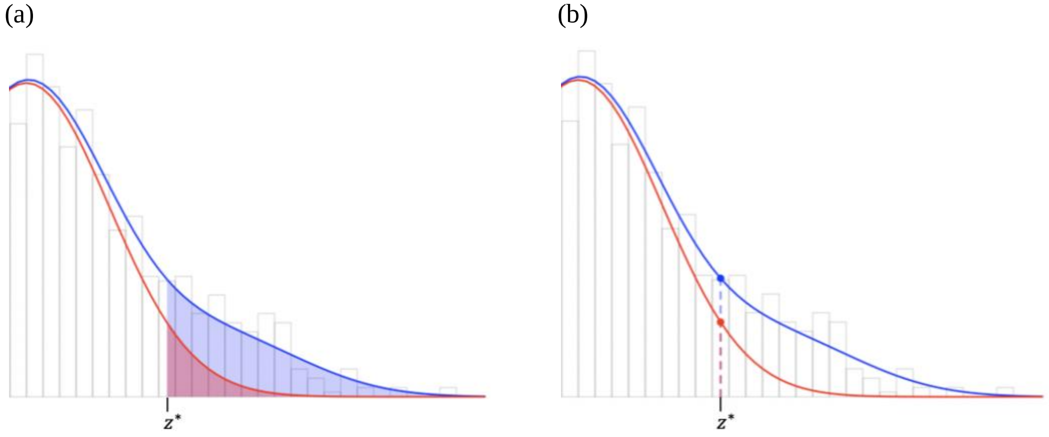
2.5.1.14 Remark 2.N: Global vs local FDR illustration

The distinction between a global tail-area false discovery rate and the local false discovery rate in terms of the empirical Bayes estimation is illustrated conceptually in Supplemental Figure 2.3.

2.5.1.15 Remark 2.O: Further technical details on the ECDF estimator

The simplest form of the upper-tail area pFDR estimate, as in Equation (2.9), results when $z_1, \dots, z_m$ are defined to be the reverse order statistics such that $z_1 \geq z_2 \geq \cdots \geq z_m$, i.e., the z_i are in decreasing order. In this case, $\#\{j: z_j \geq z_i\} = i$. We will also make use of $\#\{j: z_j \leq z_i\} = m - \#\{j: z_j > z_i\} = m - \#\{j: z_j \geq z_{i-1}\} = m - (i - 1)$.

Recall that the probabilistic form of the pFDR from Equation (2.6) is $\phi(\Gamma) = \pi_0 \cdot \Pr\,(Z \in \Gamma|\mathrm{H} = 0)/\Pr(Z \in \Gamma)$. For one-sided upper-tail area pFDRs, this can be written as $\phi\big([z, \infty)\big) = \pi_0 \cdot \Pr(Z \geq z|\mathrm{H} = 0)/\Pr(Z \geq z)$, with an estimator $\hat{\phi}\big([z_i, \infty)\big) = 1 \cdot \big(1 - \Phi(z_i)\big)/\widehat{Pr}(Z \geq z_i)$ assuming $\pi_0 = 1$ and that the theoretical $N(0,1)$ null holds.



Supplemental Figure 2.3 Illustration of global pFDR vs local FDR. The red line (and shaded area) is the null distribution, and the blue line (and shaded area) is the mixture distribution. (a) The global pFDR as a ratio of tail areas, (b) the local FDR as a ratio of densities at the observed summary measure.

We can use either the formal ECDF estimate $\hat{F}(c)$ or the generalized ECDF-type estimate $\hat{F}(\Gamma) = \#\{j: z_j \in \Gamma\}/m$ to derive the denominator $\widehat{Pr}(Z \geq z_i)$. For using the formal ECDF, the pFDR denominator can be written as $\widehat{Pr}(Z \geq z_i) = 1 - \widehat{Pr}(Z < z_i) = 1 - \widehat{Pr}(Z \leq z_{i+1})$, with the last equality holding because we are utilizing a stepwise estimate. The formal ECDF estimate is $\hat{F}(z_i) = \#\{j: z_j \leq z_i\}/m = \big(m - (i-1)\big)/m = 1 - (i-1)/m$. So, we have that $\widehat{Pr}(Z \geq z_i) = 1 - \widehat{Pr}(Z \leq z_{i+1}) = 1 - \hat{F}(z_{i+1}) = 1 - \Big(1 - \big((i+1) - 1\big)/m\Big) = i/m$.

For using the generalized ECDF-type estimator, the denominator is $\widehat{Pr}(Z \geq z_i) = \hat{F}\big([z_i, \infty)\big) = \#\big\{j: z_j \in [z_i, \infty)\big\}/m = \#\{j: z_j \geq z_i\}/m = i/m$, with the last equality holding again because the z_i are in decreasing order. Either approach gives us $\widehat{Pr}(Z \geq z_i) = i/m$, such that

$$\hat{\phi}\big([z_i, \infty)\big) = \frac{1 \cdot \big(1 - \Phi(z_i)\big)}{i/m}. \qquad \blacksquare$$

Note that because we can estimate the pFDR for any value of z, not only those observed z-values in the data, we technically need to adjust the estimate $\hat{F}\big([z, \infty)\big)$ to account for values of z outside of the observed range of z-values. Otherwise, the denominator of the pFDR estimate in Equation (2.9) would be 0, resulting in an undefined pFDR estimate. We can do this by defining

$$\hat{F}\big([z, \infty)\big) = \begin{cases} i/m\,, & \text{if } z \leq \max(z_1, \dots, z_m) \\ 1/m\,, & \text{if } z > \max(z_1, \dots, z_m) \end{cases}.$$

More generally, we could also write the modification as $\hat{F}(\Gamma) = \max\big(\#\{j: z_j \in \Gamma\}, 1\big)/m$.

2.5.1.16 Remark 2.P: Derivation of Benjamini-Hochberg adjusted p-value equality to estimated FDR

The equivalence of the Benjamini-Hochberg adjusted p-value to the pFDR, estimated as in Section 2.3.5.1, with the ECDF, holds for any p-value type (lower tail, upper tail, or two-sided), but we will show the derivation only for the upper-tail setting that we have been working with.

For Equation (2.9) we have defined the $z_1, \dots, z_m$ to be the reverse order statistics such that $z_1 \geq \cdots \geq z_m$. The corresponding one-sided upper tail p-values are thus in increasing order, i.e., they are the order statistics $p_{(1)} \leq \cdots \leq p_{(m)}$, using the conventional notation. To be clear, $p_{(1)}$ is the p-value corresponding to z_1 (the smallest p-value and largest z-value), $p_{(2)}$ is the p-value corresponding to z_2 (the second smallest p-value and second largest z-value), and so on with $p_{(i)} = 1 - \Phi(z_i)$.

We can write the estimated q-value as $\hat{q}(z_i) = \min\limits_{c \leq z_i} \hat{\phi}\big([c, \infty)\big) = \min\limits_{k: z_k \leq z_i} \hat{\phi}\big([z_k, \infty)\big)$, taking estimates across only the observed set of z-values rather than the whole continuous z-value interval. The set $\{k: z_k \leq z_i\}$ can be simplified to $k \geq i$ because the z_i are in decreasing order, so we end up with $\hat{q}(z_i) = \min\limits_{k \geq i} \hat{\phi}\big([z_k, \infty)\big)$. Substituting the estimator of Equation (2.9) in, $\hat{q}(z_i) = \min\limits_{k \geq i} \big(1 - \Phi(z_k)\big)/(k/m)$. Note that the numerator is equal to the p-value in this case, thus $\hat{q}(z_i) = \min\limits_{k \geq i} p_{(k)}/(k/m)$, which equals the adjusted Benjamini-Hochberg p-value $\tilde{p}_{(i)}^{BH} = \min\limits_{k \geq i}\big(p_{(k)}/(k/m)\big)$ defined in Section 2.2.3. Therefore, the BH adjusted p-value is equal to the q-value, a modified pFDR quantity, in the following setting: we assume $\pi_0 = 1$, it is assumed that the theoretical null of $N(0,1)$ holds, and the ECDF is used to non-parametrically estimate the mixture distribution.

2.5.1.17 Remark 2.Q: Local FDR non-parametric mixture density approximation estimate

The local FDR requires the density rather than the distribution function, therefore the simple ECDF estimate described in Section 2.3.5.1 is not sufficient for a comparable simple local FDR estimate. Efron (2010b) utilizes the following approach: discretize the observed z-values $z_1, \dots, z_m$ according to bins of width 2δ (with $\delta > 0$) such that the histogram has a corresponding number of bins K to accommodate the range of the observed z-values. Note that alternately, we could fix K and then δ will be determined by the range of the observed z-values. Define x_k to be the midpoint of bin k and note that the bounds of bin k are defined by $\mathcal{Z}_k = [x_k - \delta, x_k + \delta)$. The estimator of $f(z)$ based on the observed z-values is defined as

$$\hat{f}(z) = \frac{\#\big\{j: z_j \in [x_{k'} - \delta, x_{k'} + \delta)\big\}}{2\delta \cdot m}, \text{with } k' = \{k: z \in \mathcal{Z}_k\}.$$

Note that this quantity is approximately equal to $\Big(\hat{F}(x_{k'} + \delta) - \hat{F}(x_{k'} - \delta)\Big)/2\delta$ with $\hat{F}$ being the ECDF as in Section 2.3.5.1. The same approximation is applied to the numerator $f_0(z)$, and we get a simple estimator for the local FDR as

$$\hat{\phi}_l(z_i) = \frac{\pi_0 \cdot \big(F_0(x_{k'} + \delta) - F_0(x_{k'} - \delta)\big)/(2\delta)}{\hat{f}(z)} = \frac{1 \cdot m\big(F_0(x_{k'} + \delta) - F_0(x_{k'} - \delta)\big)}{\#\{j: z_j \in [x_{k'} - \delta, x_{k'} + \delta)\}},$$

with $k' = \{k: z_i \in \mathcal{Z}_k\}$. The bin width 2δ will contribute to the noise and instability of the estimates – large numbers of bins will result in less biased but more noisy non-parametric estimates of the local false discovery rates $\hat{\phi}_l(z_i)$. Local FDR estimates for various values of K are shown in Supplemental Figure 2.4(a)-(d) for the simple simulation example. We use the simple simulation instead of the underdispersed simulation because it provides a clearer illustration of how varying K might affect the estimate. While the exact nature of the curve varies quite a bit, the overall shape of the curve is roughly the same for each value of K in this case, and lines up pretty well with the true FDR curve.

2.5.1.18 Remark 2.R: Lindsey’s method

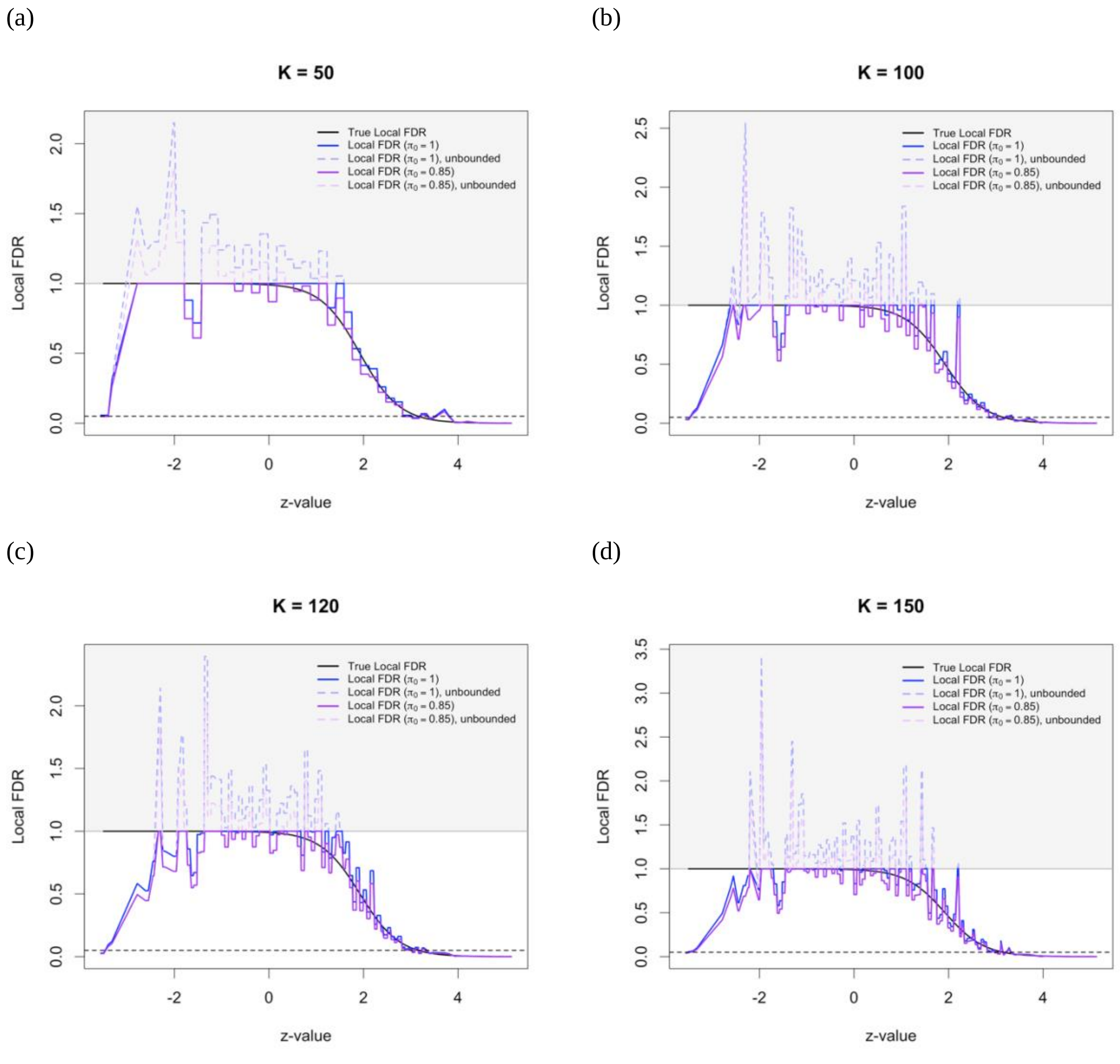
Discretize the observed z-values $z_1, \dots, z_m$ according to bins of width 2δ (with $\delta > 0$) such that the histogram has a corresponding number of bins K to accommodate the range of the observed z-values. Alternately, fix the number of bins K such that δ will be determined by the range of the observed z-values. Define x_k to be the midpoint of bin k and note that the bounds of bin k are defined by $\mathcal{Z}_k = [x_k - \delta, x_k + \delta)$. Define y_k as the count in each bin (i.e., the histogram height), $y_k = \#\{j: z_j \in \mathcal{Z}_k\}$. Model y_k as a flexible function of x_k (e.g., using a J-order polynomial or spline function) assuming a Poisson model. For example, fit the model

fit = glm(yk ~ splines::ns(xk, df = 7), family = “poisson”)

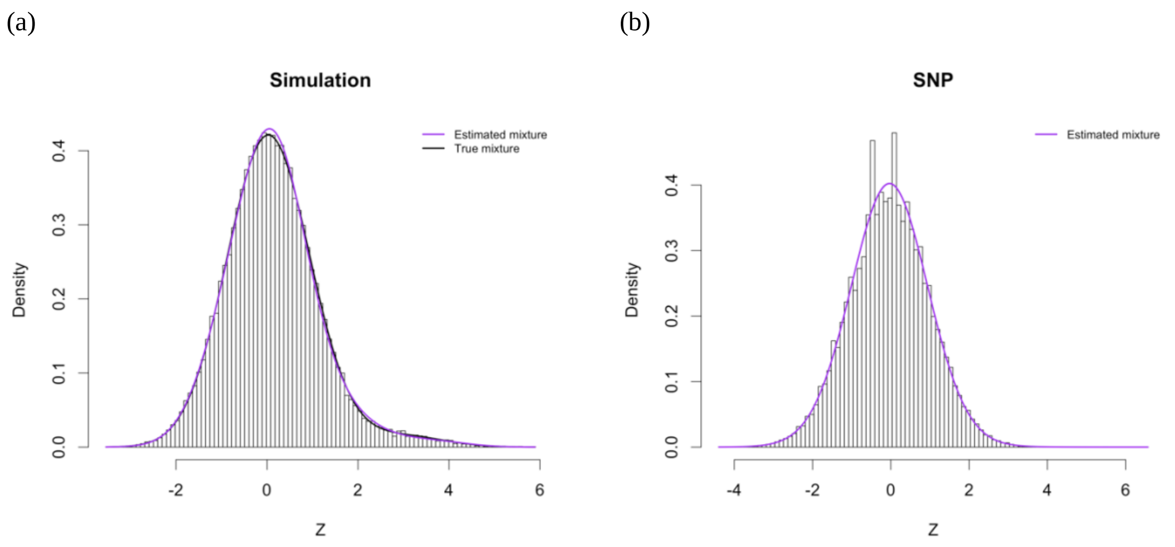
in R. See the code in the Github repository (https://github.com/weltybiostat/FDRprimer) for more detail. The estimated density is then given by $\hat{f}(z) = \hat{y}_k/(2\delta \cdot m)$, where $\hat{y}_k$ is the fitted value of y_k from the Poisson regression model evaluated at $x_k = z$. This approach is described in (Efron and Tibshirani 1996) and is based off of work in (Lindsey 1974).

2.5.1.19 Remark 2.S: Lindsey’s method fits

Supplemental Figure 2.5(a)-(b) shows $\hat{f}(z)$, the estimated mixture for the underdispersed simulation using $J = 7$ and $\delta = 0.048$ (from discretizing the z-values into 100 bins), and the SNP example using $J = 5$ and $\delta = 0.056$ (also from discretizing the z-values into 100 bins). The mixture estimates fit the z-value curves very well and provide an almost exact estimate of the true mixture for the underdispersed simulation as seen in Supplemental Figure 2.5(a). For the SNP example, the default value of $J = 7$ resulted in an improperly estimated density function where the right tail did not converge to 0. While the default of $J = 7$ may work in many cases, this issue or others such as overfitting may arise and so the parameter J may need to be adjusted.



Supplemental Figure 2.4 True and estimated local FDR as given by the estimate which uses the theoretical null and the empirical density approximation of the mixture distribution for the simple simulation, for both $\pi_0 = 1$ and the true null proportion, $\pi_0 = 0.85$. Both the unbounded and bounded FDR estimates are shown. (a) $K = 50$, (b) $K = 100$, (c) $K = 120$, and (d) $K = 150$.

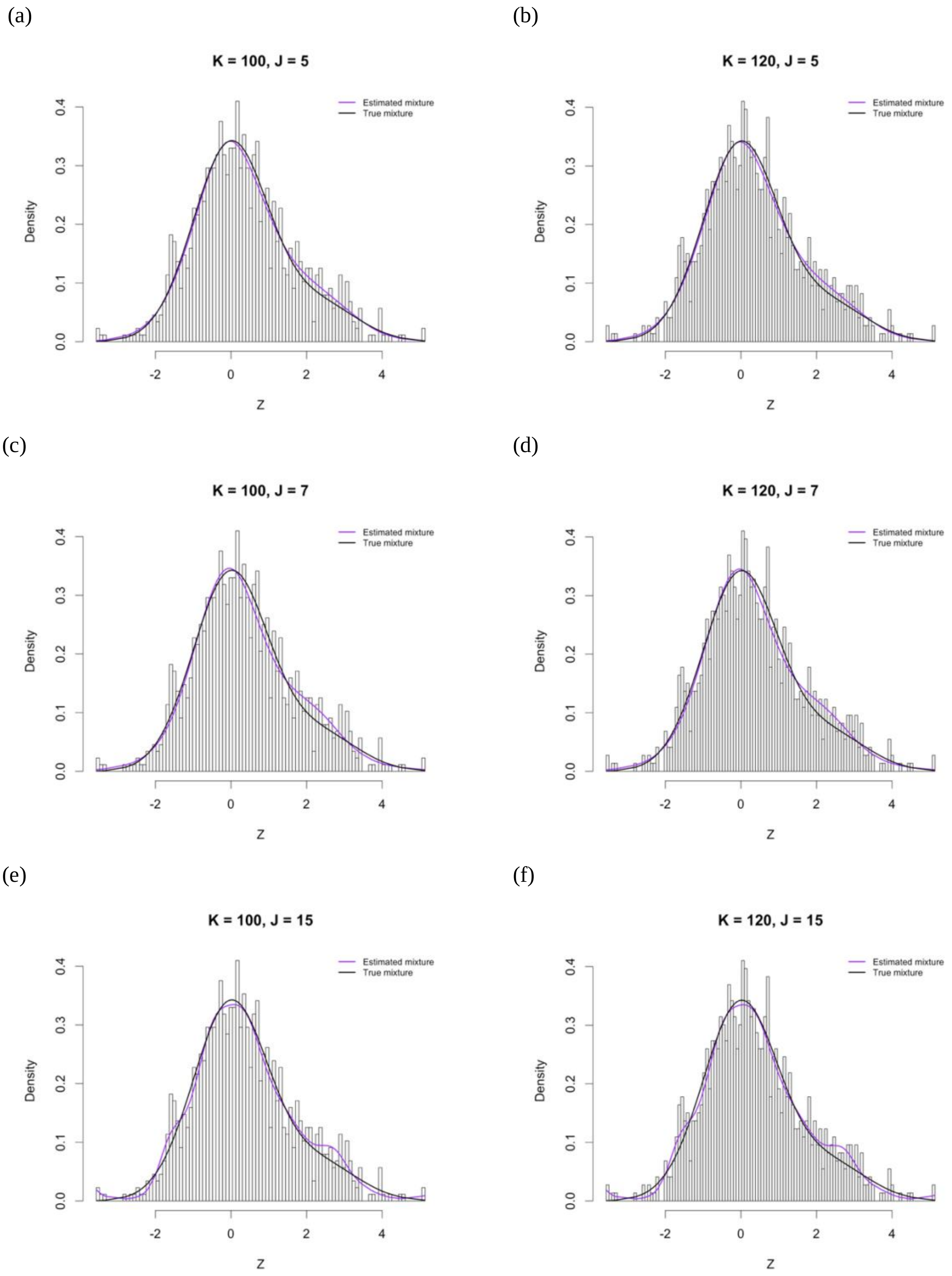


Supplemental Figure 2.5 Lindsey's method estimates for the mixture distribution. (a) The mixture distribution estimated by Lindsey's method for the underdispersed simulation, with $J = 7$ and $K = 100$ (purple line). The black line is the true mixture. (b) The mixture distribution estimated by Lindsey's method for the real-world SNP example, with $J = 5$ and $K = 100$ (purple line).

The fitted densities for varying values of J and K for the simple simulation are given in Supplemental Figure 2.6(a)-(f). There are noticeable issues with overfitting in Supplemental Figure 2.6(e)-(f) for $J = 15$, where too much flexibility allows this to happen. For both $K = 100$ and $K = 120$, the right tail of the density does not converge to zero for all $J \geq 9$. This improperly estimated density issue mentioned above prevents us from estimating the pFDR and serves as another note of caution for examining the resulting Lindsey's method fit for the chosen value of J, and adjusting as necessary.

2.5.1.20 Remark 2.T: Examples of π_0 estimation

One method described by Efron relies on the "zero assumption", which assumes that the middle proportion of the histogram of z-values contain only those from the null distribution (Efron 2010b). Of course, unless all alternative z-values correspond to extremely large effects, there are likely to be some alternative z-values in this central proportion $\mathcal{A}_0$. However, the assumption is reasonable in large-scale inference where π_0 is usually large and therefore an overwhelming majority of z in $\mathcal{A}_0$ will be null. Mathematically, the assumption is that $f_1(z) = 0$ for $z \in \mathcal{A}_0$, where $\mathcal{A}_0$ contains the middle $\alpha_0 * 100\%$ of the observed z-values. The proportion α_0 is specified ahead of time, while the exact bounds of $\mathcal{A}_0$ will depend on observed z-value distribution. Efron argues that there is no methodical approach to selecting α_0, but that differences in the π_0 estimation do not typically have a meaningful effect and so the choice is not critical. As we have seen prior in Figure 2.3(a) and Figure 2.5(a), the difference between $\widehat{pFDR}$ for $\pi_0 = 1$ and $\pi_0 = 0.9$ is quite large near $z = 0$ but is small for extreme values of z, where the bias is most relevant.



Supplemental Figure 2.6 Mixture density estimated by Lindsey's method for varying values of J and K for the simple simulation. The black line is the true mixture, and the purple lines are the mixture density estimates. For values of $K = 100$ we have $\delta = 0.044$ and for values of $K = 120$ we have $\delta = 0.0366$. (a) $K = 100$ and $J = 5$, (b) $K = 120$ and $J = 5$, (c) $K = 100$ and $J = 7$, (d) $K = 120$ and $J = 7$, (e) $K = 100$ and $J = 15$, and (f) $K = 120$ and $J = 15$.

If we use this method in the underdispersed simulation and SNP examples with $\alpha_0 = 0.50$, we obtain estimates of $\hat{\pi}_0^{Ef} = 1$ for both. Other methods are similar in spirit to Efron's, but operate differently, for example that of (Storey and Tibshirani 2003) and the (Murray and Blume 2021) "last histogram height" algorithm. These two methods assess the histogram of p-values and rely on the expectation that the histogram of null p-values should be uniform, and that large p-values are very likely to correspond to null results. For these two methods, we obtain estimates of $\hat{\pi}_0 = 1$ for both methods for the underdispersed simulation and $\hat{\pi}_0^{ST} = 1$ and $\hat{\pi}_0^{LHH} = 0.96$, respectively, for the SNP example. The last histogram height algorithm is easily implemented in R with the package `FDRestimation` with the function `get.pi0`, and Storey and Tibshirani's method with the use of the function `pi0est` in the R package `qvalue`.

2.5.1.21 Remark 2.U: Empirical null estimation methods

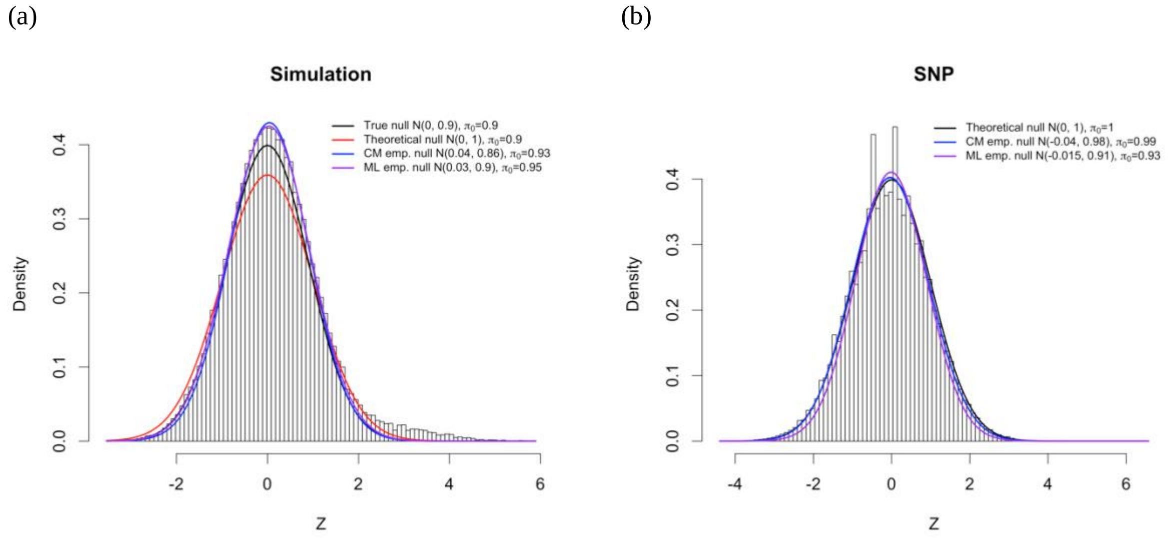
The two proposed empirical null estimation methods are described in further detail in (Efron 2004, 2007b, 2010b). Briefly, the central matching approach makes the "zero assumption" for $z \in \mathcal{A}_0$ (see Remark 2.T), such that $\log f(z)$ equals $\log \pi_0 f_0(z)$ in the central region of the observed z-values. Assume also that $\log f(z)$ has a quadratic form within $\mathcal{A}_0$ such that $\log f(z) = \beta_0 + \beta_1 z + \beta_2 z^2$ and fit with least squares to the estimate $\log \hat{f}$ from Lindsey's method. Therefore, $\beta_0 + \beta_1 z + \beta_2 z^2 = \log f(z) = \log \pi_0 f_0(z)$ and we can find a mapping between $(\beta_0, \beta_1, \beta_2)$ and $(\delta_0, \sigma_0, \pi_0)$ to obtain the estimates $(\hat{\delta}_0, \hat{\sigma}_0, \hat{\pi}_0)$. The maximum likelihood approach constructs the likelihood for z_0, the collection of z-values that fall in $\mathcal{A}_0$, and obtains maximum likelihood estimates for the parameters $(\delta_0, \sigma_0, \pi_0)$. The likelihood is

$$f_{\delta_0,\sigma_0,\pi_0}(z_0) = \left[\binom{m}{N_0}\theta^{N_0}(1-\theta)^{m-N_0}\right]\left[\prod_{I_0}\frac{\varphi_{\delta_0,\sigma_0(z_i)}}{H_0(\delta_0,\sigma_0)}\right]$$

where $I_0 = \{i\colon z_i \in \mathcal{A}_0\}$, $N_0 = \#\{I_0\}$, $H_0(\delta_0, \sigma_0) = \int_{\mathcal{A}_0} \varphi_{\delta_0,\sigma_0}(z)dz$ with $\varphi_{\delta_0,\sigma_0}$ the density for $N(\delta_0, \sigma_0)$, and $\theta = \pi_0 H_0(\delta_0, \sigma_0) = \Pr(z_i \in \mathcal{A}_0)$.

2.5.1.22 Remark 2.V: Empirical null distribution fits

The estimated null densities are shown in Supplemental Figure 2.7 for the underdispersed simulation and the SNP example. The true null distribution along with the theoretical $N(0,1)$ null for $\pi_0 = 0.9$ is shown for the underdispersed simulation and the theoretical null assuming $\pi_0 = 1$ for the SNP example. The overdispersion in the theoretical null can be seen somewhat in the tails of the distribution for the simulation, but the overestimates of π_0 from the true value of 0.9 mask this somewhat. In the case of the SNP example, the empirical null estimates are not much different than the theoretical null, which appears to fit the observed z-value distribution quite well. The unusual spikes in the histogram do not appear to cause much of an issue in the null estimation.



Supplemental Figure 2.7 Estimates of the empirical null in the examples. (a) Null densities estimated from the central matching and maximum likelihood approaches for the underdispersed simulation and (b) Null densities estimated from the central matching and maximum likelihood approaches in the real-world SNP data.

2.5.1.23 Remark 2.W: Accompanying code

The code used to generate the simulation data, figures, and tables is available at the following Github repository: https://github.com/weltybiostat/FDRprimer.

CHAPTER 3

The Positive False Discovery Rate for Second-Generation p-Values

Valerie F. Welty, Jeffrey D. Blume

3.1 Introduction

In many modern applications, the original problem of multiple comparisons (Tukey 1953), has grown vastly in scale to include hundreds of thousands, or even multiple millions of numbers of tests. The original solution, control of the family wise error rate (the probability of making one or more false rejection) (Tukey 1953; Ryan 1959) is no longer satisfactory in such cases, where a 5% chance of a single rejection of a null test is an exceedingly strict criteria (Brown and Russell 1997; Perneger 1998). Benjamini and Hochberg (1995) were the first to formalize an alternative proposal, the false discovery rate (FDR). However, as we have discussed extensively in Chapter 2, their formal definition of the FDR quantity does not answer the natural scientific question: what is the expected rate of false, i.e., null, discoveries when we observe a set of rejected tests? This question is answered instead by the positive false discovery rate, and equivalent to the Bayes FDR, which is written generally as $P(\text{null true} \mid \text{test rejected})$ (Efron, Storey, et al. 2001; Efron, Tibshirani, et al. 2001; Storey 2001a, 2001b, 2002, 2003). Various approaches to empirical estimation of this quantity continue to develop, with the vast majority for rejection procedures centered around the classical p-value (Storey and Tibshirani 2003; Tang et al. 2007; Efron 2010a; Efron 2010b).

The classical p-value remains the most common statistical measure used, despite its well-known shortcomings and misinterpretations (Greenland et al. 2016; Wasserstein and Lazar 2016). One of the most important shortcomings of the p-value is that it does not account for scientific relevance with the use of the point null hypothesis. A very small p-value could correspond to an effect size that is either indistinguishable from the null due to equipment tolerance/measurement error, or is not clinically meaningful in practice (e.g., a lengthy invasive procedure expected to extend a patient’s life by 15 minutes). However, an important reason that the p-value remains popular is that it’s seen as providing a single-number summary of the results of a scientific experiment. Indeed, it is often helpful to have a single number that indicates how “interesting” or relevant the results of an inferential test are, particularly for multiple testing scenarios, and especially large-scale inference settings, such as is in genetic, imaging, or microbiome studies. With one test or a small number of tests, an examination of estimated effects and confidence interval values together to assess real impact of significant effects may be completed. However, it is not feasible to do so for millions or many thousands of tests. Thus, a helpful single-number summary to filter or assess results, while accounting for the scientific context, remains key. The second-generation p-value (SGPV), an alternative to the p-value proposed by Blume et al. (2018), fits in well for this role. The SGPV is based on a pre-specified definition of what effect sizes would be considered clinically relevant or interesting, as would be done in equivalence testing or composite null hypothesis testing (Schuirmann 1987; Royall 1997; Blume 2002; Wellek 2010).

Application of false discovery rate concepts to the second-generation p-value may result in an improved inferential result, with focus on non-trivial, scientifically relevant results. This idea has been examined in a limited scope in prior work, such as (Blume et al. 2018, 2019). However, these make strong simplifying assumptions, and notably do not fully or rigorously incorporate the idea of an interval null hypothesis. In this chapter, we aim to provide a more comprehensive framework for, and examination of, the Bayesian FDR (i.e., pFDR), for second-generation p-values. The broad definitions for key SGPV FDR quantities are straightforward. However, exact specification or calculation of these is not as simple and requires further investigation. The key quantities for calculation of the positive false discovery rate are the design probabilities – which outline the probabilities of observing a rejected test under the null and alternative design hypotheses, respectively – as well as the proportion of tests which are truly null. For the SGPV, which utilizes an interval or composite null hypothesis, the null design probability is not clearly defined.

In this chapter, we present a variety of possible approaches to handle this specification, ranging from point value specifications to marginalization approaches, as well as the specification of the alternative design probabilities (which are ultimately needed to assess an analogous SGPV quantity, the false confirmation rate (FCR)). The behavior of these approaches, in terms of their impact on the false discovery rate quantities, are examined in Section 3.3.4. They vary in terms of complexity and ultimate usefulness. Of note, we find that the SGPV Bayes FDR will converge to 0, for all sensible choices of methods used for the null design probability and alternative design probability. The SGPV Bayes FCR will converge to 0 for most methods, except for those which place all of the weight exactly on the null boundary (in the latter case, the FCR will converge instead to $\alpha/2/(\alpha/2 + r^{-1})$, where r is the prior odds). For the majority of considered null and alternative distributions, the SGPV FDR is smaller than that of the classical unadjusted p-value, with the exception being a setting where null effects are spread across Θ_0, and all alternative effects are right at the null boundary – understandably, a challenging scenario.

In practice, to avoid specification of, and thus possible misspecification of, these design probabilities, some empirical Bayes estimation approaches are studied. If the empirically estimated mixture probability for second-generation p-values is substituted in the denominator of the Bayes FDR and FCR quantities, then two things are accomplished: 1) only one of the design probabilities must be specified for each quantity, and 2) it is possible to calculate the upper bound of the quantities, rather than needing to specify the null proportion. However, we find that, while these are useful to a degree, more extensive empirical approaches are needed to reliably estimate the false discovery rate for second-generation p-values in practice. The upper bound estimator may be quite conservative, and the remaining necessary design probability specification can still have appreciable impacts on the resulting Bayes FDR quantities.

Overall, we illustrate how second-generation p-values may be able to provide an improvement over the classical p-value in terms of false discovery rates for large-scale inference, by reducing the rate of both null and practically null effects in the set of rejected tests. However, we see that this false discovery rate may still be larger than desired for small sample sizes. Further, obtaining an estimate of the false discovery rate should accompany usage in practice. While this chapter establishes this framework for SGPVs and studies preliminary empirical Bayes estimates, they will require further development before finding utility in real studies. In particular, reliable empirical estimation of the null proportion and the relevant design probabilities for second-generation p-values would allow for practical, impactful use.

3.2 Second-generation p-value background

To calculate a second-generation p-value (SGPV), an interval null hypothesis and corresponding alternative hypothesis for a parameter or value of interest, θ, is specified as

$$\begin{aligned} &H_0: \theta \in \Theta_0 = [\theta_0^-, \theta_0^+] \\ &H_1: \theta \notin \Theta_0. \end{aligned} \tag{3.1}$$

We could also write the alternative hypothesis H_1 as $H_1: \theta \in \Theta_1 = (-\infty, \theta_0^-) \cup (\theta_0^+, \infty)$. This interval is a pre-specified range of values that are deemed to be scientifically not meaningful. For example, it might contain effect sizes that are indistinguishable from the point null hypothesis due to measurement error of instruments. It might also be comprised of effect sizes that are not clinically meaningful, such as an odds ratio between 0.95 and 1.05 for the probability of hospital re-admission for a new hospital protocol, or an extended life expectancy of ± 15 minutes for a new drug with non-minimal side effects. This null interval hypothesis is also sometimes referred to as an "indifference zone".

While not necessarily required, there are many circumstances where despite an interval null being specified, there is a specified point null hypothesis of interest, denoted by $H_{00}: \theta = \theta_0$. In one of the above examples, the point null would be that the change in life expectancy is $\theta_0 = 0$ minutes, while the interval null hypothesis of non-clinically meaningful changes in life expectancy is $\Theta_0 = [-15{,}15]$ minutes (or whatever length of time is deemed to be not clinically relevant, particularly when balanced with the risks or side effects of the new drug). It is also important to note that the interval null does not need to be symmetric; for example, we could set $\Theta_0 = [-15{,}60]$ minutes, indicating that an increased life expectancy of less than 1 hour is not clinically meaningful, but any more than 15 minutes of decreased life expectancy is.

The other necessary component of the second-generation p-value is an interval estimate, I, for the parameter of quantity of interest θ. In general, the SGPV can be calculated for any type of interval, such as a confidence interval, likelihood support interval, bootstrap percentile interval, or Bayesian credible interval. Define the bounds of the interval estimate to be $I = [\hat{\theta}_l, \hat{\theta}_u]$. The second-generation p-value, denoted by p_δ, is defined as

$$p_\delta := \frac{|I \cap \Theta_0|}{|I|} \times \max\left(\frac{|I|}{2|\Theta_0|}, 1\right). \tag{3.2}$$

It essentially measures the fraction of overlap between the null hypothesis and the interval estimate, with a small-sample correction factor (the maximum term). See Blume et al. (2018, 2019) for more details. For a symmetric null hypothesis, δ is the half-length of the interval null (i.e., $\Theta_0 = [\theta_0 - \delta, \theta_0 + \delta]$). However as noted above, the null

Table 3.1 Description of three important outcomes for second-generation p-value.

Outcome:	How we get it:	What it indicates:
$p_\delta = 0$	$\lvert I \cap \Theta_0 \rvert = 0$, i.e., the interval estimate is entirely outside of the interval null	All hypotheses supported by the interval estimate are scientifically meaningful
$0 < p_\delta < 1$	The interval estimate I has some overlap with Θ_0, but is not contained entirely inside Θ_0	There are many scenarios covered by this "inconclusive" setting; generally, it means that the interval estimate supports both scientifically meaningful and trivial effects, to varying degree indicated by the magnitude of p_δ
$p_\delta = 1$	$I \subseteq \Theta_0$, i.e., the interval estimate is entirely inside of the interval null	All hypotheses supported by the interval estimate are scientifically trivial or not meaningful

hypothesis need not be symmetric, and so we will keep things general with the notation of $\Theta_0 = [\theta_0^-, \theta_0^+]$. Unlike the traditional p-value, the support of p_δ is $[0,1]$ (inclusive of 0 and 1). There are three important "zones" related to the sgpv: equal to 0, in between 0 and 1, and equal to 1. These are summarized in Table 3.1.

It is important to note that the second-generation p-value is not invariant to the scale on which it is calculated, because it involves interval lengths and overlap sizes. In general, if a symmetric scale exists, this should be used in place of the original scale (such as in the case of an odds ratio, hazard ratio, etc.). For example, if the underlying parameter of interest is an odds ratio, the SGPV should be calculated on the log odds ratio scale, using the interval estimate I and the interval null hypothesis Θ_0 defined in terms of the log odds ratio.

3.2.1 Operational characteristics

The distribution function $F_\theta(p_\delta)$ for the second-generation p-value has not been described, however a few summaries of it have been derived, for interval estimates I used in the SGPV calculation that are exactly or asymptotically Normal (Blume et al. 2018). We assume that an estimator for θ is obtained as $\hat{\theta} \coloneqq \hat{\theta}_n$ with sampling distribution $\hat{\theta}_n \overset{A}{\sim} N(\theta, V_n)$, and construct a $100(1-\alpha)\%$ confidence interval for θ as $I = [\hat{\theta}_l, \hat{\theta}_u] = [\hat{\theta}_n - z_{\alpha/2}\sqrt{V_n}\ ,\ \hat{\theta}_n + z_{\alpha/2}\sqrt{V_n}]$ where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and Φ is the standard normal cumulative distribution function. For this chapter, we will use the simplifying assumption that the true variance is known and used to calculate the confidence interval.

In Supplement 1 (S1) of (Blume et al. 2018), the following three probabilities are provided assuming that a symmetric null hypothesis of $\Theta_0 = [\theta_0 - \delta, \theta_0 + \delta]$ is used. Here, we describe these probabilities of falling into each of the three major "zones" for the SGPV, but for a general null hypothesis $\Theta_0 = [\theta_0^-, \theta_0^+]$ instead. Given the assumptions described above, we have the following operational characteristics or probability distribution summaries:

1) The probability of observing data that are compatible with the alternative,

$$\begin{aligned} P_\theta(p_\delta = 0) &= P(p_\delta = 0; \theta, V, n, \Theta_0) \\ &= \Phi\left[\frac{\theta_0^- - \theta}{\sqrt{V/n}} - z_{\alpha/2}\right] + \Phi\left[\frac{\theta - \theta_0^+}{\sqrt{V/n}} - z_{\alpha/2}\right] \end{aligned} \tag{3.3}$$

2) The probability of observing data that are compatible with the null,

$$\begin{aligned} P_\theta(p_\delta = 1) &= P(p_\delta = 1; \theta, V, n, \Theta_0) \\ &= \begin{cases} \Phi\left[\frac{\theta_0^+ - \theta}{\sqrt{V/n}} - z_{\alpha/2}\right] - \Phi\left[\frac{\theta_0^- - \theta}{\sqrt{V/n}} + z_{\alpha/2}\right] & \text{if } |I| \leq |\Theta_0| \\ 0 & \text{if } |I| > |\Theta_0| \end{cases} \end{aligned} \tag{3.4}$$

3) The probability of observing data that are inconclusive,

$$\begin{aligned} P_\theta(0 < p_\delta < 1) &= P(0 < p_\delta < 1; \theta, V, n, \Theta_0) \\ &= \begin{cases} 1 - \Phi\left[\frac{\theta - \theta_0^+}{\sqrt{V/n}} - z_{\alpha/2}\right] - \Phi\left[\frac{\theta_0^- - \theta}{\sqrt{V/n}} - z_{\alpha/2}\right] & \text{if } |I| \leq |\theta_0| \\ \quad -\Phi\left[\frac{\theta_0^+ - \theta}{\sqrt{V/n}} - z_{\alpha/2}\right] + \Phi\left[\frac{\theta_0^- - \theta}{\sqrt{V/n}} + z_{\alpha/2}\right] & \\ 1 - \Phi\left[\frac{\theta_0^- - \theta}{\sqrt{V/n}} - z_{\alpha/2}\right] - \Phi\left[\frac{\theta - \theta_0^+}{\sqrt{V/n}} - z_{\alpha/2}\right] & \text{if } |I| > |\theta_0| \end{cases} \end{aligned} \tag{3.5}$$

Note that because we are assuming a fixed variance of V, the length of the interval estimate $|I|$ is then fixed as well, giving the piecewise definitions above. To further simplify the demonstration of FDR calculations, we will set the variance as $V = |\Theta_0|^2$ (i.e., the standard deviation equals the width of the null hypothesis). Note that these properties are a function of only the ratio of the standardized effect size, i.e., of $\theta/\sqrt{V_n}$ and the standardized null interval bounds $\theta_0^-/\sqrt{V_n}$ and $\theta_0^+/\sqrt{V_n}$. This means that if the variance is known and sample size fixed, we could re-parameterize the entire effect space to the standardized effect space such that these properties are not directly a function of V_n (although still indirectly dependent on V_n). Alternately, we could re-parameterize the effect space by only the standard deviation $SD = \sqrt{V}$ and allow n to vary. In this and other illustrations throughout the chapter, we will use a null hypothesis of $\Theta_0 = [-0.1, 0.1]$, with variance $V = |\Theta_0|^2 = 0.04$. One possible interpretation for this setting is that it roughly represents null odds ratios between 0.9 and 1.11, with a variance of approximately 1 (on the odds ratio scale).

Figure 3.1(a)-(d) illustrates the probabilities in Equations (3.3)-(3.5), as a function of the standardized effect size $\theta/\sqrt{V}$ for one sample size n^* where $|I| > |\Theta_0|$, and another sample size n^{**} where $|I| \leq |\Theta_0|$. These illustrate some important properties of the SGPV compared to classical p-values, established in (Blume et al. 2018). Unlike with the classical p-value, the SGPV power curve at $\theta_0 = 0$ is not fixed at α, rather it is bounded above by α, with the exact value dependent on the null width $|\Theta_0|$, the variance V, and the sample size n. Further, $P_\theta(p_\delta = 0) \leq \alpha$ for all $\theta \in \Theta_0$, not only $\theta = 0$. In the limit, $P_\theta(p_\delta = 0)$ converges to 0 as $n \to \infty$ for $\theta = 0$. We note that this convergence happens as well as for $\theta \in (\theta_0^- + \varepsilon, \theta_0^+ - \varepsilon)$, i.e., the entirety of the null zone excluding the boundaries. At the null boundaries, e.g., at θ_0^+, we have that $\alpha/2 \leq P_{\theta_0^+}(p_\delta = 0) \leq \alpha$ and $P_{\theta_0^+}(p_\delta = 0) \to \alpha/2$ as $n \to \infty$.

Notably, $P_\theta(p_\delta = 0) \leq P_\theta(p \leq \alpha)\ \forall\theta$, regardless of the sample size or null width. This means that while type I error is reduced for the SGPV, the power is as well, for finite n. Essentially, the SGPV requires the lower bound $\hat{\theta}_l > \theta_0^+$ to be deemed significant rather than only requiring that $\hat{\theta}_l > 0$ (focusing on the positive effect space). Therefore, we are more likely to get inconclusive results with the SGPV than with the classical p-value (and also why we are much less likely to observe type I errors). However, we still see that $P_\theta(p_\delta = 0) \to 1$ as the effect size magnitude increases to infinity, as well as $\forall\theta \in (-\infty, \theta_0^- - \varepsilon) \cup (\theta_0^+ + \varepsilon, \infty)$ as $n \to \infty$.

Some examples of $P_\theta(p_\delta = 1)$ and $P_\theta(0 < p_\delta < 1)$ curves were given originally in Supplement 1 (S1) of (Blume et al. 2018), and some additional examples illustrated in Figure 3.1(a)-(d). Here, we outline some key properties of each. The probability of $p_\delta = 1$ is non-zero only for θ in Θ_0 and in a small area around it (i.e., $(\theta_0^- - \varepsilon, \theta_0^+ + \varepsilon)$), and only when V is small enough or n large enough such that $|I| \leq |\Theta_0|$ (otherwise, $P_\theta(p_\delta = 1) = 0\ \forall\theta$, as described in Equation (3.4)). As $n \to \infty$, $P_\theta(p_\delta = 1) \to 1$ for Θ_0 except for the boundaries, i.e., $\theta \in (\theta_0^- + \varepsilon\ , \theta_0^+ - \varepsilon)$. We find that at the boundaries of the null zone, $0 \leq P_\theta(p_\delta = 1) \leq \alpha/2$ across all sample sizes, where $P_\theta(p_\delta = 1) = 0$ at the smallest n such that $|I| \leq |\Theta_0|$, and $P_\theta(p_\delta = 1) \to \alpha/2$ as $n \to \infty$.

The probability of inconclusive results, $P_\theta(0 < p_\delta < 1)$, is either unimodal or bimodal as a function of θ, depending on the sample size, and appears to be non-zero for θ within double the width of the null bound (i.e., for $\theta \in (2\theta_0^- - \varepsilon, 2\theta_0^+ + \varepsilon)$), although this requires further work to confirm. For very small n, this probability as a function of θ is unimodal with a maximum at $\theta = 0$. At some point, as n becomes larger, $P_\theta(0 < p_\delta < 1)$ becomes bimodal, with maxima at the boundaries of Θ_0, and a local minimum within Θ_0 at $\theta = 0$. As $n \to \infty$, $P_0(0 < p_\delta < 1) \to 0$, as well as for all $\theta \in (\theta_0^- + \varepsilon\ , \theta_0^+ - \varepsilon)$. In general, $1 - \alpha \leq P_\theta(0 < p_\delta < 1) \leq 1 - \alpha/2$ across all sample sizes, and $P_\theta(0 < p_\delta < 1) \to 1 - \alpha$ as $n \to \infty$.

Note that all of the statements here are dependent on the assumption that the confidence interval estimate truly has the stated coverage of α. More generally, these will be factors of the true coverage – for example, $P_\theta(p_\delta = 0)$ will be bounded above by $1 - \omega$, where ω is the coverage probability of the interval estimate I used in the SGPV calculation; we will assume the apparent coverage is the truth in the entirety of the chapter.

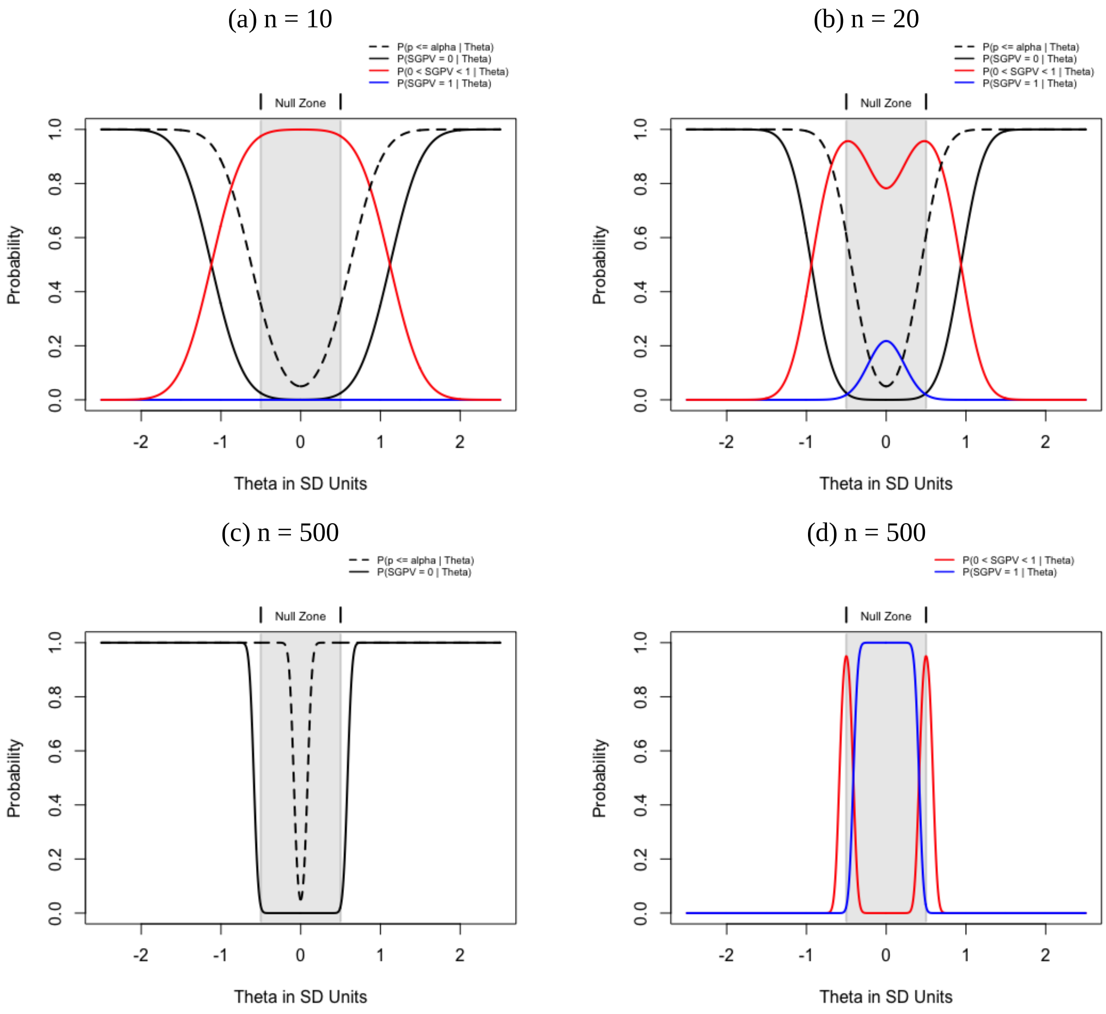


Figure 3.1 Operational characteristics of the second-generation p-value. (a)-(b) Second generation p-value operational probabilities $P_\theta(p_\delta = 0)$ (black), $P_\theta(p_\delta = 1)$ (blue), and $P_\theta(0 < p_\delta < 1)$ (red), as well as the classical p-value power curve $P_\theta(p \leq \alpha)$ (black dashed line). The null is $\Theta_0 = [-0.1, 0.1]$ with variance $V = 0.04$ (such that $\delta = |\Theta_0|/2 = \sqrt{V}/2$). The x-axis is given in standard deviation units, $\theta/\sqrt{V}$. (a) $n = n^*$, where n^* is such that $|I| > |\Theta_0|$ (thus $P_\theta(p_\delta = 1) = 0$), here $n^* = 10$. (b) $n = n^{**}$, where n^{**} is such that $|I| \leq |\Theta_0|$, here $n^* = 20$. (c)-(d) Large n, illustrating to some degree what happens to these probability curves as $n \to \infty$, here $n = 500$.

Design probabilities

In addition, there are several different design probabilities – generalized error rates or power type calculations – that may be of interest when conditioning on the general hypotheses H_0 and H_1. These are all based on the three "zones" for the SGPV, presuming that $p_\delta = 0$ is taken as a significant finding (in support of the alternative hypothesis), $p_\delta = 1$ is taken as a significant finding (in support of the null hypothesis), and $0 < p_\delta < 1$ is taken as an inconclusive result. Table 3.2 provides an overview of these main design probabilities when utilizing these three SGPV-based outcome regions.

Table 3.2 Summary of design probabilities for SGPV inference.

	$\boldsymbol{p_\delta = 1}$ (Call for Null)	$\boldsymbol{0 < p_\delta < 1}$ (Inconclusive)	$\boldsymbol{p_\delta = 0}$ (Call for Alternative)
$\boldsymbol{H_0}$ **True**	$P(p_\delta = 1 \vert H_0)$ Power	$P(0 < p_\delta < 1 \vert H_0)$ Undesirable outcome	$P(p_\delta = 0 \vert H_0)$ Error
$\boldsymbol{H_1}$ **True**	$P(p_\delta = 1 \vert H_1)$ Error	$P(0 < p_\delta < 1 \vert H_1)$ Undesirable outcome	$P(p_\delta = 0 \vert H_1)$ Power

The probability $P(p_\delta = 0|H_0)$ is general version of the typical type I error, declaring a significant finding in support of the alternative when the null hypothesis is actually true. The probability $P(p_\delta = 0|H_1)$ is the comparable power probability for the second-generation p-value. The probability $P(0 < p_\delta < 1|H_1)$ is an undesirable result, failing to correctly declare a significant finding in support of the alternative. It could be thought of as a sort of type II error. However, note that because $p_\delta = 1$ is not seen as an inconclusive result in this setting, this probability is not the complement of the power (i.e., power $\neq (1 - \text{type II})$), unlike in classical significance testing. As these quantities all condition on general composite hypotheses, they are not well defined as written. In later sections, we examine ways that these design probabilities might be defined.

Because $p_\delta = 1$ is taken as a finding in support of the null hypothesis, there is an additional set of design probabilities or error rates that can be defined. The probability $P(p_\delta = 1|H_1)$ is the null version of a type I error, the probability $P(p_\delta = 1|H_0)$ is the null version of a general power type probability, and $P(0 < p_\delta < 1|H_0)$ is another undesirable, inconclusive result (a null analogue to the type II error). All the null and alternative design probabilities can also be written generally as $P(p_\delta \in \Gamma|H_k)$, for regions $\Gamma \in (\{0\}, (0,1), \{1\})$.

For classical p-values, the null design probability (type I error) is calculated at $\theta_0 = 0$ and is fixed at α, i.e., $P(p \leq \alpha|H_0) = P_0(p \leq \alpha) = \alpha$. As described above, the point null type I error of the SGPV is less than that for the classical p-value, i.e., $P_0(p_\delta = 0) \leq P_0(p \leq \alpha) = \alpha$. In the setting relating to Figure 3.1(a), with only $n = 10$, this point null design probability is $P_0(p_\delta = 0) = 0.0004$, much lower than the classical p-value 0.05 type I error. However, because H_0 is an interval rather than a point null when using the SGPV, there are several different ways that the quantity could be defined; these will be explored in detail later sections. Although, regardless of method chosen to calculate the SGPV null design probability, we will find that $P(p_\delta = 0|H_0) \leq P(p \leq \alpha|H_0)$ for any n (due to $P_\theta(p_\delta = 0) \leq \alpha$ for any $\theta \in \Theta_0$).

For design probabilities conditioned on H_1 we might also select a single point value $\theta_a \in \Theta_1$, to calculate $P(p_\delta \in \Gamma|H_1) = P(p_\delta \in \Gamma|\theta = \theta_a)$. In classical null hypothesis significance testing, this alternative value is chosen somewhat arbitrarily, or the entire power curve is examined. In settings where the null hypothesis Θ_0 is chosen to represent clinically non-meaningful values, we might select for example $\theta_a = \theta_0^+ + \varepsilon$, just above the upper boundary of the null region (or use $\theta_a = \theta_0^+$ for simplicity), as the smallest effect size that would be clinically meaningful (in the positive effect space). For example, in the scenario relating to Figure 3.1(a), the point alternative design probabilities are $(p_\delta = 0|\theta = \theta_0^+) = 0.025$ and $(p \leq \alpha|\theta = \theta_0^+) = 0.35$ for second-generation and classical p-values, respectively. This unadjusted classical p-value rejection region has more than 10-fold the power of the SGPV, for the reasons described previously, i.e., due to a much more relaxed rejection threshold. Other approaches to specifying these null and alternative design probabilities will be described in Section 3.3.2.

3.3 False discovery rate quantity definitions

The design probabilities described in Section 3.2 are most relevant for cases where data has yet to be observed, namely study design. However, after the data are observed, the study characteristics are relevant but not the key quantities of interest. Specifically, we might be interested in some measure of reliability of the results. For example, if we observe $p \leq 0.05$ (or $p_\delta = 0$) and interpret this as a significant finding, what is the probability that this is in error (i.e., that the null hypothesis is in fact true)? These types of posterior probabilities, conditional on a statistic such as a p-value or second-generation p-value, are commonly referred to as Bayesian false discovery rates (Efron, Storey, et al. 2001; Efron, Tibshirani, et al. 2001; Storey 2001a, 2001b, 2002, 2003). While these were originally conceptualized in the context of classical p-values and z-values, the framework can be made general to accommodate any type of statistic used.

In this framework, the truth of the null hypothesis is regarded as a random variable. Define H as a Bernoulli indicator variable that the alternative hypothesis is true, such that $H = 0$ means that H_0 is true and $H = 1$ means that H_1 is true. The probability $P(H = 0)$ is denoted by π_0. Assume also that we are working with a general statistic $T(x)$, and specify a "two-group model" that is the conditional distribution $T(x)|H \sim (1 - H) \cdot F_0 + H \cdot F_1$, where F_0 is the distribution of $T(x)$ under the null hypothesis, and F_1 is the distribution of $T(x)$ under the alternative hypothesis. Thus, the "mixture distribution" for the statistic is $T(x) \sim \pi_0 \cdot F_0 + (1 - \pi_0) \cdot F_1$.

As stated, we might be interested in posterior probabilities $P(H = k|T(x) \in \Gamma)$ for $k \in \{0,1\}$. We will also use the shorthand $P(H_k|T(x) \in \mathit{\Gamma})$ interchangeably. Specifically, the Bayes false discovery rate is

$$FDR \coloneqq P(H_0|T(x) \in \Gamma) = P(H = 0|T(x) \in \Gamma), \tag{3.6}$$

where Γ is a region defined to be representative of a rejected null hypothesis, or some defined level of support for the alternative hypothesis. The key false discovery rate quantities based on classical p-values are summarized in Table 3.3(a). For example, for an unadjusted significance test based on classical p-values, the Bayes FDR is $P(H_0|p \in (0, \alpha])$. The corresponding summarization for second-generation p-values is given in Table 3.3(b), which is a simple extension of the two reliability probabilities described in (Blume et al. 2018) – which are the FDR and FCR. For the second-generation p-value, the Bayes FDR is best defined as $P(H_0|p_\delta = 0)$. Because the SGPV can also indicate support for the interval null, another quantity of interest is the Bayes false confirmation rate (FCR) for intervals that support the null, i.e., $FCR = P(H_1|p_\delta = 1)$. This is essentially a null analogue to the FDR.

We note that the reliability posterior probabilities in Table 3.3(a) are connected to the foundational false discovery rate work in multiple testing, as described in Chapter 2. When $T(x)$ is a classical p-value (or z-value), the Bayes false discovery rate, $P(H_0|p \in \Gamma)$, is equal to a quantity called the "positive false discovery rate" (pFDR) (Storey 2003; Efron 2010b). For a multiple testing scenario where m hypotheses are being tested, with R the number

Table 3.3 Summary of false discovery quantities for classical p-values and second-generation p-values. (a) Probabilistic (Bayes) false discovery quantities and terminology for classical p-values. (b) Probabilistic (Bayes) false discovery quantities and terminology for second-generation p-values.

(a) Classical p-value:

	Status:	
Result:	H_0 True	H_0 False
Reject Null	$P(H_0 \| p \leq 0.05)$ "false discovery rate" (FDR)	$P(H_1 \| p \leq 0.05)$ "true discovery rate" (TDR)
Fail to Reject Null (Inconclusive)	$P(H_0 \| p > 0.05)$	$P(H_1 \| p > 0.05)$ "false non-discovery rate" (FNR)

(b) Second-generation p-value:

	Status:	
Result:	H_0 True	H_0 False
Reject Null	$P(H_0 \| p_\delta = 0)$ "false discovery rate" (FDR)	$P(H_1 \| p_\delta = 0)$ "true discovery rate" (TDR)
Inconclusive	$P\left(H_0 \middle\| p_\delta \in (0,1)\right)^*$	$P\left(H_1 \middle\| p_\delta \in (0,1)\right)^*$
Support Null	$P(H_0 \| p_\delta = 1)$ "true confirmation rate" (TCR)	$P(H_1 \| p_\delta = 1)$ "false confirmation rate" (FCR)

* Included for completeness, but not how we might prefer to define non-discovery or non-confirmation rates.

of hypotheses that were rejected by a significance testing procedure, and V the number of these rejections corresponding to null hypotheses, the pFDR is defined as $pFDR \coloneqq E[V/R \,|R > 0]$. That is, the expected proportion of rejections that are made in error, for assessing at least one rejection. The pFDR is one component of the original Benjamini-Hochberg definition of the false discovery rate, which is $FDR \coloneqq E[V/\max(R,1)] = E[V/R \,|R > 0] \cdot \Pr(R > 0) = pFDR \cdot \Pr(R > 0)$ (Benjamini and Hochberg 1995). In the present chapter, we focus only on the Bayes/positive false discovery rate and ignore the BH definition of FDR. Therefore, as is common in the literature in the field, we will often use the acronym "FDR" alone to refer to the probabilistic FDR quantity $P(H_0|T(x) \in \Gamma)$.

In classical p-value testing, we may also define the Bayes false non-discovery rate (FNR) as $P\left(H_1 \middle| p \in (\alpha, 1)\right)$ for an unadjusted significance test (Genovese and Wasserman 2002). The term "false non-discovery rate" for classical p-values is a bit of a misnomer because it relies on a common misinterpretation of $p > 0.05$ as a result supporting the null hypothesis. This quantity is sometimes referred to in the diagnostic testing literature as the "false omission rate", which is somewhat of an improvement. We argue, however, that $p > 0.05$ is not technically a "false" conclusion for truly non-null hypotheses; rather, it is an inconclusive result – i.e., an "undesirable outcome". A more fitting term might be "missed discovery rate", because it assesses at what rate is the alternative hypothesis true for inconclusive (fail to reject) results. Regardless, we use the standard terminology for the chapter to avoid any confusion.

Defining an analogous FNR quantity for second-generation p-values is not straightforward. The most general quantities using the entire inconclusive range for second-generation p-values would be $P\big(H_1 \big| p_\delta \in (0,1)\big)$ or $P\big(H_0 \big| p_\delta \in (0,1)\big)$. These are included in Table 3.3(b) for completeness. However, these are too general to be useful because $p_\delta \in (0,1)$ covers such a wide range of inconclusive results. We could perhaps define a second-generation p-value false non-discovery rate as $P\big(H_1 \big| p_\delta \in (0,0.5)\big)$ and a false non-confirmation rate as $P\big(H_1 \big| p_\delta \in (0.5,1)\big)$. However, the full probability distribution function for p_δ has not been described (only the summary probabilities of Equations (3.3)-(3.5)), and therefore design probabilities related to SGPV regions such as $\Gamma = (0,0.5)$ or $\Gamma = (0.5,1)$ are presently unable to be calculated. We leave these topics for further study in future work.

Using Bayes' rule and the two-group specification, any of these posterior probabilities of interest can be re-expressed in two forms:

$$P(H_k | T(x) \in \Gamma) = \frac{P(H_k) \cdot P(T(x) \in \Gamma | H_k)}{P(T(x) \in \Gamma)} \tag{3.7}$$

$$= \frac{\pi_k \cdot P(T(x) \in \Gamma | H_k)}{\pi_0 \cdot P(T(x) \in \Gamma | H_0) + (1 - \pi_0) \cdot P(T(x) \in \Gamma | H_1)}, \tag{3.8}$$

for $k \in \{0,1\}$. Thus, to calculate or estimate any of these, we need to know the prior probability of the null π_0, and the relevant design probabilities of $P(T(x) \in \Gamma | H_0)$ and $P(T(x) \in \Gamma | H_1)$. Therefore, if these design probabilities and the prior probability are all known, or perhaps estimated from the observed data, it is straightforward to calculate any false discovery quantity desired.

3.3.1 Example: simple null and simple alternative hypotheses

The FDR and FCR rates for the second-generation p-value, when the simple approaches described in Section 3.2.1 are used (assuming $\theta = 0$ for the null hypothesis and some point alternative $\theta = \theta_a$) and setting the prior probability π_0 at 0.5, is the approach proposed and examined originally in (Blume et al. 2018). We illustrate their approach in Figure 3.2 as a function of θ_a in standard deviation units, i.e., $\theta_a/\sqrt{V}$, along with the analogous calculations for the FDR and FNR calculated based on the classical p-value. Note that the curves are drawn for the entire range of θ_a, including values inside the interval null hypothesis. This implies calculations $P(T(x) \in \Gamma | H_1) = P(T(x) \in \Gamma | \theta = \theta_a)$ for θ_a not in the alternative region, which do not make sense in the context of our interval null hypothesis framework. Really, we may be interested in the curves only for $\theta_a \in \Theta_1 = (-\infty, \theta_0^-) \cup (\theta_0^+, \infty)$.

Blume et al. (2018) provide discussion of some notable properties of each, for finite sample sizes and asymptotically as a function of θ_a and n, which can be observed as well in the Figure 3.2 illustration. For finite sample sizes, this simple SGPV FDR is less than the analogous classical p-value FDR, for all possible θ_a. The behavior of the FDR is driven by the size of the alternative design probability relative to the null design probability; for all θ_a, this ratio for the SGPV is larger than that of the classical p-value. The p-value FDR converges to and is

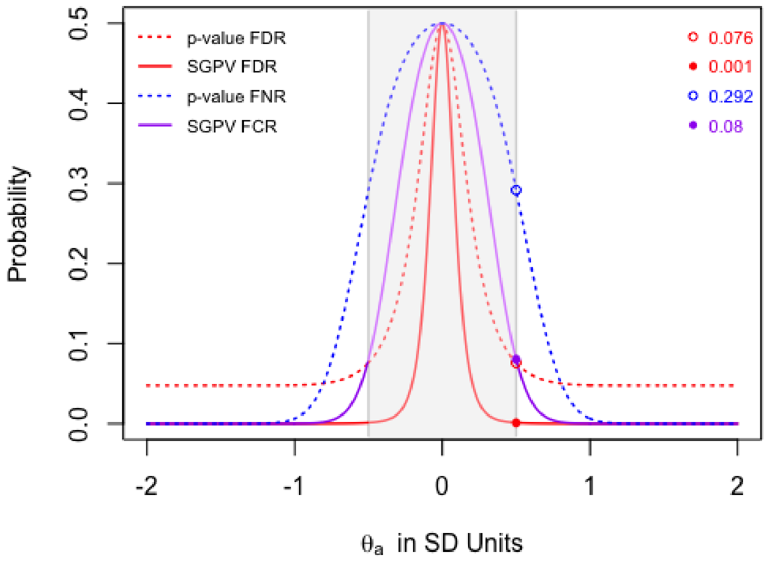
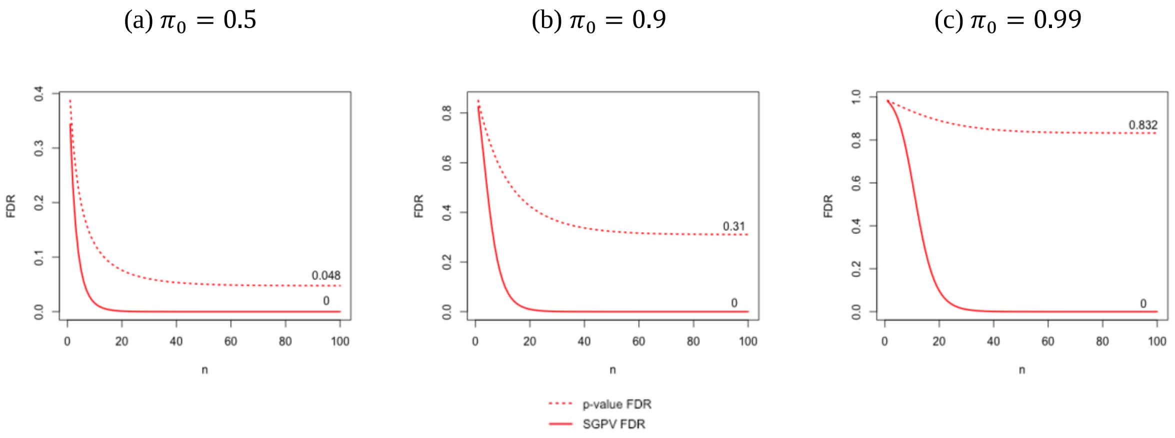


Figure 3.2 Comparison of FDR and FCR for SGPV to FDR and FNR for the classical p-value. The FDR calculation uses a point null at 0 and point alternative at θ_a, which is varied over the x-axis. Here, the null is $\Theta_0 = [-0.1,0.1]$ with variance $V = 0.04$ (such that $\delta = |\Theta_0|/2 = \sqrt{V}/2$), $\pi_0 = 0.5$, and the sample size is $n = 20$. The x-axis is given in standard deviation units, $\theta_a/\sqrt{V}$.

bounded below by $\alpha/(\alpha + r)$, where $r = (1 - \pi_0)/\pi_0$ (Wacholder et al. 2004). This is the p-value power $P_{\theta_a}(p \leq \alpha) \to 1$ but the type I error is fixed at α, thus the p-value FDR approaches this lower limit.

On the other hand, the convergence of the simple second-generation p-value point null design probability $P_0(p_\delta = 0) \to 0$ results in the (Blume et al. 2018) version of the SGPV FDR converging to 0 as $n \to \infty$. Thus, they make the argument that multiple comparisons adjustments are not needed for second-generation p-values due to this "natural" control. However, for finite (small) sample sizes, note that the FDR may not be controlled, and possibly greater than the FDR for procedures which use a multiple comparisons adjusted p-value (such as by a Bonferroni or Benjamini-Hochberg correction). This general idea is examined in further detail in Chapter 4. See also (Blume et al. 2018), Supplement 1, for a discussion of power comparisons for Bonferroni adjusted approaches. Regardless of what metric (p-value or SGPV) is used, the false discovery rates are bounded above by π_0, and the FNR and FCR are bounded above by $1 - \pi_0$.

To get a single number estimate of these rates, we might set the alternative design probabilities at the bound of the null interval, as described in Section 3.2.1. These calculations are marked on Figure 3.2 with a round point. The FDR estimated at the upper bound of the null interval, $\theta_a = \theta_0^+$ (which should indicate the smallest clinically meaningful effect size) is $P(H_0|p_\delta = 0) = 0.001$ for the SGPV compared to $P(H_0|p \leq 0.05) = 0.076$ for the p-value. These calculations for varying sample size n are given in Figure 3.3(a) for $\pi_0 = 0.5$.



(a) $\pi_0 = 0.5$ (b) $\pi_0 = 0.9$ (c) $\pi_0 = 0.99$

Figure 3.3 Simple FDR as a function of sample size for varying values of the proportion of tests which are null. Comparison of FDR for SGPV (solid red line) to FDR for the classical p-value (dotted red line) for varying n (for varying π_0). Note the differences in the y-axis ranges for (a)-(c); the axes for smaller π_0 are reduced, to be able to fully examine the differences in the classical p-value and SGPV FDRs.

Another important note to emphasize is that the limiting behavior for the p-value FDR is dependent upon the null prior probability π_0. That is, no matter how much data we observe, it is impossible to fully overcome the prior probability. These limiting behaviors are illustrated in Figure 3.3(b)-(c), in contrast with the convergence of the SGPV FDR to zero. If $\pi_0 = 0.5$, the p-value FDR converges to a reasonable value of 0.048. However, if we assume that the majority of the tests correspond to truly null hypotheses, such as $\pi_0 = 0.9$, then the p-value FDR can be no smaller than 0.31. In a possibly extreme example where $\pi_0 = 0.99$, the FDR converges to 0.83. Although, this large value of π_0 may actually not be uncommon in large-scale inference contexts.

3.3.2 Composite hypotheses approaches

Making the assumptions of point null and alternative values provides simple calculations for the false discovery rate quantities. However, one of the key aspects of the second-generation p-value is incorporating an interval null hypothesis. Naturally, we would want to account for this in our computations; we might want to marginalize across the set of interval null (or even alternative) regions. However, this needs to be done in the design probability space, rather than the false discovery rate space. The easiest numerical approaches which account for the interval hypotheses would be to take a minimum or maximum, or the simple average across the null zone. More advanced techniques using weighted marginalization may however provide more robust measures.

In general, we might marginalize over $P_\theta(p_\delta = 0)$ with a prior distribution or weighting function that corresponds to the respective hypothesis of interest. For null design probabilities, this would be calculated as $P(p_\delta \in \Gamma | H_0) = \int_{\Theta_0} P_\theta(p_\delta \in \Gamma) \cdot g_0(\theta) d\theta$ for a weighting function $g_0(\theta)$ that has a support restricted to Θ_0. For alternative design probabilities, this would be calculated as $P(p_\delta \in \Gamma | H_1) = \int_{\Theta_1} P_\theta(p_\delta \in \Gamma) \cdot g_1(\theta) d\theta$ for a weighting function $g_1(\theta)$ that has a support restricted to Θ_1 (or a subset of it). When these marginalization approaches are used for both design probabilities, the general false discovery quantity calculations then look like

$$P(H_k|T(x) \in \Gamma) \\ = \frac{\pi_k \cdot \int_{\Theta_k} P_\theta(T(x) \in \Gamma) \cdot g_k(\theta) d\theta}{\pi_0 \cdot \int_{\Theta_0} P_\theta(T(x) \in \Gamma) \cdot g_0(\theta) d\theta + \pi_1 \cdot \int_{\Theta_1} P_\theta(T(x) \in \Gamma) \cdot g_1(\theta) d\theta}, \tag{3.9}$$

for $k \in \{0,1\}$ (depending on whether we are interested in the false discovery, non-discovery, or confirmation rate). For simplicity, for the alternative design probabilities we will focus on the positive effect space (θ_0^+, ∞). That is, we calculate $P(p_\delta \in \Gamma|H_1) = \int_{\theta_0^+}^{\infty} P_\theta(p_\delta \in \Gamma) \cdot g_1(\theta) d\theta$ and $g_1(\theta)$ has support $(\theta_0^+, \infty) \subset \Theta_1$. However, note that for a symmetric null hypothesis and symmetric $P_\theta(T(x) \in \Gamma)$, this approach is equivalent to using a symmetric prior $g_1(\theta)$ that covers the entire range of Θ_1 (i.e., puts equal weight on effects of equal magnitude in the negative and positive alternative effect size space).

One straightforward approach to getting a marginalized null probability would be to use an unweighted average across the null zone, i.e., $g_0(\theta) \sim Unif(\theta_0^-, \theta_0^+)$. This prior function is straightforward and does not require the specification of any hyperparameters. Alternative types of prior functions that could be considered are a unimodal distribution (with either small or large variance), or perhaps a U-shaped prior function. This latter option would give a result similar to the frequentist approach to composite null design probabilities, which is to take the maximum value of the type I error over Θ_0. That is, for second-generation p-values, calculate $P(p_\delta \in \Gamma|H_0) = P_{\theta^m}(p_\delta \in \Gamma)$ where $\theta^m \coloneqq \mathrm{argmax}_{\theta \in \Theta_0} P_\theta(p_\delta = 0)$. For second-generation p-values, the maximum value of $P_\theta(p_\delta = 0)$ occurs at the boundaries of Θ_0. If we use this method in combination with the alternative design probability $P_{\theta_0^+}(p_\delta \in \Gamma)$ calculated at the smallest scientifically interesting effect (as in Section 3.3.1), the false discovery rate simplifies to π_0 and the FCR simplifies to $1 - \pi_0$ for any sample size n. It doesn't make much sense to use this estimator in practice; instead, we will focus on options which use the point null or the marginalization approaches.

Choosing a marginalization function for the alternative design probabilities is not so straightforward, particularly because the space is unconstrained and non-continuous. As mentioned above, we will focus on the positive alternative effect space to address the latter issue. In this setting we might want to choose a prior function that is unimodal near the edge of this range (as discussed above, because θ_0^+ would represent the smallest clinically meaningful effect size), and that converges to zero at an appropriate rate relative to the convergence of $P_\theta(T(x) \in \Gamma)$. What is "appropriate" is not clearly defined, but one key aspect is that if g_1 converges too slowly, the resulting design probability $P(p_\delta = 0|H_1)$ would be approximately equal to 1 regardless of sample size (because too much weight is put on values of θ where the power curve has already converged to 1). On the other hand, if $g_1(\theta)$ converges very quickly, then the majority of the weight is put on $\theta \in (\theta_0^+, \theta_0^+ + c)$ for some small c (this may however be the desired result in some cases). In the present chapter, we consider a linearly shifted Exponential distribution $g_1(\theta) \sim Exp(\lambda, a)$, where λ is the standard rate parameter of the exponential distribution and a is the shift parameter. Other priors for $g_1(\theta)$ might be useful; for example, if there is another, larger effect size $\theta_a > \theta_0^+$ of interest (i.e., θ_0^+ is the smallest detectable effect, and θ_a is the effect size that scientists are hoping to observe), then we might use a truncated Normal distribution with mean $\mu = \theta_a$ and support $\theta \in (\theta_0^+, \infty)$.

Before exploring example FDR calculations, we provide some alternate notation. When using the two-group model setup described in Section 3.3, the false discovery quantities can be written as

$$P(H_k|T(x) \in \Gamma) = \begin{cases} \left(1 + r\psi_{\Gamma,\mathrm{T}}\right)^{-1} & \text{if } k = 0 \\ \left(1 + \left(r\psi_{\Gamma,\mathrm{T}}\right)^{-1}\right)^{-1} & \text{if } k = 1 \end{cases} \tag{3.10}$$

$$= \left(1 + \left(r\psi_{\Gamma,\mathrm{T}}\right)^{1-2k}\right)^{-1}, \tag{3.11}$$

for $k \in \{0,1\}$, where $r = \pi_1/\pi_0$ is the prior odds for the alternative versus the null, and

$$\psi_{\Gamma,\mathrm{T}} = \frac{P(T(x) \in \Gamma|H_1)}{P(T(x) \in \Gamma|H_0)} \tag{3.12}$$

is the likelihood ratio (LR) for the alternative versus null hypotheses. When using the marginalization methods as in Equation (3.9), this looks more like a Bayes factor (Jeffreys 1935; Kass and Raftery 1995); however, we will make sure to use $g_0(\theta)$ and $g_1(\theta)$ that do not overlap, and thus is still a proper likelihood ratio (Royall 1997; Blume 2002). Note that quantities such as the FDR and FNR use different regions Γ, therefore the likelihood ratio $\psi_{\Gamma,\mathrm{T}}$ in calculating the FDR is different than the likelihood ratio used in calculating the FNR.

3.3.3 Common scenarios

Example 1

In Section 3.3.1, we calculated the false discovery quantities based on the point null assumption $P(T(x) \in \Gamma|H_0) \coloneqq P_0(T(x) \in \Gamma)$ and the point alternative assumption $P(T(x) \in \Gamma|H_1) \coloneqq P_{\theta_0^+}(T(x) \in \Gamma)$; we refer back to this approach as Example 1. Recall that this is a specification of the (Blume et al. 2018) approach, setting $\theta_a = \theta_0^+$.

Example 2

Using the simple unweighted average of $P_\theta(T(x) \in \Gamma)$ across the interval null region, which is equivalent to using the prior function $g_0(\theta) \sim Unif(\theta_0^-, \theta_0^+)$, with the point alternative assumption gives the likelihood ratio

$$\psi_{\Gamma,\mathrm{T}} = \frac{P_{\theta_0^+}(T(x) \in \Gamma)}{\int_{\Theta_0} P_\theta(T(x) \in \Gamma) \cdot g_0(\theta) d\theta} \tag{3.13}$$

for $g_0(\theta) = 1/(\theta_0^+ - \theta_0^-)$. This approach provides a balance between using the minimum type I error (at the point null) and the maximum type I error.

Example 3

It might be the case that we have specified an interval null hypothesis, but that the point null θ_0 is the least interesting effect (“most null”), while values of θ near the boundaries of Θ_0 are questionably of interest. To account for this in the calculation of the design probabilities for H_0, we might use a weighted average rather than the unweighted, uniform distribution approach employed above. A Normal distribution centered at point null θ_0 and truncated to cover only Θ_0 can give several flexible forms for $g_0(\theta)$; this distribution is denoted either by $tNorm(\mu, \sigma, \mathrm{a}, \mathrm{b})$, with a and b the truncation bounds, or by $tNorm(\mu, \sigma, \Omega)$, where Ω is the truncation range of the distribution (such that $\Omega = (\mathrm{a}, \mathrm{b})$). In this example, will use $g_0(\theta) \sim tNorm(\mu = 0, \sigma = |\Theta_0|/5\ , a = \theta_0^-, b = \theta_0^+)$, which truncates the Normal distribution at ± 5 standard deviations from 0 and results in a bell-shape type curve for g_0.

Using this approach with the point alternative assumption, the likelihood ratio quantities $\psi_{\Gamma,\mathrm{T}}$ have the same form as in Equation (3.13), but with prior distribution

$$g_0(\theta) = \begin{cases} \varphi\left(\frac{\theta - 0}{|\Theta_0|/5}\right) \cdot \left[\int_{\Theta_0} \varphi\left(\frac{\theta - 0}{|\Theta_0|/5}\right) d\theta\right]^{-1} & \text{for } \theta \in \Theta_0 \\ 0 & \text{for } \theta \notin \Theta_0 \end{cases} \tag{3.14}$$

where φ is the standard normal density $N(0,1)$. This approach places most of the weight on values close to 0, but still accounts for all θ in the null region. We could consider alternate variance parameters for this truncated Normal distribution, such as $\sigma = |\Theta_0|$ to get a relatively flat prior or $\sigma = |\Theta_0|/20$ to get a strong prior with the vast majority of the weight right around 0.

Example 4

In this example, we will utilize the shifted exponential marginalization approach for calculating the alternative design probability $P(T(x) \in \Gamma | H_1)$. Combined with the point null assumption, the likelihood ratio looks like

$$\psi_{\Gamma,\mathrm{T}} = \frac{\int_{\theta_0^+}^{\infty} P_\theta(T(x) \in \Gamma) \cdot g_1(\theta) d\theta}{P_{\theta_0}(T(x) \in \Gamma)} \tag{3.15}$$

for $g_1(\theta) \sim Exp(\lambda, a = \theta_0^+)$. In our examples, we set λ to have a default value of $\lambda = (\theta_0^+ + 1)/(V_n)^{0.1898}$. This function was chosen because it provided sensible values for power calculations for both the classical p-value and SGPV for various Θ_0 and sample sizes. Note that we have not derived a closed-form solution for the marginalized quantities; rather, we rely on numerical calculations using the `integrate` function in R.

Example 5

Combining the uniform marginalization approach for the null design probability with the shifted exponential alternative marginalization, the likelihood ratio has the form

$$\psi_{\Gamma,\mathrm{T}} = \frac{\int_{\theta_0^+}^{\infty} P_\theta(T(x) \in \Gamma) \cdot g_1(\theta) d\theta}{\int_{\Theta_0} P_\theta(T(x) \in \Gamma) \cdot g_0(\theta) d\theta} \tag{3.16}$$

for $g_1(\theta) \sim Exp(\lambda, a = \theta_0^+)$ and $g_0(\theta) \sim Unif(\theta_0^-, \theta_0^+)$.

Example 6

Lastly, we will use the truncated Normal for the null, with the shifted exponential for the alternative. This likelihood ratio has the same form as in Equation (3.16), but with g_0 as in Equation (3.14).

3.3.4 Numerical illustrations and asymptotic behavior

The classical p-value based false discovery rate ($P(H_0|T(x) \in \Gamma)$ for $T(x) = p$ and $\Gamma = (0,0.05]$), and the second-generation p-value based false discovery rate ($P(H_0|T(x) \in \Gamma)$ for $T(x) = p_\delta$ and $\Gamma = \{0\}$) are provided in Table 3.4(a) for each of the example approaches described above, with varying sample sizes. The p-value FDRs are only calculated for methods which use the point value for the null design probability (Examples 1 and 4). Note that for the p-value FDR using the averaged alternative design probability, we use the same shifted exponential distribution as for the SGPV (that is, shifted to $a = \theta_0^+$) for purposes of comparison. The p-value false non-discovery rate ($P(H_1|T(x) \in \Gamma)$ for $T(x) = p$ and $\Gamma = (0.05,1)$) for the examples where a point null is used are given in Table 3.4(b). The second-generation p-value false confirmation rate ($P(H_1|T(x) \in \Gamma)$ for $T(x) = p_\delta$ and $\Gamma = \{1\}$) calculations are in Table 3.4(c).

For approaches which use the same method for calculating the null design probabilities for $p_\delta = 0$, the SGPV FDR using the exponential marginalization method will be smaller for finite values of n than the point alternative approach. For finite n, using the truncated normal marginalized null provides a balance between the minimum null (at the point 0) and the uniform marginalization approach. For almost all the considered approaches, the p-value FDR is larger than the SGPV FDR across all sample sizes. The exception is when the uniform null is used in combination with the point alternative at θ_0^+ (Example 2) for small sample sizes (e.g., 10 – 100). However, in this setting, the SGPV FDR becomes smaller for $n > 228$ as the ratio of the alternative to null design probability becomes larger for the second-generation p-value.

We see that the choices of null and alternative method only impact the second-generation p-value FDR calculations for finite n; the limiting behavior is the same for all considered approaches, although the convergence rate differs. As $n \to \infty$, the SGPV FDRs converge to 0, as opposed to the p-value FDR which converges to $\alpha/(\alpha + r)$ (as noted in Section 3.3.1 (Wacholder et al. 2004)). As long as the null design probability converges to zero, and the alternative design probability converges to a non-zero constant, the SGPV FDR converges to zero. As described in Section 3.2.1, $P_\theta(p_\delta = 0) \to 0$ as $n \to \infty$ for $\theta \in (\theta_0^- + \varepsilon, \theta_0^+ - \varepsilon)$, such that any approach – aside from the maximization method where the null design probability is calculated at the null boundary, or perhaps a weighting method which places some point mass at the null boundaries – will result in $P(p_\delta = 0|H_0) \to 0$. More details are provided in Remark 3.A in the Appendix.

In the illustrative examples in Table 3.4(a), we see that SGPV FDR has converged for the point null method by $n = 100$ or $n = 20$ (for the point alternative and shifted exponential alternatives, respectively). However, the null marginalization methods do not converge even by $n = 500{,}000$ for several of the example approaches. Therefore, while the asymptotic behavior is the same, even large-sample behavior can differ between approaches,

Table 3.4 False discovery, non-discovery, and confirmation rate quantity results for classical p-value significance testing and second-generation p-value based cutoffs. Assumes $\pi_0 = 0.5$, i.e., $r = 1$ with a null region $\Theta_0 = [-0.1,0.1]$ and variance $V = |\Theta_0|^2 = 0.04$.

(a) p-Value $FDR = P(H_0|p \leq 0.05)$ and SGPV $FDR = P(H_0|p_\delta = 0)$

	Ex. 1		**Ex. 2**	**Ex. 3**	**Ex. 4**		**Ex. 5**	**Ex. 6**
	p-value	SGPV	SGPV	SGPV	p-value	SGPV	SGPV	SGPV
n = 10	0.1242	0.0157	0.1929	0.0831	0.053	0.0006	0.0086	0.0033
n = 20	0.0759	0.0011	0.1445	0.0424	0.0495	0	0.0057	0.0015
n = 100	0.0477	0	0.0703	0.0107	0.0476	0	0.0023	0.0003
n = 500,000	0.0476	0	0.0034	0.0003	0.0476	0	0.0001	0
∞	0.0476 $\left(\frac{\alpha}{\alpha+r}\right)$	0	0	0	0.0476 $\left(\frac{\alpha}{\alpha+r}\right)$	0	0	0

(b) p-Value $FNR = P(H_1|p > 0.05)$

	Ex. 1	**Ex. 4**
n = 10	0.4053	0.2454
n = 20	0.2917	0.2149
n = 100	0.0012	0.1514
n = 500,000	0	0.0279
∞	0	0

(c) SGPV $FCR = P(H_1|p_\delta = 1)$

Ex. 1	**Ex. 2**	**Ex. 3**	**Ex. 4**	**Ex. 5**	**Ex. 6**
*	*	*	*	*	*
0.0803	0.1366	0.1033	0.0052	0.0094	0.0069
0.0244	0.0396	0.0290	0.0009	0.0015	0.0011
0.0244	0.0248	0.0244	0	0	0
0.0244 $\left(\frac{\alpha/2}{\alpha/2+\frac{1}{r}}\right)$	0.0244 $\left(\frac{\alpha/2}{\alpha/2+\frac{1}{r}}\right)$	0.0244 $\left(\frac{\alpha/2}{\alpha/2+\frac{1}{r}}\right)$	0	0	0

which further emphasizes the need for care to be taken with selecting the weighting method. More rigorous derivation of convergence rates for various approaches for calculating the FDRs are left as an area for future work.

The choice of method for the alternative design probability can affect the SGPV FCR convergence. In our examples, the marginalization approach with the shifted exponential (Examples 4-6) converges to 0, while the approach of setting the alternative design probability at the edge of the null region (Examples 1-3) will have $FCR \rightarrow (\alpha/2)/(\alpha/2 + r^{-1})$ in the limit. Briefly, this is because $P_\theta(p_\delta = 1) \rightarrow 0$ for $\theta \in (-\infty, \theta_0^- - \varepsilon) \cup (\theta_0^+ + \varepsilon, \infty)$, such that choices for the alternative design probability which do not place point mass at the null boundaries will converge to zero. For details on this and the asymptotic results for Examples 1-6, refer to Remark 3.B.

The calculations in Table 3.4 provide the true FDR for when each specified null and alternative marginalization functions match the true distribution of underlying effects. If we wish to take a full conservative approach for calculating the FDR, we might consider the specifications as outlined in Example 2, with the uniform null function and the calculation of the alternative design probability at the boundary of the null region. Other choices could result in either less conservative, or underestimated FDRs, depending on whether these choices correctly specify the distribution of effects. As much as possible, we would prefer to avoid specification of these probabilities, and thus the issue of misspecification. Empirical methods may offer a solution, and we take a step in this direction in Section 3.4.

3.4 Large-scale inference: empirical estimation

To calculate the false discovery rate for some inferential statistic, such as the classical p-value or second-generation p-value, we have seen that the two main design probabilities, $P(T(x) \in \Gamma|H_0)$ and $P(T(x) \in \Gamma|H_1)$, must be specified in some manner. This may be done either via a selection of a single point, utilizing a maximization approach, or specifying a marginalization function across null and/or alternative effects. If these specifications are not correct, the resulting FDRs may be over- or under-estimated. This presents a challenge, because the specification is generally arbitrary. For inference based on classical p-values, empirical Bayes estimation methods have been used to avoid or minimize the requirement of such specifications, by leveraging the information available from a large number of simultaneous tests (such as Efron 2004; Tang et al. 2007; Efron 2010b, to name just few). Additional discussion of and reference to such methods can also be found in Chapter 2. Such methods have not been examined for inference based on second-generation p-values. In this section, we examine a first step towards achieving such an FDR quantity estimator in large-scale inference.

3.4.1 Empirical Bayes mixture estimates

In the setting of FDR estimation for classical p-values or z-values, a simple approach for empirical estimation of false discovery rate quantities replaces the two-group form of the mixture distribution with a direct non-parametric estimate of it, using the empirical cumulative distribution function (ECDF) of the p-values. It is defined for m statistical tests with p-values $p^1, \dots, p^m$ as $\hat{F}(c) = \#\{j: p^j \leq c\}/m$, i.e., the observed proportion of p-values less than or equal to c. We can generalize this to estimate the probability that p falls in any general region Γ as $\hat{P}(p \in \Gamma) = \#\{j: p^j \in \Gamma\}/m$. The second-generation p-value analogue would be to estimate $\hat{P}(p_\delta \in \Gamma) = \#\left\{j: p_\delta^j \in \Gamma\right\}/m$ for m statistical tests with second-generation p-values $p_\delta^1, \dots, p_\delta^m$. In our contexts, we are interested in

$$\hat{P}(p_\delta = 0) = \#\left\{j: p_\delta^j = 0\right\}/m \tag{3.17}$$

and

$$\hat{P}(p_\delta = 1) = \#\left\{j: p_\delta^j = 1\right\}/m. \tag{3.18}$$

These empirical mixture distribution or probability estimates can then be substituted for $P(p_\delta = 0)$ or $P(p_\delta = 1)$ in the denominator of false discovery quantities as described in Equation (3.7). Using this empirical mixture estimate $\hat{P}(p_\delta = 0)$ eliminates the need for specification of the alternative design probability for the SGPV FDR, but not for the SGPV FCR.

We illustrate numerical results for this empirical mixture distribution estimate with a simple simulation scenario. Estimated effects $\hat{\theta}_n^1, \dots, \hat{\theta}_n^m$ were directly generated from a simple 2-group model with a single common null mean at 0 and single common alternative mean at the boundary of the null hypothesis zone. The number of truly null effects ($m_0 = \pi_0 \cdot m$) and number of truly alternative effects ($m_1 = (1 - \pi_0) \cdot m$) were kept as fixed quantities, rather than random variables. The null region is $\Theta_0 = [-0.1,0.1]$, and the variance is $V = |\Theta_0|^2 = 0.04$. The alternative mean μ_1 was set at $\mu_1 = \theta_0^+ = 0.1$, and we examined two values for π_0, 0.5 and 0.9. Note that while have focused on $\pi_0 = 0.5$ up until this point, it is common in large-scale inference settings to expect a majority of the hypotheses tested to be null (e.g., $\pi_0 \geq 0.9$ is sometimes cited as a general rule of thumb (Efron 2010b). Thus, $\pi_0 = 0.9$ represents a setting that is more likely to be encountered in practice. For this setup, the underlying mixture probabilities are written generally as

$$P(T(x) \in \Gamma) = \pi_0 \cdot P_0(T(x) \in \Gamma) + (1 - \pi_0) \cdot P_{\mu_1}(T(x) \in \Gamma). \tag{3.19}$$

Note that these quantities are also reliant on the null zone, variance, and sample size such that we should more explicitly write the mixture probability as $P(T(x) \in \Gamma; V, n, \Theta_0)$; however, we will stick with the shorthand notation for convenience.

Examples of the empirical mixture estimates $\hat{P}(p \in \Gamma)$ and $\hat{P}(p_\delta \in \Gamma)$ for various sample sizes n (10, 20, 100, and 50,000) and number of tests m (100, 1,000, and 10,000) are given in Table 3.5(a) for $\pi_0 = 0.5$ and Table 3.5(b) for $\pi_0 = 0.9$. Note that these examples are simply for illustration, as a single simulated data set does not represent the overall behavior of the estimator. Assessments such as the bias and variance of these estimators is left for future work. The true mixture probabilities, calculated by Equation (3.19), are also given for each sample size and asymptotically.

Firstly, we notice that the mixture probabilities for all quantities vary widely with sample size n, except perhaps for second-generation p-value discoveries, which generally have a low probability of occurring. In the provided example simulated data sets, the empirical mixture probability estimates are generally close to the true mixture probabilities. This was true for both $\pi_0 = 0.5$ and $\pi_0 = 0.9$. For smaller number of tests, such as $m = 100$, we did observe some estimates that were off by a quite a bit. The magnitude of error appears to be worse for second-generation p-values than for classical p-values for both values of π_0. This is likely because it is more difficult to get a second-generation p-value rejection than a classical p-value rejection, so this relatively rare event is more difficult to estimate well. The error was much smaller for larger values of m, but still noticeable. However, we cannot say conclusively from this single example estimate.

Note that for $n = 10$, the sample size is too small relative to the indifference zone and variance to be able to observe a confirmation with the second-generation p-value (0 rejections observed, by definition). In our settings, we require a sample size of at least $n \geq 15.4$ for $P_\theta(p_\delta = 1) > 0$ to be possible. Additionally, in some settings, we did not observe any $T(x) \in \Gamma$ in our simulated example data. For example, in two instances we did not observe any $p_\delta = 0$, therefore the empirical mixture estimate equals zero (for $n = 10$, $m = 100$, as well as for $n = 50{,}000$, $m = 1{,}000$ in Table 3.5(b)). When this happens, the empirical FDR estimate is undefined; this is sensible, because we would have no interest in estimating the probability that the null is true for rejected hypotheses, when there aren't any observed. This was much more likely to occur for small m, but not impossible for larger m. Generally, the smaller m_1 is (the number of truly non-null hypotheses), the more likely this is to occur. Overall, it appears that the sample size n does not have much of an impact on how close the empirical mixture estimates are to the truth.

Table 3.5 True mixture probabilities and empirical estimates for various values of null proportion, n, and m. The true mixture probabilities are $\mathrm{True}(\mathrm{n}) = P(T(x) \in \Gamma; n)$, and the empirical estimates are $\mathrm{Est}_{\mathrm{i}}(\mathrm{n}, \mathrm{m}) = \widehat{P}(T(x) \in \Gamma; n, m)$. The asymptotic true mixture probabilities are also provided.

(a) $\pi_0 = 0.5$

		p-value	SGPV	p-value non-	SGPV
		discovery	discovery	discovery	confirmation
		$P(p \leq 0.05)$	$P(p_\delta = 0)$	$P(p > 0.05)$	$P(p_\delta = 1)$
n = 10	$\mathbf{True(n)}$	**0.2013**	**0.0127**	**0.7987**	**0**
m = 100	$\mathrm{Est_i(n, m)}$	0.2500	0.0300	0.7500	0.0000
m = 1,000		0.2120	0.0120	0.7880	0.0000
m = 10,000		0.2039	0.0122	0.7961	0.0000
n = 20	$\mathbf{True(n)}$	**0.3294**	**0.0125**	**0.6706**	**0.1183**
m = 100	$\mathrm{Est_i(n, m)}$	0.3200	0.0300	0.6800	0.1100
m = 1,000		0.2960	0.0060	0.7040	0.1150
m = 10,000		0.3354	0.0119	0.6646	0.1210
n = 100	$\mathbf{True(n)}$	**0.5244**	**0.0125**	**0.4756**	**0.5113**
m = 100	$\mathrm{Est_i(n, m)}$	0.5200	0.0200	0.4800	0.4900
m = 1,000		0.5230	0.0100	0.4770	0.5140
m = 10,000		0.5237	0.0127	0.4763	0.5115
n = 50,000	$\mathbf{True(n)}$	**0.525**	**0.0125**	**0.475**	**0.5125**
m = 100	$\mathrm{Est_i(n, m)}$	0.5400	0.0100	0.4600	0.5100
m = 1,000		0.5210	0.0130	0.4790	0.5110
m = 10,000		0.5254	0.0129	0.4746	0.5132
∞	$\mathbf{True(\infty)}$	**0.525**	**0.0125**	**0.475**	**0.5125**
		$1 - (1 - \alpha) \cdot \pi_0$	$(1 - \pi_0) \cdot \alpha/2$	$(1 - \alpha) \cdot \pi_0$	$\pi_0 + (1 - \pi_0) \cdot \alpha/2$

(b) $\pi_0 = 0.9$

		p-value	SGPV	p-value non-	SGPV
		discovery	discovery	discovery	confirmation
		$P(p \leq 0.05)$	$P(p_\delta = 0)$	$P(p > 0.05)$	$P(p_\delta = 1)$
n = 10	$\mathbf{True(n)}$	**0.0803**	**0.0029**	**0.9197**	**0**
m = 100	$\mathrm{Est_i(n, m)}$	0.0900	0.0000	0.9100	0.0000
m = 1,000		0.0870	0.0020	0.9130	0.0000
m = 10,000		0.0767	0.0034	0.9233	0.0000
n = 20	$\mathbf{True(n)}$	**0.1059**	**0.0025**	**0.8941**	**0.1977**
m = 100	$\mathrm{Est_i(n, m)}$	0.1200	0.0100	0.8800	0.2300
m = 1,000		0.0980	0.0020	0.9020	0.2090
m = 10,000		0.1073	0.0025	0.8927	0.1977
n = 100	$\mathbf{True(n)}$	**0.1449**	**0.0025**	**0.8551**	**0.9004**
m = 100	$\mathrm{Est_i(n, m)}$	0.1400	0.0100	0.8600	0.8900
m = 1,000		0.1570	0.0040	0.8430	0.8940
m = 10,000		0.1484	0.0029	0.8516	0.8997
n = 50,000	$\mathbf{True(n)}$	**0.145**	**0.0025**	**0.855**	**0.9025**
m = 100	$\mathrm{Est_i(n, m)}$	0.1300	0.0100	0.8700	0.9000
m = 1,000		0.1450	0.0000	0.8550	0.9030
m = 10,000		0.1457	0.0020	0.8543	0.9037
∞	$\mathbf{True(\infty)}$	**0.145**	**0.0025**	**0.855**	**0.9025**
		$1 - (1 - \alpha) \cdot \pi_0$	$(1 - \pi_0) \cdot \alpha/2$	$(1 - \alpha) \cdot \pi_0$	$\pi_0 + (1 - \pi_0) \cdot \alpha/2$

3.4.2 False discovery rate estimates

Using Equation (3.7), the general form of an empirical Bayes second-generation p-value FDR estimate, utilizing the empirical mixture probability estimates of Section 3.4.1, is

$$\hat{P}(H_0|p_\delta = 0) = \frac{\pi_0 \cdot P(p_\delta = 0|H_0)}{\#\{j: p_\delta^j = 0\}/m} \tag{3.20}$$

and the second-generation p-value FCR estimate is

$$\hat{P}(H_1|p_\delta = 1) = \frac{\pi_1 \cdot P(p_\delta = 1|H_1)}{\#\{j: p_\delta^j = 1\}/m}. \tag{3.21}$$

The classical p-value FDR and FNR forms parallel this. Note that we have not yet dealt with the unknown null and alternative proportions, π_0 and π_1. One benefit of the empirical mixture approach is that we are able to set the unknown π_0 to 1 in the FDR calculation and π_1 to 1 in the FCR (or FNR) calculation, resulting in upper bound estimates for each:

$$\hat{P}(H_0|p_\delta = 0) \leq \frac{1 \cdot P(p_\delta = 0|H_0)}{\#\{j: p_\delta^j = 0\}/m} \tag{3.22}$$

$$\hat{P}(H_1|p_\delta = 1) \leq \frac{1 \cdot P(p_\delta = 1|H_1)}{\#\{j: p_\delta^j = 1\}/m}. \tag{3.23}$$

This is not possible when the mixture is calculated directly from mixture of the two design probabilities, otherwise the false discovery quantities will always simplify to 1. We use the same simulation setup from Section 3.4.1 to illustrate example calculations for these empirically estimated false discovery quantities.

The results for the example simulated data sets with $\pi_0 = 0.5$ and $\pi_0 = 0.9$ are given in Table 3.6(a) and Table 3.6(b). While the denominators of the FDR quantities (the mixture distribution) are estimated empirically, the numerator design probabilities must still be specified. Each of the design probability calculation methods from Section 3.3.2 are considered, for the numerators of the FDR and FCR (or FNR). As before, we do not include any marginalization approaches for the null design probabilities for the classical p-value FDR. We include both the false discovery rate quantity estimates from using either i) the true π_0 and π_1 values (as in Equations 3.20 and 3.21), and ii) the upper bound estimates (Equations 3.22 and 3.23). The former are the "oracle" estimator (i.e., if the true π_0 and π_1 were known) and are denoted by "$FDR_i^{est}(n)$: Or". The latter upper bound estimates are denoted by "$FDR_i^{est}(n)$: Ub". Recall that our true null mean is at 0, and the true alternative common mean is $\mu_1 = \theta_0^+$. Thus, the FDR quantities based on the point null represent those using the correctly specified numerator null design probabilities, and the FCR or FNR quantities based on the point alternative represent the correctly specified numerator alternative design probabilities.

It is clear that, if we know what the true π_0 and data-generating mechanism are (i.e., using design probabilities which reflect the underlying truth), then the false discovery rate quantity estimates using the empirical mixture distribution are quite close to the true FDR or FNR/FCR values. Note that there are some larger deviations for smaller m, due to less accurate estimation of the mixture distribution. However, the more important observations

Table 3.6 True false discovery quantities and estimated false discovery quantities with the empirical mixture probability estimate for the simple simulation for various values of null proportion, n, and m. The oracle estimator ($\text{FDR}_\text{i}^\text{est}$: Or) uses the true π_0 and π_1 values; the upper bound ($\text{FDR}_\text{i}^\text{est}$: Ub) sets π_0 at 1 for the FDR, and π_1 at 1 for the FNR and FCR. The results for various approaches to specifying the null and alternative design probabilities, as described in Section 3.3, are given for p-values and second-generation p-value rejection regions.

(a) $\pi_0 = 0.5$

		p-value FDR	SGPV FDR			p-value FNR		SGPV FCR	
		Point Null	Point Null	Unif. Null	T. Norm Null	Point Alt.	Shifted Exp.	Point Alt.	Shifted Exp.
n = 10	$\mathbf{FDR^{true}}$	**0.1242**	**0.0157**	-	-	**0.4053**	-	*****	-
m = 100	$\text{FDR}_\text{i}^\text{est}$: Or	0.1	0.0066	0.0996	0.0377	0.4316	0.0712	*	*
	$\text{FDR}_\text{i}^\text{est}$: Ub	0.2	0.0133	0.1991	0.0755	0.8632	0.1425		
m = 1,000		0.1179	0.0166	0.2489	0.0944	0.4108	0.0678	*	*
		0.2358	0.0332	0.4978	0.1887	0.8216	0.1356		
m = 10,000		0.1226	0.0163	0.2448	0.0928	0.4066	0.0671	*	*
		0.2452	0.0327	0.4897	0.1856	0.8132	0.1342		
n = 20	$\mathbf{FDR^{true}}$	**0.0759**	**0.0011**	-	-	**0.2917**	-	**0.0803**	-
m = 100	$\text{FDR}_\text{i}^\text{est}$: Or	0.0781	0.0005	0.0704	0.0185	0.2877	0.0297	0.0864	0.0052
	$\text{FDR}_\text{i}^\text{est}$: Ub	0.1563	0.0009	0.1408	0.0369	0.5753	0.0595	0.1727	0.0103
m = 1,000		0.0845	0.0023	0.352	0.0923	0.2779	0.0287	0.0826	0.0049
		0.1689	0.0045	0.7041	0.1846	0.5557	0.0574	0.1652	0.0099
m = 10,000		0.0745	0.0011	0.1775	0.0465	0.2943	0.0304	0.0785	0.0047
		0.1491	0.0023	0.355	0.0931	0.5887	0.0608	0.157	0.0094
n = 100	$\mathbf{FDR^{true}}$	**0.0477**	**0**	-	-	**0.0012**	-	**0.0244**	-
m = 100	$\text{FDR}_\text{i}^\text{est}$: Or	0.0481	0	0.0472	0.0067	0.0012	0	0.0255	0.0009
	$\text{FDR}_\text{i}^\text{est}$: Ub	0.0962		0.0945	0.0135	0.0025	0.0001	0.051	0.0018
m = 1,000		0.0478	0	0.0945	0.0135	0.0012	0	0.0243	0.0009
		0.0956		0.1889	0.0269	0.0025	0.0001	0.0486	0.0017
m = 10,000		0.0477	0	0.0744	0.0106	0.0012	0	0.0244	0.0009
		0.0955		0.1488	0.0212	0.0025	0.0001	0.0489	0.0017
n = 50,000	$\mathbf{FDR^{true}}$	**0.0476**	**0**	-	-	**0**	-	**0.0244**	-
m = 100	$\text{FDR}_\text{i}^\text{est}$: Or	0.0463	0	0.0042	0.0004	0	0	0.0245	0
	$\text{FDR}_\text{i}^\text{est}$: Ub	0.0926		0.0084	0.0008			0.049	
m = 1,000		0.048	0	0.0032	0.0003	0	0	0.0245	0
		0.096		0.0065	0.0006			0.0489	
m = 10,000		0.0476	0	0.0033	0.0003	0	0	0.0244	0
		0.0952		0.0065	0.0006			0.0487	
∞	$\mathbf{FDR^{true}}(\infty)$	**0.0476**	**0**	-	-	**0**	-	**0.0244**	-

(b) $\pi_0 = 0.9$

		p-value FDR	SGPV FDR			p-value FNR		SGPV FCR	
		Point Null	Point Null	Unif. Null	T. Norm Null	Point Alt.	Shifted Exp.	Point Alt.	Shifted Exp.
n = 10	$\mathbf{FDR}^{\mathbf{true}}$	**0.5607**	**0.1254**	-	-	**0.0704**	-	*	-
m = 100	$\text{FDR}_\text{i}^\text{est}$: Or	0.5	n/a	n/a	n/a	0.0711	0.0117	*	*
	$\text{FDR}_\text{i}^\text{est}$: Ub	0.5556				0.7114	0.1174		
m = 1,000		0.5172	0.1793	1	1	0.0709	0.0117	*	*
		0.5747	0.1992			0.7091	0.117		
m = 10,000		0.5867	0.1055	1	0.5994	0.0701	0.0116	*	*
		0.6519	0.1172		0.666	0.7012	0.1157		
n = 20	$\mathbf{FDR}^{\mathbf{true}}$	**0.425**	**0.0097**	-	-	**0.0438**	-	**0.0096**	-
m = 100	$\text{FDR}_\text{i}^\text{est}$: Or	0.375	0.0024	0.3802	0.0997	0.0445	0.0046	0.0083	0.0005
	$\text{FDR}_\text{i}^\text{est}$: Ub	0.4167	0.0027	0.4224	0.1107	0.4446	0.0459	0.0826	0.0049
m = 1,000		0.4592	0.0122	1	0.4983	0.0434	0.0045	0.0091	0.0005
		0.5102	0.0136		0.5537	0.4337	0.0448	0.0909	0.0054
m = 10,000		0.4194	0.0098	1	0.3987	0.0438	0.0045	0.0096	0.0006
		0.466	0.0109		0.4429	0.4382	0.0453	0.0961	0.0058
n = 100	$\mathbf{FDR}^{\mathbf{true}}$	**0.3106**	**0**	-	-	**0.0001**	-	**0.0028**	-
m = 100	$\text{FDR}_\text{i}^\text{est}$: Or	0.3214	0	0.17	0.0242	0.0001	0	0.0028	0.0001
	$\text{FDR}_\text{i}^\text{est}$: Ub	0.3571		0.1889	0.0269	0.0014		0.0281	0.001
m = 1,000		0.2866	0	0.4251	0.0606	0.0001	0	0.0028	0.0001
		0.3185		0.4723	0.0673	0.0014		0.028	0.001
m = 10,000		0.3032	0	0.5863	0.0835	0.0001	0	0.0028	0.0001
		0.3369		0.6514	0.0928	0.0014		0.0278	0.001
n = 50,000	$\mathbf{FDR}^{\mathbf{true}}$	**0.3103**	**0**	-	-	**0**	-	**0.0028**	-
m = 100	$\text{FDR}_\text{i}^\text{est}$: Or	0.3462	0	0.0076	0.0007	0	0	0.0028	0
	$\text{FDR}_\text{i}^\text{est}$: Ub	0.3846		0.0084	0.0008			0.0278	
m = 1,000		0.3103	n/a	n/a	n/a	0	0	0.0028	0
		0.3448						0.0277	
m = 10,000		0.3089	0	0.038	0.0035	0	0	0.0028	0
		0.3432		0.0422	0.0039			0.0277	
∞	$\mathbf{FDR}^{\mathbf{true}}(\infty)$	**0.3103**	**0**	-	-	**0**	-	**0.0028**	-

from Table 3.6(a) and Table 3.6(b) relate to the scenarios where the numerator design probabilities are not correctly specified. As noted in Section 3.3.4, the asymptotic behavior of the different second-generation p-value FDR estimators are the same for each of the approaches. However, we can see that for finite sample sizes, using approaches that do not align with the true data generating mechanism can also result in big differences between the estimates and the true rates, even if the true π_0 is known. Overall, in this example where the null is truly a point at 0, using a marginalization approach for the design probability will result in quite an overestimate of the FDR for the second-generation p-value. In an extreme example, for $n = 20$ and $m = 100$ with $\pi_0 = 0.9$, the true SGPV FDR is 0.0024, but is estimated to be 0.38 with the uniform null approach. The truncated normal prior is more closely aligned to the truth than the uniform prior, and thus the resulting FDR estimates are closer, although they too can overestimate the FDR to a large degree. We also observe quite a few settings, for $\pi_0 = 0.9$, where the FDR estimate is equal to 1 with the null marginalization approaches.

On the reverse side, if the alternatives are truly all at a single point, θ_0^+, then using our shifted exponential approach for the alternative design probabilities can result in a severe underestimate of the FCR or FNR. For classical p-values, the false non-discovery rate estimates are the same in the limit, however for the second-generation p-value false confirmation rate, the shifted exponential approach gives the wrong result, even in the limit. However, we note that this type of data generating setting (with all effects only at θ_0^+) is not as likely to be observed in practice. It seems more likely that the effects may be clustered near θ_0^+, with a few larger effects; in this case, the shifted exponential could then produce the correct behavior, in the limit.

If the true π_0 is unknown, as is the case in practice, then the upper bound estimate can be used. The utility of this estimate varies, however. When $\pi_0 = 0.5$, the FDR and FNR/FCR upper bound estimates are both $1/0.5 = 2$x that of the estimates calculated with the true π_0 and π_1. When $\pi_0 = 0.9$, the FDR estimates are only $1/0.9 \approx 1.1$x that of the oracle estimates, however the FNR and FCR upper bound estimates are $1/0.1 = 10$x that of the oracle estimates. Overall, may conclude that if $\pi_0 \approx 0.9$, which is commonly the case in large-scale inference settings, the upper bound estimate for the FDR may be minimally conservative; however, the same cannot be said for the false confirmation rate. For classical p-values, there are methods which estimate π_0 empirically from the m observed p-values or z-values, with varying degrees of success (see for example, Storey and Tibshirani 2003; Efron 2010b; Murray and Blume 2021). One open question remains of how such methods might perform when the null is defined as an interval range of trivial effects. Alternately, perhaps a method to estimate $\pi_0 = \#\{\theta \in \Theta_0\}$ based on the observed second-generation p-values could be developed. The often-unsatisfactory upper bound estimate behavior in our examples highlights the need for some sort of improved approach for handing π_0 in the false discovery rate quantity estimation; thus, is an important area that requires further development.

Further, the selection of the null design probability (for the FDR) or the alternative design probability (for the FCR or FNR) remains crucial. While methods to estimate these exist for standard p-values, much of the existing methodology relies on the established distributional properties of p-values. These have not been established for second-generation p-values. Therefore, we ultimately will desire some empirical approach for estimating $P(p_\delta = 0|H_0)$ and $P(p_\delta = 0|H_1)$ (as well as for $p_\delta = 1$) for second-generation p-values. This will be an important area for future work, in order to facilitate the use of FDR estimation for second-generation p-values in practice.

3.5 Closing remarks

The second-generation p-value is a metric which can incorporate scientific relevance into large-scale inference. It has some improved characteristics over the classical p-value, including improved type I error and false discovery rate properties corresponding to point null effects (Blume et al. 2018, Blume et al. 2019). An additional quantity, the false confirmation rate (FCR), is defined for the SGPV as the rate at which $p_\delta = 1$ – indicating support for the null hypothesis – occurs for alternative effects (Blume et al. 2018). To calculate such quantities as the FDR and FCR for second-generation p-values, the null and alternative design probabilities, $P(p_\delta = 0|H_0)$ and $P(p_\delta = 0|H_1)$ (and equivalently for $p_\delta = 1$), must be known, specified, or estimated. Further, the null proportion π_0 must either be known or estimated, else we can only define upper bounds on the FDR and FCR.

In this chapter, we provide a detailed outline of the framework for false discovery quantities for the second-generation p-value, define general forms for the design probability components, and propose methods for specifying these which incorporate the composite nature of both the null and alternative hypotheses. In this process, some further finite sample and asymptotic properties of SGPV frequency probabilities for certain values of the parameter space, which have not been previously described in the literature, are outlined here. Finally, we describe an approach for empirical estimation of FDR quantities for the second-generation p-value, illustrate their calculation in a simulation example, and discuss their utility in practice.

In large-scale inference based on classical p-values, the design probability $P(p \leq \alpha|H_0)$ is usually well-defined for a point null, and we often do not need $P(p \leq \alpha|H_1)$ at all (when an empirical Bayes estimate of the mixture distribution is used). Additionally, methods for estimating π_0 based on the z-values have been established when π_0 corresponds only to the point null effects. In our application of second-generation p-values to false discovery rates, we encounter a variety of challenges which must be addressed.

Blume et al. (2018) originally calculate the SGPV null design probability assuming the point null, despite the interval null being a key assumption used with second-generation p-value framework. Further, they calculate the full FDR curve, varying the point alternative effect size. However, in practice, a single FDR estimate is desired. In the present chapter, we have proposed a more general definition for the FDR and FCR, which requires a specification of distribution or weighting functions to marginalize across the interval null space and/or alternative spaces, respectively.

We find that for most sensible choices of the null weighting function (i.e., those which do not place all or the majority of the weight on the boundaries of the null region), the null design probability for the SGPV will converge to 0. However, for finite sample sizes, the design probability – and resulting FDR or FCR – may vary quite a bit depending on the weights. In our examples, placing uniform weight across the null zone results in an FDR considerably larger than the point null or the truncated Normal weighting distribution approach, with the latter providing a balance between the point null and uniform approaches. The null design probability must still be specified for the estimation approach utilizing empirical Bayes estimation, which replaces the mixture probability in the denominator of the FDR quantities with a simple estimate based on the observed SGPVs. Thus, misspecification of this null design quantity can greatly impact the resulting FDR estimate for finite n.

One benefit of the empirical Bayes mixture estimate for SGPV, however, is that it does allow for a reduction in assumptions made. For example, for the FDR, the alternative design probability does not need to be specified. Further, when the mixture distribution is directly estimated, it is feasible to obtain an upper bound on the FDR estimate without specifying π_0. The upper bound estimate overestimates the true FDR estimate (i.e., if π_0 was known) by a factor of $1/\pi_0$. Thus, if the true null proportion is large, such as 0.9 or more, the upper bound estimate is an overestimate by a factor of approximately 1.1 or less, which is reasonable. However, if the null proportion is much smaller, such as one half, the upper bound estimate is double the true FDR estimate.

For the false confirmation rate, we are still faced with specifying the alternative design probability $P(p_\delta = 1|H_1)$, however we then do not need to specify the null design probability when the empirical mixture estimator is used for the denominator. Misspecification of this design probability can also have a large impact on the FCR estimates. The shifted exponential weighting function approach we examined resulted in much smaller FCR than the conservative approach of setting the alternative at the boundary of the null zone (representing the smallest scientifically relevant effect size, and results in the minimum value for the alternative design probability). Further, this latter approach results in a FCR which does not converge to 0, rather to a lower bound of $\alpha/(\alpha + 2r^{-1})$ where $r = \pi_1/\pi_0$. This is in contrast to the shifted exponential approach FCR converging to 0 as the sample size goes to infinity. As with the FDR, using the empirical mixture estimator, we can define the upper bound on the FCR estimate, which overestimates the true FCR estimate by a factor of $1/(1 - \pi_0)$. This means that, in common scenarios with large π_0 such as 0.9, the FCR upper bound greatly overestimates the true estimate, by a factor of 10.

Thus, an estimator for π_0 would greatly improve use of FDR estimates in practice, rather than relying on the upper bounds which can be considerable overestimates. Existing methods developed for point null hypotheses may not work well in our interval null hypothesis setting, however this is left an open question for future research. Additionally, methods for estimating the design probabilities for second-generation p-values would also be desired to obtain robust estimates of false discovery rate quantities, rather than needing to specify the choice of weighting function(s). Due to the sensitivity which we have demonstrated of the FDR and FCR estimates to the specification of these quantities, we advise that further examination and development of robust estimation methods be studied before applying these in practical use. However, we expect that the work described in this chapter will aid in directing such future research and ultimately in usage of the second-generation p-value false discovery rate for large-scale inference.

3.6 Appendix A: Remarks and supplemental content

3.6.1 Remarks

3.6.1.1 Remark 3.A

The second-generation p-value Bayes false discovery rate can be written in several forms, including:

$$
\begin{aligned}
SGPV\ FDR &= \frac{\pi_0 \times \Pr(p_\delta = 0|H_0)}{\pi_0 \times \Pr(p_\delta = 0|H_0) + \pi_1 \times \Pr(p_\delta = 0|H_1)} \\
&= \frac{\Pr(p_\delta = 0|H_0)}{\Pr(p_\delta = 0|H_0) + \frac{\pi_1}{\pi_0} \times \Pr(p_\delta = 0|H_1)} \\
&= \left(1 + \frac{\pi_1}{\pi_0} \times \frac{\Pr(p_\delta = 0|H_1)}{\Pr(p_\delta = 0|H_0)}\right)^{-1}.
\end{aligned}
$$

When $\pi_0 = 1$, we directly get $SGPV\ FDR = 1$ for every sample size. When $\pi_0 = 0$, we directly get $SGPV\ FDR = 0$ for every sample size. For $0 < \pi_0 < 1$, the $SGPV\ FDR \to 0$ when $\Pr(p_\delta = 0|H_0) \to 0$ and as long as $\Pr(p_\delta = 0|H_1)$ converges to some non-zero constant. For the SGPV, $P_\theta(p_\delta = 0) > 0\ \forall\theta \in \Theta_1$ as well as at $\theta = \theta_0^-$ and $\theta = \theta_0^+$ for all sample sizes, therefore, $\Pr(p_\delta = 0|H_1) > 0$ regardless of the approach used to calculate this alternative design probability. It was established prior in (Blume et al. 2018) that using the point null approach results in $\Pr(p_\delta = 0|H_0) \to 0$ because $P_0(p_\delta = 0) \to 0$ as $n \to \infty$. If, instead, the frequentist maximum approach is used for the null design probability, then $\Pr(p_\delta = 0|H_0) \to \alpha/2$, because as we discussed in Section 3.2.1, $P_{\theta_0^+}(p_\delta = 0) \to \alpha/2$ and this will be the maximum across Θ_0. Therefore, in this case, the SGPV FDR would not converge to 0. However, we considered utilizing marginalization approaches to account for the interval null hypothesis instead. For $\theta \in (\theta_0^-, \theta_0^+)$, i.e., the entirety of the null zone excluding the boundaries, $P_\theta(p_\delta = 0) \to 0$ as $n \to \infty$ (as discussed in Section 3.2.1). Therefore, our marginalization approaches, which take a weighted average of $P_\theta(p_\delta = 0)$ across Θ_0, will result in $\Pr(p_\delta = 0|H_0) \to 0$ and therefore $SGPV\ FDR \to 0$. We expect that for all choices of weighting functions which are continuous, or which do not have any point mass at the null boundaries, this will be the case. On the other hand, potential choices which do place all (as with the frequentist maximum approach) or some point mass at either or both of θ_0^- or θ_0^+, can result in $\Pr(p_\delta = 0|H_0) \to c$ with $c > 0$ such that the SGPV FDR does not converge to 0.

3.6.1.2 Remark 3.B

As in Remark 3.A for the FDR, the second-generation p-value Bayes false confirmation rate (FCR) can be written in several forms:

$$\begin{aligned} SGPV\ FCR &= \frac{\pi_1 \times \Pr(p_\delta = 1|H_1)}{\pi_1 \times \Pr(p_\delta = 1|H_1) + \pi_0 \times \Pr(p_\delta = 1|H_0)} \\ &= \frac{\Pr(p_\delta = 1|H_1)}{\Pr(p_\delta = 1|H_1) + \frac{\pi_0}{\pi_1} \times \Pr(p_\delta = 1|H_0)} \\ &= \left(1 + \frac{\pi_0}{\pi_1} \times \frac{\Pr(p_\delta = 1|H_0)}{\Pr(p_\delta = 1|H_1)}\right)^{-1}. \end{aligned}$$

When $\pi_0 = 0$, i.e., $\pi_1 = 1$, we directly get $SGPV\ FCR = 1$ for every sample size. When $\pi_0 = 1$, i.e., $\pi_1 = 0$, we directly get $SGPV\ FCR = 0$ for every sample size. For $0 < \pi_0 < 1$, the $SGPV\ FCR \to 0$ if $\Pr(p_\delta = 1|H_1) \to 0$ and $\Pr(p_\delta = 1|H_0)$ converges to some non-zero constant. As described in Section 3.2.1, $P_\theta(p_\delta = 1) \to 1$ for $\theta \in (\theta_0^- + \varepsilon\,, \theta_0^+ - \varepsilon)$, $P_\theta(p_\delta = 1) \to \alpha/2$ for $\theta \in \{\theta_0^-, \theta_0^+\}$ (the null boundary points), and $P_\theta(p_\delta = 1) \to 0$ for $\theta \in (-\infty, \theta_0^- - \varepsilon) \cup (\theta_0^+ + \varepsilon, \infty)$. Therefore, for methods of specifying the alternative design probability which do not place point mass on the null boundary points (such as with the shifted Exponential distribution in Examples 4-6), $\Pr(p_\delta = 1|H_1) \to 0$ such that $SGPV\ FCR \to 0$. On the other extreme, for the approach which places all the point mass right at the null boundary (i.e., Examples 1-3), then $\Pr(p_\delta = 1|H_1) \to \alpha/2$ and with the approaches we used for the null design probability (i.e., which do not place any point mass on the null boundaries), $\Pr(p_\delta = 1|H_0) \to 1$, such that $SGPV\ FCR \to (\alpha/2)/(\alpha/2 + r^{-1})$ with $r = \pi_1/\pi_0$.

CHAPTER 4

False Discovery Rate Quantities for Second-Generation p-Values: Behavior, FDR Control, and Impact of Trivial Effects

Valerie F. Welty, Jeffrey D. Blume

4.1 Introduction

The field of multiple testing has evolved over time, originating with a small multiplicity of tests, where control of the probability that one or more false rejection is made (the family wise error rate) is a sensible approach (Tukey 1953; Ryan 1959). However, an alternative approach emerged, particularly as the number of tests typically considered slowly grew, originating from ideas discussed in (Spjøtvoll 1972; Berger and Sellke 1987; Sorić 1989) – i.e., considering the number or rate of false rejections. This concept was first formalized in (Benjamini and Hochberg 1995) and a simple method for an FDR controlling proposed as a modification of (Simes 1986).

Since this time, the possible number of simultaneous tests has exploded, with even several hundreds of thousands or millions of tests being considered at a time. Fields such as genomics, neuroimaging, and microbiome studies are common biological applications where large-scale inference is relevant. In such instances, we typically have identical statistical tests being performed for a large number of similar units, such as genes or SNPs (genomics), voxels from fMRI images (neuroimaging), or microbial taxa (microbiome). For example, the well-known (Golub et al. 1999) microarray paper studied the association between gene expression and leukemia for 7,129 genes. The NIH Human Microbiome Project has been used to study different scientific questions, such as variation by sex of 748 bacterial taxa in the gut (Human Microbiome Project 2012; Zhou et al. 2018). In Chapter 2, we examined a study of the association between the number of variant alleles for 224,866 SNPs and the occurrence of prostate cancer (original data from (Schaid and Chang 2005)). Therefore, the idea of controlling or minimizing the rate of false discoveries is as important as ever. However, it is also important to consider other factors, such as the power to reject true alternatives, so as to not miss potential discoveries in scientific studies, given that such large-scale inference procedures often form the foundation upon which future work is built.

Much of standard inferential procedures, particularly for large-scale inference, are developed with focus on the classical p-value. The second-generation p-value (SGPV) was introduced by Blume et al. (2018) as an alternative metric, incorporating an interval null hypothesis rather than a point null hypothesis. It is a measure of the overlap between the null hypothesis and the interval range of estimated effects (e.g., from a confidence interval), is scaled to the $[0,1]$ range and includes a small sample correction factor. For further details on the SGPV, see (Blume et al. 2018, 2019), as well as Section 3.2. The key aspect of the SGPV which we will focus on for this chapter is that the second-generation p-value, denoted by p_δ, equals 0 when the interval of estimated effects (e.g., an estimated 95% confidence interval) lies completely outside of the interval null hypothesis zone, i.e., there is no overlap between them. This serves as a natural choice for defining a procedure for rejecting null hypotheses based on the SGPV.

Blume et al. (2018) establishes the general operational probabilities for the SGPV (denoted by p_δ), including this probability $\Pr(p_\delta = 0|\theta)$, showing that it converges to zero when $\theta = 0$, and propose a simple definition of the Bayes/positive false discovery rate (Efron, Storey, et al. 2001; Efron, Tibshirani, et al. 2001; Storey 2001a, 2001b, 2002, 2003) applied to the second-generation p-value, $\Pr(H_0|p_\delta = 0)$, under a strong set of assumptions. Building on this work, in Chapter 3, we defined a general form for the SGPV positive false discovery rate and find that for most sensible approaches of calculating this quantity, it converges to 0 with sample size and is smaller than the analogous unadjusted p-value pFDR (see Section 3.3.4 for more on these theoretical results).

In the present chapter, we provide a more comprehensive examination of using SGPVs in large-scale inference. The advancements include examining a broader set of quantities such as overall FDR control and power, relaxing some assumptions made prior, and including comparison with common multiple comparisons corrective p-value procedures.

First, while the pFDR (i.e., Bayes FDR) is arguably the most important quantity for large-scale inference, it is still useful to examine the complete picture of the operating characteristics of a procedure. Control of the overall false discovery rate, defined as $pFDR \times \Pr(R > 0)$ (Benjamini and Hochberg 1995), via a reduced $\Pr(R > 0)$, may still be useful in some scenarios. We begin with a derivation of the probability of making at least one rejection for the second-generation p-value and examine its behavior as a function of sample size and number of tests. In addition to reducing the rate of false rejections, we are interested in rejecting as many of the non-null tests as possible, i.e., to increase or maximize the power of a testing procedure (Hochberg and Tamhane 1987). Thus, we also include an examination of power to understand this aspect of the SGPV in comparison to other common procedures in practice.

Secondly, the currently established results for false discovery quantities for second-generation p-values rely on a strong set of assumptions, such as fixed and known variance, as well as a common variance among all m tests. We use extensive simulations to examine a broad set of false discovery rate and power quantities for the SGPV and other methods, particularly under a wider range of settings for the variance such that we might study the robustness of these estimates. It is unlikely in practice that all tests will have the same variance. Our simulations include both settings where there is a common variance, and settings where there are more than one variance value among tests (i.e., one small and one large variance). We examined both the setting where the variance distribution is independent of effect size, and one where it is dependent on effect size. Additionally, the assumption of fixed, known variance is seldom true in practice, rather, an estimator is used. In our simulations, we replaced the true variance of the sample mean by the sample variance estimator. Thus, we examine behavior of the second-generation p-value in scenarios which should more closely represent their use in real-world practical settings.

Additionally, the unadjusted p-value is unlikely to be the p-value choice in practice for settings where a measurement such as the false discovery rate would be used. Prior work in (Blume et al. 2018) and in Chapter 3 has focused on comparison between the SGPV and unadjusted p-value (or some loose connections between SGPV and the Bonferroni correction). When we are interested in examining, or possibly controlling, the false discovery rate or pFDR of an inferential procedure in large-scale inference, it is much more likely that a procedure such as Benjamini-Hochberg (BH) designed to control the FDR would be the “gold standard” procedure (Benjamini and Hochberg 1995). As such, it is important to understand how the second-generation p-value compares to this approach. In this chapter, a comparison is made with the Bonferroni (Dunn 1961) and Benjamini-Hochberg procedures, as well as with some hybrid methods which combine multiple comparisons adjusted p-values and the effect size in some manner (such as (Goodman et al. 2019)).

Overall, with these extensions, we aim to develop a further understanding of the use second-generation p-values in practice, particularly as they compare to other prevalent choices. We conclude with a discussion of recommendations for practical implementation, and some future areas of development which may further improve the usage of second-generation p-values in large-scale inference.

4.2 The FDR and other relevant metrics for the SGPV

In a large-scale inference scenario, we commonly use the random variable R to denote the total number of tests which are rejected by a particularly define rejection procedure, such as the standard $p \leq \alpha$ or a multiple comparisons adjusted procedure. The random variable V denotes the number of rejected tests corresponding to null effects. Benjamini and Hochberg (1995) originally defined the false discovery rate as $FDR = E[Q]$, where Q is a random variable equal to either $Q = V/R$ when $R > 0$ or $Q = 0$ when $R = 0$. However, when we think of the phrase “the rate of false discoveries of a multiple comparisons inference”, a different definition might come to mind – namely, the expected value of the first component of Q, the false discovery proportion V/R specifically when $R > 0$. This quantity was formally defined in (Storey 2002, 2003) as the positive false discovery rate, i.e., $pFDR = \mathrm{E}[V/R\,|R > 0]$. The overall FDR and the pFDR can be confused in practice. Of course, there is a relationship between them: $FDR = \mathrm{E}[V/R\,|R > 0] \times \Pr(R > 0) + 0 \times \Pr(R = 0) = pFDR \times \Pr(R > 0)$. Another way to think of the FDR is that it is a weighted average of the rate of false discoveries in observed (1 or more) discoveries, and of the 0 discoveries. Under a certain model of assumptions, it has been established that the pFDR rate quantity also has the probabilistic interpretation of $pFDR = \mathrm{E}[V/R\,|R > 0] = \Pr(H_0|Reject)$ (Storey 2001a, 2001b, 2002, 2003). This quantity answers the question that is more natural to ask of a set of observed rejected hypotheses: how likely is it that these are false rejections, or what proportion of them do we expect to be false rejections?

The pFDR for second-generation p-value rejections (those with $p_\delta = 0$) has been studied in some detail in (Blume et al. 2018, 2019), and in Chapter 3. Asymptotically, the SGPV pFDR converges to zero as a result of convergence of the null design probability $\Pr(p_\delta = 0|H_0)$. However, for finite n, the pFDR can be large, although it is at least smaller than the pFDR for the classical, unadjusted p-value. On the other hand, the overall FDR of the SGPV procedure, or the quantity $\Pr(R > 0)$, has not been examined as of yet.

Throughout the chapter, we will examine these quantities in various settings relating to a simple two or three group underlying model for the effect sizes and variance. Often in large-scale inference, a two-group model is described for the p-value or z-value distributions. In the case of the second-generation p-value, we need to know both the effect size estimate $\hat{\theta}$ and its variance V_n (or an estimate). Further, we will later extend the two-group model into a three-group model to differentiate between zero, trivial, and non-trivial effect sizes (Section 4.4). However, for the time being we will remain in a scenario where there are only two true effect sizes observed in the m hypotheses: $\theta_0 = 0$ or θ_1, where θ_1 is some meaningful effect size outside of Θ_0. The vector of true effect sizes (effect size for each test i) is defined as $\underline{\theta} = (\theta_1, \dots, \theta_m)$ where $\theta_i \in \{\theta_0, \theta_1\}$. There are $m_0 = \#\{\theta_i = \theta_0\} = \sum_{i=1}^{m} I[\theta_i = \theta_0]$ number of null tests (a proportion $\pi_0 = m_0/m$ of the tests) and $m_1 = \#\{\theta_i = \theta_1\} = \sum_{i=1}^{m} I[\theta_i = \theta_1]$ number of alternative tests (a proportion $\pi_1 = m_1/m$ of the tests). These are treated as fixed quantities at the data-generating level, although we note that these are frequently modeled as random variables. In the present chapter, we will focus on $\hat{\theta}$ as an estimate of a sample mean from a Normal distribution, such that $\underline{X_i} = X_{i,1}, \dots, X_{i,n}$ with $X_{i,j} \overset{iid}{\sim} N(\theta_i, \sigma^2)$ where $\hat{\theta}_i = \bar{X}_i \sim N(\theta_i, V_n)$ and $V_n = \sigma^2/n$ (or potentially an estimator $\hat{V}_n$). This assumes a common variance σ^2 between all tests, however we will relax this in later sections.

4.2.1 Direct calculations for fixed variance

In this section, we define the probability $Pr(R > 0)$ for second-generation p-value rejections of $p_\delta = 0$, review the pFDR definition, and ultimately calculate the overall FDR. Recall that R is the number of rejected tests, i.e., $R = \#\{p_\delta = 0\} = \sum_{i=1}^{m} I\left[p_{\delta,i} = 0\right]$ for the SGPV. Thus, $\Pr(R > 0) = 1 - \Pr(R = 0) = 1 - \prod_{i=1}^{m}\left(1 - Pr_{\theta_i}(p_\delta = 0; \sigma^2)\right)$, assuming independence between tests, which simplifies to

$$\Pr(R > 0) = 1 - \prod\nolimits_k \left(1 - Pr_{\theta_k}(p_\delta = 0; \sigma^2)\right)^{m_k} \text{ for } k \in \{0,1\}, \tag{4.1}$$

for fixed m_0 and m_1. The probability $Pr_\theta(p_\delta = 0; \sigma^2, n, \delta)$ has been defined in (Blume et al. 2018, 2019). The FDR is then $FDR = pFDR \times \Pr(R > 0)$, where

$$\begin{aligned} pFDR &= \Pr(H_0 | p_\delta = 0) \\ &= \frac{\pi_0 \times Pr_{\theta_0}(p_\delta = 0; \sigma^2)}{\pi_0 \times Pr_{\theta_0}(p_\delta = 0; \sigma^2) + \pi_1 \times Pr_{\theta_1}(p_\delta = 0; \sigma^2)}, \end{aligned} \tag{4.2}$$

for this two-mean data model. Further details on the pFDR can be found in Chapter 3.

These three major false discovery rate quantities (the pFDR, $Pr(R > 0)$, and FDR) are illustrated in Figure 4.1(a) as a function of sample size n, for a setting where 90% of the tests are null. In the first figure, there are only $m = 10$ tests, in the second, there are $m = 100$ tests, and in the last figure, there are $m = 1{,}000$ tests. These test sizes are smaller than typically seen with most modern multiple comparisons or large-scale inference settings; however, these were chosen to illustrate cases where $Pr(R > 0)$ is non-zero. This positive rejection probability is affected by factors such as π_0, the alternative effect size (here, set to $\theta_1 = 1.5\delta$) and the variance (set at $\sigma^2 = 1.5 \times (2\delta)^2$), and, to a lesser degree, the null width 2δ (in this example, $\delta = 0.1$). Under these settings, and for rejections based on $p_\delta = 0$, we observe $Pr(R > 0) < 1$ only for small values of m. We can see that even by $m = 1{,}000$ tests, $Pr(R > 0) = 1$ for all sample sizes. Note that after a certain sample size, the probability $Pr(R > 0)$ is

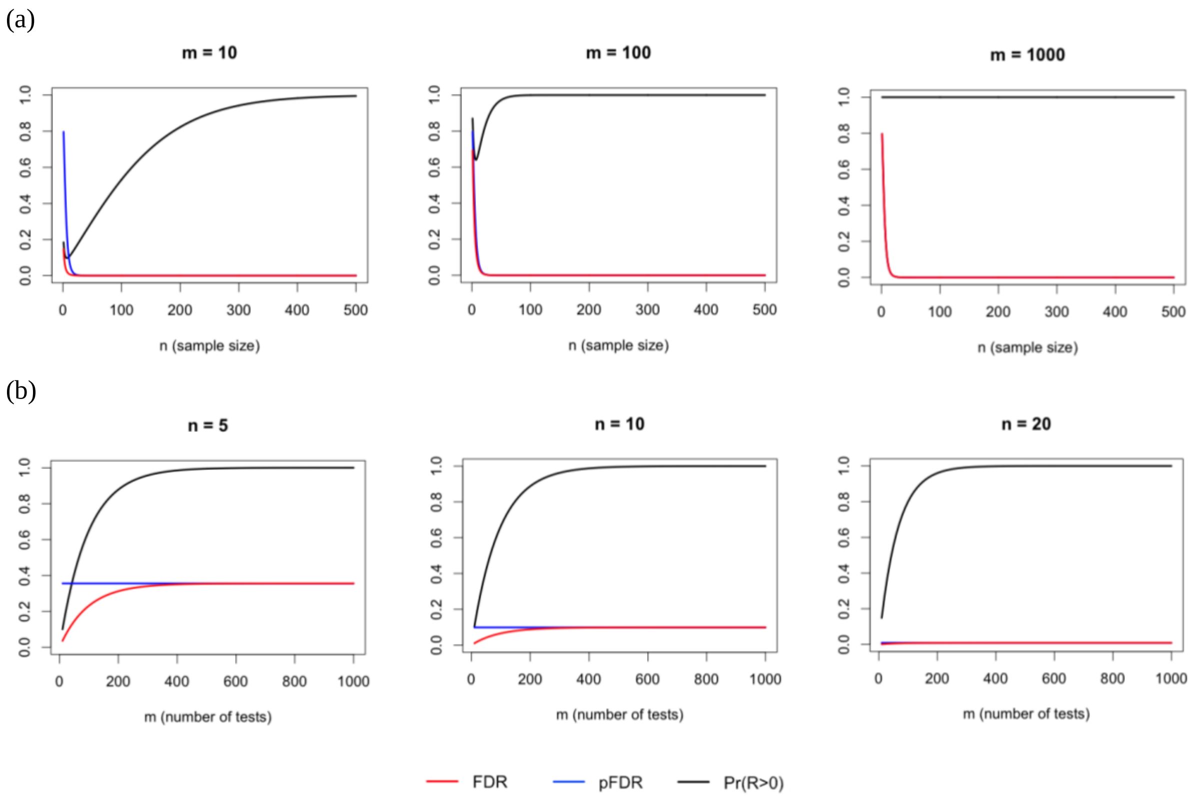


Figure 4.1 Relevant false discovery rate quantities for the second-generation p-value rejection rule. In this setting, $\sigma^2 = 1.5 \times (2\delta)^2$, $\pi_0 = 0.9$, the alternative effects have mean $\theta_1 = 1.5\delta$, and $\delta = 0.1$. (a) As a function of n, for varying values of m. (b) As a function of m, for varying values of n.

non-decreasing, because the confidence interval width shrinks and therefore $p_\delta = 0$ is more likely for non-null tests. However, notice that this quantity is not fully monotonic across all n. For extremely small n, the estimator $\hat{\theta}$ is very volatile, increasing the probability that an unusually large $\hat{\theta}$ will be observed, such that even the wide CI does not overlap with Θ_0 at all (i.e., $p_\delta = 0$). Around the inflection point, the $\hat{\theta}$ starts to become less variable, but the CI width does not decrease much. as such, a non-overlapping CI with Θ_0 is much less probable. In general, however, $Pr(R > 0) \to 1$ as $n \to \infty$ and as $m \to \infty$ for the second-generation p-value, aside from special cases such as when alternative effects are arbitrarily close to the boundary of the null region.

We can also observe in Figure 4.1(a) previously established behavior of the pFDR for the second-generation p-value (the blue line), namely that $pFDR \to 0$ as $n \to \infty$. Note that under the current set of assumptions, the pFDR for the SGPV is not affected by the number of tests, m. However, the FDR which scales the pFDR by $Pr(R > 0)$, can become much smaller for small n and small m due to $Pr(R > 0)$ being much less than 1. That is, the FDR is much lower than the pFDR for these small n and m, although it is not necessarily strictly controlled at the 0.05 level. The interrelationships between these quantities are illustrated further in Section 4.2.2.

Figure 4.1(b) shows a different perspective, varying the number of tests m on the x-axis for a selection of sample sizes. The constant nature of the pFDR as a function of m is more easily seen as a horizontal line in each figure. Because the sample size is fixed for each figure, so too is $Pr_{\theta_i}(p_\delta = 0; \sigma^2)$, given that σ^2 is presently assumed to be a known constant among all tests. As such, the probability $Pr(R > 0)$ is monotonically increasing as a function of m. As noted above, $Pr(R > 0) \to 1$ and consequently $FDR \to pFDR$ as $m \to \infty$. Overall, we see that the SGPV FDR is not generally controlled, and when it is controlled, we typically see that $FDR \approx pFDR \leq 0.05$ for the more common sample sizes or number of tests. The pFDR converges relatively quickly, although this will depend on factors such as the variance, whether the known variance or an estimate is used for SGPV calculations, and the alternative effect size.

4.2.2 Discussion: FDR control and pFDR minimization

The distinction between the FDR and pFDR is important. In this section, we provide further illustrations of their interrelationships, as well as discussion of the roles of each in large scale inference settings. Recall that the FDR is the expected value of the random variable Q, which is defined as either the observed proportion of false rejections (V/R) when rejections are observed ($R > 0$), or as zero when no rejections are observed ($R = 0$). This is in some ways a sensible definition. If no rejections were made at all, then we know with certainty that no false rejections were made, and as such it might be a natural choice to set the quantity at zero. On the other hand, this also means that a procedure that is FDR controlling (i.e., $FDR \leq 0.05$), will tell us nothing about how many incorrect rejections to expect in a set of observed rejected tests. For illustration, consider the four scenarios in Table 4.1 below:

Table 4.1 Example scenarios for controlled FDR.

	FDR ($E[Q]$)	pFDR ($E[V/R\,|R > 0]$)	$\Pr(R > 0)$
Scenario 1	0.05	0.05	1
Scenario 2	0.05	0.1	0.5
Scenario 3	0.05	0.5	0.1
Scenario 4	0.05	0.99	0.051

In each of these scenarios, the FDR is equal to 0.05. However, the manner in which it is controlled is very different. In Scenario 1, the FDR is controlled naturally, via control of the pFDR. On the other hand, in Scenario 2, 10% of observed rejections are incorrect on average, yet rejections are only made half of the time, thus the expected value of Q is brought back to 0.05. Scenario 3 is a more extreme example of this behavior, with 50% of observed rejections expected to be false, but observing any rejections only 10% of the time. Control of the FDR regardless of its component values may be intuitive when viewed from a frequentist lens. However, viewing it from a more practical lens, applying this procedure to real-life data will give us only one single data point of V and R. In the case of Scenario 3, we would either 1) reject nothing (90% of the time), or 2) we obtain a set of rejected tests which are expected to be 50% wrong.

It is important to re-emphasize that if we observe a set of rejected tests based on a multiple testing procedure which has been shown to control the FDR, we have no idea whether we are in Scenario 1, 2, 3, or 4 (or anything in between). For this reason, we will generally prefer a procedure for which $pFDR \leq 0.05$. Strict pFDR control is not possible in general (Benjamini and Hochberg 1995; Storey 2002, 2003), however we might aim to at least minimize the pFDR, if not control it. Note that it has been previously argued in the literature that in practice, large-scale inference experiments typically have $\Pr(R > 0) \approx 1$, such that this distinction is not crucial (Storey 2003; Efron 2010b). However, this may not always be the case, particularly if the variability of the measurements is very large, or the sample size is very small. We see some examples of this in Sections 4.3 and 4.4, where $\Pr(R > 0)$ is examined for some common multiple testing procedures.

Procedures which control the FDR "artificially" via reduction in $\Pr(R > 0)$ may be less desirable in common contexts where multiple testing procedures are used. For example, large-scale inference settings are frequently viewed as exploratory (Storey 2002; Goeman and Solari 2011). In such cases, extremely low power, and in particular the likelihood of observing $R = 0$, is of particular concern. This is due to rejections being more so seen as "signals of interest", to inspire further inferential or laboratory studies, rather than final establishments of associations. If no rejections are observed, the next scientific step is left unclear. In such contexts, we might prefer a procedure which minimizes the pFDR, with higher power, rather than a procedure which strictly controls the FDR but is also likely to provide no rejected tests. Essentially, we might prefer to cast a wider net at a "first-stage" inference procedure, then refine results with further study in the future.

FDR control may still be the top priority if the inference is in a context where consequences of false rejections are higher. Certainly, a study which results in no rejections could be viewed as a waste of money. However, if outcomes of a large-scale inference study will have larger, direct impacts, such as directly affecting patient safety or immediately changing clinical practice, then in such a case, we might prefer to still employ a procedure that artificially controls the FDR, with a reduced probability of calling any signals as significant. However, it is still of great importance to get an estimate of the pFDR, to know, if significant associations are observed, how many of these resulting signals we can expect to be erroneous, and guide decisions about future implementation.

4.3 Simulations for sample variance

In Section 4.2.1, we examined the probability Pr(R>0) and the overall FDR control (or lack thereof) related to the second-generation p-value for a single, known value of variance. Next, we will examine simulation estimates for false discovery quantities in a wider array of settings, and with the sample variance estimator in place of the true variance. Note that by comparing these relaxed scenarios to the prior ones, we may study the robustness of the FDR quantities to violations in these original assumptions.

Further, we will compare the performance of the SGPV against several well-known multiple testing control methods, namely Bonferroni (albeit for the purpose of controlling the FWER, not FDR) (Dunn 1961) and the Benjamini-Hochberg method (Benjamini and Hochberg 1995). The Benjamini-Hochberg procedure is an adaptive procedure, meaning that the rejection rule is a function of observed data, unlike Bonferroni where the procedure rejects any p-value below a pre-determined cutoff of α/m.

Direct calculation of the pFDR, Pr(R>0) and FDR for the SGPV is possible when the variance is fixed. It is more complex to derive these quantities in scenarios when the variance is non-common, and an estimator is used in place of the known variance. Here, we use simulation methods to obtain estimates of the quantities of interest. The simulated data are generated under the same type of data-generating framework as described in Section 4.2, but with a non-common variance such that $\underline{X_i} = X_{i,1}, \ldots, X_{i,n}$ with $X_{i,j} \overset{iid}{\sim} N(\theta_i, \sigma_i^2)$ where $\hat{\theta}_i = \bar{X}_i = \sum_{j=1}^{n} X_{i,j}/n \sim N(\theta_i, V_{n,i})$ and $V_{n,i} = \sigma_i^2/n$. To calculate the p-values and confidence intervals for the second-generation p-value, we will use the estimated sample variance for the sample mean estimate $\hat{\theta}_i$ (and as a result, use a t-statistic in place of a z-statistic, and a t critical value in place of a z critical value).

For the simulations, we examined settings with:

i) common variance of σ^2 among all tests,

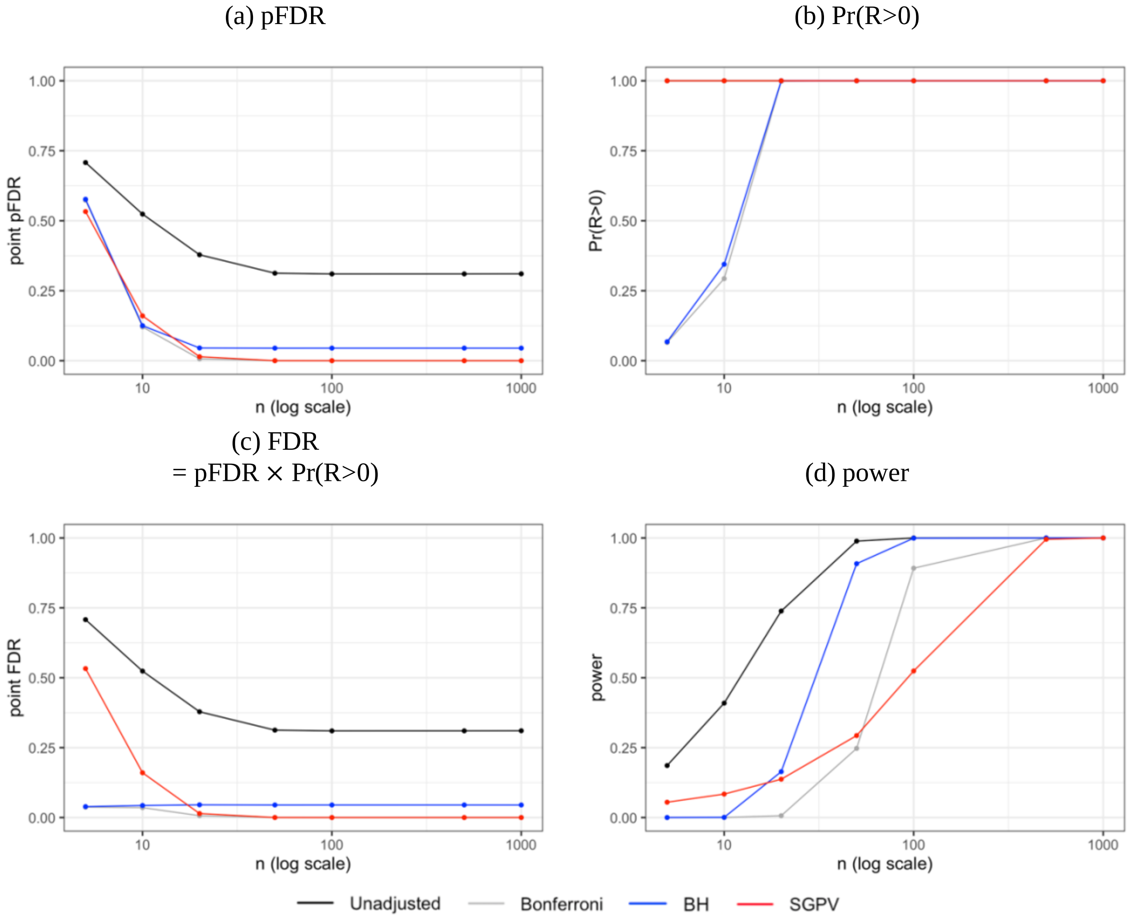
ii) two variance values, σ_s^2 and σ_l^2, with equal probability of being observed for both null and alternative tests (i.e., $\Pr(\sigma^2 = \sigma_s^2 | \theta = \theta_k) = \Pr(\sigma^2 = \sigma_l^2 | \theta = \theta_k) = 0.5$ for $k \in \{0,1\}$), and

iii) as in (ii), but with all null effects having the small variance σ_s^2, and all alternative effects having the large variance σ_l^2 (i.e., $\Pr(\sigma^2 = \sigma_s^2 | \theta = \theta_0) = 1$ and $\Pr(\sigma^2 = \sigma_l^2 | \theta = \theta_1) = 1$).

We also examined various values of number of tests m, proportion of null tests π_0, and size of alternative effect θ_1. The pFDR, probability of observing positive effects Pr(R>0), the resulting FDR, and the estimate of observed power (i.e., the rejection rate of alternative tests with mean θ_1, $E[\#\{\text{rejected with } \theta = \theta_1\}/m_1]$) are given in Figures 4.2 – 4.4 for a selection of settings. These settings were selected to illustrate several different behaviors of the SGPV in comparison to the Benjamini-Hochberg procedure.

Figure 4.2(a)-(d) is a setting where the pFDR of the SGPV and BH procedures is similar for finite sample sizes. Figure 4.3(a)-(d) is a setting where the SGPV pFDR is quite a bit larger than the pFDR for the BH procedure for small sample sizes. Lastly, Figure 4.4(a)-(d) is a setting where the pFDR of the SGPV is quite smaller than the BH procedure pFDR, particularly in small sample sizes. In each of these figures, there are $m = 10{,}000$ tests, where $\pi_0 = 0.9$ of these are null, and the alternative effect magnitude is 1.5 times the minimum effect size of scientific relevance (i.e., $\theta_1 = 1.5\delta$). The only difference among each setting is the variance magnitudes and/or distributions among tests. Some other behaviors for each of the quantities are more consistent across various settings. The key false discovery rate and power quantities for a wider variety of settings (including other variance settings, number of tests, and alternative effect sizes θ_1) are also given in Supplemental Figure 4.2(a)-(d).

Examining Figures 4.2 – 4.4 and Supplemental Figure 4.2(a)-(d), we see the following overall trends. The Benjamini-Hochberg procedure does successfully control the FDR around 0.05 across all considered settings. However, for small sample sizes, the pFDR can be much larger than 5% (with FDR control due to small $\Pr(R > 0)$, even as small as 0.1). In many cases, the BH $pFDR$ has reduced to 5% by a sample size of $n = 20$, and in nearly all of the examined settings by $n = 100$. For small n, the pFDR for the SGPV can be quite large, ranging between 20% and 80% in our examined settings for $n \in (5,10,20)$. For $m = 10{,}000$ (the case in Figures 4.2 – 4.4) and



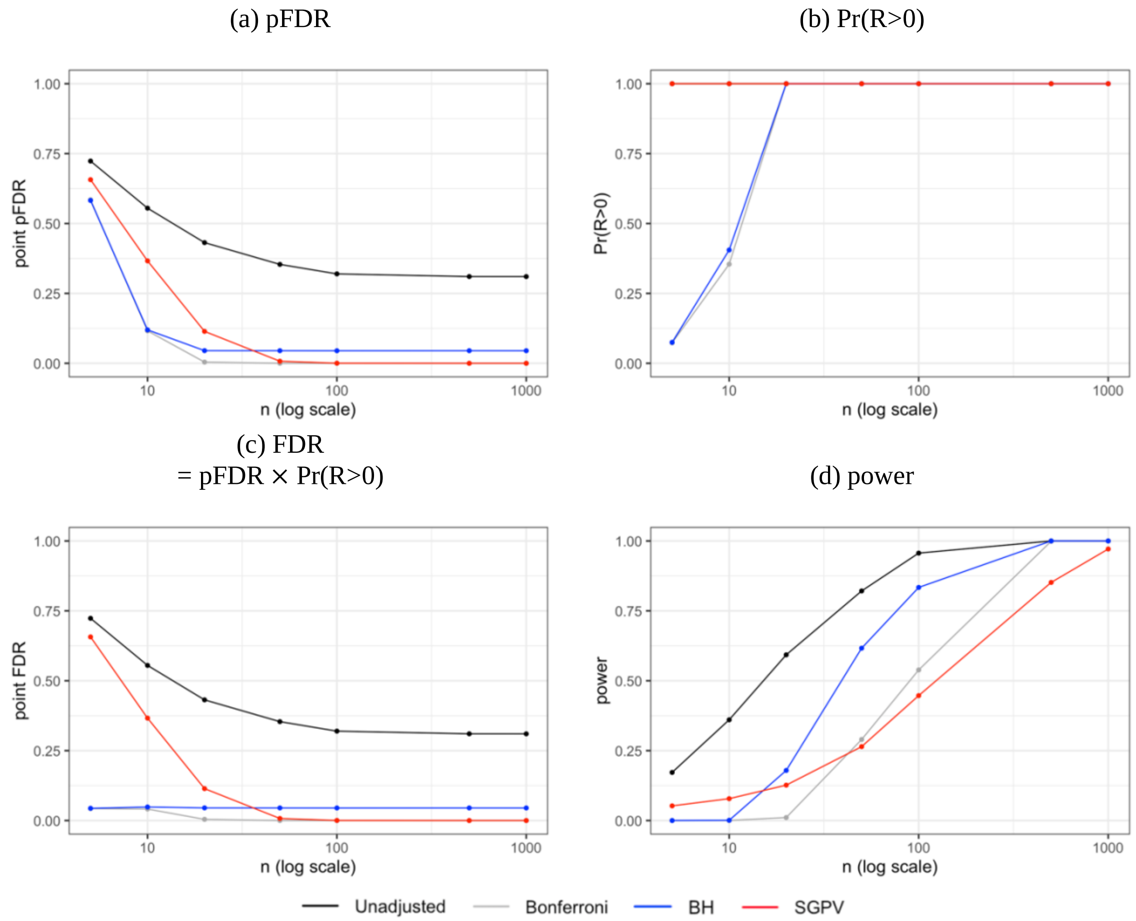
(a) pFDR

(b) Pr(R>0)

(c) FDR = pFDR × Pr(R>0)

(d) power

Figure 4.2 Simulation estimates of false discovery rate quantities for a setting with only null and non-trivial alternative effects, and with tests having a common variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.9, \pi_1 = 0.1, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and there is a common variance of $\sigma^2 = 1.5 \times (2\delta)^2$ for all tests.



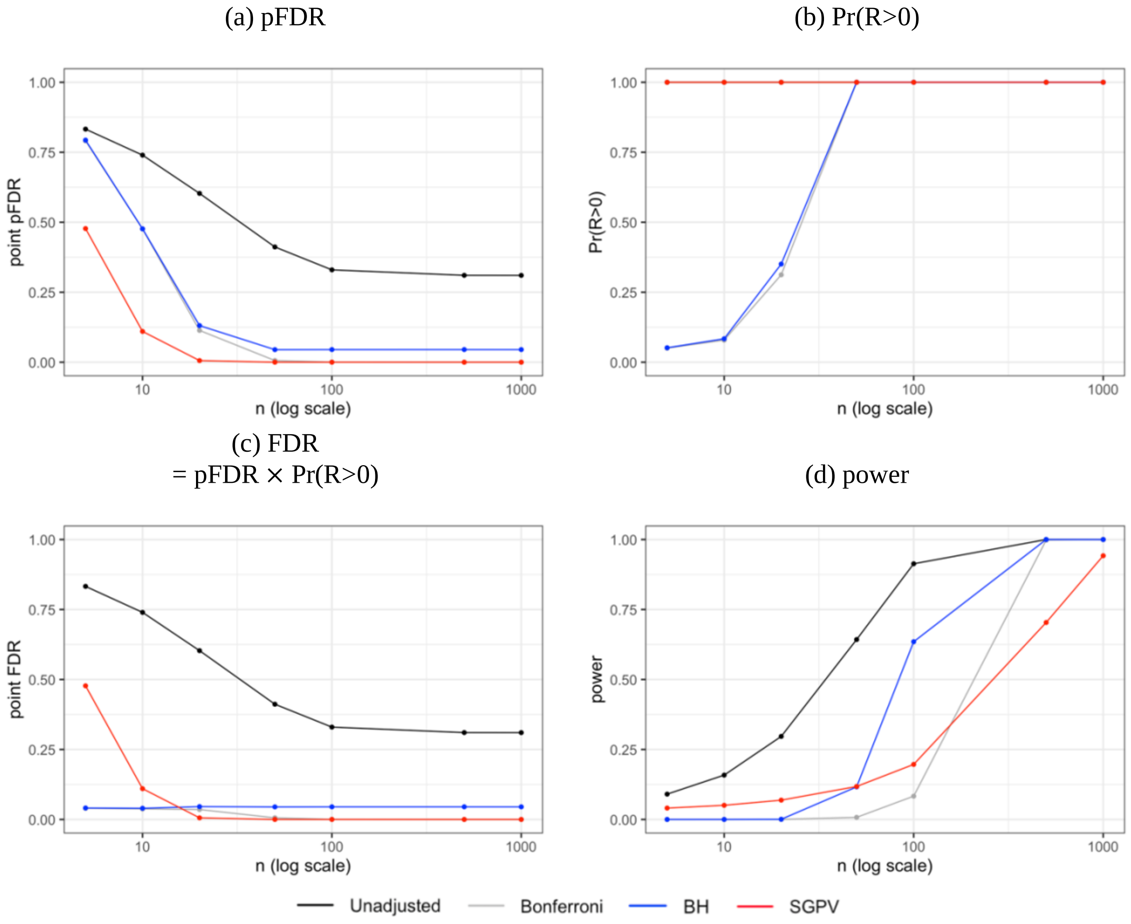
(a) pFDR

(b) Pr(R>0)

(c) FDR
= pFDR × Pr(R>0)

(d) power

Figure 4.3 Simulation estimates of false discovery rate quantities for a setting with only null and non-trivial alternative effects, and with two possible variance values distributed randomly among tests. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.9, \pi_1 = 0.1, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and $\sigma^2 \in \{1 \times (2\delta)^2, 5 \times (2\delta)^2\}$, with random distribution among tests (null and alternative tests have 0.5 probability of having each variance value).



(a) pFDR

(b) Pr(R>0)

(c) FDR
= pFDR × Pr(R>0)

(d) power

Figure 4.4 Simulation estimates of false discovery rate quantities for a setting with only null and non-trivial alternative effects, and with null tests having smaller variance and alternative tests having larger variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.9, \pi_1 = 0.1, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and $\sigma^2 \in \{1 \times (2\delta)^2, 5 \times (2\delta)^2\}$, with all null tests having variance of $1 \times (2\delta)^2$ and all non-trivial alternative tests having variance of $5 \times (2\delta)^2$.

$m = 1{,}000$ (seen in Supplemental Figure 4.2(a)-(d)), the SGPV $\Pr(R > 0) \approx 1$ for all sample sizes, effect sizes, and variance magnitudes. For a very small number of tests ($m = 100$), the quantity $\Pr(R > 0)$ for SGPV rejections is commonly less than 1 for these small sample sizes, however, it is not small enough to control the FDR, only to reduce it somewhat (i.e., $\Pr(R > 0) > 0.05/pFDR$).

When extremely small sample sizes are observed (e.g., $n = 5$), the pFDR for the SGPV procedure is mostly comparable to that of BH procedure. For other small sample sizes (e.g., $n = 10$ or $n = 20$), the difference between the SGPV pFDR and BH pFDR varies, dependent on the distribution and magnitude of the variances, and the magnitude of the alternative effect size. In cases where all or some of the tests have very large variance (such as $\sigma^2 = 10 \times (2\delta)^2$, with either common variance or randomly distributed variance), the BH pFDR is much smaller than that of the SGPV pFDR for small n. In other settings, where the variance of the alternative effects is larger and the variance for null effects is small, the SGPV pFDR is significantly lower than that of BH, including for $n = 5$.

Of course, for medium-to-large sample sizes, the SGPV pFDR converges to 0, and the BH pFDR converges to $\pi_0 \times 0.05$.

Interestingly, the SGPV procedure has slightly larger power than BH for small n, although both have very poor power (≤ 0.2 in many cases). This is primarily due to the increased rate of observing $R > 0$ for SGPV compared to BH, and in many cases a tradeoff with FDR control. The SGPV procedure loses power in the medium range of sample sizes, but eventually the power converges to 1 (somewhere in the range of $n = 100$ and $n = 1{,}000$ for most settings, depending on variance and effect size), along with the power of the BH procedure. In settings where the effect size is very large, e.g., $\theta_1 = 3\delta$, the SGPV nearly always has comparable or better power than the BH procedure, converging to 1 at a faster rate.

4.4 Rejections of trivial effects for an interval null hypothesis

A key benefit of the second-generation p-value is that it accounts for trivial effects, defined by the pre-specified interval null hypothesis or “indifference zone”, which consists of the range of effects deemed to be scientifically uninteresting or not impactful. Therefore, with the SGPV, we hope to focus on effects that **would** be scientifically meaningful. This is in contrast to the classical p-value, which only examines effects in relation to their corresponding standard error. With small enough variance, or a large enough sample size, trivial effects can frequently be rejected by standard or even adjusted p-value procedures such as Bonferroni. If these trivial effects truly represent effects which are uninteresting, then a procedure which controls the rate of discovering effects with mean exactly equal to zero (the point null) but has a large rate of trivial effect discoveries, will not be useful.

Here, we extend the prior data model to include effects which are non-zero but fall within the pre-defined interval null hypothesis Θ_0. Mathematically, one approach to incorporating such effects is to keep with the two-group model as described in Section 4.2 and expand the null component of the overall mixture into another mixture of zero and non-zero trivial effects. The alternative component of the mixture will include only the non-trivial effects outside of Θ_0. A second approach is to define a three-group model, which consists of 1) zero null effects, 2) non-zero trivial null effects, and 3) non-trivial alternate effects. This approach is useful to distinguish between the zero and non-zero trivial effects, particularly to illustrate drivers of overall behavior. Note that these two approaches are not necessarily distinct from a data-generating perspective, rather the important differences are from a notational and mathematical framework perspective. We will implement the latter approach with a three-mean model where some of the tests are null with true mean $\theta_0 = 0$, some of the tests are non-zero but trivial with true mean $\theta_{tv} \in \Theta_{tv} = [\theta_0^-, 0) \cup (0, \theta_0^+]$, and the rest are alternative with a non-trivial true mean of $\theta_1 \in \Theta_1 = (-\infty, \theta_0^-) \cup (\theta_0^+, \infty)$. The vector of true means is $\underline{\theta} = (\theta_1, \dots, \theta_m)$ with $\theta_i \in \{\theta_0, \theta_{tv}, \theta_1\}$. The quantities m_0 and m_1 are defined as before, and $m_{tv} = \#\{\theta_i = \theta_{tv}\} = \sum_{i=1}^{m} I[\theta_i = \theta_{tv}]$ is the number of trivial tests (a proportion $\pi_{tv} = m_{tv}/m$ of the tests).

With an expanded model, we need to consider additional sets of the false discovery rate quantities. For the pre-specified interval null hypothesis region, the positive false discovery rate is $pFDR_{\{comb\}} = \Pr(\theta \in \Theta_0 | p_\delta = 0)$, and the corresponding FDR is $FDR_{\{comb\}} = pFDR_{\{comb\}} \times \Pr(R > 0)$. This combined, or interval null, pFDR can be decomposed into the two components related to zero and non-zero trivial effects. That is,

$$pFDR_{\{comb\}} = pFDR_{\{pt\}} + pFDR_{\{tv\}}, \tag{4.3}$$

where $pFDR_{\{pt\}} = \Pr(\theta = 0 | p_\delta = 0)$ is the classical, "point null" pFDR and $pFDR_{\{tv\}} = \Pr(\theta \in \Theta_{tv} | p_\delta = 0)$ is the pFDR corresponding to rejections of effects with trivial but non-zero mean (in our examples, $\theta = \theta_{tv}$). The corresponding FDRs are $FDR_{\{pt\}} = pFDR_{\{pt\}} \times \Pr(R > 0)$, and $FDR_{\{tv\}} = pFDR_{\{tv\}} \times \Pr(R > 0)$, respectively.

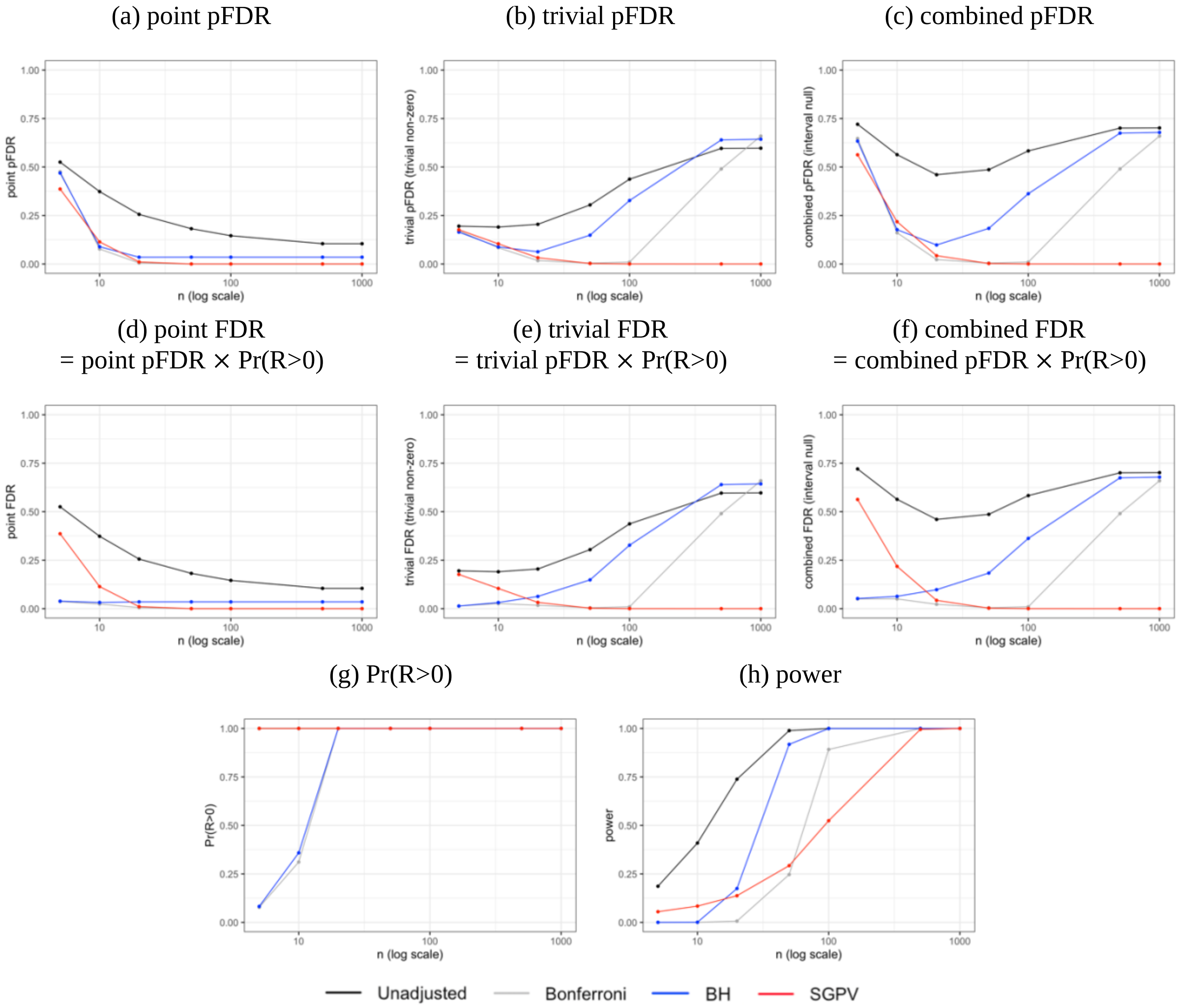
4.4.1 Results

The full set of false discovery rate quantities and power to reject non-trivial alternate effects are provided for one example setting in Figure 4.5. As in the prior examples, there are $m = 10{,}000$ tests, an alternative effect size of 1.5 times that of the minimum scientifically relevant value (i.e., $\theta_1 = 1.5\delta$), and 10% of the tests having this alternative, non-trivial mean. However, rather than all other tests having mean exactly equal to zero, 20% of the tests have a trivial non-zero effect size 0.5 times that of the minimum scientifically relevant value. That is, $\pi_0 = 0.7$ of the tests have $\theta_0 = 0$ and $\pi_{tv} = 0.2$ of the tests have $\theta_{tv} = 0.5\delta$, for a combined interval null proportion of 90%. All tests have a common variance of $V = 1.5 \times (2\delta)^2$ in this chosen setting. Here, we can see that the trivial and combined pFDR and FDR quantities for all classical p-value methods do not decrease to a lower bound, as with the SGPV, where both $pFDR_{\{tv\}} \to 0$ and $pFDR_{\{comb\}} \to 0$ as $n \to \infty$. These simulation results support the theoretical results of Chapter 3 (i.e., from Section 3.3.4), that the interval null pFDR/Bayes FDR will converge to 0 for sensible null effect distributions. Procedures such as BH may have $FDR_{\{tv\}} \leq 0.05$ and/or $FDR_{\{comb\}} \leq 0.05$ in small samples due simply to the low probability of observing rejections. However, $Pr(R > 0) \to 1$ as $n \to \infty$, and thus ultimately the BH has undesirable behavior in both the pFDR and FDR. These results hold true in general across all examined settings, as seen in Supplemental Figure 4.6.

4.4.2 Illustration: which type of tests rejected by SGPV vs. p-value approaches

To illustrate some examples of SGPV behavior as compared to classical p-value approaches, the estimated 95% confidence intervals from a single simulated data set, having $m = 100$ tests with $n = 20$ observations for each test, are plotted in Figure 4.6(a). These intervals are in no particular order, other than being grouped by their true mean – intervals 1:70 have true mean $\theta_0 = 0$, intervals 71:90 have true mean $\theta_{tv} = 0.5\delta$, and intervals 91:100 have true mean $\theta_1 = 1.5\delta$. Each interval has either a smaller true underlying variance (equal to the squared null width) or a larger true variance value (5 times the squared null width), with equal probability. The grey shaded horizontal rectangle represents the interval null/indifference zone.

Intervals which are colored in red are those where $p_\delta = 0$, i.e., intervals which are rejected based on SGPV criteria. Intervals in blue are those with $p_\delta \neq 0$ (where light blue corresponds to $0 < p_\delta < 1$, and dark blue corresponds to $p_\delta = 1$). Above each interval, where observed, are symbols which signify a rejection by a classical p-value approach: the circular dots correspond to unadjusted p-value rejections, the triangles correspond to Bonferroni procedure rejections, and the diamonds correspond to BH procedure rejections. It is important to note



(a) point pFDR

(b) trivial pFDR

(c) combined pFDR

(d) point FDR
= point pFDR × Pr(R>0)

(e) trivial FDR
= trivial pFDR × Pr(R>0)

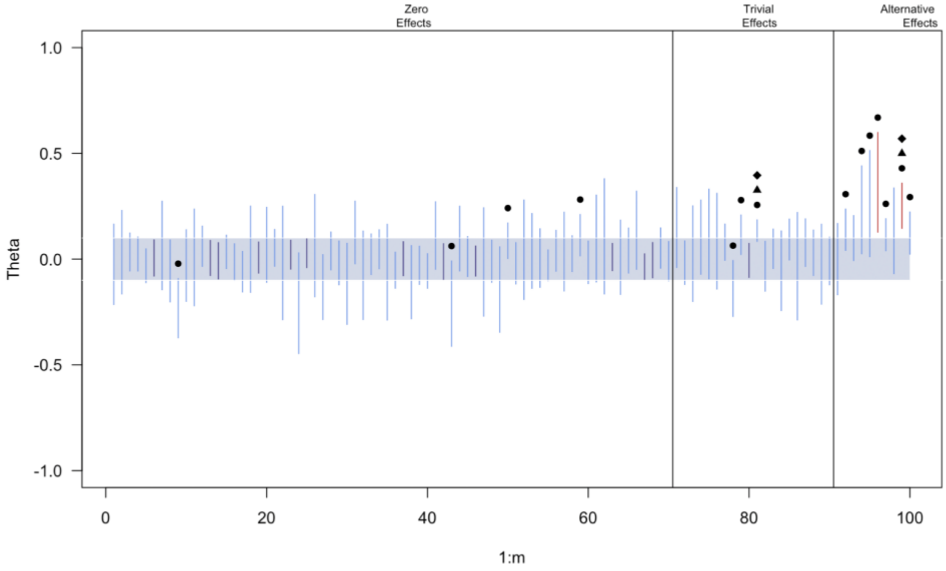
(f) combined FDR
= combined pFDR × Pr(R>0)

(g) Pr(R>0)

(h) power

Figure 4.5 Simulation estimates of false discovery rate quantities for a setting with null, trivial, and non-trivial alternative effects, and with tests having a common variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1, \theta_{tv} = 0.5\delta, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and there is a common variance of $\sigma^2 = 1.5 \times (2\delta)^2$ for all tests.

(a) n = 20



(b) n = 100

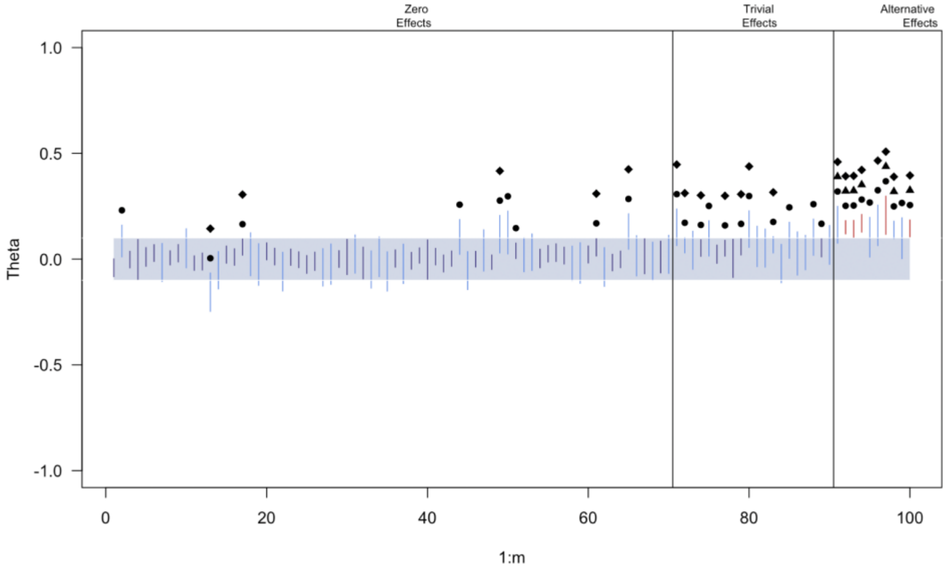


Figure 4.6 Illustration of rejected intervals for each testing method in two example settings. The colored vertical interval line segments denote each of the m = 100 confidence intervals (CIs). The CIs are grouped by their true effect size θ (the intervals are in no particular order otherwise; recall that there are only 3 unique effect sizes). The vertical black lines denote the boundary between each effect size θ. That is, the first $m_0 = 70$ intervals have true $\theta = 0$ (zero null effects), the next $m_{tv} = 20$ intervals have true $\theta = 0.5\delta$ (trivial non-zero effects), and the final $m_1 = 10$ intervals have true $\theta = 1.5\delta$ (alternative, non-trivial effects). Each test has either a small or large true underlying variance, with random distribution among all effect sizes. The shaded blue/grey rectangle represents the interval null hypothesis $\Theta \in [-0.1, 0.1]$. The color of the CIs represents the SGPV result. CIs colored in light blue represent $p_\delta \in (0,1)$, colored in dark blue/purple represent $p_\delta = 1$, and colored in red represent $p_\delta = 0$ (i.e., SGPV rejections). Intervals rejected by the unadjusted p-value threshold have a black circle above them, intervals rejected by the Bonferroni procedure have a black triangle above them, and intervals rejected by the Benjamini-Hochberg procedure have a black diamond above them. (a) $n = 20$, (b) $n = 100$.

that this setting and the simulated data were not selected randomly. Rather, they were intentionally chosen to illustrate behavior that is sometimes observed where the second-generation p-value excels above the other methods.

We see that, as expected, unadjusted p-values reject many more tests than other methods, which corresponds in this case to improved power to reject true alternatives (power = 0.7), but also a decent number of rejections of null and trivial intervals (point FDP = 0.057, combined FDP = 0.078). Here, the rejections by Bonferroni and BH are identical, although this is not generally the case. However, some key differences are seen between the BH and SGPV methods. While both methods have an observed (point) false discovery proportion of $FDP = 0$, the BH (and Bonferroni) method reject a single trivial interval (combined FDP = 0.011). We can see that this interval had an estimated effect size $\hat{\theta}$ outside of the indifference zone and a small variance, leading to an adjusted p-value small enough to meet both rejection thresholds. Further, in this instance, the SGPV has improved power: it rejects an additional second, wider interval, that was not picked up by the other adjusted methods due to having a larger observed variance (SGPV power = 0.2 vs BH power = 0.1).

Figure 4.6(b) shows a single set of simulated intervals with all the same settings as in Figure 4.6(a), except with an increased sample size of $n = 100$. In this case, we can see that the BH procedure has a higher rate of rejections overall – true effect rejections are increased (better power), however the zero and trivial effects are also rejected at an increased rate (0.071 and 0.35, respectively). The SGPV correctly excludes all zero and trivial effects from rejections, although it does have lower power: only 5 of the 10 true alternative tests are rejected, as opposed to 8 of the 10 with BH. From results in prior sections, we know that eventually the power of the SGPV procedure does also converge to 1 (usually anywhere from n = 100 to n = 500 in our examined settings of variance and effect size). However, Figure 4.6(b) also serves to illustrate how the increase occurs at a slower rate than the other considered procedures.

4.4.3 Hybrid procedures

In this section, we briefly examine two procedures that utilize both traditional p-values and the effect size directly in some manner. First, we examine a procedure which only counts as rejected tests those that are rejected both by the Benjamini-Hochberg procedure and by the SGPV ($p_\delta = 0$). That is, we take the intersection of BH and SGPV rejections (thus, will refer to this test by the shorthand "BH-SGPV"). The hope with defining this procedure is that we might take advantage of the best behavior from each method (e.g., FDR control of BH, improved pFDR of the SGPV for certain settings, and of BH for others, and the low rates of inclusion of trivial effect sizes in rejections).

Additionally, we examine a variation of a hybrid procedure described in (Goodman et al. 2019), the "minimum effect size plus p-value" (MESP) approach. The original procedure rejects tests only when the unadjusted p-value rejects ($p \leq \alpha$) and the point effect estimate $\hat{\theta}$ falls outside of Θ_0. We will instead focus on a modification which uses the BH adjusted p-value instead of the standard unadjusted p-value threshold. We denote this procedure by "BH-MESP". Note that this procedure is very similar to the BH-SGPV procedure, but with a less stringent requirement related to effect size: only the point estimate needs to be outside of Θ_0 (for BH-MESP), rather than the entire confidence interval (as with BH-SGPV).

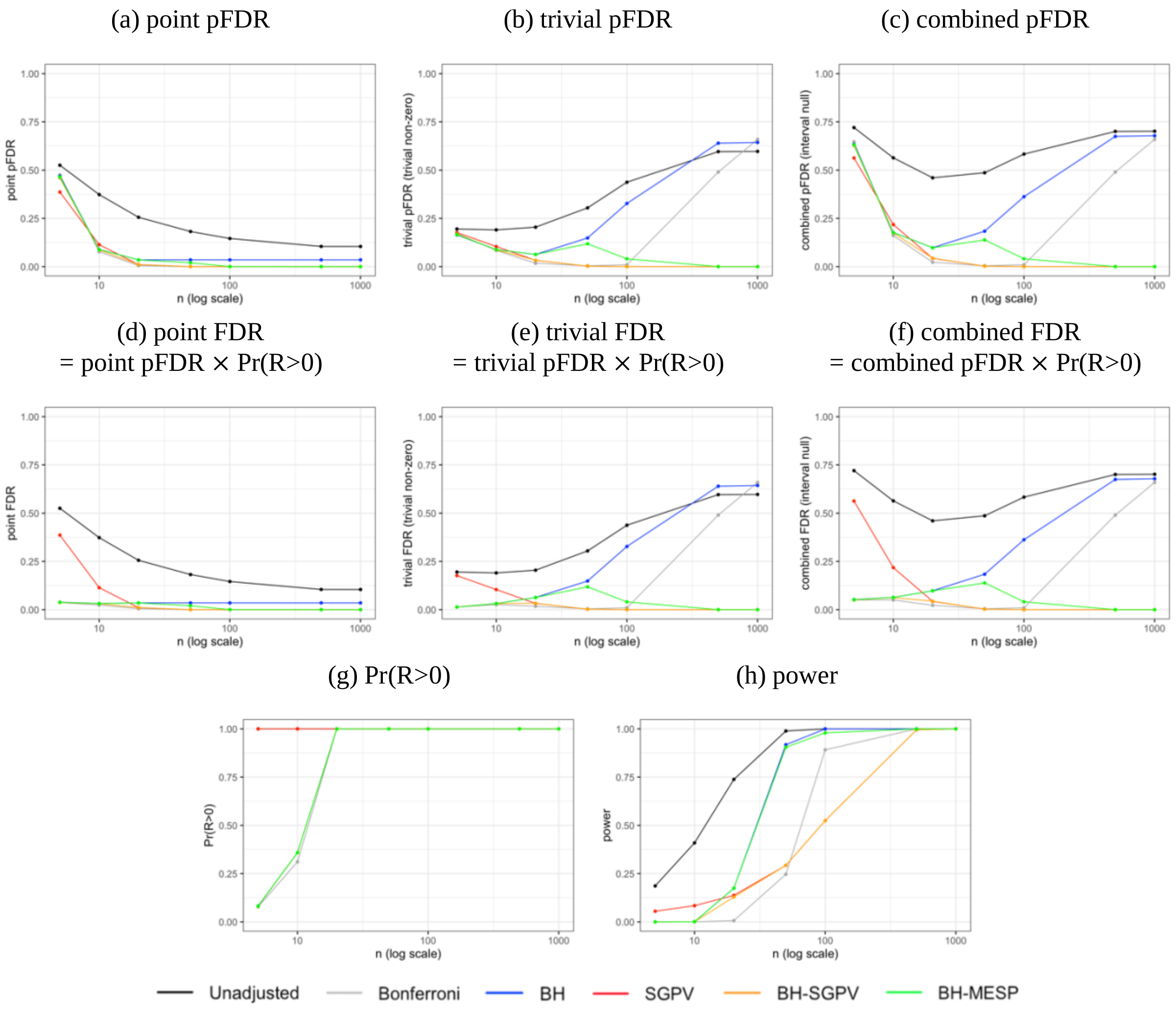
Figure 4.7 shows the same results as in Figure 4.5 (where the true underlying variance is common among all tests), but with these two hybrid approaches added. We can see that both procedures have $pFDR_{\{tv\}} \rightarrow 0$ and $pFDR_{\{comb\}} \rightarrow 0$ as $n \rightarrow \infty$, although there is an interesting case of one sample size (i.e., $n = 50$) for which the BH-MESP has an increased trivial and combined pFDR than for other sample sizes. Notably, both hybrid approaches do also achieve control of the point pFDR for small sample sizes. The trivial and combined pFDR are also controlled, or nearly controlled, across sample sizes (other than the case of $n = 50$ for BH-MESP referenced above). In this setting, the power of BH-MESP is nearly equivalent to the power of the BH procedure alone. Note that the power is calculated only for the non-trivial alternative effects (not including trivial effect sizes). The BH-SGPV procedure however does lose power as compared to SGPV alone for small sample sizes (although it is comparable to BH and BH-MESP, with all having power equal to zero for $n = 5$ and $n = 10$).

Figure 4.8 provides the results for the example setting where the variance values are $\sigma^2 \in \{1 \times (2\delta)^2, 5 \times (2\delta)^2\}$, where all null tests have the smaller variance, all alternative tests have the larger variance, and trivial effects have one of the variance values each with probability 0.5. This is the setting described in Figure 4.3, where the SGPV excels (in terms of the point pFDR being lower). However, we can see that this improved $pFDR_{\{pt\}}$ for small sample sizes does not hold for the two hybrid procedures, and instead have point pFDR comparable to the BH procedure alone. For the trivial and combined pFDRs, the behavior of BH-SGPV and BH-MESP are similar to BH for small sample sizes, intermediate between SGPV and BH approaches for moderate sample sizes, and similar to SGPV for larger sample sizes. The behavior of the power of the two hybrid procedures follows the same patterns as described for the prior example setting.

Overall, we see that the hybrid approaches are an improvement only in particular cases of sample size or variance magnitudes and distributions. For very small sample sizes (e.g., $n = 2$ to $n = 20$), they can maintain FDR control over both exactly zero and trivial effect sizes. However, as with BH, this FDR control comes from the Pr(R>0) being less than 1, not from a reduction in the pFDR. In some circumstances, the SGPV approach has a significantly lower pFDR than that of the hybrid methods for small and moderate sample sizes; in many other circumstances, the hybrid methods have comparable pFDRs, and in some cases the pFDR of the SGPV is significantly worse (Supplemental Figure 4.9).

4.5 Discussion

The positive false discovery rate for rejections based on the second-generation p-value has been examined previously in a limited scope of settings, such as in (Blume et al. 2018) and in Chapter 3. In this chapter, we have performed a more holistic examination of the second-generation p-value behavior in large-scale inference, including comparisons with other prevalent methods and the study of additional quantities such as the overall false discovery rate (pFDR scaled by $Pr(R > 0)$) and the power to reject scientifically meaningful effects. Additionally, our use of



(a) point pFDR

(b) trivial pFDR

(c) combined pFDR

(d) point FDR = point pFDR × Pr(R>0)

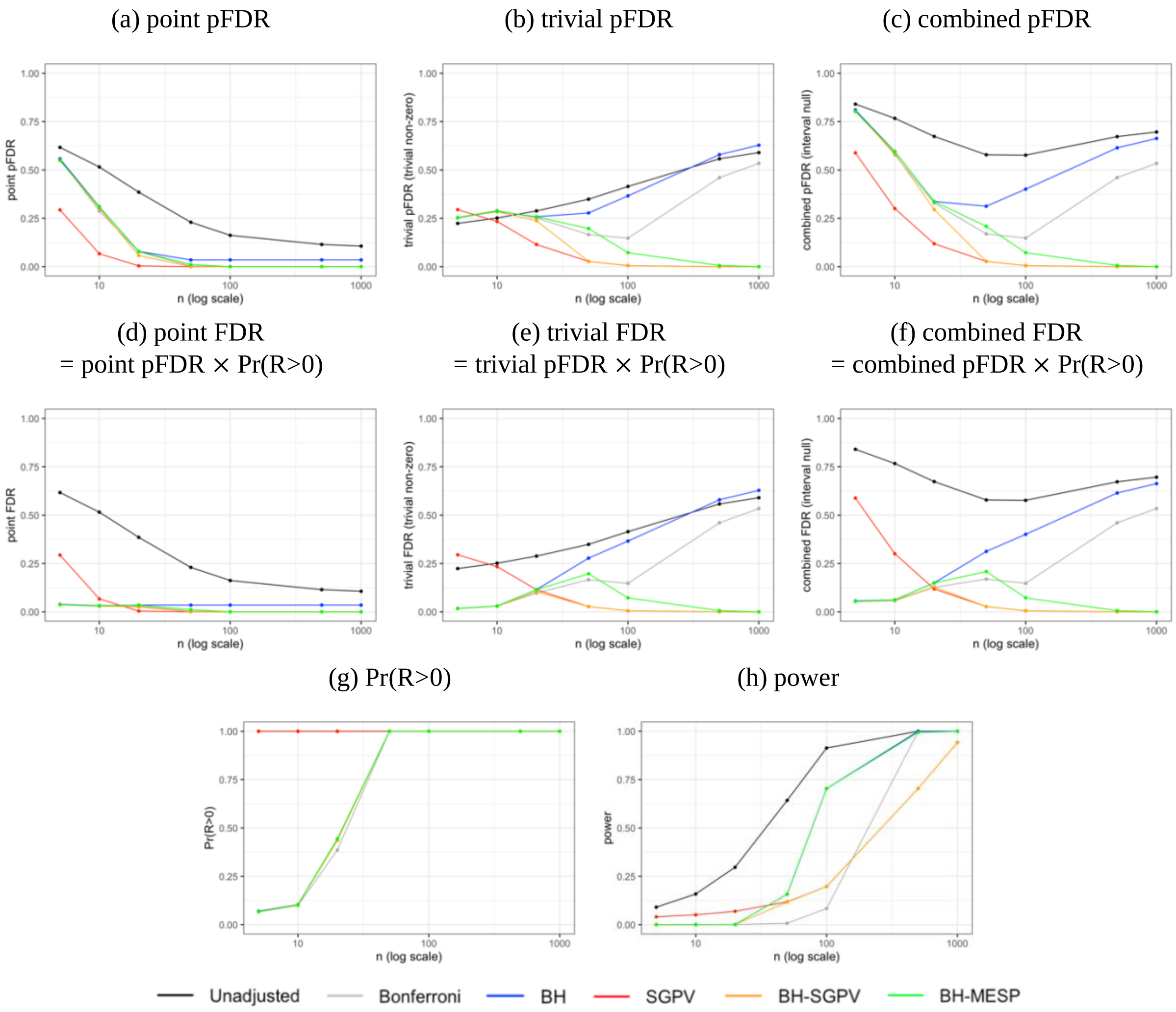
(e) trivial FDR = trivial pFDR × Pr(R>0)

(f) combined FDR = combined pFDR × Pr(R>0)

(g) Pr(R>0)

(h) power

Figure 4.7 Simulation estimates of false discovery rate quantities, including results from the hybrid procedures, for a setting with null, trivial, and non-trivial alternative effects, and with tests having a common variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1, \theta_{tv} = 0.5\delta, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and there is a common variance of $\sigma^2 = 1.5 \times (2\delta)^2$ for all tests.



(a) point pFDR

(b) trivial pFDR

(c) combined pFDR

(d) point FDR
= point pFDR × Pr(R>0)

(e) trivial FDR
= trivial pFDR × Pr(R>0)

(f) combined FDR
= combined pFDR × Pr(R>0)

(g) Pr(R>0)

(h) power

Figure 4.8 Simulation estimates of false discovery rate quantities, including results from the hybrid procedures, for a setting with null, trivial, and non-trivial alternative effects, and with null tests having smaller variance and alternative tests having larger variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1, \theta_{tv} = 0.5\delta, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and $\sigma^2 \in \{1 \times (2\delta)^2, 5 \times (2\delta)^2\}$, with all null tests having variance of $1 \times (2\delta)^2$, all non-trivial alternative tests having variance of $5 \times (2\delta)^2$, and random distribution of the variance values among the trivial tests.

the estimated sample variance in place of the known variance, and examination a wider range of scenarios for the variance of tests, further elucidates practical use of the SGPV.

First, we have derived the probability $Pr(R > 0)$ for the second-generation p-value. For the number of tests m which are now more common, i.e., in the hundreds of thousands or even millions, this probability for the SGPV is 1, and therefore the FDR is not controlled unless the pFDR itself is controlled. We find that $Pr(R > 0) < 1$ for SGPV only when the number of tests is small – relative to modern large-scale inference standards – in combination with a small sample size. However, regardless of the number of tests, we find that this probability does converge to 1 with increasing sample size. In general, what this indicates is that for most commonly encountered settings of large-scale inference, the SGPV does not generally control the FDR across all sample sizes. Rather, it is naturally controlled when the pFDR itself is below the threshold, such as 0.05, for large enough sample size.

Next, we studied a range of simulation scenarios to examine the pFDR, the probability $Pr(R > 0)$, the resulting overall FDR, and the power of the SGPV procedure in comparison with standard approaches such as the unadjusted p-value, Bonferroni, and Benjamini-Hochberg procedures, with the assumption of known variance relaxed (i.e., the sample variance estimator used). To build concepts, we began by ignoring the issue of trivial effects. Our results from these settings demonstrated several things. One finding that may be of interest to frequent users of the BH procedure is that – although the FDR for point null effects is indeed controlled at $\pi_0 q$ across all sample sizes – for small n, the control comes from a lower probability of rejections. That is, the pFDR may be quite large, i.e., $\pi_0 q \leq BH\ pFDR \leq \pi_0$, where we observed some pFDR values close to the π_0 upper boundary, in settings with very small sample size. For example, with $\pi_0 = 0.9$, we may see many more than $\pi_0 q = 4.5\%$ of the tests rejected by BH to be in fact null findings; even as many as 90% of them in some cases such as with $n = 5$ or $n = 10$. Thus, in practice, particularly with small sample size, rejections from the BH procedure must be interpreted with extra care. If the set of rejections is erroneously interpreted as having an expected rate of false rejections of 5%, when in reality it could be ten times that or more, this may be very misleading and contribute to future wasted scientific endeavor.

In this scenario of no trivial effects, for smaller sample sizes, the pFDR of the SGPV is variable in comparison to that of popular multiple testing procedures such as Bonferroni and Benjamini-Hochberg. In particular, compared to BH the SGPV pFDR is sometimes lower, sometimes higher, and sometimes comparable. We find that this varies based on factors such as the true alternative effect size, the magnitudes and distributions of the underlying variance for each test, and the magnitude of sample size. In the limit, however, we observed the same behavior established under the more rigid assumptions of known and common variance, namely that both SGPV and BH result in a controlled pFDR, with the BH converging to the lower bound of $\pi_0 q$ (as $pFDR \rightarrow FDR$) and with the SGPV pFDR converging to zero (aligning with results established in (Blume et al. 2018)).

Understanding the behavior of the second-generation p-value and other common procedures under the consideration of the point null hypothesis is helpful for establishing concepts. However, the main advantage of the SGPV emerges when we are concerned with trivial effects, as the traditional p-value methods do not account for the magnitude of effects; an effect size such as 0.001 is considered an alternative effect under the standard point null paradigm, even though it may be practically indistinguishable from 0. Therefore, our next step was to examine settings with trivial effects, and to extend the definition of the false discovery rate quantities to account for the interval null hypothesis (similarly to Chapter 3).

In these circumstances, classical p-value methods, regardless of whether they are adjusted for multiple comparisons, increasingly reject these practically null effects when the sample size is large enough. We observed in our simulations, as expected, that procedures based on the classical p-value lack FDR control for combined zero and trivial effects for virtually all sample sizes, underlying variance, or true alternative effect size. Even in the limit, classical p-value methods, regardless of whether they are adjusted for multiple comparisons, increasingly reject these practically null effects.

Instead, methods which account for absolute effect size (rather than standardized effect size, as with the p-value) are needed for appropriate behavior of the false discovery rates. The SGPV is one such procedure. The interval null pFDR for the second-generation p-value, which measures the rate of rejections of both exactly null **and** practically null (trivial) effects, however, does converge to 0. These results support the theoretical result derived in Chapter 3 (i.e., in Section 3.3.4).

Additionally, we considered two others – referred to as hybrid procedures – which incorporate both the BH procedure and an effect size criterion such as the SGPV or the estimated effect. The BH-SGPV hybrid approach takes the intersection of those two tests, while the BH-MESP hybrid approach is a modification of (Goodman et al. 2019) which requires that the BH test reject and that the estimated effect lies outside of the interval null. These approaches can provide some improvements, although they are not necessarily a universal solution. The hybrid methods were found to not fully control the interval null FDR for finite sample sizes, however they do result in a considerable reduction. For many settings with very small sample size, the interval null FDR for the hybrid methods ranged from approximately 0-0.12, compared to 0-0.85 for the SGPV alone. However, as with the BH procedure, much of this FDR control or attenuation was observed to result from a reduction in the probability of making any rejections, rather than a reduction in the pFDR itself. This means that the same tradeoff as discussed with the BH procedure, between FDR control and minimization of the rate of false rejections in observed rejections (pFDR), must be considered. In many settings, the interval null positive false discovery rates of the hybrid methods are comparable to that of the SGPV or BH procedures by themselves; in others, the pFDRs are found to be worse than the SGPV method (for some finite sample sizes, and for some settings of variance distribution among tests). Although, the SGPV and the hybrid methods are all seen to have an interval null pFDR (and thus FDR) which converge to zero in our simulation results.

So far, we have focused our discussion on false discovery quantities. Ideally, a procedure would control the FDR and/or minimize the pFDR, while also maximizing power. It has been previously established in (Blume et al. 2018) that the SGPV has reduced power compared to the classical p-value, which is intuitively clear, as the SGPV requires that the confidence interval excludes the entire null region, not only zero. Further, in their Supplement 1, Blume et al. (2018) provide one example for fixed sample size where the SGPV power is observed to be greater than that of the Bonferroni procedure for large number of tests. In the present chapter, we examine the power of the

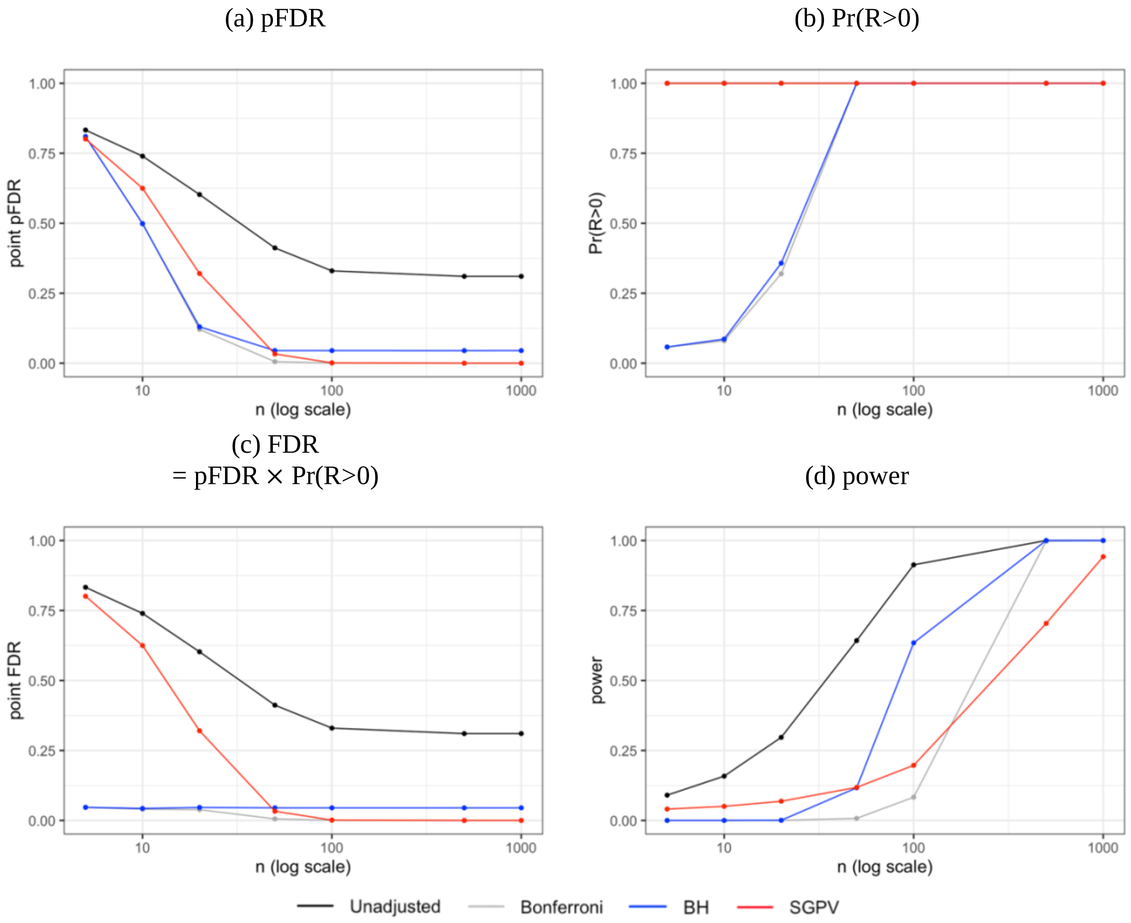
SGPV, Bonferroni, BH, and hybrid procedures across a wider range of settings. We find that for moderate sample sizes, the BH has greater power than the SGPV procedure and converges to 1 at a faster rate, although as n becomes larger (e.g., between n = 500 and n = 1,000), the SGPV and BH procedures both converge to 1 (although often the BH procedure converges much earlier, e.g., between n = 100 and n = 500). The comparison with Bonferroni is similar, although the discrepancy between the SGPV and Bonferroni power was often smaller. For most sample sizes, the BH adjusted MESP hybrid procedure was found to have similar power to the BH procedure alone, and the BH-SGPV hybrid procedure was found to have similar power to the SGPV procedure alone. At times, the power of the hybrid procedures was lower, but not by much. Interestingly, for small sample sizes, the SGPV procedure power was found to be the greatest of all considered – this is likely because, unlike with BH or the hybrid procedures where frequently the number of rejections was 0 (i.e., $\Pr(R > 0) < 1$) to achieve FDR control, the SGPV procedure nearly always rejected at least one test. This further exhibits the tradeoff between quantities, e.g., between strict FDR control and power.

Overall, which procedure is best depends on 1) the main priorities of the study (control of the FDR, minimized pFDR, statistical power, etc.), 2) the importance of excluding trivial effects, and 3) factors such as sample size, number of tests, and the true magnitudes and distributions of effect sizes and variances. If strict FDR control of discovery of effects which are exactly zero is the main priority, then an approach such as BH may still be preferred. On the other hand, if FDR control is the key goal but trivial effects are of concern (to reduce their inclusion in the set of rejections), our methods demonstrate that a hybrid method such as BH-SGPV or BH-MESP would achieve these goals the best in most scenarios, although with a possible reduction in power. If, however, pFDR minimization for combined zero and trivial effects is the most important priority, then the best choice – from either the SGPV alone or a hybrid method – depends on the sample size and expected distributions and magnitudes of variances. Of course, in practice, these last quantities are not known, and therefore there is not always a clear choice.

In this chapter, we have examined only a subset of many potential multiple testing approaches; however, it is not likely that any another particular method will be universally best in all scenarios, either. As with many statistical approaches, tradeoffs are likely necessary. It is clear, however, that when trivial effects are taken into consideration, traditional methods are unsatisfactory, and some form of method which accounts for effect size is needed. Regardless, after an approach is chosen based on the priorities and circumstances of the study, a reliable method to estimate the pFDR remains the important next step.

One promising area for future work lies in a more localized or ranking based assessment using second-generation concepts. The general idea is that, following a large-scale testing procedure, the next step in the scientific process might be to follow-up on the significant findings in more depth, such as with laboratory studies. However, if a large number of tests initially met the significance criteria, it is not feasible to follow up on all of them, and thus we need a method to prioritize them in some way. Essentially, we wish to define a ranking for associations, by which to sort and determine the top set of “interesting” findings. In the specific setting of genetics, candidate genetic variants may often initially be ranked by their raw p-value; further specialized prioritization methods have also been developed to incorporate information from other sources or modalities (Cantor et al. 2010, Doncheva et al. 2012, Tam et al. 2019). In more general settings, focusing only on the statistical information at hand, we would benefit from 1) prioritizing tests which have lower false discovery rates, and 2) prioritizing tests which are likely to have truly meaningful effects. Utilizing the second-generation p-value framework in place of classical p-values may be helpful in this way. One potential approach would be to use the SGPV and the respective delta-gap, which is an accompanying measure to the SGPV, defined in (Blume et al. 2018), which measures the scaled distance between the null and interval estimate boundaries. This metric provides a unique measure for each $p_\delta = 0$, and if used as a ranking of tests, would focus attention on the those supporting larger effect sizes. One possible modification to the delta-gap, i.e., additionally standardizing by the estimated standard error, could also be considered. Alternately, a more local FDR quantity or estimate for the second-generation p-value might be defined, as discussed in (Blume et al. 2019), with the smallest estimated local FDRs guiding which tests to focus follow-up efforts on. There is some similarity in these approaches compared to those proposed for genetic applications which incorporate both p-values and effect size, such as in (Tusher et al. 2001, Xiao et al. 2014). Overall, combining the ideas of effect size and localized FDRs, possibly via second-generation p-values, with the incorporation of an interval null hypothesis, could result in a clear path forward for future studies, and one which accounts for scientific relevance.

4.6 Appendix A: Remarks and supplemental content



(a) pFDR

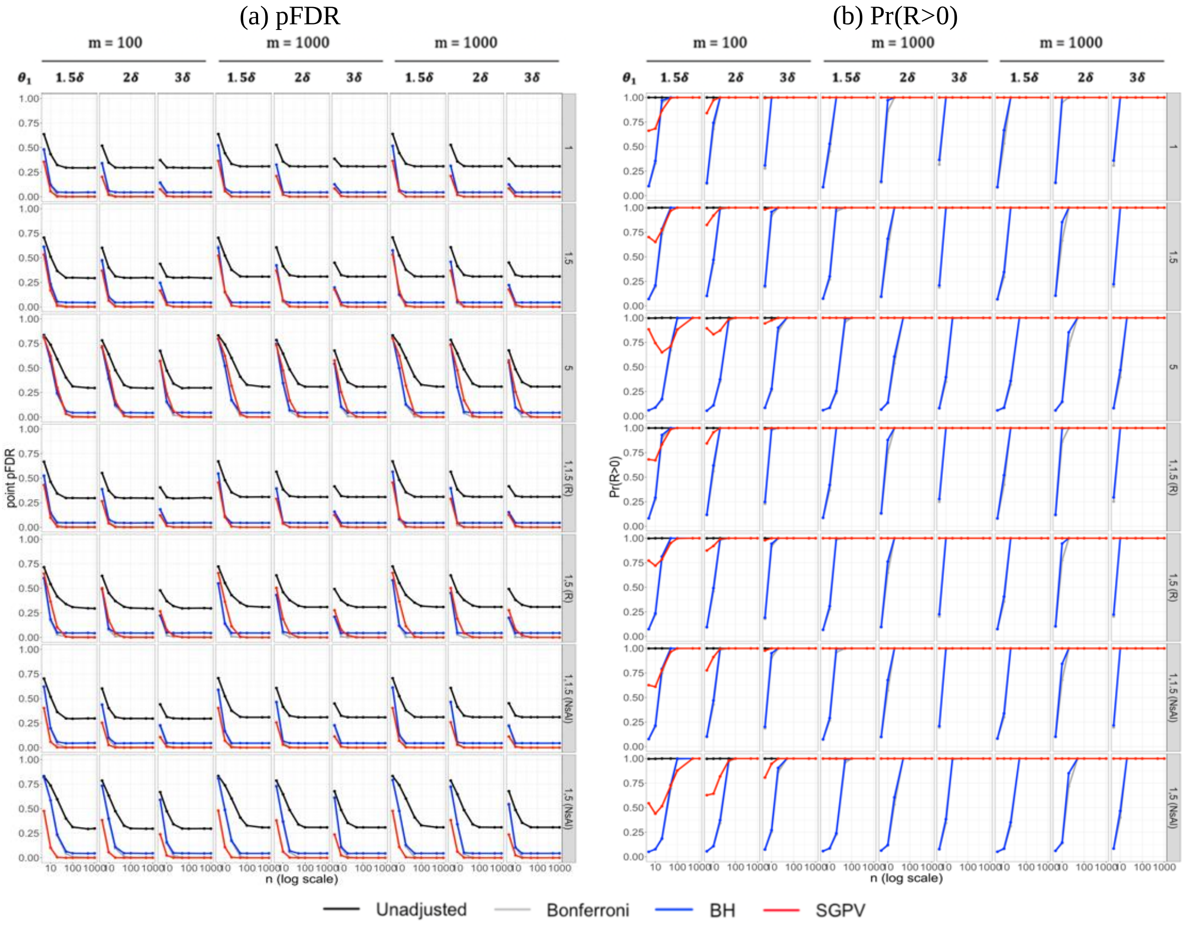
(b) Pr(R>0)

(c) FDR = pFDR × Pr(R>0)

(d) power

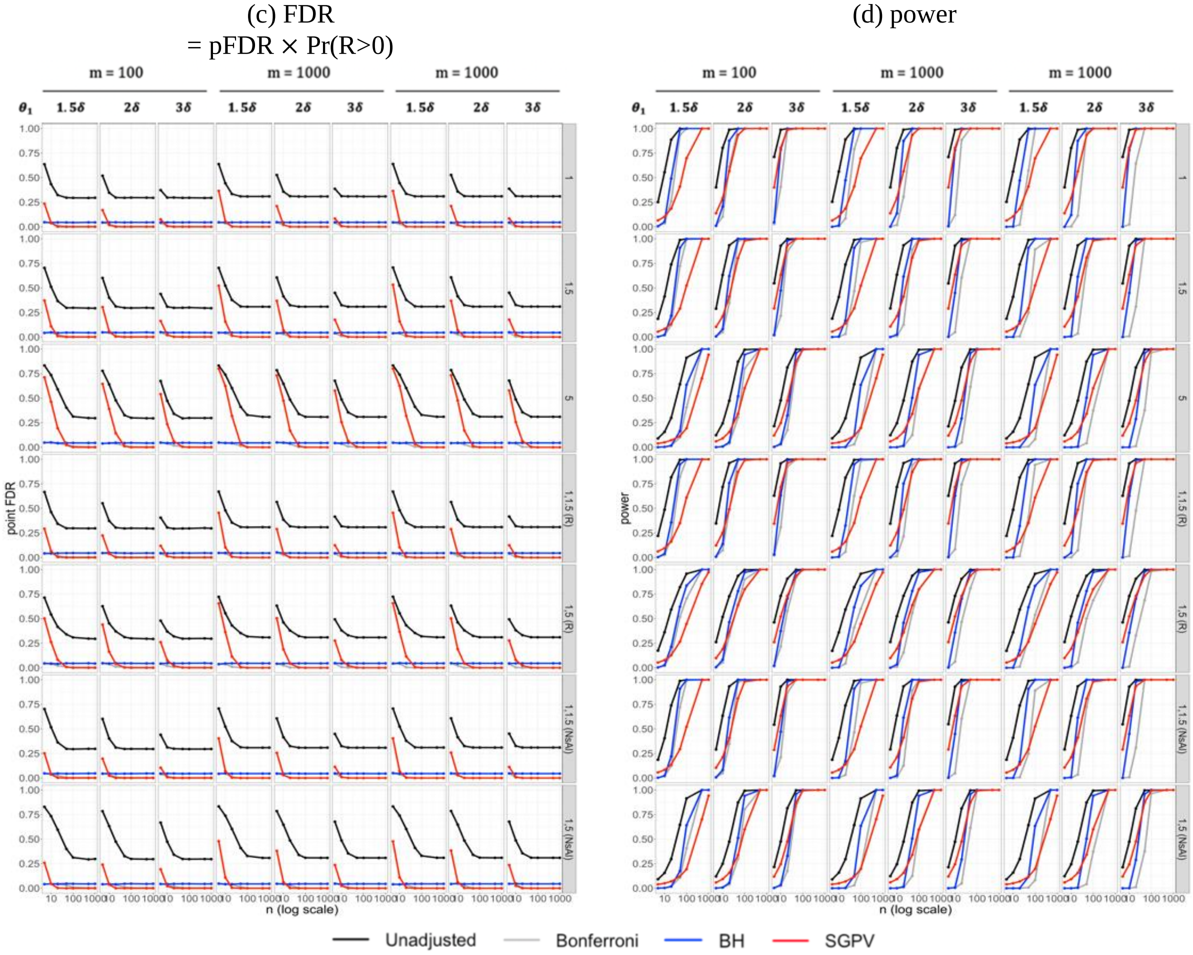
Supplemental Figure 4.1 Simulation estimates of false discovery rate quantities for a setting with only null and non-trivial alternative effects, and with tests having a large common variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.9, \pi_1 = 0.1, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and there is a common variance of $\sigma^2 = 5 \times (2\delta)^2$ for all tests.

Supplemental Figure 4.2 Simulation estimates of false discovery rate quantities across all examined combinations of alternative effect size, numbers of tests and variance settings, with only null and non-trivial alternative effects. Here, we are examining the setting of $\pi_0 = 0.9$ and $\delta = 0.1$. All considered settings of number of tests m (first-level grouping along the horizontal axis), alternative effect size (second-level sub-grouping along the horizontal axis), and variance (along the vertical axis) are provided. The variance labels shown are in terms of scale of the variance compared to the baseline of $\sigma^2 = (2\delta)^2$. That is, the values shown are ω, where $\sigma^2 = \omega \times (2\delta)^2$. In these variance labels, "(R)" denotes the scenarios where each test has an equal probability of observing each of the two variance values, and "(NsAl)" denotes the scenario where all null tests have the smaller variance value and all alternative tests have the larger variance value. (a) pFDR, (b) Pr(R>0), (c) FDR, (d) Power.

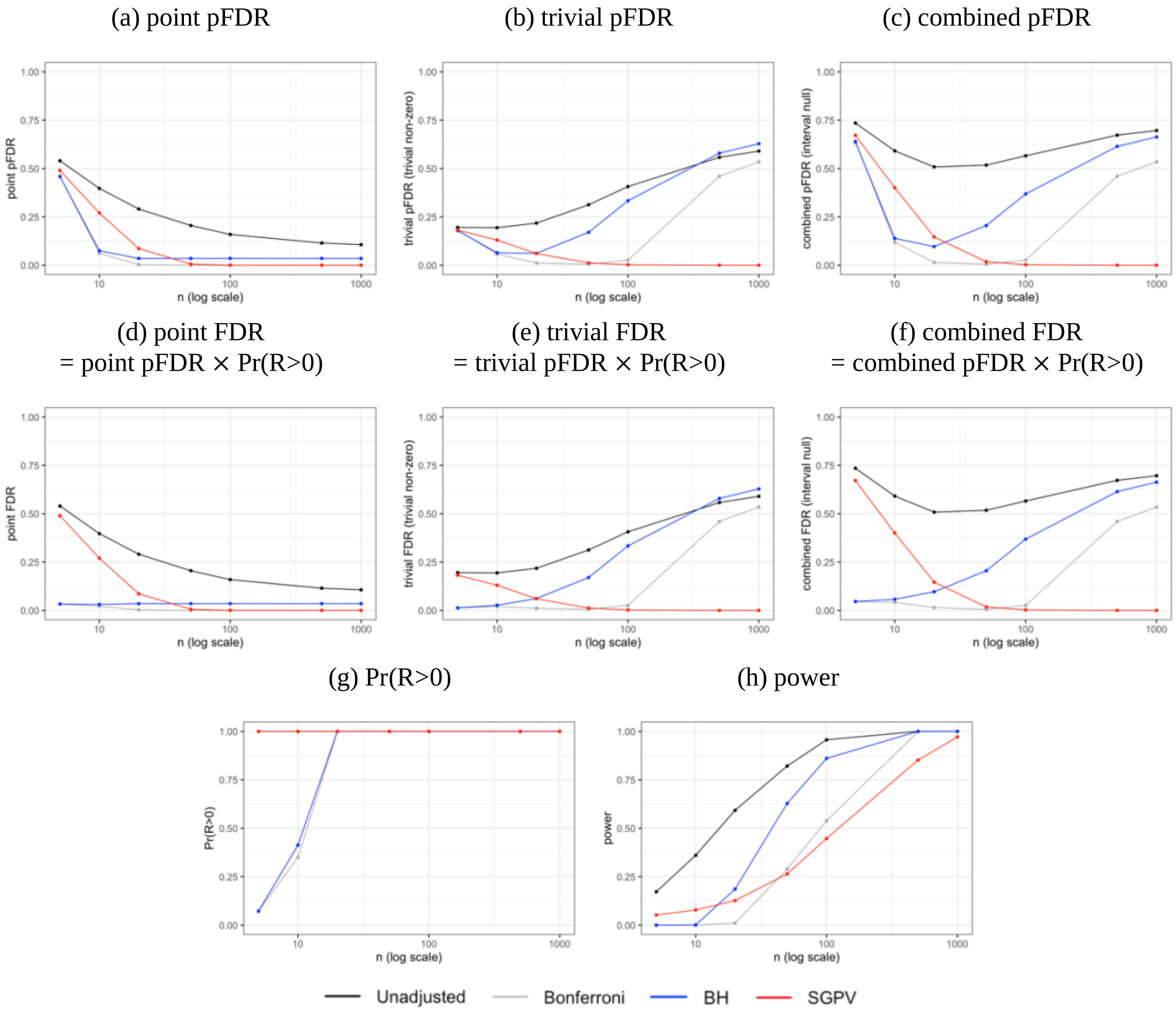


(a) pFDR

(b) Pr(R>0)


(c) FDR
= pFDR × Pr(R>0)

(d) power



(a) point pFDR

(b) trivial pFDR

(c) combined pFDR

(d) point FDR
= point pFDR × Pr(R>0)

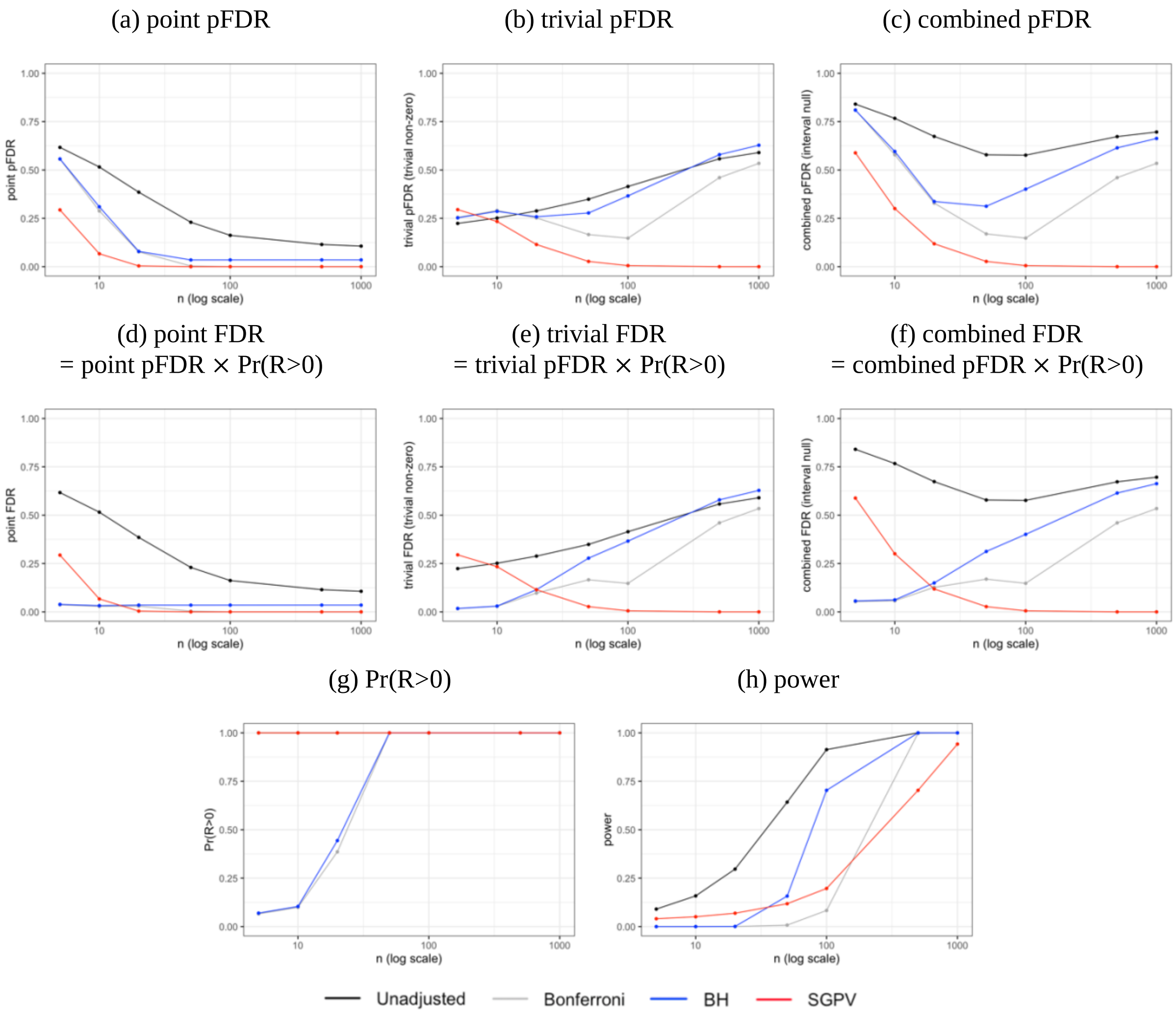
(e) trivial FDR
= trivial pFDR × Pr(R>0)

(f) combined FDR
= combined pFDR × Pr(R>0)

(g) Pr(R>0)

(h) power

Supplemental Figure 4.3 Simulation estimates of false discovery rate quantities for a setting with null, trivial, and non-trivial alternative effects, and with two possible variance values distributed randomly among tests. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1, \theta_{tv} = 0.5\delta, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and $\sigma^2 \in \{1 \times (2\delta)^2, 5 \times (2\delta)^2\}$, with random distribution among tests (tests have 0.5 probability of having each variance value).


(a) point pFDR

(b) trivial pFDR

(c) combined pFDR

(d) point FDR
= point pFDR × Pr(R>0)

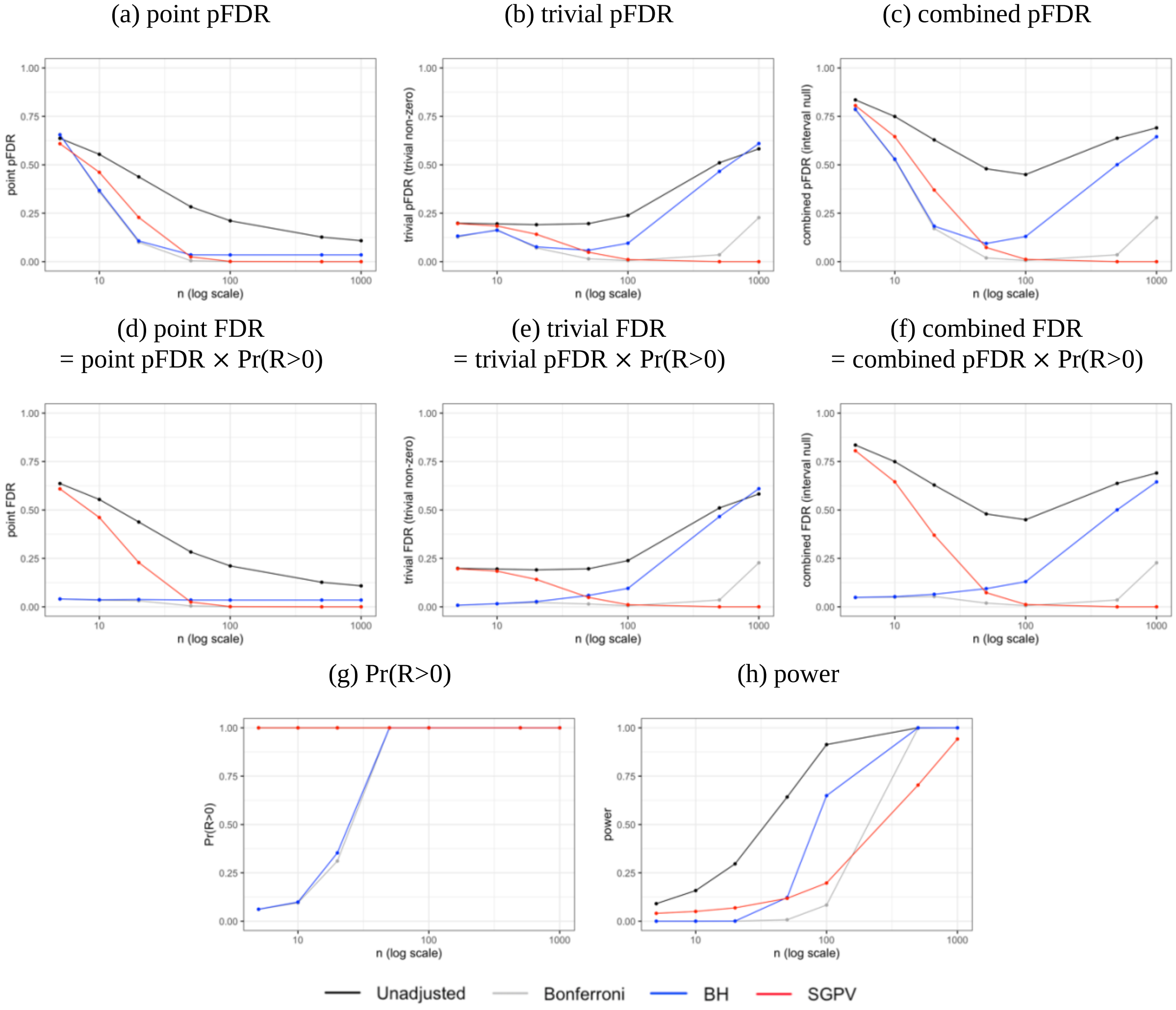
(e) trivial FDR
= trivial pFDR × Pr(R>0)

(f) combined FDR
= combined pFDR × Pr(R>0)

(g) Pr(R>0)

(h) power

Supplemental Figure 4.4 Simulation estimates of false discovery rate quantities for a setting with null, trivial, and non-trivial alternative effects, and with null tests having smaller variance and alternative tests having larger variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1, \theta_{tv} = 0.5\delta, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and $\sigma^2 \in \{1 \times (2\delta)^2, 5 \times (2\delta)^2\}$, with all null tests having variance of $1 \times (2\delta)^2$, all non-trivial alternative tests having variance of $5 \times (2\delta)^2$, and random distribution of the variance values among the trivial tests.



(a) point pFDR

(b) trivial pFDR

(c) combined pFDR

(d) point FDR
= point pFDR × Pr(R>0)

(e) trivial FDR
= trivial pFDR × Pr(R>0)

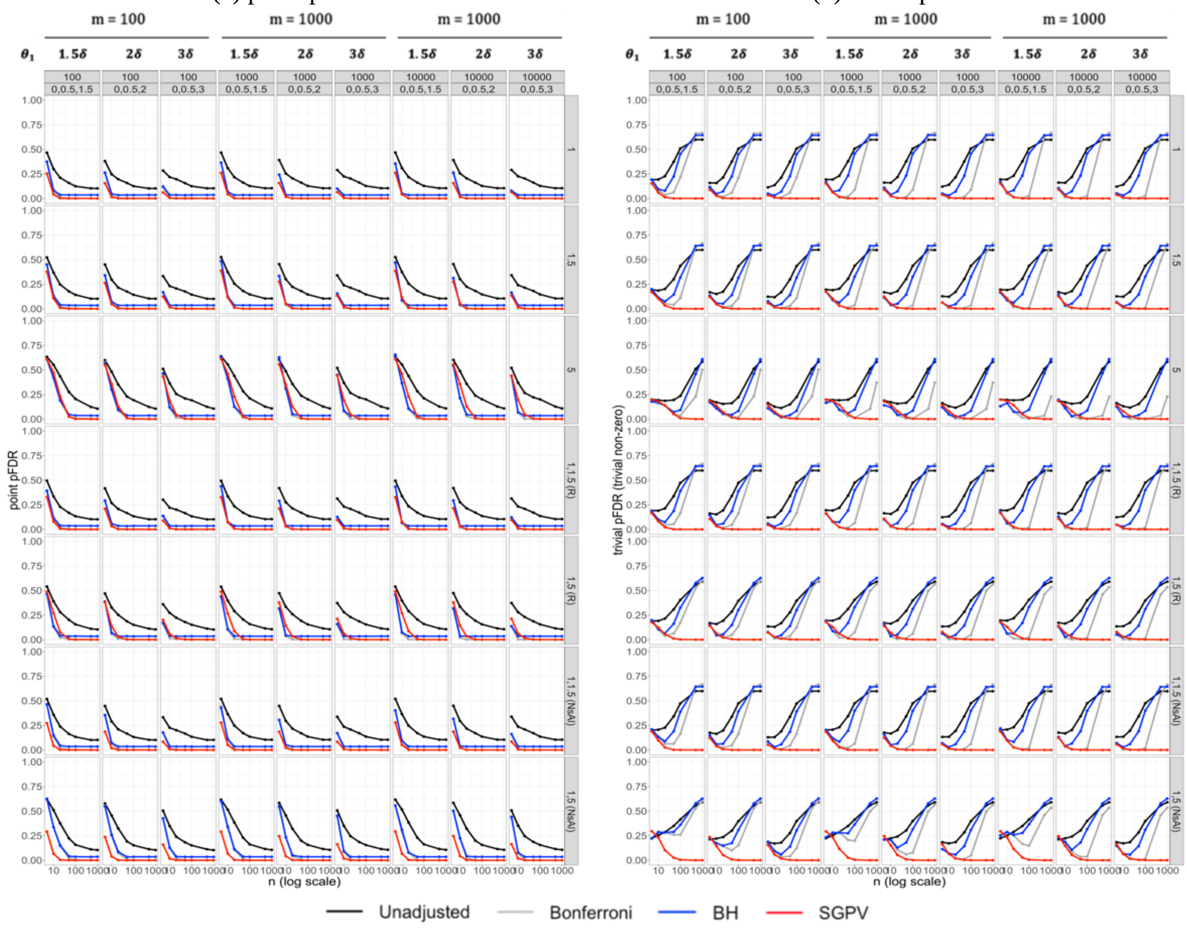
(f) combined FDR
= combined pFDR × Pr(R>0)

(g) Pr(R>0)

(h) power

Supplemental Figure 4.5 Simulation estimates of false discovery rate quantities for a setting with null, trivial, and non-trivial alternative effects, and with tests having a large common variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1, \theta_{tv} = 0.5\delta, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and there is a common variance of $\sigma^2 = 5 \times (2\delta)^2$ for all tests.

Supplemental Figure 4.6 Simulation estimates of false discovery rate quantities across all examined combinations of alternative effect size, numbers of tests and variance settings, with null, trivial, and non-trivial alternative effects. Here, we are examining the setting of $\pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1$, and $\delta = 0.1$. All considered settings of number of tests m (first-level grouping along the horizontal axis), alternative effect size (second-level sub-grouping along the horizontal axis), and variance (along the vertical axis) are provided. The variance labels shown are in terms of scale of the variance compared to the baseline of $\sigma^2 = (2\delta)^2$. That is, the values shown are ω, where $\sigma^2 = \omega \times (2\delta)^2$. In these variance labels, "(R)" denotes the scenarios where each test has an equal probability of observing each of the two variance values, and "(NsAl)" denotes the scenario where all zero null tests have smaller variance, all alternative tests have larger variance, and trivial effects have one of the two variance values with probability 0.5 for each. (a) point pFDR, (b) trivial pFDR, (c) combined pFDR, (d) point FDR, (e) trivial FDR, (f) combined FDR, (g) Pr(R>0), (h) power.

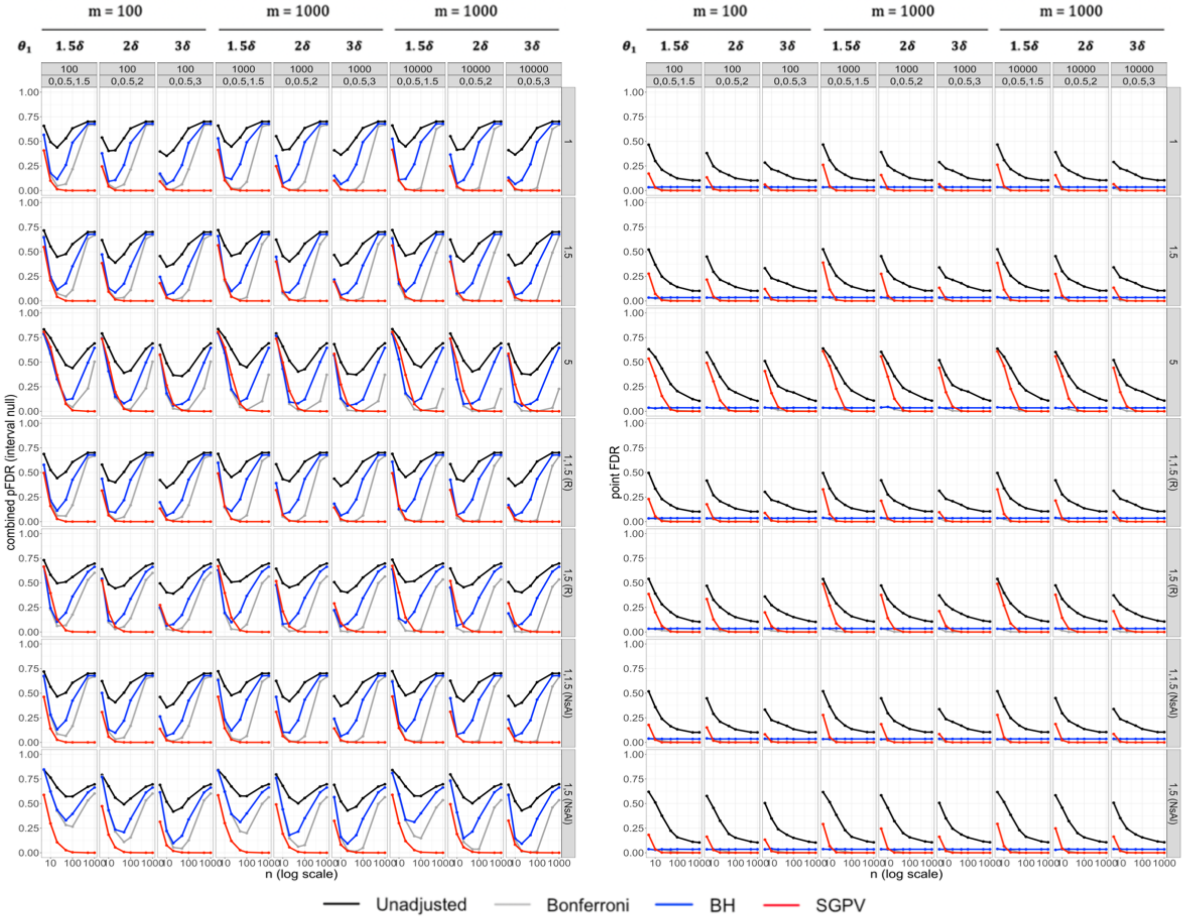


(a) point pFDR

(b) trivial pFDR

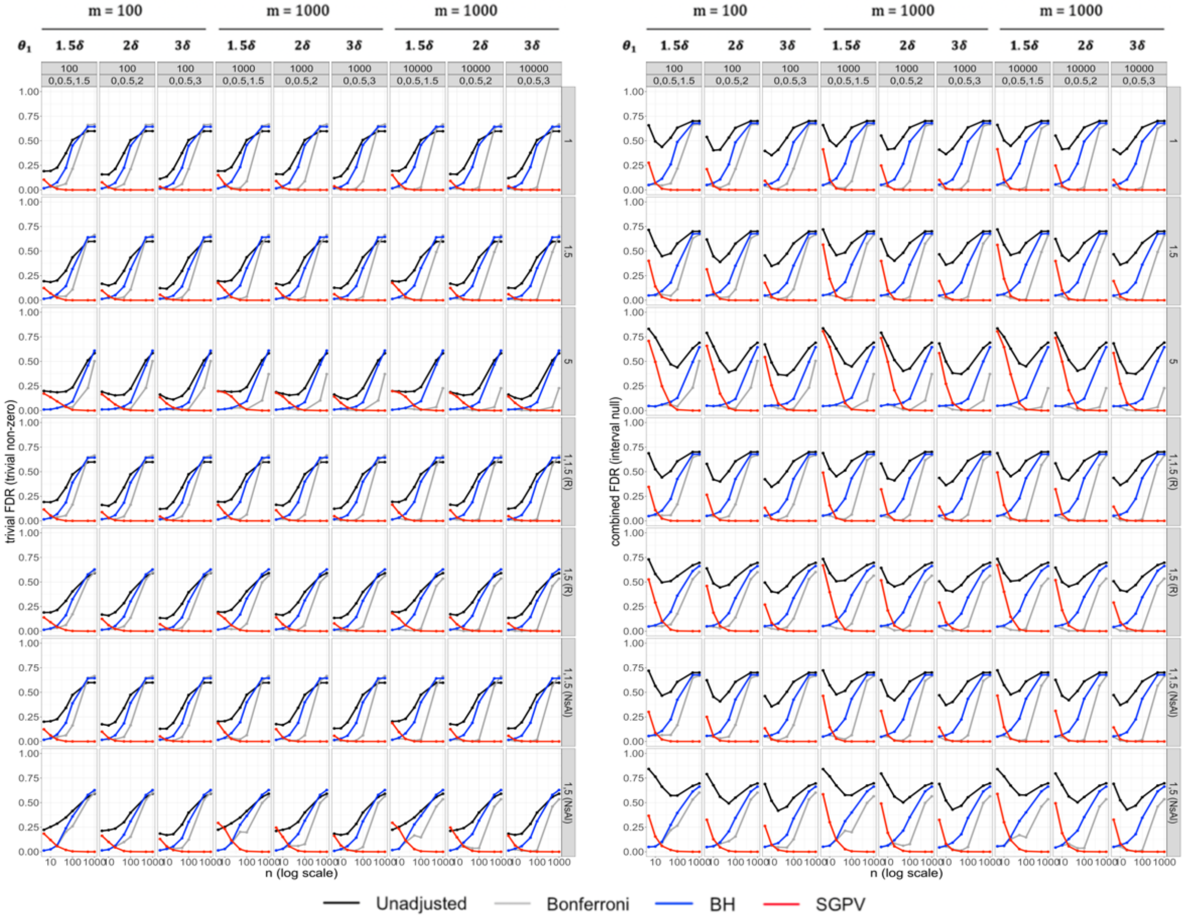
(c) combined pFDR

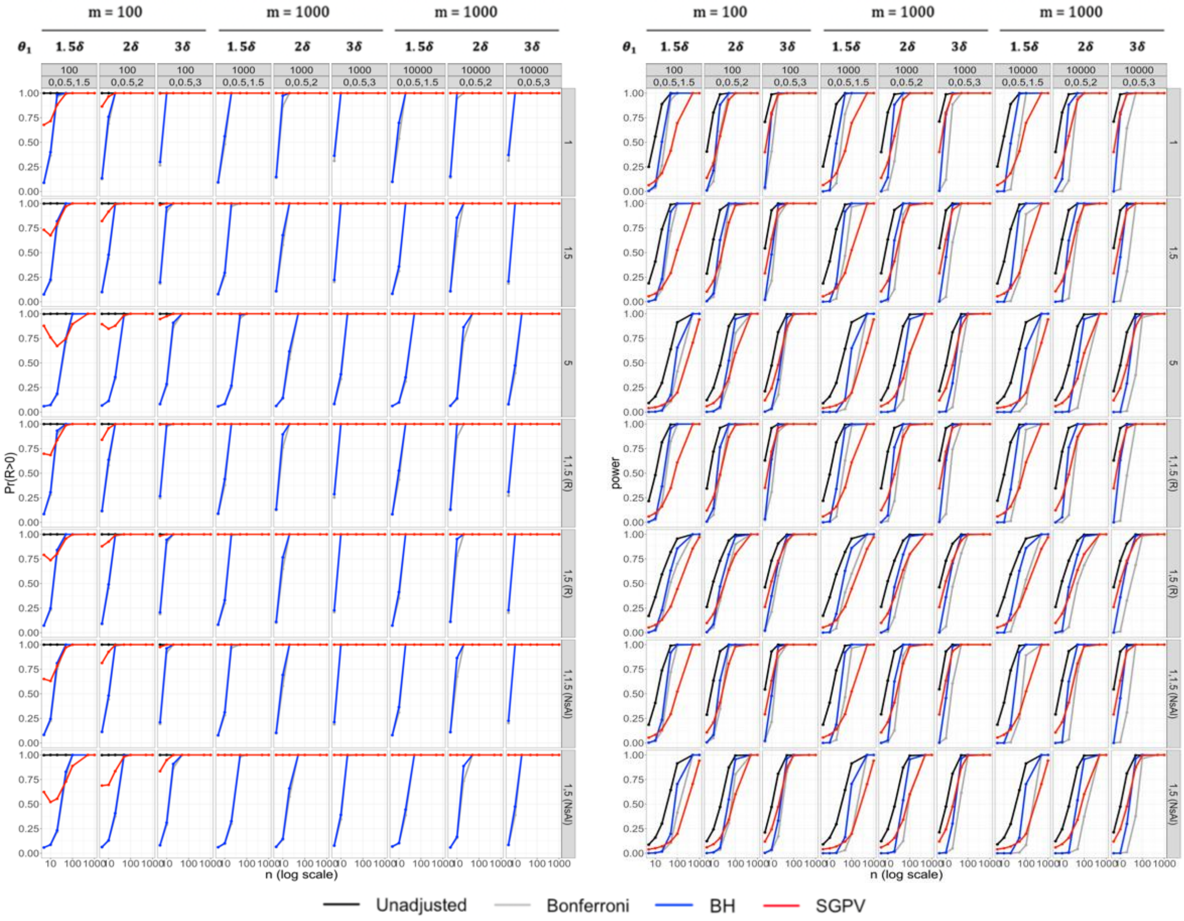
(d) point FDR
= point pFDR × Pr(R>0)



(e) trivial FDR
= trivial pFDR × Pr(R>0)

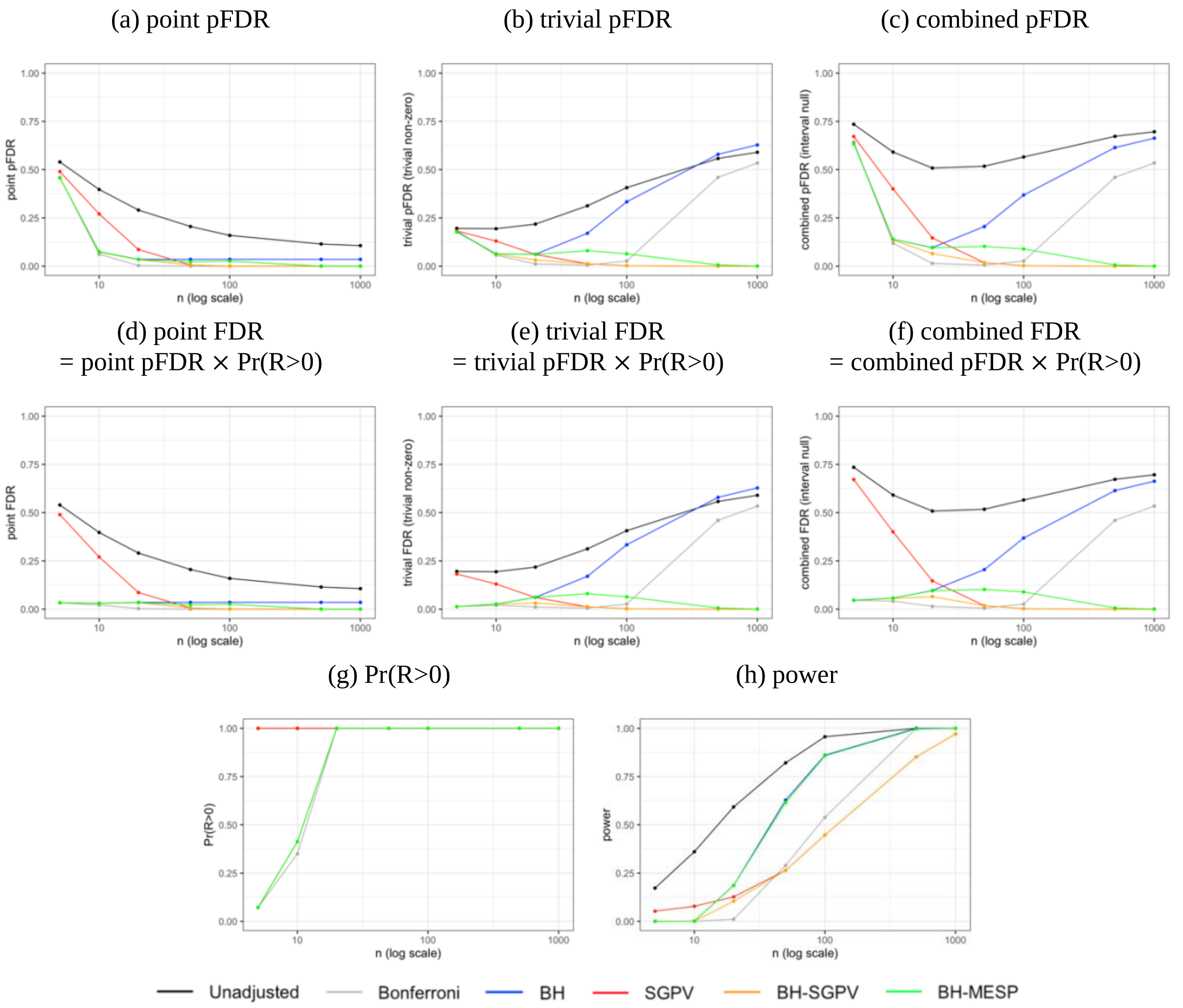
(f) combined FDR
= combined pFDR × Pr(R>0)





(g) Pr(R>0)

(h) power



(a) point pFDR

(b) trivial pFDR

(c) combined pFDR

(d) point FDR
= point pFDR × Pr(R>0)

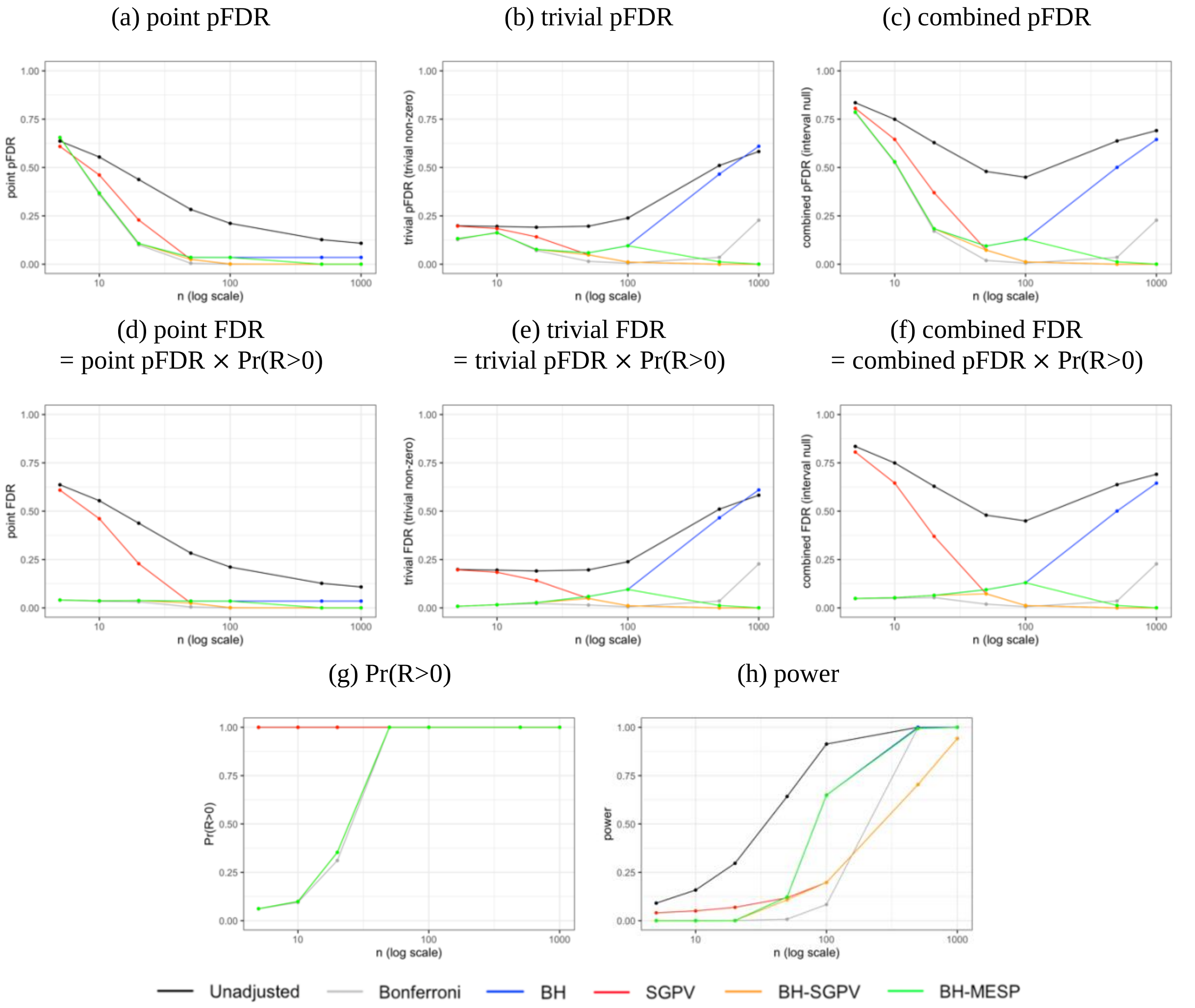
(e) trivial FDR
= trivial pFDR × Pr(R>0)

(f) combined FDR
= combined pFDR × Pr(R>0)

(g) Pr(R>0)

(h) power

Supplemental Figure 4.7 Simulation estimates of false discovery rate quantities, including results from the hybrid procedures, for a setting with null, trivial, and non-trivial alternative effects, and with two possible variance values distributed randomly among tests. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1, \theta_{tv} = 0.5\delta, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and $\sigma^2 \in \{1 \times (2\delta)^2, 5 \times (2\delta)^2\}$, with random distribution among tests (null, trivial and alternative tests each have 0.5 probability of having each variance value).



(a) point pFDR

(b) trivial pFDR

(c) combined pFDR

(d) point FDR
= point pFDR × Pr(R>0)

(e) trivial FDR
= trivial pFDR × Pr(R>0)

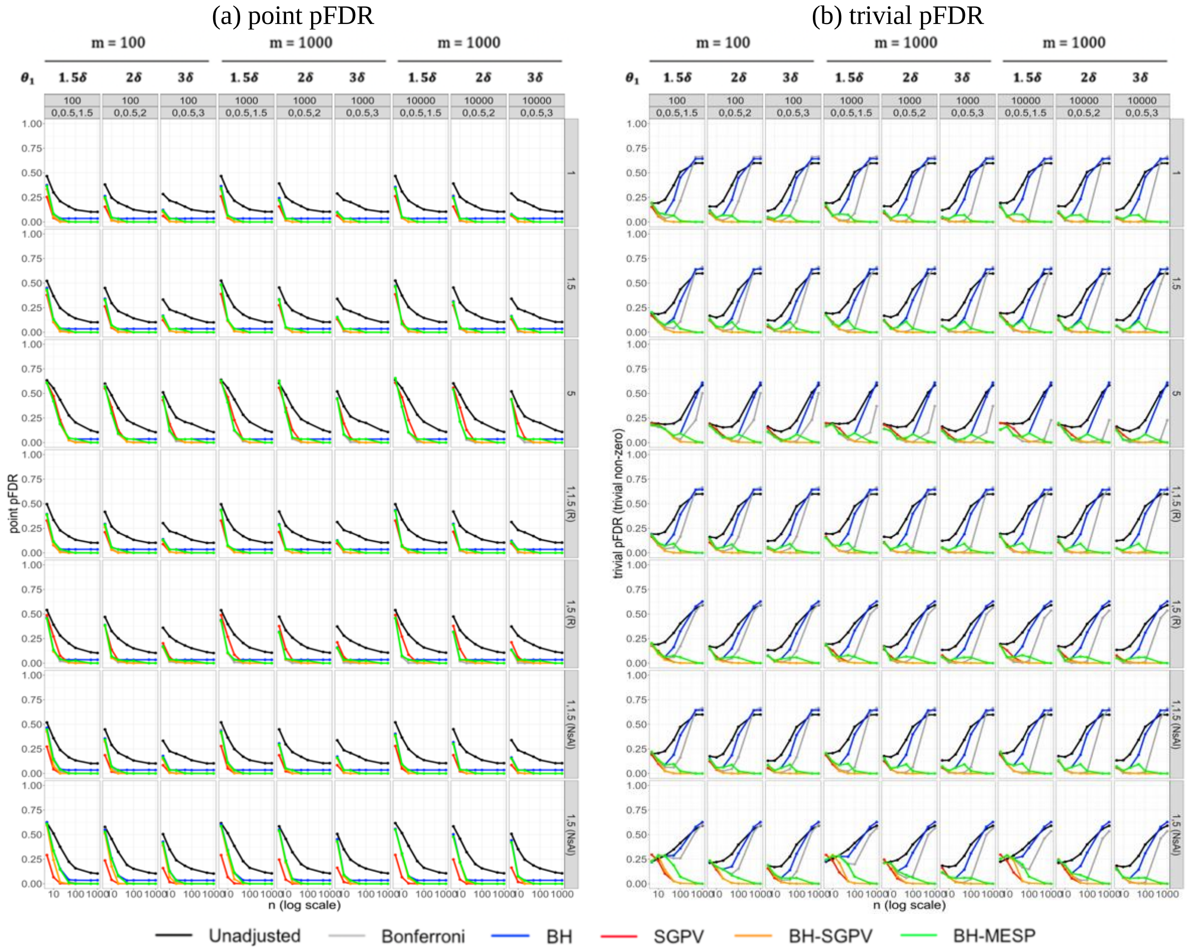
(f) combined FDR
= combined pFDR × Pr(R>0)

(g) Pr(R>0)

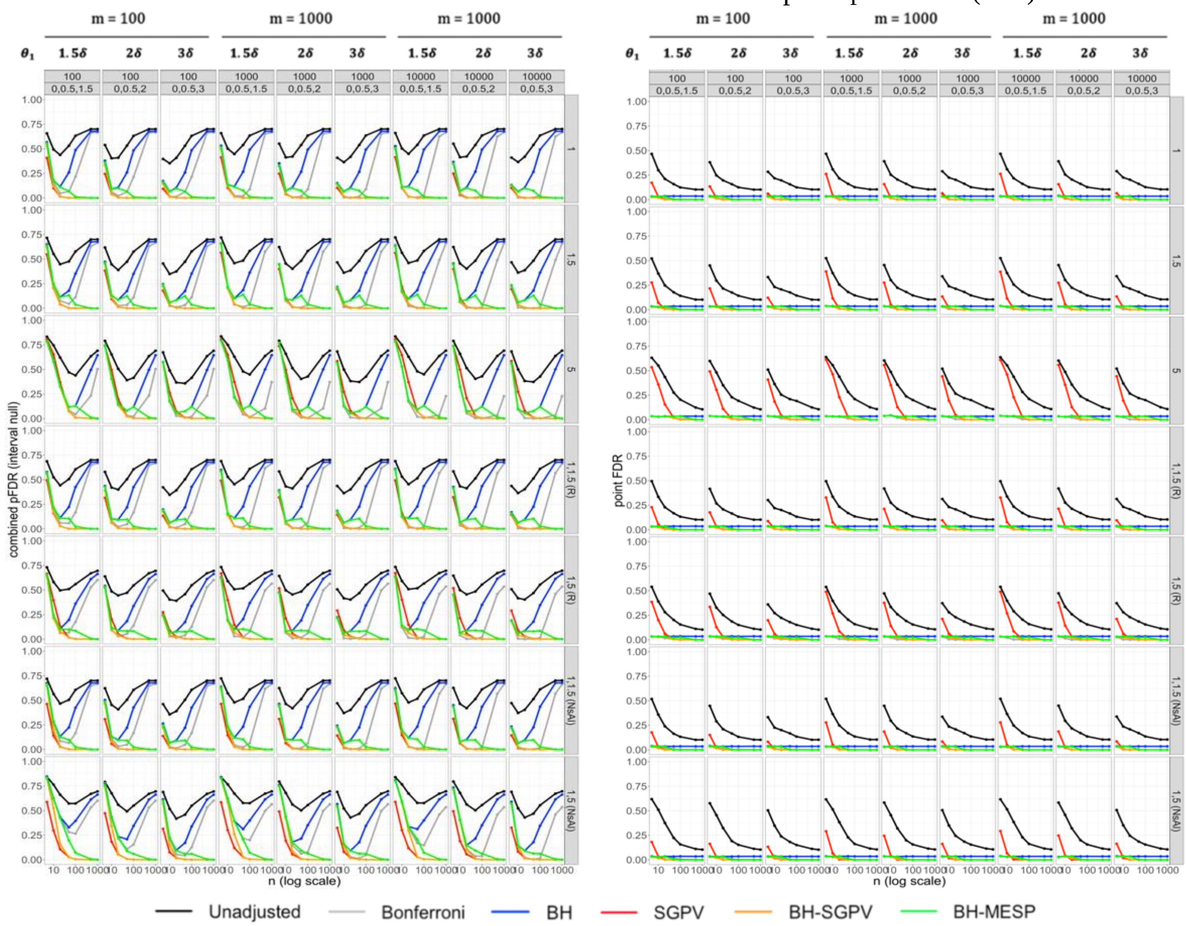
(h) power

Supplemental Figure 4.8 Simulation estimates of false discovery rate quantities, including results from the hybrid procedures, for a setting with null, trivial, and non-trivial alternative effects, and with tests having a large common variance. Specifically, the setting is $m = 10{,}000, \pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1, \theta_{tv} = 0.5\delta, \theta_1 = 1.5\delta$, with $\delta = 0.1$ and there is a common variance of $\sigma^2 = 5 \times (2\delta)^2$ for all tests.

Supplemental Figure 4.9 Simulation estimates of false discovery rate quantities, including results from the hybrid procedures, across all examined combinations of alternative effect size, numbers of tests and variance settings, with null, trivial, and non-trivial alternative effects. Here, we are examining the setting of $\pi_0 = 0.7, \pi_{tv} = 0.2, \pi_1 = 0.1$, and $\delta = 0.1$. All considered settings of number of tests m (first-level grouping along the horizontal axis), alternative effect size (second-level sub-grouping along the horizontal axis), and variance (along the vertical axis) are provided. The variance labels shown are in terms of scale of the variance compared to the baseline of $\sigma^2 = (2\delta)^2$. That is, the values shown are ω, where $\sigma^2 = \omega \times (2\delta)^2$. In these variance labels, "(R)" denotes the scenarios where each test has an equal probability of observing each of the two variance values, and "(NsAl)" denotes the scenario where all zero null tests have smaller variance, all alternative tests have larger variance, and trivial effects have one of the two variance values with probability 0.5 for each. (a) point pFDR, (b) trivial pFDR, (c) combined pFDR, (d) point FDR, (e) trivial FDR, (f) combined FDR, (g) Pr(R>0), (h) power.


(a) point pFDR

(b) trivial pFDR

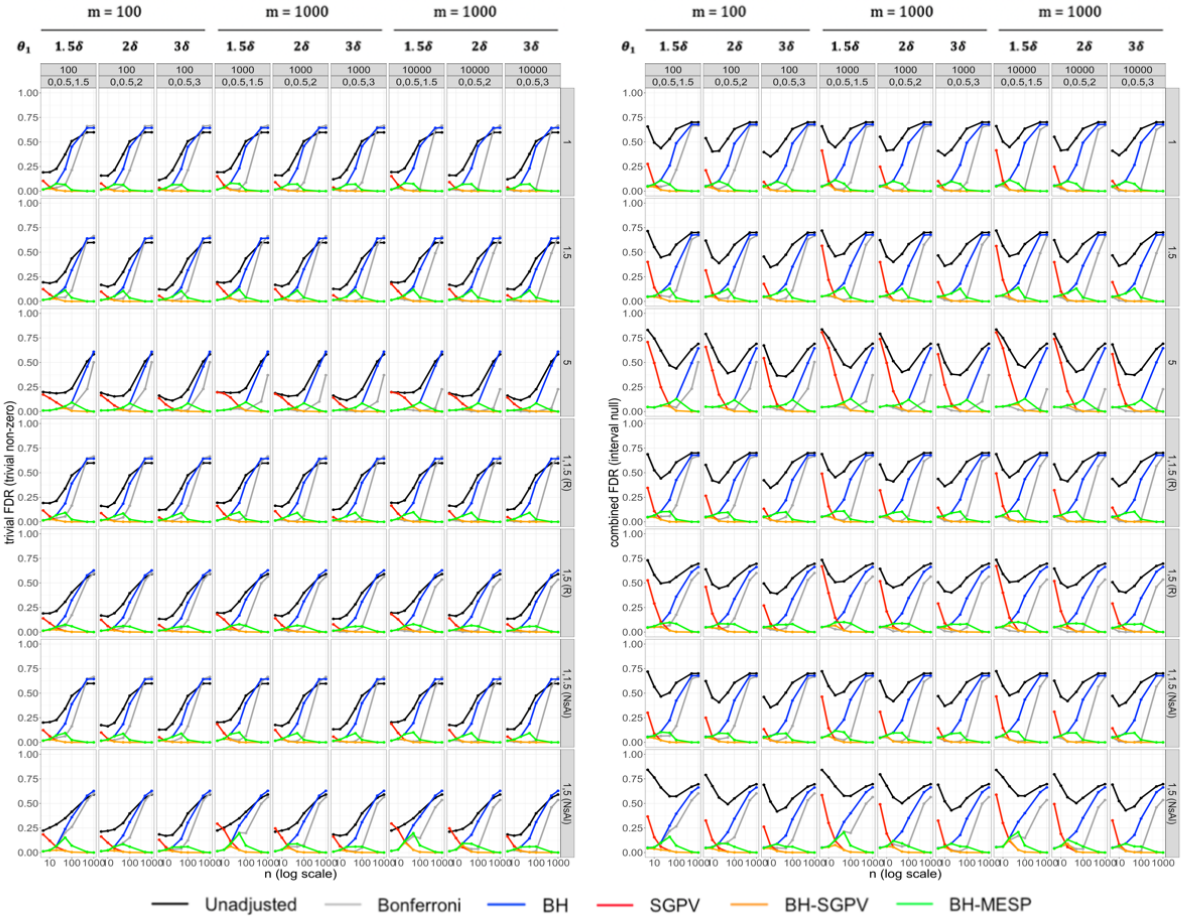


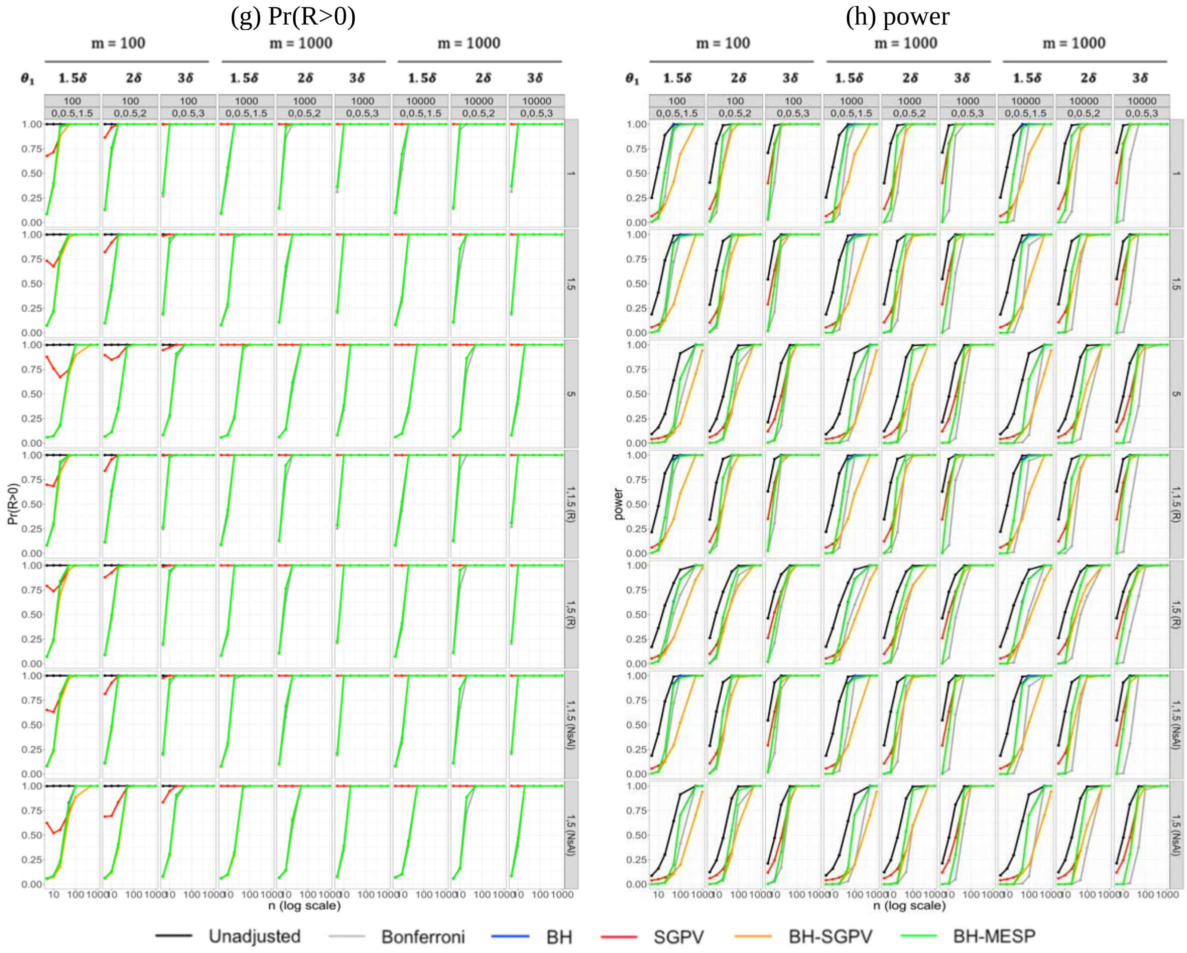
(c) combined pFDR

(d) point FDR
= point pFDR × Pr(R>0)

(e) trivial FDR
= trivial pFDR × Pr(R>0)

(f) combined FDR
= combined pFDR × Pr(R>0)



(g) Pr(R>0)
(h) power
m = 100
m = 1000
m = 1000
θ_1
1.5δ
2δ
3δ
100
0,0.5,1.5
0,0.5,2
0,0.5,3
1000
10000
1
1.5
5
1,1.5 (R)
1.5 (R)
1,1.5 (NsAI)
1.5 (NsAI)
Pr(R>0)
power
0.00
0.25
0.50
0.75
1.00
10
100
1000
n (log scale)
Unadjusted
Bonferroni
BH
SGPV
BH-SGPV
BH-MESP

CHAPTER 5

Conclusion

Statistical inference on a large scale is common in modern times with high throughput methodologies seen in the fields of genetics, imaging, and microbiome studies. While these studies require further consideration of controlling and measuring errors, they also allow for leveraging this multiplicity of information via empirical Bayes estimation. Incorporating an interval null hypothesis – which formally accounts for scientific significance in addition to statistical significance – into the false discovery rate framework requires additional considerations. Overall, this dissertation addresses these topics in various ways.

We first elucidated the foundations of false discovery rates and related quantities. Establishing this unified framework allows for a clearer exposition of the relationship between the various important quantities. Of particular note is the distinction between the FDR as originally defined by Benjamini and Hochberg (1995), and the pFDR, referenced by Benjamin and Hochberg but formally established in (Storey 2003). While they are clearly defined in the foundational FDR literature, in practice, however, these two quantities are often conflated or the distinction between them understated. Strict FDR control may be desired in some scenarios, but the pFDR, which measures the propensity for observed results to be misleading, is the main scientific quantity of interest. Several common approaches for empirical estimation in the setting of classical p-values are examined, illustrating their utility in large-scale inference.

Next, we explicate the calculation and estimation of the positive false discovery and confirmation rate for the second-generation p-value, a metric which incorporates an interval null range denoting null and practically null effects, first proposed in (Blume et al. 2018). We expand on the prior work of (Blume et al. 2018, 2019) in this area, by developing a more general definition of SGPV false discovery quantities which fully addresses the interval null hypothesis. Proposed approaches include marginalization across the null and alternative parameter spaces, with a variety of approaches for defining weighting distributions, for each respective design probability (the probability of observing particular results under each hypothesis). For the most part, we find that the choice of weighting function does not influence large-sample properties, but small-sample estimates may be greatly impacted by these choices. This highlights the need for empirical methods. One empirical approach is proposed, with the estimation of the mixture probability in the denominator of the FDR quantities. However, this does not offer a complete solution, because the respective numerator design probability (null for the FDR, and alternative for the FCR) must still be specified. The specification of the null proportion may be avoided with an upper bound estimate on the FDR quantities; however, it may be an appreciable overestimate of the true quantity in some cases, as we demonstrate with numerical arguments and simulation examples. Therefore, estimation of this quantity, along with the design probabilities, remains an important area for future work prior to practical implementation in real-world studies.

Finally, we take a more comprehensive look at the behavior of the second-generation p-value in large-scale inference, expanding our view to include overall FDR control and the power to reject meaningful effects. Extensive simulations were used to estimate these quantities for the SGPV, in addition to commonplace multiple comparisons adjusted p-value procedures, under a relaxed set of assumptions. This represents an important first step in understanding how the second-generation p-value operates under real-world large-scale inference conditions. When focusing only on exactly null effects, the SGPV and Benjamini and Hochberg procedures are observed to have different strengths, dependent on the underlying setting. Of particular note is that for very small sample sizes, the Benjamini Hochberg procedure provides overall FDR control, but the rate of null rejections (when one more rejections is made, i.e., the pFDR) is frequently very large.

When the importance of rejecting practically null, i.e., trivial effects, is considered, classical p-value methods falsely reject at an increased rate for large sample sizes. We find that the SGPV in general does not control the overall FDR, primarily only falling below α when the pFDR itself is less than α. While this "natural" control of the false discovery rate represents intuitive statistical behavior, strict FDR control may still at times be desired in practice. In such a case, we examine some hybrid methods (one newly proposed, and one adapted from (Goodman et al. 2019)) and find that these can be useful for small sample sizes for controlling the FDR. However, there may be a tradeoff with lower power or a higher propensity for misleading results in some settings. Overall, we have demonstrated some benefits and drawbacks of using the second-generation p-value in large-scale inference. We provide some recommendations for use in practice, although the best choice often depends on unknown factors such as the underlying variance among tests. We conclude with discussion of some other uses of the second-generation p-value in large-scale inference, such as with ranking tests.

REFERENCES

Alishahi, K., Ehyaei, A., & Shojaie, A. (2016). A generalized Benjamini-Hochberg procedure for multivariate hypothesis testing. *ArXiv:1606.02386*.

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., & Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, **39**(1), 1–20.

Aubert, J., Bar-Hen, A., Daudin, J.-J., & Robin, S. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics*, **5**(1), 125.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 405–416.

Benjamini, Y., & Braun, H. (2002). John W. Tukey's contributions to multiple comparisons. *The Annals of Statistics*, **30**(6), 1576–1594.

Benjamini, Y., & 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.

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, **25**(1), 60–83.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**(4), 1165–1188.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence. *Journal of the American Statistical Association*, **82**(397), 112.

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, **21**(17), 2563–2599.

Blume, J. D. (2008). How often likelihood ratios are misleading in sequential trials. *Communications in Statistics – Theory and Methods*, **37**(8), 1193–1206.

Blume, J. D. (2011). Likelihood and its evidential framework. In Bandyopadhyay, P. S. & Forster, M. R. (Eds.), *Handbook of The Philosophy of Science: Philosophy of Statistics* (Vol. 7, pp. 493–511). *North Holland*.

Blume, J. D., Greevy, R. A., Welty, V. F., Smith, J. R., & Dupont, W. D. (2019). An introduction to second-generation p-values. *The American Statistician*, **73**(sup1), 157–167.

Blume, J. D., McGowan, L., Dupont, W. D., & Greevy, R. A. (2018). Second-generation p-values: improved rigor, reproducibility, & transparency in statistical analyses. *PLOS ONE*, **13**(3), e0188299.

Boca, S. M., & Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, e6035.

Broberg, P. (2004). A new estimate of the proportion unchanged genes in a microarray experiment. *Genome Biology*, **5**(5), P10.

Broberg, P. (2005). A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, **6**(1), 199.

Brown, B. W., & Russell, K. (1997). Methods correcting for multiple testing: operating characteristics. *Statistics in Medicine*, **16**(22), 2511–2528.

Cabras, S. (2010). A note on multiple testing for composite null hypotheses. *Journal of Statistical Planning and Inference*, **140**(3), 659–666.

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *American Journal of Human Genetics*, **86**(1), 6–22.

Chen, X., Robinson, D. G., & Storey, J. D. (2019). The functional false discovery rate with applications to genomics. *Biostatistics*, **22**(1), 68–81.

Chi, Z. (2010). Multiple hypothesis testing on composite nulls using constrained p-values. *Electronic Journal of Statistics*, **4**, 271–299.

Chong, E. Y., Huang, Y., Wu, H., Ghasemzadeh, N., Uppal, K., Quyyumi, A. A., Jones, D. P., & Yu, T. (2015). Local false discovery rate estimation using feature reliability in LC/MS metabolomics data. *Scientific Reports*, **5**(1), 17221.

Döhler, S., & Roquain, E. (2020). Controlling the false discovery exceedance for heterogeneous tests. *Electronic Journal of Statistics*, **14**(2), 4244–4272.

Doncheva, N. T., Kacprowski, T., & Albrecht, M. (2012). Recent approaches to the prioritization of candidate disease genes. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **4**(5), 429–442.

Dudoit, S., van der Laan, M. J., & Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology*, Article 13.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, **56**(293), 52–64.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**(465), 96–104.

Efron, B. (2007a). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, **102**(477), 93–103.

Efron, B. (2007b). Size, power and false discovery rates. *The Annals of Statistics*, **35**(4), 1351–1377.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, **23**(1), 1–22.

Efron, B. (2010a). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association*, **105**(491), 1042–1055.

Efron, B. (2010b). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press*.

Efron, B., Storey, J. D., & Tibshirani, R. (2001). Microarrays, empirical Bayes methods, and false discovery rates. Technical Report. Stanford University

Efron, B., & Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *The Annals of Statistics*, **24**(6), 2431–2461.

Efron, B., & Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**(1), 70–86.

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

Genovese, C. R., & Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 499–517.

Genovese, C. R., & Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, **32**(3), 1035–1061.

Genovese, C. R., & Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, **101**(476), 1408–1417.

Goeman, J. J., & Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, **26**(4), 584–597.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Goodman, W. M., Spruill, S. E., & Komaroff, E. (2019). A proposed hybrid effect size plus p-value criterion: empirical evidence supporting its use. *The American Statistician*, **73**(sup1), 168–185.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, **31**(4), 337–350.

Heesen, P., & Janssen, A. (2015). Inequalities for the false discovery rate (FDR) under dependence. *Electronic Journal of Statistics*, **9**(1), 679–716.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**(4), 800–802.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. *John Wiley*, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**(2), 65–70.

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486**(7402), 207–214.

Javanmard, A., & Montanari, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, **46**(2), 526–554.

Jeffreys, H. (1935). Some Tests of Significance, Treated by the Theory of Probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, **31**(2), 203–222.

Jiang, H., & Doerge, R. W. (2008). Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer Informatics*, **6**, 117693510800600000.

Kang, J. (2020). Comparison of methods for the proportion of true null hypotheses in microarray studies. *Communications for Statistical Applications and Methods*, **27**(1), 141–148.

Kass, R. E. & Raftery, A. E. (1995). Bayes factor. *Journal of the American Statistical Association*, **90**(430), 773–795.

Korn, E. L., Troendle, J. F., McShane, L. M., & Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, **124**(2), 379–398.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., & Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, **20**, 118.

van der Laan, M. J., Dudoit, S., & Pollard, K. S. (2004). Multiple Testing. Part III. Procedures for Control of the Generalized Family-Wise Error Rate and Proportion of False Positives. U.C. Berkeley Division of Biostatistics Working Paper Series 1140, Berkeley Electronic Press.

Langaas, M., Lindqvist, B., & Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(4), 555–572.

Lehmann, E., & Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, **33**(3), 1138–1154.

Lei, L., & Fithian, W. (2018). AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**(4), 649–679.

Liao, J. G., Lin, Y., Selvanayagam, Z. E., & Shih, W. (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, **20**(16), 2694–2701.

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

Meinshausen, N. (2006). False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, **33**(2), 227–237.

Miller, R. G. (1981). *Simultaneous statistical inference (2nd ed.)*. *Springer-Verlag*, New York.

Murray, M. H., & Blume, J. D. (2021). FDRestimation: flexible false discovery rate computation in R. *F1000Research*, **10**, 441.

Nettleton, D., Hwang, G. J., Caldo, R. A., & Wise, R. P. (2005). Estimating the number of true null hypotheses from a histogram of p values. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**(3), 337.

Pawitan, Y., Murthy, K. R., Michiels, S., & Ploner, A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, **21**(20), 3865–3872.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, **316**, 1236.

Ploner, A., Calza, S., Gusnanto, A., & Pawitan, Y. (2006). Multidimensional local false discovery rate for microarray studies. *Bioinformatics*, **22**(5), 556–565.

Pounds, S., & Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics*, **20**(11), 1737–1745.

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

Ramdas, A., Zrnic, T., Wainwright, M., & Jordan, M. (2018). SAFFRON: an adaptive algorithm for online control of the false discovery rate. *Proceedings of the 35th International Conference on Machine Learning*, **80**, 4286–4294.

Robertson, D. S., & Wason, J. M. (2018). Online control of the false discovery rate in biomedical research. *ArXiv:1809.07292*.

Romano, J. P., & Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, **35**(4), 1378–1408.

Royall, R. M. (1997). *Statistical evidence: a likelihood paradigm*. *Chapman and Hall/CRC*.

Ryan, T. A. (1959). Multiple comparison in psychological research. *Psychological Bulletin*, **56**(1), 26–47.

Ryan, T. P., & Woodall, W. H. (2005). The most-cited statistical papers. *Journal of Applied Statistics,* **32**(5), 461–474.

Sarkar, S. K. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *The Annals of Statistics*, **35**(6), 2405–2420.

Schaid, D. J., & Chang, B. (2005). Description of the international consortium for prostate cancer genetics, and failure to replicate linkage of hereditary prostate cancer to 20q13. *The Prostate*, **63**(3), 276–290.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657–680.

Schweder, T., & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, **69**(3), 493–502.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, **46**(1), 561–584.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62**(318), 626.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**(3), 751–754.

Sorić, B. (1989). Statistical 'discoveries' and effect-size estimation. *Journal of the American Statistical Association*, **84**(406), 608–610.

Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics,* **43**(2), 398–411.

Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics (Oxford, England),* **18**(2), 275–294.

Storey, J. D. (2001a). A new approach to false discovery rates and multiple hypothesis testing. Technical Report. Stanford University.

Storey, J. D. (2001b). The false discovery rate: a Bayesian interpretation and the q-value. Technical Report. Stanford University.

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

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, **31**(6), 2013–2035.

Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(1), 187–205.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**(16), 9440–9445.

Sun, W., & McLain, A. C. (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association*, **107**(498), 673–687.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, **20**(8), 467–484.

Tang, Y., Ghosal, S., & Roy, A. (2007). Nonparametric Bayesian estimation of positive false discovery rates. *Biometrics*, **63**(4), 1126–1134.

Tsai, C., Hsueh, H., & Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, **59**(4), 1071–1081.

Tukey, J. W. (1953). The problem of multiple comparisons. In *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948–1983* (pp. 1–300). *Chapman and Hall*.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, **6**(1), 100–116.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(9), 5116–5121. Erratum in: *Proceedings of the National Academy of Sciences of the United States of America*, **98**(18), 10515.

Wacholder, S., Chanock, S., Garcia-Closas, M., El-ghormli, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidmiology studies. *Journal of the National Cancer Institute*, **96**(6), 434–442.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, **70**(2), 129–133.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority (2nd ed.)*. *Chapman and Hall/CRC*.

Xiao, Y., Hsiao, T. H., Suresh, U., Chen, H. I., Wu, X., Wolf, S. E., & Chen, Y. (2014). A novel significance score for gene selection and ranking. *Bioinformatic*s, **30**(6), 801–807.

Zhang, M. J., Xia, F., & Zou, J. (2019). Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature Communications*, **10**(1), 3433.

Zhou, Y. H., Brooks, P., & Wang, X. (2018). A two-stage hidden Markov model design for biomarker detection, with application to microbiome research. *Statistics in Biosciences*, **10**(1), 41–58.