On second-generation *p*-values for equivalence testing and study planning, and flexible false discovery rate computation for classical *p*-values

By

Megan Hollister Murray

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 12th, 2022

Nashville, Tennessee

Approved:

Dandan Liu, Ph.D.

Jeffrey D. Blume, Ph.D.

Thomas G. Stewart, Ph.D.

Melinda C. Aldrich, MPH, Ph.D.

This dissertation is dedicated to my loving husband, Paul, and to my supportive parents, Chris and Kathy.

I cannot express in words how thankful I am for each of you.

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Jeffrey. You have given me so much of your time over the last 4 years! You saw potential in me before I had done any research at Vanderbilt. I will forever be grateful for your guidance in this process.

I would like to thank each of my committee members, Dandan, Tom, and Melinda. You cared about my research ideas and wanted to help me to succeed. Your feedback greatly improved my arguments.

I would like to thank my Hollister family, Chris, Kathy, Daniel, A.J., and Abby. Mom and Dad, you have always encouraged me to reach my highest academic potential. You believed in me when I didn't. Thank you for everything!

I would like to thank my Murray family, Steve, Neva, Phil, Paul, Tiff, and Kali. You have accepted me into the family and supported me along every step in this process.

I would like to thank my cohort and other students in the Vanderbilt Biostatistics department. This program had its challenges and I couldn't have gotten through it without you all.

Lastly, I would like to thank all other friends and professors that have helped me on this journey!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Abstract

Second-generation p-values (SGPVs) have been proposed and discussed in the literature as an alternative inferential statistic to indicate when the data support the null or alternative hypothesis, or when the data are inconclusive. As they are inferentially non-denominational, SGPVs can be used with frequentists, likelihood or Bayesian methods. This dissertation compares the behavior of SGPVs to classical equivalence tests, explores the operational characteristics of SGPV for study planning, and presents a tool for flexible false discovery rate computations for classical p-values.

First, we derive the mathematical relationship between SGPVs and traditional equivalence testing. We provide a conceptual framework for comparing the two approaches and conclude that the flexibility of the second-generation p-value framework offers notable advantages including ease of use, clear interpretation, and improved statistical properties. Second, we investigate different ways to specify the "interval null hypothesis" or indifference zone that is critical to the second-generation p-value (SGPV). We propose allowing the indifference zone to shrink as the sample size grows as a way of mitigating collaborator uncertainty about the indifference zone. Shrinking the indifference zone can balance the power and errors in a classical sense, but it is only practically useful in certain settings. Third, we introduce a new user-friendly R package for estimating FDRs and computing adjusted p-values for FDR control. A key contribution of this package is that it distinguishes between these two quantities while also offering several refined algorithms for estimating them. In conclusion, this work identifies the most flexible and easy-to-use method for establishing equivalence, proposes a new concept to adjust for collaborator uncertainty in SGPV methods, and creates a new user-friendly package for computing FDRs and adjusted p-values. The motivation behind these contributions is that the reporting of second-generation p-values and false discovery rates greatly improves the dissemination, transparency, and accessibility of statistical analyses.

1.2 Summary of Work

A statistician's role is to choose an analysis method appropriate for dataset and the collaborator's hypothesis and then be able to explain this method and the results to the collaborator. Complexities in the data and inherent communication gaps within research teams can make this a difficult process. We want to explore ways to improve transparency and accessibility of statisticians' analyses. This is a very broad goal so in this dissertation we focus on two specific statistical areas, interval null hypotheses and false discovery rates.

These two methods have been discussed in academic literature since the 1980s and have been widely used in practice.

Traditionally, an analysis compares data to a single value statistic like a mean. For example, the reported average height of American women is 5 foot 4 inches (Fryar et al., 2018). When a random sample's average height is compared to the reported national average this is called a point null hypothesis. Almost never do we find absolute equivalence between the data and the point null hypothesis. Therefore sometimes it is beneficial to change this point null hypothesis so that when the data is "close enough" we still conclude equivalence. We call this a region of practical equivalence or an interval null.

Two methods that use the interval null framework are equivalence tests and second-generation p-values. The Journal of Pharmacokinetics and Biopharmaceutics first introduced equivalence tests as a way to establish bioequivalence in 1984 (Hauck and Anderson, 1984). There are many popular versions of equivalence tests but we will focus on the frequentist Two-One Sided t-Tests (TOST) (Schuirmann, 1987). The TOST reports when a data driven confidence interval is contained within the interval null. In 2018 second-generation p-values (SGPVs) were proposed to indicate when the data support the null or alternative hypothesis, or when the data are inconclusive (Blume et al., 2018). SGPVs are inferentially non-denominational, as they use any data driven interval; a frequentist confidence interval, a likelihood support interval, or a Bayesian credible interval.

In the first paper we identify the most flexible and easy-to-use method for establishing statistical equivalence with an interval null hypothesis. A thorough investigation is done to compare the behavior of second-generation p-values to classical equivalence tests. This includes the derivation of the mathematical relationship between SGPVs and traditional equivalence testing. We also compare the large sample size behavior, Type 1 error, and power between these methods. We conclude that the flexibility of the second-generation p-value framework offers notable advantages including ease of use, clear interpretation, and improved statistical properties. This detailed comparison clarifies the properties and difference between SGPVs and equivalence tests. This paper will help statisticians to decide what method to use to establish equivalence.

In the second paper we propose a technique to address the complexities of study planning with a collaborator. The collaborator can be confident, uncertain, or unable to make a hypothesis; or in the SGPV context, to identify an indifference zone. We investigate how different ways of specifying the interval null hypothesis in second-generation p-value (SGPV) analyses change the statistical properties of the test. Then after examining these results we propose allowing the indifference zone to shrink as the sample size grows as a way of mitigating collaborator uncertainty about the indifference zone. Shrinking the indifference zone can balance the power and errors in a classical sense. However, we only suggest shrinking of the indifference zone in certain settings when the collaborator is uncertain in the indifference zone. When the collaborator

is confident in a hypothesized indifference zone then the statistician should use the collaborator's specified interval. Our recommendations given for different levels of collaborator uncertainty allow the statistician to obtain the most accurate test results and conclusions.

Another commonly discussed area of statistics is multiple comparisons or multiple testing in large-scale datasets. Classical p-values can be adjusted to maintain control of the family-wise error rate (FWER) (Tukey, 1953). However in large-scale inference this FWER control can come at the cost of Type II Error rate inflation. Instead it has become common practice to control the false discovery rate (FDR) instead of the FWER in these settings because its Type II Error rate inflation is much less severe (Benjamini and Hochberg, 1995). The FDR is the propensity for an observed result to be mistaken and FDR estimates should accompany observed results to help the user contextualize the impact of findings. It is also important to note that in practice methods for controlling the FDR are often confused with the methods used to provide an estimate of the FDR for a particular result.

In our third and final paper we present a tool for flexible and transparent false discovery rate computation for classical p-values. This new user-friendly R package is titled "FDRestimation" and can be found on CRAN. It can be used to estimate FDRs and compute adjusted p-values for FDR control. This package clearly distinguishes between these two quantities while also offering several adjustment methods like Benjamini-Hochberg, Benjamini-Yekutieli, Bonferroni, Sidak, and others. In addition, this package can be used to estimate the null proportion, which is a value used in FDR computation, for a given dataset using many previously proposed methods. In this paper we also propose a new method for null proportion estimation called "Last Histogram Height". In conclusion we strongly encourage more transparent reporting of false discovery rates for observed findings.

To summarize, in this dissertation we identify the most flexible and easy-to-use method for establishing statistical equivalence with an interval null hypothesis, we propose a technique to address the complexities of study planning for SGPVs with collaborator uncertainty, and we present a tool for flexible and transparent false discovery rate computation for classical p-values. These contributions greatly improve statistical practice and the transparency of results being communicated with collaborators.

CHAPTER 2

Establishing Statistical Equivalence: A Comparison of Second-generation *p*-values and Equivalence Tests

2.1 Abstract

Equivalence testing has been well-established and used widely in psychology and clinical trial research since 1980s. More recently Blume and collaborators proposed second-generation *p*-values (SGPVs) to address the known flaws of the classical *p*-value. The framework employed in both of these methods uses an interval null or indifference zone, which indicates a region of practical equivalence, instead of the traditional use of a single point null. Given that statisticians choose which the methods they apply, a thorough comparison of the two approaches is warranted. In this paper we establish the conceptual and technical relationship between these two methods. We have derived the direct mathematical connection between the reported *p*-values for SGPV and equivalence tests. This derivation has not been shown in previous literature and we feel it is essential in understanding the difference in the two methods. We have also compared large sample size behavior, Type 1 error, and power between these methods. We conclude that while connections between the methods do exist, the flexibility of the second-generation *p*-value framework offers notable advantages including ease of use, clear interpretation, and desirable statistical properties.

Keywords

Second-generation *p*-values, equivalence tests, Two One Sided Tests, interval null, evidence, bioequivalent

2.2 Introduction

In a typical work flow, statisticians analyze a dataset to see if it matches some assumed or expected behavior. This analysis commonly characterizes the data with a single value, such as mean value or odds ratio. In practice, absolute equivalence between the data and a single value or "point null hypothesis" is rarely found. Some practitioners may think an estimated parameter is "close enough" to the point null in order to form their conclusions. This region of practical equivalence must be identified by a collaborator before looking at the data. Two methods that use this framework are equivalence tests and second-generation p-values.

Equivalence tests were first introduced in the Journal of Pharmacokinetics and Biopharmaceutics in 1984 (Hauck and Anderson, 1984). The main purpose of these tests was to establish bioequivalence; for example when a pharmaceutical company is testing for drug approval they compare the new drug's performance to an older approved drug's performance (Kirkwood and Westlake, 1981; Rogers et al., 1993). There are many different proposed versions of equivalence tests. The most well-known tests include the Kirkwood and Westlake's test of mean equivalence using a confidence interval for the difference in two means (Kirkwood and Westlake, 1981), the frequentist Two-One Sided t-Tests (TOST) (Schuirmann, 1987), the test of Anderson and Hauck (Hauck and Anderson, 1984), the test of Berger and Hsu(Berger and Hsu, 1996), and even the Bayesian Region of Practical Equivalence (ROPE) (Kruschke, 2018). In this paper we will focus our comparison between SGPV and equivalence tests specifically on the TOST method. Some literature has criticized equivalence tests for their poor statistical behavior and cautioned users to avoid it (Berger and Hsu, 1996; Ennis and Ennis, 2010; Perlman and Wu, 1999; Meyners, 2012). In this paper we will address and clarify some of these concerns.

The second-generation p-value (SGPV) method was proposed to use a region of practical equivalence instead of a point null hypothesis. The SGPV method was first introduced in 2018 as a way to measure "the proportion of data-supported hypotheses that are null" (Blume et al., 2018, 2019). This proportion is included in the reported p-value and final inference outcome as a measure of overlap between the data interval and the indifference zone. The reported p-values from the SGPV method, contained in the unit interval, indicate when the data are consistent with an alternative hypothesis (~ 0), when the data are inconclusive (~ 0.5), or when the data are consistent with a null hypothesis (~ 1). Even though this is a fairly new method it has generated great interest and has a variety of uses.

2.2.1 Inference Outcomes

Both SGPVs and equivalence tests compare a data driven interval to a pre-specified scientifically relevant indifference zone or equivalence range. It is of interest to know exactly how these two methods are related to one another.

		SGPV Outcomes		
		Consistent with the alternative (SGPV near 0)	**Inconclusive** (SGPV near ½)	**Consistent with the null** (SGPV near 1)
Equivalence Tests Outcomes	**Consistent with the alternative** (*p*-value is unable to indicate this)	Not applicable **A**	Not applicable **B**	Not applicable **C**
	Inconclusive (*p*-value is non-significant)	Can occur **D**	Can occur **E**	Never occurs **F**
	Consistent with the null (*p*-value is significant)	Never occurs **H**	Can occur in small samples **I**	Can occur **J**

Table 2.1: Inference outcome comparison table between equivalence tests and SGPVs. Equivalence testing yields two evidential outcomes while SGPVs yield three outcomes. This is shown with the first row, "Consistent with the alternative", not being applicable for equivalence *p*-values. This is a critical distinction in favor of SGPVs interpretation.

To begin this comparison, we must answer the question whether equivalence tests provide the same information as SGPVs. The answer to this question is visualized in Table 2.1. For inference purposes we assume the null is that the data agrees or is practically equivalent with the indifference zone. The three outcomes include data being consistent with the null, the data being inconclusive, and the data being consistent with the alternative. In Table 2.1 we see that it is possible for the SGPV to report all three outcomes, whereas the equivalence tests can only report two of these outcomes: data being consistent with the null or inconclusive. For the user this means when using an equivalence test, for example the TOST, they can never conclude or report that the data is consistent with the alternative. The user can also never conclude non-equivalence or a difference with the interval null.

Next, to help visualize this comparison in the *p*-value space Figure 2.1 shows a simple simulated example. Data was simulated 500 times under the true point null, N(0,1), with sample size of n=6 and tested against the indifference zone $[\theta^-, \theta^+]$=[-0.375,0.375]. SGPVs were computed using the "sgpv" R package Welty et al. (2020) and TOST reported *p*-values were computed using the "TOSTER" R package (Lakens and Caldwell, 2022). In Figure 2.1 the majority of *p*-values occur in three of the zones in this six-zone grid.

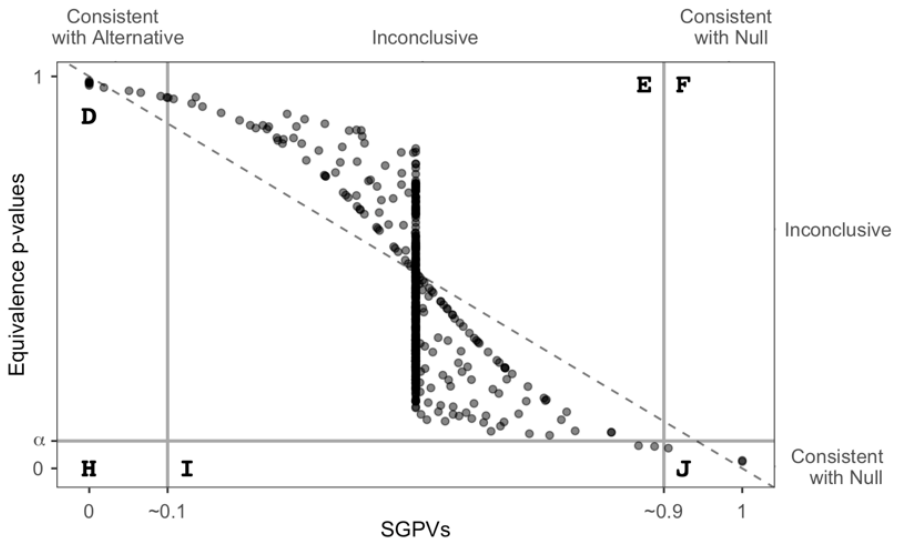

Figure 2.1: Graph of simulated SGPVs versus reported TOST p-values. In this example there are 500 iterations of sample size of n=6 for random data generated under the point null, N(0,1), and tested against the indifference zone $[\theta^-, \theta^+]$=[-0.375,0.375].

These most common zones include the zone labelled "D" in which is SGPV consistent with the alternative and equivalence tests inconclusive, the "E" zone which both SGPV and equivalence tests are inconclusive, and the "J" zone which is both SGPV and equivalence tests consistent with the null. In zone "D" the SGPV and TOST tests differ in their inference outcomes. Having an additional inference outcome of data being consistent with the alternative is a major conceptual advantage for SGPV over equivalence tests. This differentiation in number of possible outcomes makes a level comparison theoretically impossible. We continue on with the comparison even with this inconsistency to ensure thorough investigation before presenting our final opinions and recommendations.

2.3 Background/Methods

2.3.1 Setup and Notation

To clarify notation the user is interested in learning about the difference between two populations, this difference is represented by parameter θ. For example, θ is often the difference in population means $\theta = \mu_1 - \mu_2$ or relative risk $\theta = p_1/p_2$. The user starts by collecting data $x = (x_1, ..., x_n)$, from which an uncertainty interval for the population difference, θ, is constructed, this will be denoted as $I_x = (I_x^-, I_x^+)$. For example, I_x might be a confidence interval, credible interval or likelihood support interval for the difference in population means. For the ease of exposition, in this paper we will assume that I_x is a 95% confidence interval for the difference in means. The null hypothesis can be written as $H_0 : \theta = \theta_0$. For example, when considering equivalence

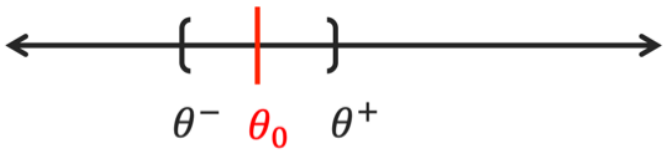

Figure 2.2: Line graph showing the indifference zone, equivalence range or interval null.

between the two populations, for the difference in population means we would typically set $\theta_0 = 0$ or for the relative risk $\theta_0 = 1$. This is called a "strong" null hypothesis because it states that the population parameters must be exactly equal to be considered null.

However, it is often the case that we only care if the populations means are "close enough", in the sense that θ is near θ_0, which we write as $\theta^- \leq \theta \leq \theta^+$. Here $[\theta^-, \theta^+]$ is an indifference region or a region of practical equivalence around the point null of θ_0. In this context if the difference in the data populations is within this zone, we treat the populations as if they were equal. Thus, what we really have is an interval null hypothesis that states that $H_0 : \theta^- \leq \theta \leq \theta^+$ and we want to report or measure how consistent the data are with this null interval. Figure 2.2 illustrates this interval null or indifference zone. The area outside $[\theta^-, \theta^+]$ is considered the alternative hypothesis or the range of non-equivalence (difference).

An important wrinkle for statistical procedures is that when hypothesis testing investigators often want to measure the evidence for the null hypothesis. Of course, this is a hypothesis testing no-no (Fisher, 1959). It is strictly not allowed, as large p-values indicate that the data are inconclusive, and small p-values indicate that the data different from the null or are consistent with the alternative. p-values can never be interpreted as representing evidence for the null hypothesis. Equivalence testing methods came up with a creative solution to this problem. They flip the null and implicit alternative hypotheses. As such, we now have $H_0 : (\theta < \theta^-$ or $\theta > \theta^+)$ so that the new null is the old implicit alternative and the new implicit alternative is the old null; $H_1 : (\theta \geq \theta^-$ and $\theta \leq \theta^+)$, or more clearly $H_1 : \theta^- \leq \theta \leq \theta^+$. Hence rejecting an equivalence test because of a small or significant p-value now means we have evidence that data are equivalent to the indifference zone. One specific equivalence test that flips the null and the implicit alternative is the TOST procedure. However, there are other methods, such as the SGPV, available where the null hypothesis is not flipped but can still conclude evidence for data equivalence.

2.3.2 Second-generation *p*-values

The aim of the second-generation p-value method (SGPV) is to measure the fraction of data-supported effect sizes that are within the indifference zone (Blume et al., 2018, 2019). To do this the uncertainty interval, I_x, is compared to the indifference zone, $[\theta^-, \theta^+]$ by computing the amount of overlap between these intervals.

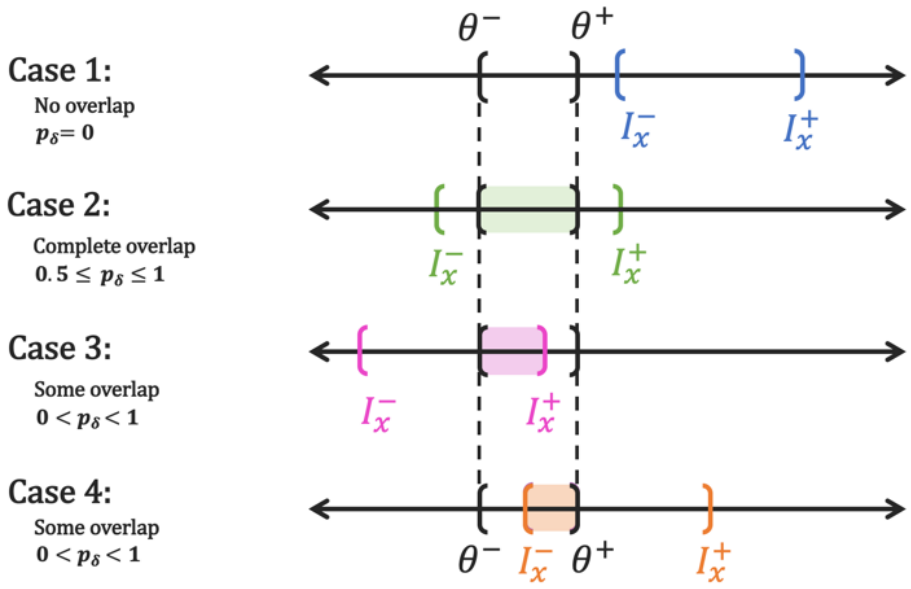

Figure 2.3: Illustration of the four cases of overlap between the indifference zone and the uncertainty interval and their corresponding SGPVs.

Equation 2.1 shows the reported second-generation p-value formula where $H_0 = [\theta^-, \theta^+]$ and $(I_x \cap H_0)$ is the intersection or overlap of the two intervals. The function denoted with $f(x) = |x|$ returns the length of the interval or section of overlap.

$$p_\delta = \frac{|I_x \cap H_0|}{|I_x|} \times \max\left\{\frac{|I_x|}{2|H_0|}, 1\right\} \tag{2.1}$$

In Equation 2.1 the fraction of length of overlap to length of indifference zone is multiplied by a correction factor. This correction factor applies when $|I_x| > 2|H_0|$ meaning the uncertainty interval is very wide in comparison to the null interval. When the data are "sufficiently precise" or the uncertainty interval length is $|I_x| \leq 2|H_0|$, the SGPV is simply the fraction of overlap, $p_\delta = \frac{|I_x \cap H_0|}{|I_x|}$ (Blume et al., 2018).

There are three inference interpretations for p_δ when data are inconclusive we have $0 < p_\delta < 1$, when data are consistent with the alternative we have $p_\delta = 0$, and when data are consistent with the null or indifference zone we have $p_\delta = 1$. For SGPV there are four possible cases for overlap between $I_x = (I_x^-, I_x^+)$and $H_0 = [\theta^-, \theta^+]$, which is illustrated in Figure 2.3. These include no overlap, complete overlap, partial overlap from the left, and partial overlap from the right.

As shown the indifference zone, $H_0 = [\theta^-, \theta^+]$, remains fixed while different data uncertainty intervals are tested; this is how the method should be pictured in practice. For Case 1, or no overlap, the interpretation is that the data are always consistent with the alternative. For Case 3 and 4 the interpretation is that the

data are inconclusive, $(0 < p_\delta < 1)$. For Case 2 the interpretation is that either the data are inconclusive, $(0 < p_\delta < 1)$, or that the data are consistent with the null, $(p_\delta = 1)$. Complete overlap can be pictured in 2 different ways, either $H_0 \subset I_x$ or $I_x \subset H_0$. In Figure 2.3 we see the first way, or Case 2a, where $0 < p_\delta < 1$. For Case 2b where $I_x \subset H_0$ the SGPV is always $p_\delta = 1$, or the data are consistent with the null. These cases help to understand what p_δ measures and reports in practice.

2.3.3 Equivalence Tests

Equivalence tests were created to establish a similarity or practical equivalence between new data and an established dataset or equivalence range (Kirkwood and Westlake, 1981; Mandallaz and Mau, 1981; Schuirmann, 1987; Seaman and Serlin, 1998). The most popular method introduced in 1987 is the Two One-Sided Tests (TOST) procedure (Schuirmann, 1987). As mentioned before, the TOST procedure flips the null and alternative hypotheses, and tests if the data $I_x = (I_x^-, I_x^+)$are outside the equivalence range $[\theta^-, \theta^+]$ (Schuirmann, 1987). This is the opposite of what the null hypothesis is testing in the SGPV method. It is flipped so that the user can statistically conclude bioequivalence when the procedure rejects. Because of this flip, the SGPV the reported p-values, p_δ, should be compared to 1 minus the reported TOST p-value, $1 - p_T$. This also means when plotted against one another the reported p-value the line of equality will not be $y = x$ but instead $y = 1 - x$ for the unit interval $x, y \in [0, 1]$. This concept will be clear in the graphs shown below.

One thing to note here is that there is no measure of overlap or length of the intervals included in this procedure. Instead of using a measure of overlap, it must test the if the data is outside the indifference zone twice with two different one-sided tests. First the data is tested against the lower bound of the indifference zone $(H_{0_1} : \theta < \theta^-)$ and then against the upper bound of the indifference zone $(H_{0_2} : \theta > \theta^+)$. The tests are ordinary, one-sided, α-level t-tests. The composite null hypothesis for the TOST is that $H_0 : (\theta < \theta^- or \theta > \theta^+)$. The conclusion if both one-sided tests reject is that we reject the null (H_0) and accept the alternative $H_1 : (\theta \geq \theta^- and \theta \leq \theta^+)$ or $H_1 : \theta \in [\theta^-, \theta^+]$. We then conclude that our evidence is contained in the equivalence range,$I_x = (I_x^-, I_x^+) \subset [\theta^-, \theta^+]$; for example in standard practice this means the two populations are equivalent (Schuirmann, 1987). The reported p-value, p_T, associated with the TOST procedure is the p-value of largest magnitude from the two one-sided tests, as seen in Equation 2.2.

$$p_T = \max\{p_{T_1}, p_{T_2}\} \tag{2.2}$$

As mentioned above there are only two inference interpretations for this method because the reported TOST p-value has the same interpretation as a traditional p-value. Either both tests reject at some α signifi-

cance level and we conclude the data is consistent with the equivalence range or we don't reject both tests and conclude the data is inconclusive. When one of the tests rejects this means one side of the data uncertainty interval is inside the equivalence range, or partial overlap. However, if only one test rejects the method still concludes inconclusive and reports the larger p-value. The information of partial overlap with the indifference zone is never reported to the user. The user has no reason to believe their dataset has any overlap with the indifference zone when seeing just the reported p-value. This reported conclusion is highly misleading.

It is important to note as originally introduced by Schuirmann if each of the one-sided tests are testing for significance at an $\alpha - level$ then the conclusion is that an uncertainty interval, I_x, of $(1-2\alpha)\%$ is contained within the equivalence range (Schuirmann, 1987). For example, if each test is testing at an 5% level then we conclude that the 90% uncertainty interval is contained within the equivalence range. Therefore, when we compare the TOST and the SGPV using the same α (5%) for the tests we would draw conclusions about different uncertainty intervals. The TOST is concluding a smaller interval, I_x, of $(1-2\alpha)\%$ or 90%, is within the indifference zone compared to SGPV, I_x, of $(1-\alpha)\%$ or 95%. These are very different statements that affect the user's choice and interpretation of α. To be clear in this paper we will use the same α in all tests for the following examples.

2.3.4 Relationship between SGPV and TOST

To see the extent of the relation, we simulated SGPV and TOST p-values and plotted them against one another. This example has 500 iterations of data generated under the true point null, which is θ_0=0, for the mean using a standard normal distribution with sample size n=6. Here the indifference zone,$[\theta^-, \theta^+]$, is set to [-0.375,0.375], which is the middle $\approx$ 30% of a standard normal distribution. The points have different shapes and shading to distinguish between cases of overlap for the indifference zone and the uncertainty interval. Here the uncertainty interval I_x is a 95% confidence interval for the mean from a standard t-test.

In Figure 2.4 we see that TOST reported p-values and SGPVs are not one to one especially in this scenario with a small sample size, but that there is some interesting behavior occurring. If we were to split the graph into quadrants using x=0.5 and y=0.5, quadrants 2 and 4 have lots of data points and quadrants 1 and 3 don't. To when $0 \leq p_\delta \leq 0.5$ then the TOST is restricted to $0.5 \leq p_T \leq 1$ and when $0.5 \leq p_\delta \leq 1$ then the TOST reported p-value is restricted to $0 \leq p_T \leq 0.5$. Also notice that the cases have patterns or trends. For example, Case 1, no overlap, is clustered around $p_\delta = 0$ and $p_T = 1$. For Case 2 points are stuck on the $p_\delta = 0.5$ line or in the bottom right hand quadrant. The data in Cases 3 and 4 generally follow a curved line that can identify. Because of these observed patterns we have derived the technical or mathematical connection between SGPV and TOST reported p-values in the next section.

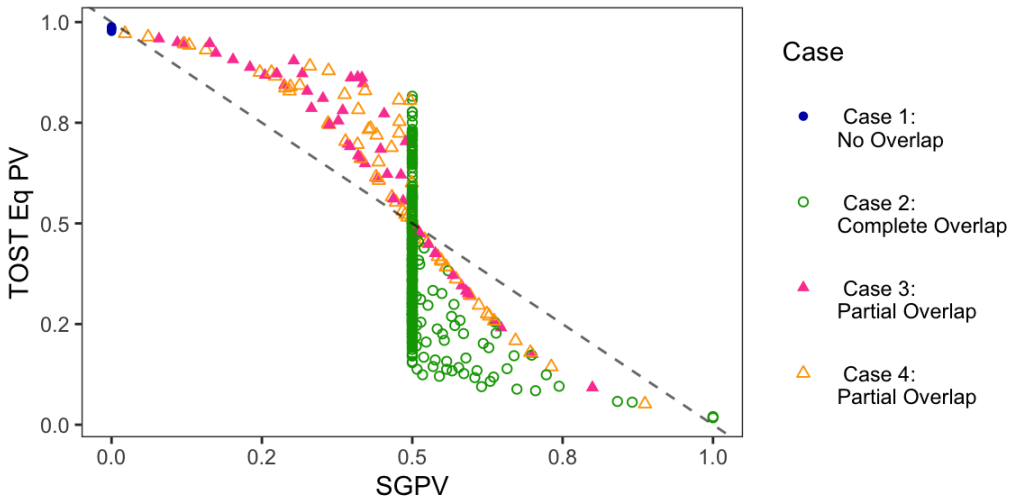

Figure 2.4: Graph of same simulated data as shown in Figure 2.1 but here highlighting differences in cases of overlap. The four cases of overlap are represented by different shapes and shading of points.

2.4 Technical Derivation

In order to truly understand the connection between the TOST and the SGPV reported p-values a mathematical link must be established. To start this derivation, we look at the second-generation p-value formula in Equation 2.1. As a reminder the data derived uncertainty interval is denoted as $I_x = (I_x^-, I_x^+)$ and the prespecified indifference zone is denoted as $H_0 = [\theta^-, \theta^+]$. Here we can replace all of the length computations with exact differences seen in Equation 2.4. However, the "overlap length" in the numerator must remain because the exact difference varies with different cases of overlap. We can see this more clearly by splitting Figure 2.4 up into separate cases of overlap, like in Figure 2.5.

$$p_\delta = \frac{|I_x \cap H_0|}{|I_x|} \times \max\left\{\frac{|I_x|}{2|H_0|}, 1\right\} \tag{2.3}$$

$$= \frac{\text{overlap length}}{(I_x^+ - I_x^-)} \times \max\left\{\frac{(I_x^+ - I_x^-)}{2(\theta^+ - \theta^-)}, 1\right\} \tag{2.4}$$

2.4.1 Case 1 (no overlap)

The first case of overlap, when there is no overlap between the uncertainty interval and indifference zone, can be seen on the number line in Figure 2.6. Here overlap length=0 and Equation 2.4 becomes $p_\delta = 0$ no matter what the reported TOST p-value, p_T, is. For the TOST test if there is no overlap then one of the one-sided tests will return p_T very large and potentially close to 1. For the specific example shown in Figure 2.6 below the one-sided test for $H_{0_1} : \theta > \theta^+$ will return a very high p-value because the uncertainty interval is to the

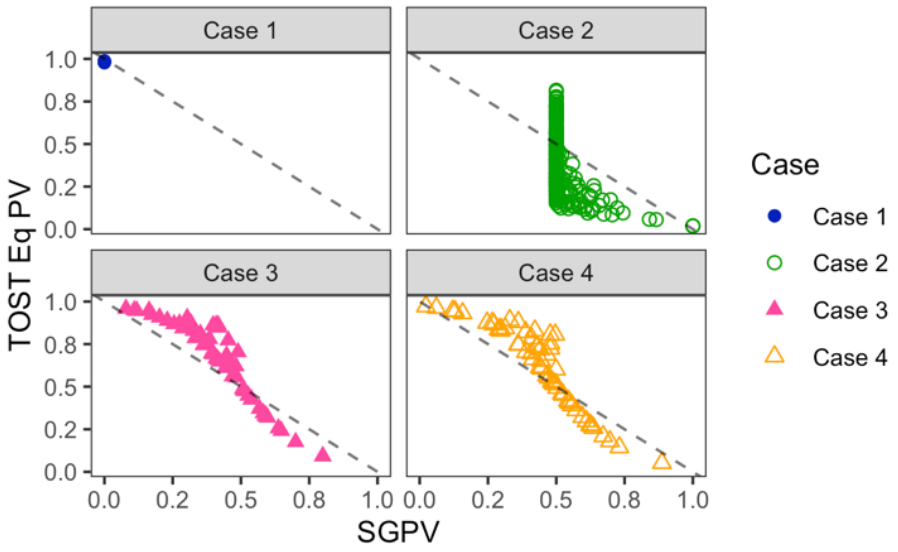

Figure 2.5: Graph of same simulated data as shown in Figure 2.1 but here split into 4 separate plots by overlap case. This is shown to emphasize the mathematical patterns within each case.

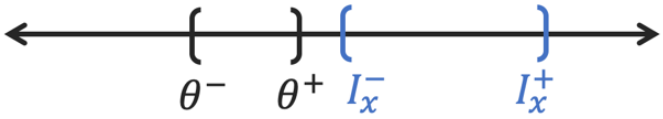

Figure 2.6: Line graph for Case 1 of overlap, or no overlap.

right of the equivalence range. Therefore, for Case 1 when $p_\delta = 0$ then p_T is close to 1 and SGPV concludes data is consistent with the alternative and TOST concludes data is inconclusive with indifference zone.

2.4.2 Case 2 (complete overlap)

Next we look at Case 2, when there is complete overlap between the uncertainty interval and indifference zone, shown on the line graph in Figure 2.7. There are two subcases of complete overlap; Figure 2.7 shows when the indifference zone is contained inside the uncertainty interval $H_0 \subset I_x$, we will call this Case 2a. The other version is when the uncertainty interval is contained inside the indifference zone $I_x \subset H_0$, for reference Case 2b. For Case 2b the overlap length=$I_x^+ - I_x^-$ and so $p_\delta = 1$ always. Here the TOST from both one-sided tests will return p_T close to 0. Both SGPV and TOST conclude for Case 2b that the data is consistent with the null.

On the other hand, for Case 2a the derived behavior is Equation 2.5 below. The correction factor causes the SGPV to range to be $0.5 \leq p_\delta \leq 1$ depending on the lengths of the intervals. Here the TOST varies from $0 \leq p_T \leq 1$ because I_x^- and I_x^+ are outside of $\theta^- and \theta^+$ and only the largest p-value is reported. For Case 2a

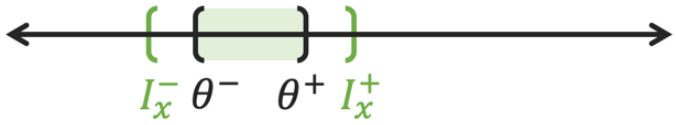

Figure 2.7: Line graph for Case 2a of overlap, or complete overlap.

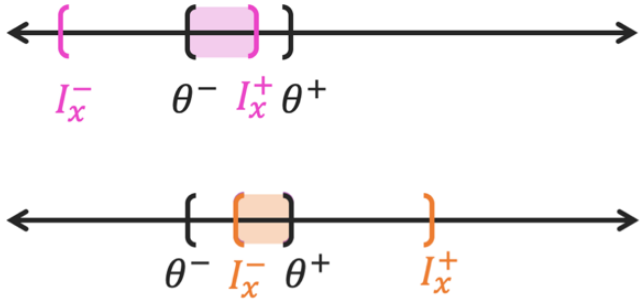

Figure 2.8: Line graph for Case 3 (on top) and Case 4 (on bottom) of overlap, or partial overlap.

there is no exact mathematical connection between the TOST and SGPV.

$$p_\delta = \frac{(\theta^+ - \theta^-)}{\frac{2c_\alpha S}{\sqrt{n}}} \times \max\left\{\frac{\frac{c_\alpha S}{\sqrt{n}}}{(\theta^+ - \theta^-)}, 1\right\} \tag{2.5}$$

2.4.3 Cases 3 and 4 (partial overlap)

Finally, with the Cases 3 and 4 in Figure 2.8 we see partial overlap between the uncertainty interval and indifference zone. The overlap length= $I_x^+ - \theta^-$ in Case 3 and overlap length $= \theta^+ - I_x^-$ in Case 4.

We know by definition of a simple t-test that $p_{T_1} = P(\theta < \theta^- | H_0) = 1 - F_n\left(\frac{\bar{x}-\theta^-}{S/(\sqrt{n})}\right)$ and $p_{T_2} = P(\theta > \theta^+ | H_0) = F_n\left(\frac{\bar{x}-\theta^+}{S/(\sqrt{n})}\right)$. Using this information the derived mathematical connection between the reported TOST p-value and the SGPV can be seen in Equation 2.6. This specific equation can be seen as the curved trend line that is present in the simulated examples shown in Figures 4 and 5.

$$p_\delta = \left[\frac{1}{2c_\alpha} F_n^{-1}(1 - p_T) + \frac{1}{2}\right] \times \max\left\{\frac{\frac{c_\alpha S}{\sqrt{n}}}{(\theta^+ - \theta^-)}, 1\right\} \tag{2.6}$$

These derived connections can help us to compare the SGPV to the TOST in any given scenario. We know the amount of overlap can limit the values $p_\delta and p_T$ can take, which in turn influences the inference outcomes that the method reports.

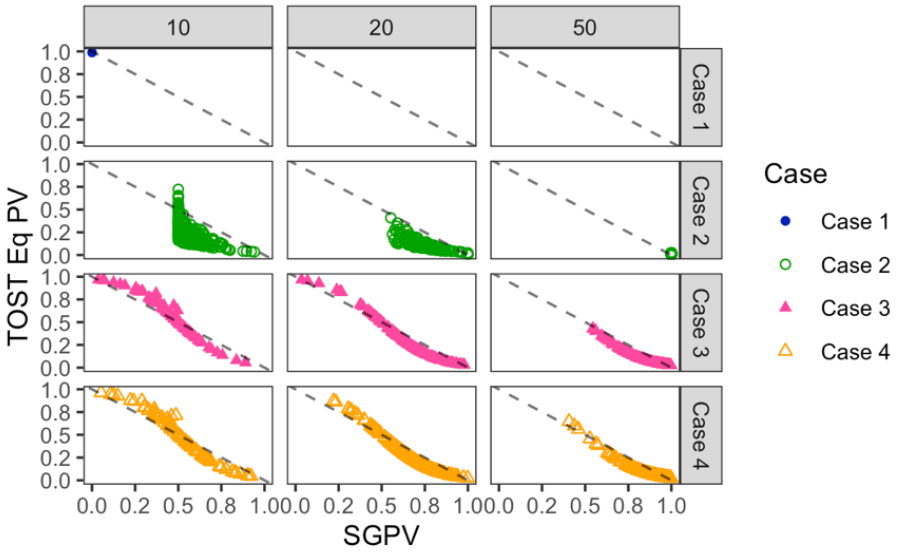

Figure 2.9: Graph of simulated SGPVs versus reported TOST p-values. In this example there are 500 iterations repeated for three different sample sizes of $n = 10, 20$, and 50 for data generated under the null, N(0,1), and tested against the indifference zone $[\theta^-, \theta^+]$=[-0.375,0.375]. Here the plots are separated by case and by sample size to show the convergence over sample size in each of the cases.

2.4.4 Limiting Behavior

The relationship between SGPV and TOST is most complicated in small sample sizes. To visualize how sample size affects this relationship SGPVs and reported TOST p-values were simulated under the true point null, $N(0,1)$, for 500 iterations, repeated for three different sample sizes of $n = 10, 20$, and 50and tested against an indifference zone of $[\theta^-, \theta^+]$=[-0.375,0.375]. In Figure 2.9 we see that for a sample size of n=10 the points lie the farthest from the line of equality, y=1-x. As the sample size increases we see the points begin to converge closer to equality, on a curved line which is identified in Equation 2.6 above, and to the location where data is consistent with the null, $p_T = 0$ and $p_\delta = 1$. Notice as the sample size increases there are less outliers from the curved line. This can be mathematically explained when the uncertainty interval shrinks smaller than 2 times the indifference zone length, so that the correction factor, or the $max\{\}$ function, in Equation 2.6 goes away.

For data generated under the point null, when the sample size increases the uncertainty interval in all cases narrows naturally and is more likely to be centered closer to the point null. This means the uncertainty interval is either contained in or has overlap with the indifference zone. Each of the cases of overlap have different limiting probabilities under the true point null. For Case 1 under the null as $n \to \infty$ the $P(\text{Case 1}) \to 0$. For Case 2a because the uncertainty interval shrinks smaller than the indifference zone and $P(\text{Case 2a}) \to 0$. Next for Case 2b which is when $I_x \subset H_0$, under the true point null this is becomes most likely when the uncertainty

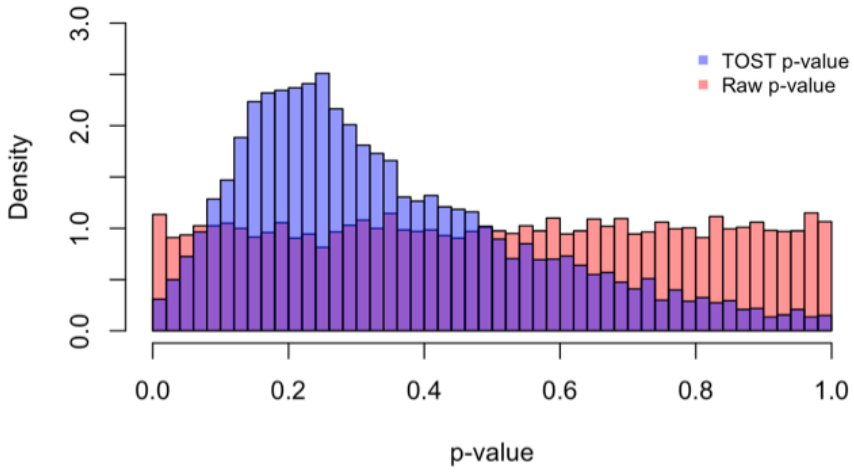

Figure 2.10: Histogram of simulated raw p-values and TOST p-values under the null. In this example there are 10,000 iterations of sample size of n=6 for data generated under the null, N(0,1), and tested against the indifference zone $[\theta^-, \theta^+]$=[-0.5,0.5].

interval shrinks, $P(\text{Case 2b}) \to 1$. Finally, for Cases 3 and 4 of partial overlap $P(\text{Case 3 or Case 4}) \to 0$.

2.5 Error Comparison

2.5.1 Type I Error

Type I Error is the probability under the null that a test or method "rejects the null". In order to compare how both procedures control the Type 1 Error, we must first understand the intention behind the reported p-values. In the traditional p-value space when $p > 0.05$ the null is rejected. The SGPV rejects the null when data is consistent with the alternative or $p_\delta = 0$. TOST rejects the null when the data in inconclusive and $p_{T_1} > 0.05$ and $p_{T_2} > 0.05$. We ran a quick simulated example of data under the null for sample size $n = 6$ with 10,000 iterations to generate raw p-values, TOST p-values and SGPV p-values. Figure 2.10 shows a histogram comparing these results for raw p-values and TOST reported p-values. We see that the raw p-values are uniformly distributed over the range from 0 to 1, which is expected when data is generated under the null (Hung et al., 1997). However, the TOST p-values are not uniform; they are right skewed with a peak at probability 0.2. We also found that as the sample size increases the TOST reported p-values converge to $p_T = 0$ which is not a stable distribution. When data is generated under the true point null each one-sided test does not have null or uniform behavior. Therefore, the max or larger reported TOST p-value distribution is also non-uniform.

The user must change their interpretation of TOST p-values because the distribution is non-uniform. When reporting results a traditional p-value threshold or cut-off of $\alpha = 0.05$ is used when approximately 5% of the p-values under the null occur between 0 and 0.05. For the TOST in the example shown in Figure 2.10

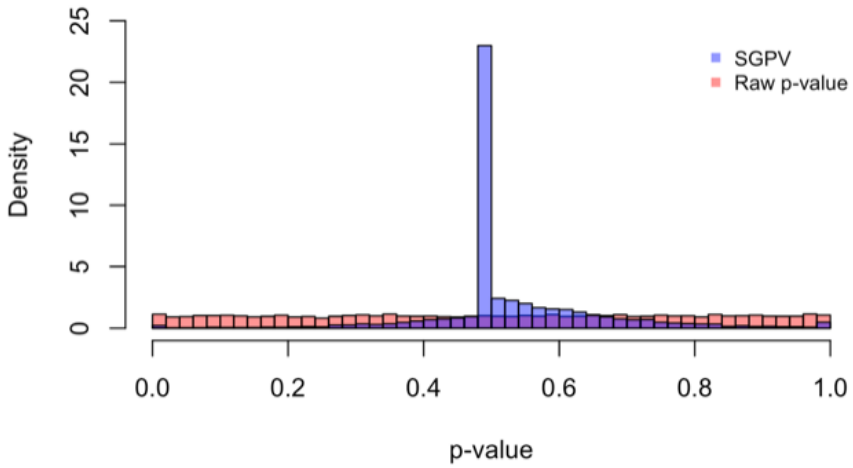

Figure 2.11: Histogram from the same simulated data as shown in Figure 2.10. Here we see the raw p-values and SGPVs generated under the null.

if we used a cutoff of $\alpha = 0.05$ we only have 2.28% of the results, p_T, in the range 0 to 0.05. In order to truly capture 5% of the results the p-value threshold would have to be recalculated under simulation, and here it would be set at $\alpha = 0.0817$. This is a major problem for users of TOST p-values who want to interpret them as universal scales of evidence against the null hypothesis. It is very plausible, that in one study a TOST p-value of 0.0817 is actually less significant than a TOST p-value of 0.023 simply because of the different underlying null p-value distributions.

This discrepancy is not taken into consideration when reporting the Type I Error for TOST in the original published procedure (Schuirmann, 1987). In this case, one can not report the p-value as a measure of evidence against the null, but rather one must report the simulated quantile of the p-value under the null. The users should be aware of this difference and hire a statistician to run simulations and recalculate the 5% threshold for analyses using TOST method. This is a very subtle point about p-values.

$$p_T = |p_{T_1} - p_{T_2}| \tag{2.7}$$

One solution to this is instead of taking the maximum of the p-values from the two tests, take the absolute value of the difference shown in Equation 2.7. This has been shown to have better p-value behavior under the null, more uniformly distributed under the null, but this is not a perfect fix (Hauck and Anderson, 1984).

For comparison the SGPVs performed as expected under the null with a large spike at 0.5 as seen in Figure 2.11 (Blume et al., 2018). The SGPVs spike at 0.5 is when the uncertainty interval is very large and

	Type 1 Error:	Power:
TOST	0.163	0.727
SGPV	0.164	0.826
SGPV fair (removed all $p_\delta = 0.5$)	0.197	0.780

Table 2.2: Comparison of power between TOST and SGPV methods with similar Type I Errors. This example was for a one sample test for proportion with 10,000 iterations of sample size of n=22 for data generated under the null, θ_0=0.1, and tested against the indifference zone $[\theta^-, \theta^+]$=[0.05,0.15].

contains the indifference zone, Case 2a. Here the length correction factor in Equation 2.1 applies. SGPV still requires a threshold to identify Type 1 Error but if this threshold remains than the Type 1 Error rate converges to 0 as the sample size increases. However, unlike the TOST *p*-value the SGPV is not a tail area probability but rather a proportion, this is an important conceptual difference. This can be useful in multiple comparisons as the SGPV is considered a measure of statistical evidence. The TOST *p*-value cannot be interpreted this way.

2.5.2 Type II Error and Power profile of tests (ROC)

It has been previously claimed that the TOST "always has higher power than SGPV" (Lakens and Delacre, 2018). This statement is false because there are many counterexamples available. Here we show one counterexample to disprove this statement. We used 10,000 simulations of one sample exact tests for proportion where$\theta^- = 0.05$and $\theta^+ = 0.15$ and n=22. Here we tested if the proportion of successes was θ_0=0.1, so data was simulated using a random discrete binomial distribution and then tested with an exact binomial test instead of a t-test. In Table 2.2 we show one example where it is clear that SGPV has slightly higher power than the TOST method. It is obvious that power of $0.826 > 0.727$ in this example with similar Type I error. Even if we removed the $p_\delta = 0.5$ we get higher power $0.780 > 0.727$. Further, it has been proven in other literature for many different methods that the TOST is not uniformly more powerful (Ennis and Ennis, 2010).

We conducted a brief analysis of receiver operating characteristics, ROC, of the two tests using the same data from Table 2.2. In this analysis shown in Figure 2.12 the best predictive performance is when a method maximizes the area under the curve, AUC (Hanley and McNeil, 1982). This reduces the number of false positives while increasing the number of true positives. For the sake of interpretation, we have labeled the

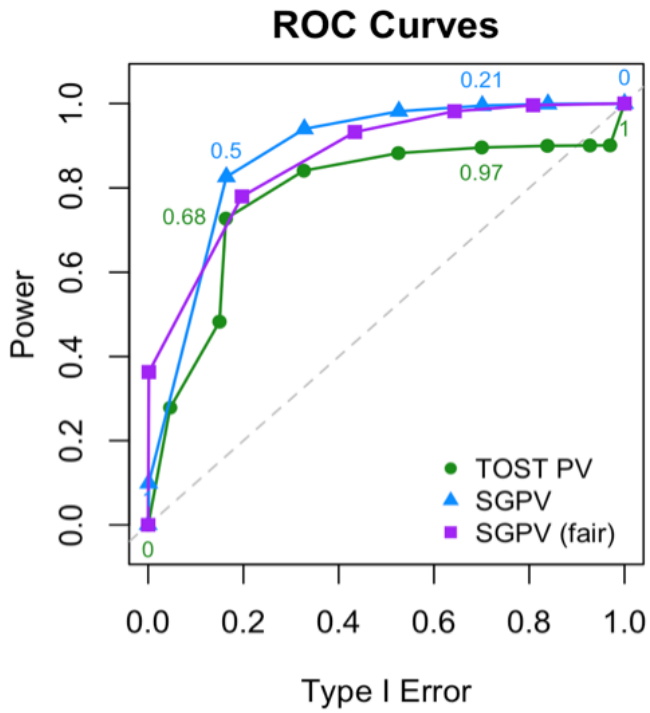

Figure 2.12: ROC curves for the same simulated data as shown in Table 2.2. Here we compare the TOST, shown with circles, SGPV, shown with triangles, and SGPV fair, shown with squares.

Figure 2.12 x-axis "Type 1 Error" instead of false positives and the y-axis "Power" instead of true positives. In this example SGPV, shown as the line with triangles, has higher AUC (0.882) than the TOST(0.779), the line with circles. This means SGPV has more predictive power for this simulated test for proportions.

In Figure 2.12 we kept all SGPVs even those that were equal to 0.5 for the line with the triangles. It is important to note that the interpretation of SGPVs close to 0.5 as "inconclusive" and this value can have a high frequency in simulation because of the correction factor in Equation 2.1. When we removed all values where $p_\delta = 0.5$, labeled as "SGPV fair", the line with the squares in Figure 2.12 resulted. Here SGPV fair still had an AUC (0.870) that was greater than the TOST(0.779). This behavior was similar in multiple other cases we tried, however we will not provide a formal proof for which test is uniformly more powerful.

2.6 Discussion and Comments

Previous publications have considered the benefits and flaws of equivalence p-values. Specifically, Berger and Hsu thoroughly proved that the TOST does not always assert bioequivalence at the α-level(Berger and Hsu, 1996). This severely undermines the claims and interpretation of the TOST reported p-values because even when a specific α is used, the conclusions of the test are for a smaller or more precise confidence interval. It was also stated by Berger and Hsu that "the TOST is highly biased with power much less than

0.05 for moderate and large (standard deviation)" (Berger and Hsu, 1996). This is largely due to the fact that the length or amount of overlap is not used when computing results. Along the same lines, Perlman and Wu noted that when the sample variance of the data is too large the TOST can never conclude data is consistent with the alternative and should then advise the user to collect more observations (Perlman and Wu, 1999). These flaws have been used to warn against the TOST method for at least the last 20 years, yet this method is still in practice especially in psychology research.

Both the TOST reported *p*-value and the power calculations ignore the fact that two tests were conducted and only use information from one of the one-sided tests even though the conclusion of null comes from both of the one-sided tests' results. This discrepancy between one and two test information needs to be corrected. This could be corrected with a length or overlap correction factor. The power and reported *p*-values should use information from both sides of the indifference zone, like the SGPV, for it to be an appropriate measure. In this paper we focused on TOST but for future research we propose this comparison be extended to non-inferiority tests and the Bayesian ROPE methods.

After seeing the results from this comparison, the SGPV has superior properties and the TOST can be misleading in practice. First, the SGPV's addition of a third inference interpretation of data being consistent with the alternative allows the user to conclude statistical difference. This is almost never done in practice because most methods have limited inference capabilities. Second, the amount of overlap in the two intervals should be reported as it is very valuable to the user. Third, the TOST conclusions are limited to a $(1-2\alpha)\%$ confidence interval, where the SGPV can be used for any user-specified data uncertainty interval. After this investigation we can conclude that SGPVs are more valuable, more informative, and more flexible. Only when users are given complete information about their specific hypotheses can they make appropriate recommendations for future science.

CHAPTER 3

Adjusting for Collaborator Uncertainty in the Indifference Zone for Second-Generation *p*-value Applications

3.1 Abstract

In this paper we investigate different ways to incorporate collaborator uncertainty into statistical analysis in order to identify the best practice for future use. Specifically, we focus on collaborator uncertainty in identifying the indifference zone for use in a second-generation *p*-value (SGPV) analysis. Traditionally, SGPV uses a fixed bound that remains unchanged after being identified by a collaborator. We propose a new concept of shrinking the indifference zone as sample size increases in the case of collaborator uncertainty. Our results show that when an uncertain but wide small sample interval is identified, shrinking the indifference zone balances the errors between behaviors of a fixed zone and a point null. It increases power, $P_{\theta_1}(p_\delta = 0)$, when compared to a fixed zone and increases probability of true nulls, $P_{\theta_0}(p_\delta = 1)$ when compared to a single point. This identified trade-off and improved behavior will change future analyses and will benefit communication between statisticians and collaborators.

Keywords

interval null, uncertainty, evidence, shrinking, test errors, second-generation *p*-values

3.2 Introduction

When starting a statistical analysis, the analyst looks to a field expert for an assumed hypothesis. This given hypothesis is then treated as ”truth” in order to properly interpret results. However, in practice the true underlying distribution is never known. The field expert or collaborator can only make an informed guess and this guess will always have some uncertainty. Although rarely done in practice, the amount of uncertainty in the assumed hypothesis should be discussed and influence the statistician's analysis plan.

Consider a fictional field expert in biochemistry, Dr. Jane Doe, who is currently collecting data to measure adult response in body temperature to minimal exposure of chemical ”X”. She hypothesizes that the mean body temperature will remain stable or at 98.6 degrees Fahrenheit (F). However, there is no previous literature to serve as a basis for her hypothesis. She consults a statistician, who recommends that Dr. Jane Doe instead identify a range of body temperatures that are practically equivalent to 98.6 F.

This range can be treated as an interval null or indifference zone. This concept of changing a point null into an interval null has changed how researchers think about hypothesis testing (Cohen, 2021). Hypothesis tests do not always have to test for strict equality. When using a null indifference zone, an equivalence test or second-generation *p*-value (SGPV) should be used for the analysis (Blume et al., 2018, 2019).

Even though Dr. Jane Doe is more confident in specifying a range of null temperatures than a single point, the lack of previous literature gives her concern. When expressing this concern to the statistician, the statistician replied that this range will be treated as ”truth” and all results will be interpreted in this context. Current statistical practice does not have the flexibility to account for this concern or uncertainty. In this paper, we propose a better way to conduct SGPV analyses with hypothesis uncertainty. We show our investigation of the different ways in which the analysis can be modified in light of this concern to improve results. To conclude, we share the results for the best methods we found to account for different levels of collaborator uncertainty in SGPV analyses.

3.2.1 Uncertainty Framework

Collaborator uncertainty can be put into three categories; unable, uncertain, and confident. First, when a collaborator is unable to identify an indifference zone this is usually because there is no expertise or little previous literature. In this case we suggest the SGPV method should not be used because there is no basis to form the pre-specified hypothesis. Then, statisticians would be forced to randomly guess and check the errors before deciding on a null hypothesis. This ”double-dipping” of the data to determine a null hypothesis and to compute the results is bad statistical practice. It introduces data bias and will have misleading results. Next, in the case where the collaborator is confident in a pre-specified indifference zone the statistician should treat this as the true null and conduct the normal analysis. Finally, when the collaborator is uncertain of the pre-

specified null indifference zone they may have a hesitant guess or a wide small sample estimate. However, not enough previous literature is available in order for them to have full confidence in the interval. Statisticians should not treat a "likely" true interval the same as a "confidently" true interval.

One way to improve the analysis for likely true intervals is to make the indifference zone more precise. This idea is to reduce the width of the interval between the point null and the wide small sample interval bounds by a factor proportional to the standard deviation. Our research investigates how narrowing this distance affects test errors and performance. Another idea is to use the wide small sample interval but then shrink this interval at a set rate as sample size increases, for example $1/\sqrt{n}$. Here the underlying theory is that the null zone will shrink away from any possible alternative points. As the sample size increases the interval null becomes more precise and is more likely to contain only practically null values. This relieves collaborators' concerns knowing different sized datasets are compared to different levels of precision.

In this paper we will first outline the properties of the SGPV method and show how they relate to the indifference zone. Then, we will examine different scenarios of collaborator uncertainty and how the solutions of narrowing and shrinkage affect the test properties. Finally, we will interpret our findings and present our recommendations for future research.

3.3 Background

3.3.1 Notation

Here we set the notation for use in all the following examples. The collaborator chooses a null hypothesis that the parameter of interest,θ, practically null or within a specified indifference zone, $\hat{\theta}H_0$. In Figure 3.1 we can see an illustration of the indifference zone where $H_0 = [\theta^-, \theta^+] = [\theta_0 - \delta, \theta_0 + \delta]$; where the length of the null zone is 2δ. The indifference zone contains and is centered around the point null or θ_0.

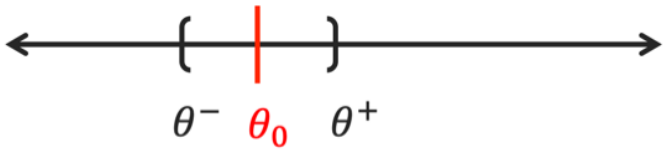

Figure 3.1: Line graph showing the indifference zone or interval null.

The investigator then collects data $x = (x_1, ..., x_n)$, to test this hypothesis. From the dataset an uncertainty interval for the parameter of interest, θ, is constructed. In this paper we call that interval $I(x) = (I_x^-, I_x^+)$. This uncertainty interval might be a confidence interval, credible interval or likelihood support interval for the mean. To simplify, in this paper we will assume that I(x) is a 95% confidence interval for the mean. This technically can be defined in Equation 3.1 and here the length of the uncertainty interval is $2 \times Z_{\alpha/2} \frac{\sqrt{V}}{\sqrt{n}}$.

$$I(x) = \left(\hat{\theta} - Z_{\alpha/2}\frac{\sqrt{V}}{\sqrt{n}}, \hat{\theta} + Z_{\alpha/2}\frac{\sqrt{V}}{\sqrt{n}}\right) \tag{3.1}$$

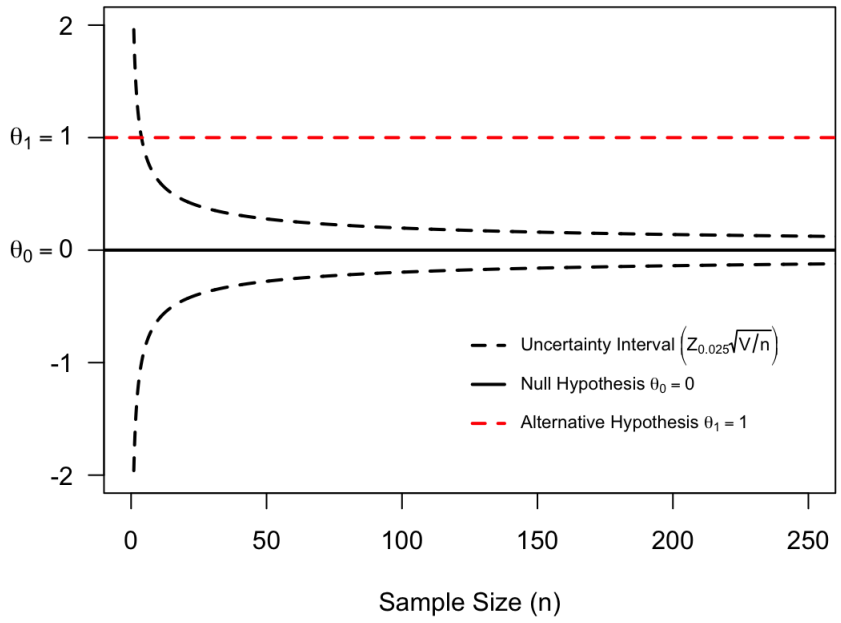

Figure 3.2: Plot showing data driven 95% confidence interval or uncertainty interval for the mean changing over sample size.

It is commonly known that the data driven uncertainty interval, specifically a confidence interval for a parameter like the mean, shrinks as the sample size grows (Blume et al., 2018; Robbins, 1970). For all of the following examples we have used a 95% confidence interval for the mean with $\theta_0 = \hat{\theta} = 0, \theta_1 = 1$ and $x_i \sim N(0,1)$. In Figure 3.2 the uncertainty interval is shown from sample size in the range $n \in [0,250]$. This figure confirms that the uncertainty interval narrows closer to the point null, or θ_0, as the sample size increases. Now we must identify the test and hypothesis that is used to analyze this data.

3.3.2 Second-generation *p*-values

The second-generation *p*-values (SGPV) method was created to measure the fraction of data-supported effect sizes that are within the indifference zone (Blume et al., 2018, 2019). The uncertainty interval, I(x), is compared to the indifference zone, $H_0 = [\theta^-, \theta^+]$. Specifically the amount of overlap between these two

intervals is used to decide the inference outcome. Equation 3.2 shows the reported p-value or SGPV where $(I \cap H_0)$ is the intersection or overlap of the two intervals and the function $f(x) = |x|$ is the length of that interval or overlap:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2|H_0|}, 1\right\} \tag{3.2}$$

It is clear in Equation 3.2 that the length of overlap is used to determine p_δ. This amount of overlap can be condensed into four different cases, which shown in Figure 3.3. Each of these cases of overlap corresponds to one of three inference outcomes. In Case 1 $p_\delta = 0$ and there is no overlap between the intervals. This means the outcome concludes the data is consistent with the alternative. In Case 2 when the outcome can either conclude the data is consistent is inconclusive or the data is consistent with the null. Finally, in Cases 3 and 4 then $0 < p_\delta < 1$, or there is some overlap between the intervals the outcome concludes the data is inconclusive with the null.

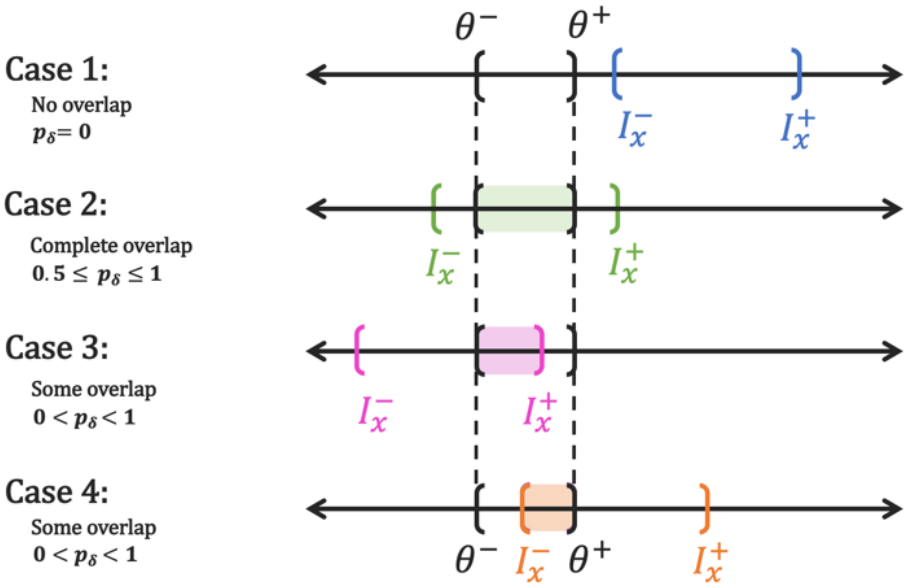

Figure 3.3: Illustration of the four cases of overlap between the indifference zone and the uncertainty interval for SGPVs.

3.3.3 Properties and Errors

In order to define the errors or properties of the SGPV method we must first identify all possible outcomes. The three inference outcomes of the SGPVs include data being consistent with the null, data being consistent with the alternative, and data being inconclusive. This means that the errors will be defined first by whether the data are truly null, H_0, or alternative, H_1, and secondly by the probability of being null, probability of

being alternative and the probability of being inconclusive.

3.3.3.1 Probability that data are consistent with the alternative

This probability found in Equation 3.4 comes directly from Supplement Equation 3.5 from Blume's paper in 2018 (Blume et al., 2018). This inference outcome only occurs when $p_\delta = 0$ is the data consistent with the alternative.

$$\beta = P_\theta(p_\delta = 0) \tag{3.3}$$

$$= \Phi\left[\frac{\sqrt{n}(\theta_0 - \delta) - \sqrt{n}\theta}{\sqrt{V}} - Z_{\alpha/2}\right] + \Phi\left[-\frac{\sqrt{n}(\theta_0 + \delta) + \sqrt{n}\theta}{\sqrt{V}} - Z_{\alpha/2}\right] \tag{3.4}$$

Power

When data is generated under an alternative or $\hat{\theta} = \theta_1 \neq \theta_0$ then Equation 3.2 becomes $\beta_1 = P_{\theta_1}(p_\delta = 0)$. Traditionally this is known as the Power or (1- Type II Error) for a test.

Type I Error

When data is generated under the point null, $\hat{\theta} = \theta_0$ Equation 3.4 reduces to Equation 3.5. Traditionally this is identified as the Type I Error for a test.

$$\beta_0 = P_{\theta_0}(p_\delta = 0) = 2\Phi\left[-\frac{\sqrt{n}\delta)}{\sqrt{V}} - Z_{\alpha/2}\right] \tag{3.5}$$

Probability that data are consistent with the null

This probability found in Equation 3.7 comes directly from Supplement Equation 3.8 from Blume's paper in 2018 (Blume et al., 2018). This inference outcomes only occurs when $p_\delta = 1$ and consequently when the uncertainty interval is smaller than the indifference zone; $|I_x| < |H_0| or \delta > Z_{\alpha/2}\sqrt{\frac{V}{n}}$.

$$\omega = P_\theta(p_\delta = 1) \tag{3.6}$$

$$= \Phi\left[\frac{\sqrt{n}(\theta_0 + \delta) - \sqrt{n}\theta}{\sqrt{V}} - Z_{\alpha/2}\right] - \Phi\left[\frac{\sqrt{n}(\theta_0 - \delta) - \sqrt{n}\theta}{\sqrt{V}} + Z_{\alpha/2}\right] \tag{3.7}$$

Otherwise when the uncertainty interval is wider than the indifference zone,$\delta \leq Z_{\alpha/2}\sqrt{\frac{V}{n}}$, we can never conclude the data is consistent with the null, $p_\delta \neq 1$. This means the probability will be 0 or Equation 3.8

holds. This is a form of "self-protection" in the framework of the SGPV method. The test will never conclude data is consistent with the null when the uncertainty interval cannot be contained within the indifference zone.

$$\omega = P_\theta(p_\delta = 1) = 0 \tag{3.8}$$

3.3.3.2 Probability that data are inconclusive

This probability found in Equation 3.9 comes directly from Supplement Equation 3.11 from Blume's paper in 2018 (Blume et al., 2018). This inference outcome occurs when there is some overlap between the uncertainty interval and the indifference zone. These 3 probabilities obtain the characteristic that their sum is always 1 under a specified dataset. When $\hat{\theta} = \theta_0$ holds then $\beta_0 + \omega_0 + \gamma_0 = 1$, and when $\hat{\theta} = \theta_1 \neq \theta_0$ holds then $\beta_1 + \omega_1 + \gamma_1 = 1$.

$$\gamma = P_\theta(0 < p_\delta < 1) = 1 - P_\theta(p_\delta = 0) - P_\theta(p_\delta = 1) \tag{3.9}$$

3.3.4 Connection for probabilities

Table 3.1 compares these errors or probabilities to a traditional *p*-value significance framework. For the columns under traditional *p*-value the reported *p*-values are compared to a pre-specified threshold, most commonly 0.05. Even after choosing the optimal threshold the traditional *p*-value tests cannot conclude data is consistent with the null. The 2 outcomes for traditional *p*-values are to reject the null (data are consistent with the alternative) or to fail to reject the null (data are inconclusive). The SGPVs' gain of an additional inference outcome allows users to make more specific conclusions based on the data.

In order to pick the best statistical analysis method or indifference zone technique we must balance these errors. The "correct" errors must be maximized; the probability of data being consistent with null under $\theta_0(\omega_0)$ and the probability of data being consistent with alternative under $\theta_1(\beta_1$ or power). The "incorrect" errors must be minimized; the probability of data being consistent with null or inconclusive under $\theta_1(\gamma_1$ and $\omega_1)$ and the probability of data being consistent with alternative or inconclusive under $\theta_0(\gamma_0$ and β_0 or Type I error). This perfect balance between maximizing correct errors and minimizing incorrect errors is rare in practice. Often this is the goal of a-priori study planning and conversations between the collaborator and the statistician.

		Underlying Truth			
		Null is true		Alternative is true	
		Traditional *p*-value	SGPV	Traditional *p*-value	SGPV
Inference	Data are consistent with the alternative	Type I error $P_{\theta_0}(p \leq 0.05)$	β_0 $P_{\theta_0}(p_\delta = 0)$	Power or (1 - Type II error) $P_{\theta_1}(p \leq 0.05)$	β_1 $P_{\theta_1}(p_\delta = 0)$
	Data are inconclusive	1 - Type I error $P_{\theta_0}(p > 0.05)$	γ_0 $P_{\theta_0}(0 < p_\delta < 1)$	Type II error $P_{\theta_1}(p > 0.05)$	γ_1 $P_{\theta_1}(0 < p_\delta < 1)$
	Data are consistent with the null	NA	ω_0 $P_{\theta_0}(p_\delta = 1)$	NA	ω_1 $P_{\theta_1}(p_\delta = 1)$

Table 3.1: Table linking test properties or errors for specific inference outcomes under different underlying data distributions for traditional *p*-values compared to SGPVs

3.3.5 Neyman Pearson Extension

This classical idea of trading off errors was famously addressed by Jerzy Neyman and Egon Pearson in the Neyman-Pearson lemma (Neyman et al., 1933). In the traditional two inference outcome framework they compared tests with the same Type I error, $P_{\theta_0}(p \leq 0.05)$, in order to maximize power, $P_{\theta_1}(p \leq 0.05)$. The best test which has identical Type I errors to other tests but has the highest power is not always unique but is labeled the "Most Powerful" (MP) test.

$$\text{SGPV Power}_{NP} = \frac{\beta_1}{(\beta_1 + \gamma_1)} \tag{3.10}$$

$$= \frac{(P_{\theta_1}(p_\delta = 0))}{(P_{\theta_1}(p_\delta = 0) + P_{\theta_1}(0 < p_\delta < 1))} \tag{3.11}$$

$$\text{SGPV Type I Error}_{NP} = \frac{\beta_0}{(\beta_0 + \gamma_0)} \tag{3.12}$$

$$= \frac{(P_{\theta_0}(p_\delta = 0))}{(P_{\theta_0}(p_\delta = 0) + P_{\theta_0}(0 < p_\delta < 1))} \tag{3.13}$$

As we transition to the three inference outcome framework of the SGPV method we will need to modify the Neyman-Pearson lemma. To make a direct comparison we must remove the third outcome of data being consistent with the null in the denominator as it was not in the original framework. The SGPV Neyman-Pearson equivalent for power and Type I error are shown in Equations 8 and 9. Basically we have scaled the Type I error and power by the appropriate denominators. We will use these new equations to find the "Most Powerful" test when comparing different indifference zone procedures.

3.4 Indifference Zone Procedures

The original SGPV method requires a collaborator to pre-specify a fixed indifference zone. This performs best when the collaborator is confident in these fixed values. However, there are ways to modify this procedure in light of collaborator uncertainty. In this section we will compare the properties or errors for different indifference zone procedures. First, we will look at fixed interval behavior. Next, we will simulate narrowing the indifference zone by a fixed proportion of the standard deviation. Finally, we will show results from shrinking the difference zone.

3.4.1 Fixed Intervals

To begin our analysis, we must first establish the behavior or probabilities of fixed indifference zones. The goal of an indifference zone is to identify a range of practically equivalent null values. Importantly this range must not include any alternative values that the researcher desires to identify, these should be distinct from the null. However, with natural human error sometimes the collaborator will identify the wrong indifference zone, or a range that contains an alternative value.

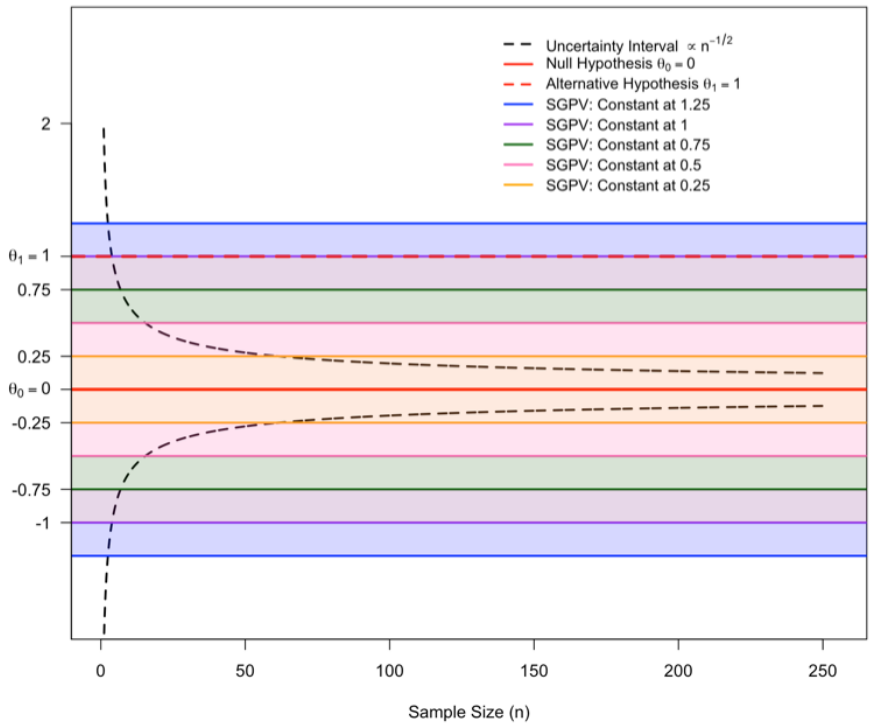

Figure 3.4: Plot showing different fixed indifference zones over sample size versus the assumed uncertainty interval. Here we can also see how they compared to the null and the alternatives.

In Figure 3.4 we see five different fixed indifference zones plotted over sample size. We see that even the smallest indifference zone [-0.25,0.25] is wider than the uncertainty interval for $n \geq 60$. Also note that an alternative point is identified at$\theta_1 = 1$ and so the two largest indifference zones, in blue and purple, include this alternative point. This will be important in understanding what happens if the collaborator identifies an indifference zone that is too wide. For computation the "sgpv" R package was used (Welty et al., 2020).

In Figure 3.5 we see the resulting six probabilities or errors over sample size. As expected the two largest indifference zones shown as the blue and purple lines have the worst statistical properties behavior. They have almost no power, β_1, and high probability of incorrectly concluding null, ω_1. Even though the blue and purple lines have the highest probability of correctly identifying the null, they cannot differentiate between null and alternative. We want an interval with appropriate width; wide enough to capture all practically null values but narrow enough to be precise and exclude alternative points. Also, important to note is the point null behavior, or the red lines. The point null as discussed before in the traditional p-values cannot identify any true nulls. However, it has the highest power. After reviewing these results, we can move forward with our investigation knowing we want to find an interval that balances errors between the point null and the widest interval $[-1.25, 1.25]$.

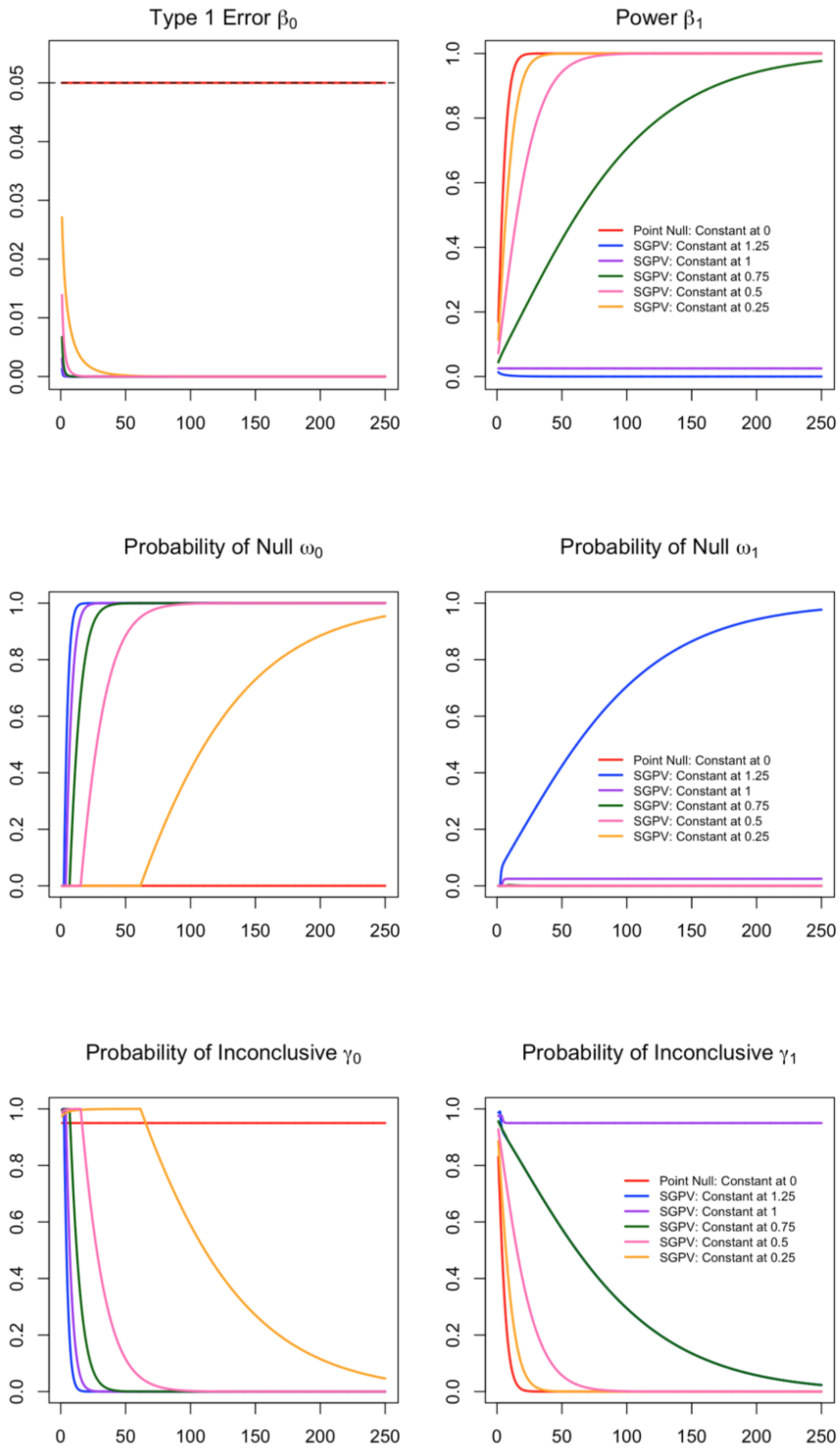

Figure 3.5: Plot showing 6 different probabilities, $\beta_0, \beta_1, \omega_0, \omega_1, \gamma_0, and \gamma_1$, over sample size for fixed SGPV indifference zones.

In order to compare which fixed bound has the best balance of all the statistical properties we first notice

Type I error, β_0, and probability of false nulls, ω_1, are low and almost identical for bounds at 0.25, 0.5, 0.75, and 1. Next we are looking maximize both the power, β_1, and the probability of true nulls, ω_0, so we plot the sum of these against sample size in Figure 3.6. The indifference zone that best maximizes this sum after $n = 30$ is $[-0.5, 0.5]$ or the pink line. The next best indifference zones are constant at 0.75 and constant at 0.25. As expected the statistical properties perform best when the indifference zone is between the point null and the closest alternative point of interest. Interestingly the best zone happens to be exactly halfway between the point null and the closest alternative point of interest. This closest alternative point of interest is not known in practice.

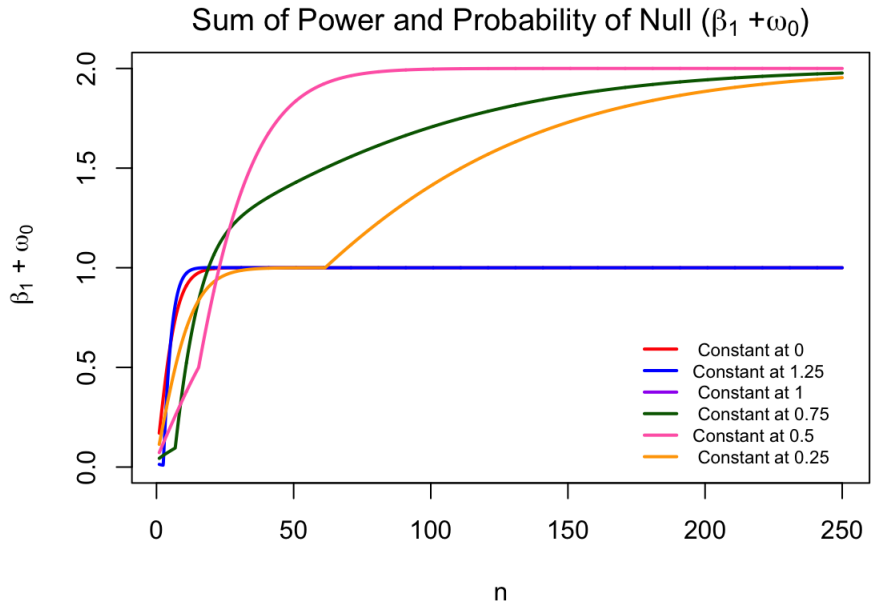

Figure 3.6: Plot comparing "correct errors" power, β_1, versus probability of true nulls, ω_0, for fixed SGPV indifference zones.

3.4.2 Narrow the Interval

After seeing the behavior in Figure 3.6, we consider what happens when the collaborator is not confident in the chosen indifference zone. We could "jump" from line to line in Figure 3.4 in order to obtain good test performance. However, we want to avoid data double-dipping in this procedure. When the collaborator is unsure of an indifference zone but is confident in the closest alternative point to the point null we can narrow the interval. Picking a distance between the known point null and hypothesized closest alternative point to narrow will allow the statistical properties to have the better performance when compared to just using the distance of the hypothesized closest alternative point. Using that interval would be too wide and could result in low power and high probability of false nulls.

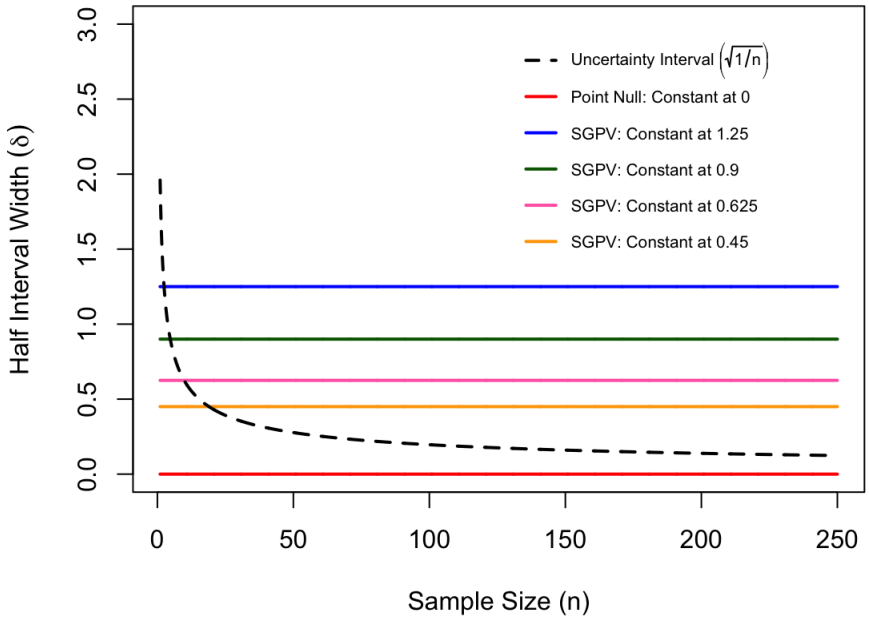

Figure 3.7: Plot showing comparing indifference zones narrows to half the distance between a point null and an alternative point. Here we are comparing the blue line at 1.25 to the pink line at 0.625 and the green line at 0.9 to the yellow line at 0.45.

After many simulations and shown in Figure 3.6 we found the ideal width to be the halfway distance between the point null and the closest alternative point. Here we use the indifference zone that is halfway between the point null, θ_0, and the collaborator identified alternative $\hat{\theta}_1$. Where $|\hat{\theta}_1 - \theta_0| = d$, the indifference zone is$[\theta_0 - \frac{d}{2}, \theta_0 + \frac{d}{2}]$. This procedure allows both correct errors, power, β_1, and probability of true nulls, ω_0, to be maximized. Also seen in Figure 3.5 when the indifference zone is [-0.5,0.5] and the true alternative is $\theta_1 = 1$, the probability of being inconclusive and the incorrect errors, β_0 and ω_1, remain relatively low when compared to other zones.

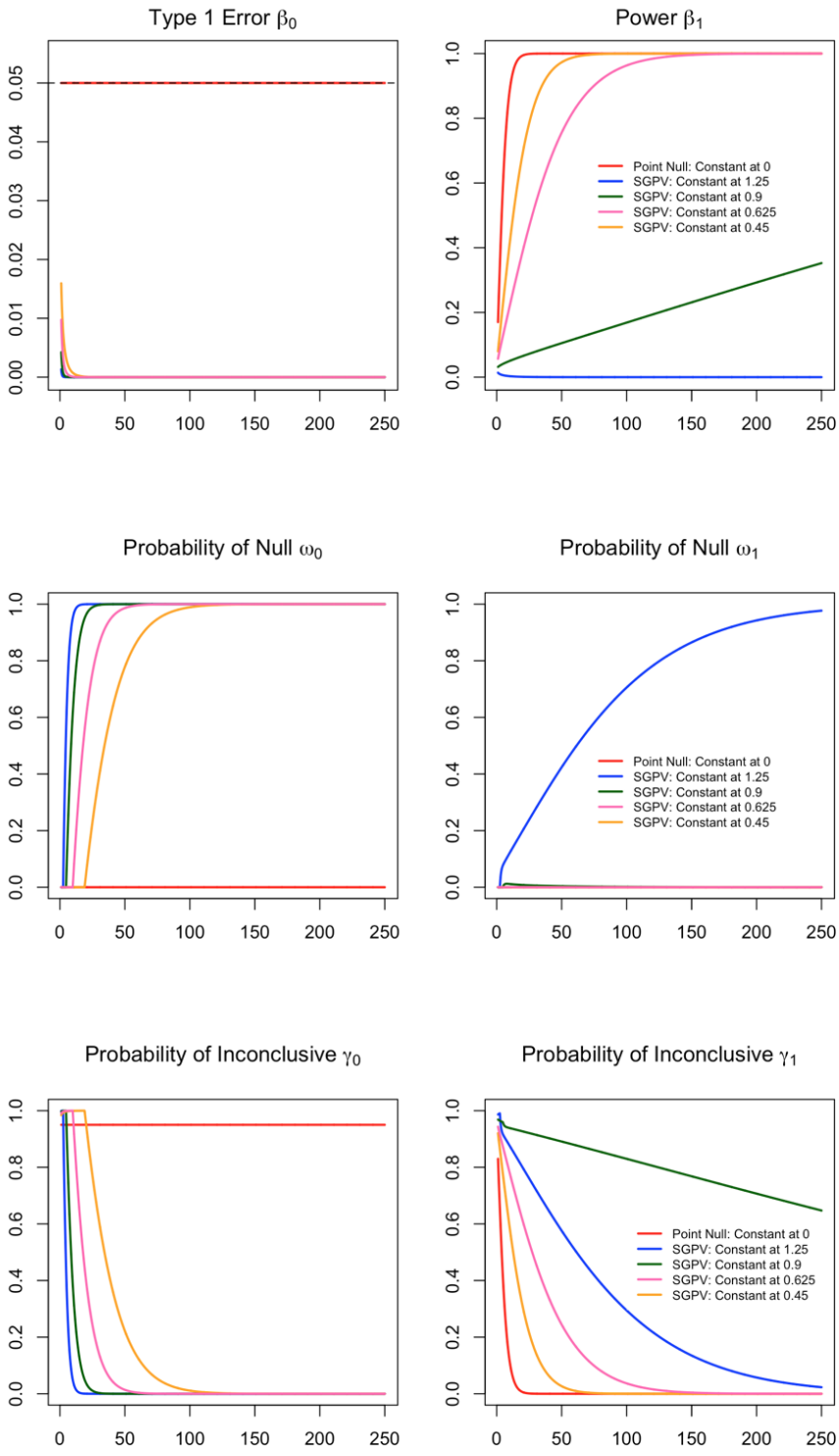

Figure 3.8: Plot showing 6 different probabilities, $\beta_0, \beta_1, \omega_0, \omega_1, \gamma_0, and \gamma_1$ over sample size for narrowed indifference zones. Here we are comparing the blue line at 1.25 to the pink line at 0.625 and the green line at 0.9 to the yellow line at 0.45.

In Figures 5 and 6 we see can see the difference in behavior for wide versus small intervals. Now let us consider what actually implementing this process, specifically when the collaborator incorrectly identifies θ_1. Here the collaborator is confident so we will assume cases where they are incorrect but close, $\hat{\theta}_1' = 1.25$ and $\hat{\theta}_1'' = 0.9$. In Figures 7 and 8 we compare the blue indifference zone at 1.25 to the narrowed pink indifference zone at 0.625 and the green indifference zone at 0.9 to the narrowed yellow indifference zone at 0.45. We see in both of these comparisons that the power, β_1, is improved and the probability of inconclusive under alternative, γ_1, is reduced. So even when the collaborator incorrectly identifies the closest alternative point narrowing the interval to half the distance always improves final results.

3.4.3 Shrinking over Sample Size

Finally, we will consider the cases where the collaborator is uncertain but can identify a wide small sample indifference zone or make a guess for the closest alternative point. We introduce a new concept of shrinking the indifference zone over sample size from a wide small sample uncertain estimate, $\left[\hat{\theta}^-, \hat{\theta}^+\right]$. This idea of shrinking the indifference zone with sample size is similar to the concept behind multiple comparisons in large data (Benjamini, 2010; Miller, 1981).

In order to choose what rate the indifference zone should shrink with sample size, we must first consider at what rate is the uncertainty interval shrinking naturally. We know that under assumptions of 95% confidence interval the rate of shrinkage is $\frac{1}{\sqrt{n}}$ as we can see in the formula for the length of I_x in Equation 3.14. For all examples we assume $V = 1$.

$$|I_x|(n) = 2Z_{\alpha/2}\frac{\sqrt{V}}{\sqrt{n}} \tag{3.14}$$

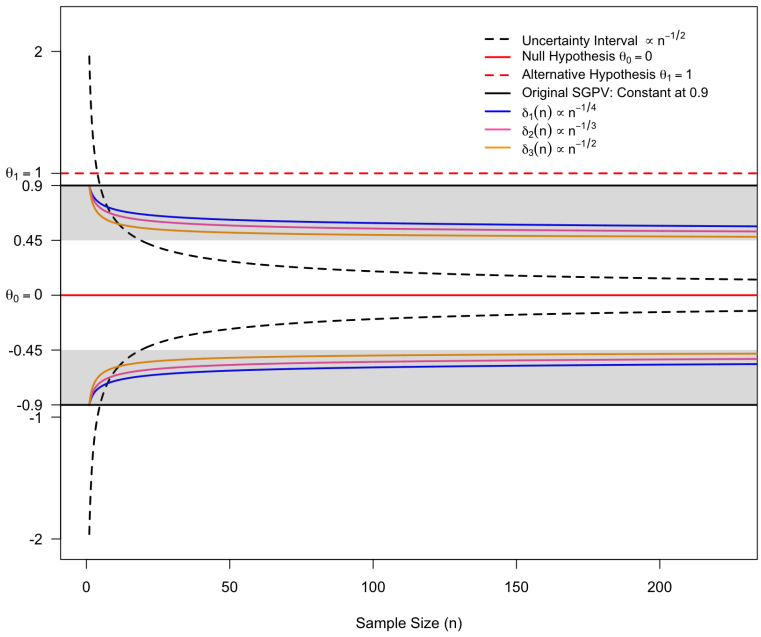

Figure 3.9: Plot showing different shrinking indifference zones over sample size versus the assumed uncertainty interval. Here we can also see how they compared to the null and the alternatives.

As defined in Blume's paper δ, or half of the length of the null zone, does not depend on the sample size and is a fixed bound (Blume et al., 2018). For our analyses we would like to change this concept and consider different δs that shrink at different rates with the sample size. In Equations 11, 12, and 13 we have chosen three rate of shrinkage for δ to analyze. These equations are proportional to because we will fix them to all start at $\left[\hat{\theta}^{-}, \hat{\theta}^{+}\right]$ when $n = 1$ and to end at $[\theta_0 - d/2, \theta_0 + d/2]$ when $n = \infty$. This fixed start and end points were determined using the information from the previous sections.

$$\delta_1(n) = \frac{\sqrt{V}}{n^{1/4}} \tag{3.15}$$

$$\delta_2(n) = \frac{\sqrt{V}}{n^{1/3}} \tag{3.16}$$

$$\delta_3(n) = \frac{\sqrt{V}}{n^{1/2}} \tag{3.17}$$

We plot these indifference zones in Figure 3.9 to see how they behave over sample size. After investigating different scenarios, we suggest the following procedure. If the collaborator is uncertain ask them for a wide small sample estimate of the indifference zone or a best guess of a wide closest alternative point. This will be an interval where we will start shrinkage from so this interval should be wider and be in context of very small sample sizes. Now as seen above the ideal behavior occurs at a point halfway between the closest alternative and point null. Therefore, we will limit the range of the shrinking zone to stop at the point halfway between the wide small sample estimate and the point null. Here when $|\hat{\theta}^- - \theta_0| = |\hat{\theta}^+ - \theta_0| = d$ then we use the indifference zone that shrinks from $\left[\hat{\theta}^-, \hat{\theta}^+\right]$ to $[\theta_0 - \frac{d}{2}, \theta_0 + \frac{d}{2}]$. If we let the zones shrink to 0 they will behave similar to a point null which never identifies true nulls; we do not want this behavior.

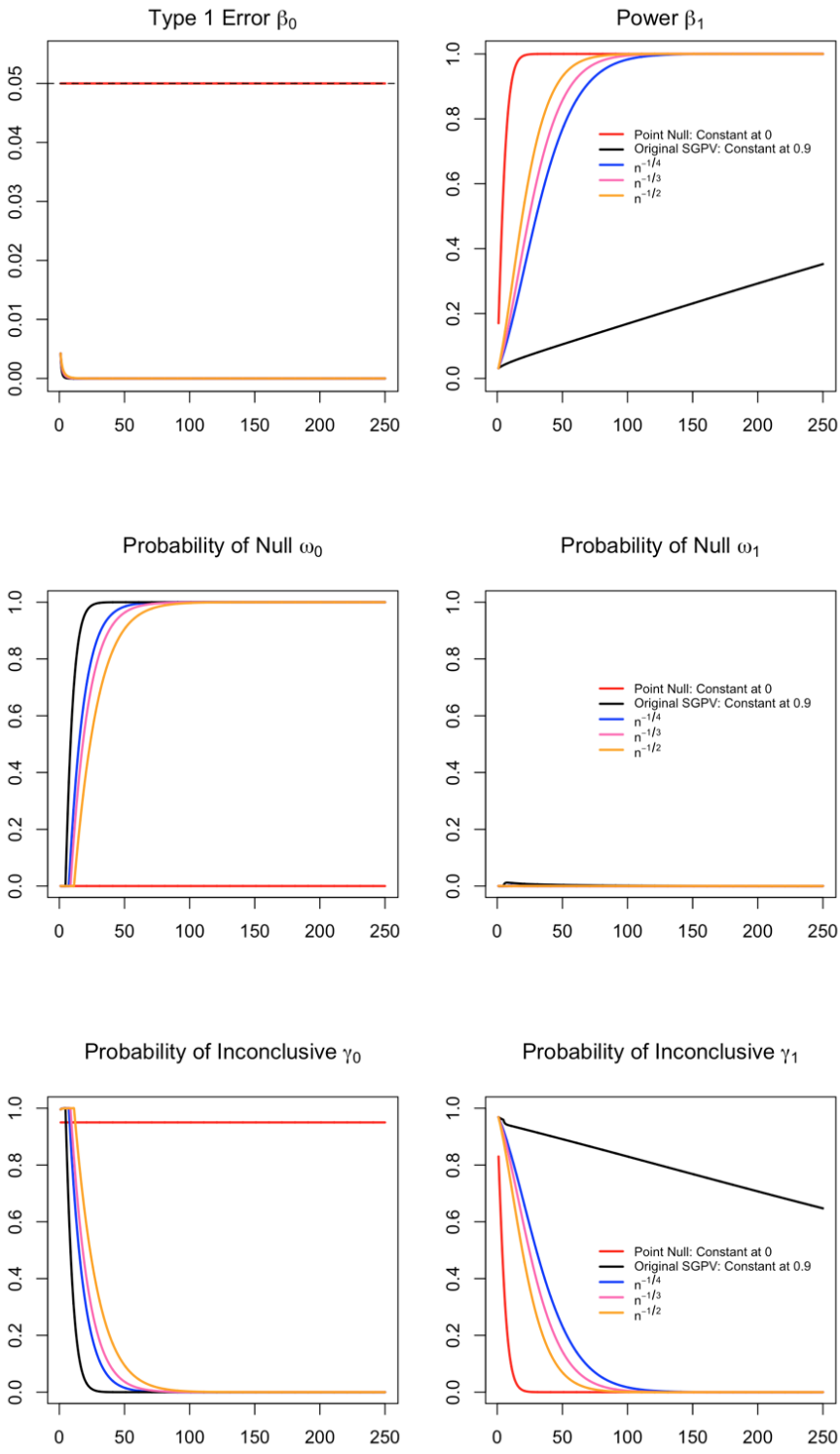

Figure 3.10: Plot showing 6 different probabilities, $\beta_0, \beta_1, \omega_0, \omega_1, \gamma_0, and \gamma_1$, over sample size for shrinking SGPV indifference zones.

For example, we will assume either the indifference zone wide small sample estimate is $\left[\hat{\theta}^-, \hat{\theta}^+\right] =$

$[-0.9, 0.9]$ or the estimate of the closest alternative point is $\hat{\theta}_1 = 0.9$. In either of these cases the following results will apply. In Figure 3.9 we see δ_1,δ_2,and δ_3 all begin at 0.9 when n=1 and when $n > 250$ shrink close to 0.45, which is the halfway point. In the plot it is clear that the uncertainty interval is shrinking to 0 and so becomes smaller than all of the indifference zone quickly. This is actually good null behavior because when the indifference zone can contain the uncertainty interval, $I_x \subset H_0$, then we can conclude the data is consistent with the null.

In Figure 3.10 we can see the six different probabilities/errors for the shrinking indifference zones. When compared to the fixed wide small sample estimate at 0.9 or the black line all of the shrinking zones have improved power β_1, less probability of being inconclusive γ_1, and almost identical Type 1 error. The slight loss in probability of true nulls $\omega_0$and gain in probability of being inconclusive γ_0 is worth the others errors benefit. This is the balance we are seeking in all of these scenarios.

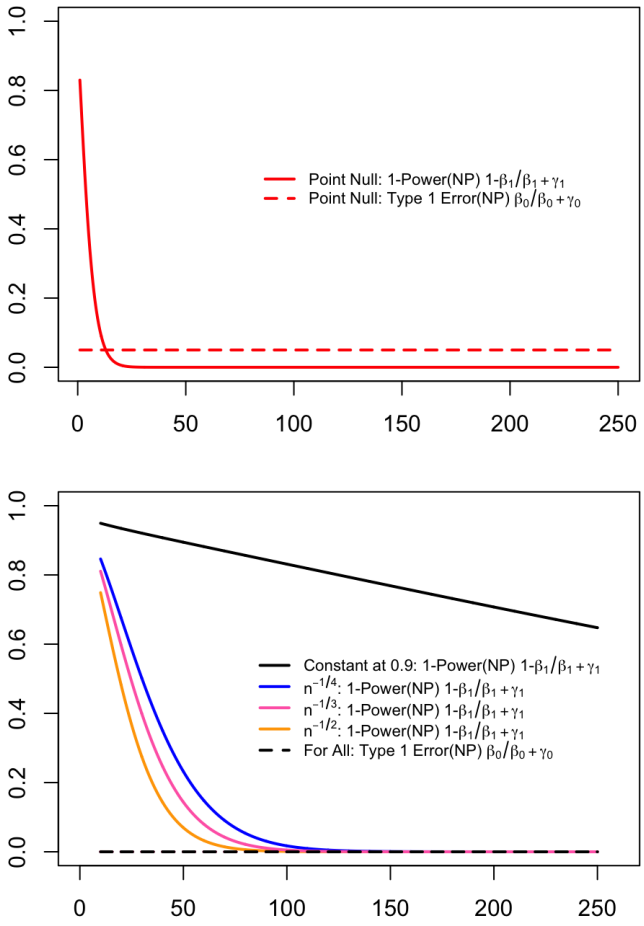

Figure 3.11: Plot showing Neyman-Pearson translation for shrinking SGPV indifference zones.

To really compare these indifference zones, we can look to our modified Neyman Pearson method de-

scribed above. As defined in Equations 8 and 9 the 1-Power and Type 1 error equivalents are plotted on both plots shown in Figure 3.11. The dashed lines are the Neyman Pearson equivalents for Type 1 Error and the solid lines are 1-Power. We have split these into two different plots because the Neyman Pearson method compares tests with equivalent Type 1 Error in order to choose the most powerful one and the point null has higher Type I error compared to the other indifference zones. The plot on the bottom only shows $n \in [10, 250]$ because this is where the Type 1 Error is equivalent for the four methods. For indifference zone hypotheses the yellow line or the zone shrinking at rate $\frac{1}{n^{1/2}}$ has the lowest 1-Power that intersects with the Type I error. This means it has best performance or is the "Most Powerful" when compared to the other zones. Even though $\frac{1}{n^{1/2}}$ was the most powerful all three shrinkage rates had very similar performance here so we leave it up to the collaborator to choose which to use.

3.5 Discussion and Comments

We propose our final recommendations for SGPV analysis with collaborator uncertainty in Table 3.2. The collaborator is either confident, uncertain or unable. The most ideal results come when the collaborator is certain. In other cases, the statistician has to choose the best analysis in order to give the best possible the results.

When the collaborator is confident their chosen indifference zone does not contain any significant alternative points we suggest using a fixed zone in a traditional SGPV analysis. When the collaborator can only confidently identify the closest alternative point, then narrowing the interval will ensure the best results. When unsure the collaborator should be asked to identify their best guess of the wide small sample estimate for the indifference zone or alternative point. Shrinking from this wide small sample estimate to the halfway point over sample size is the best approach. This approach balances errors between a point null and fixed null wide small sample approach. When the collaborator is completely unable to identify an indifference zone or alternative point then a point null hypothesis should be used instead.

In Table 3.3 we present results from a real dataset in different levels of collaborator uncertainty. The ACTIV-6 clinical trial recently released resulted from the ivermectin trial (Interventions et al., 2022). In this trial 1591 patients with mild to moderate cases of COVID-19 were randomized to ivermectin treatment or placebo. Patients reported each day their symptoms and severity, any health care visits, and medications taken. The endpoint of interest between the two groups is mean time spent unwell and was estimated using a longitudinal ordinal regression model. The results published are that "the difference in the amount of time spent feeling unwell with COVID was estimated to be 0.49 days in favor of ivermectin" with a 95 credible interval of (0.15, 0.82) (Interventions et al., 2022).

The data is especially interesting because time spent unwell is measured in discrete number of days be-

Collaborator Hypothesis		Suggested SGPV Analysis	Outcome
Confident	Confident in a pre-specified scientifically relevant indifference zone, $[\widehat{\theta^-}, \widehat{\theta^+}]$.	Use this indifference zone, $[\widehat{\theta^-}, \widehat{\theta^+}]$.	Ideal
	Confident in a pre-specified scientifically relevant alternative point, $\widehat{\theta_1}$, closest to the point null.	Use the indifference zone that is halfway between the point null, θ_0, and $\widehat{\theta_1}$. Where $\lvert\widehat{\theta_1} - \theta_0\rvert = d$, the indifference zone is $\left[\theta_0 - \frac{d}{2}, \theta_0 + \frac{d}{2}\right]$.	Great
Uncertain	Uncertain of but can identify a conservative pre-specified scientifically relevant indifference zone, $[\widehat{\theta^-}, \widehat{\theta^+}]$.	Use the indifference zone that shrinks from $[\widehat{\theta^-}, \widehat{\theta^+}]$ to $\left[\theta_0 - \frac{d}{2}, \theta_0 + \frac{d}{2}\right]$, with $\lvert\widehat{\theta^-} - \theta_0\rvert = \lvert\widehat{\theta^+} - \theta_0\rvert = d$.	Good
	Uncertain of but can identify a conservative pre-specified scientifically relevant alternative point, $\widehat{\theta_1}$, closest to the point null.	Use the indifference zone that shrinks from $[\theta_0 - d, \theta_0 + d]$ to $\left[\theta_0 - \frac{d}{2}, \theta_0 + \frac{d}{2}\right]$, with $\lvert\widehat{\theta_1} - \theta_0\rvert = d$.	Ok
Unable	Cannot identify a pre-specified scientifically relevant indifference zone; any alternative no matter how close to the point null is meaningful.	In this case do not use SGPV, instead a point null hypothesis should be used. If forced, random indifference zones can be tested until errors are acceptable.	Poor

Table 3.2: Final recommendations for different levels of collaborator uncertainty in SGPV analyses.

Collaborator Hypothesis		Second-generation p-value results	Inference Outcome
Confident	Confident in a pre-specified scientifically relevant indifference zone, $\left[\widehat{\theta^-}, \widehat{\theta^+}\right] = [-1, 1]$.	$p_\delta = 1$	Data are consistent with the null.
	Confident in a pre-specified scientifically relevant alternative point, $\widehat{\theta_1} = 1$ day, closest to the point null. Then $\left[\widehat{\theta^-}, \widehat{\theta^+}\right] = [-0.5, 0.5]$.	$p_\delta = 0.522$	Inconclusive
Uncertain	Uncertain of a wide small sample pre-specified scientifically relevant indifference zone, $\left[\widehat{\theta^-}, \widehat{\theta^+}\right] = [-1.5, 1.5]$. Shrinking at a rate of $n^{-1/2}$ for the sample size $\left[\widehat{\theta^-}, \widehat{\theta^+}\right] = [-0.753, 0.753]$	$p_\delta = 0.900$	Data are almost fully consistent with the null.
	Uncertain of a wide small sample scientifically relevant alternative point, $\widehat{\theta_1} = 0.75$ day or 18 hours, closest to the point null. Shrinking at a rate of $n^{-1/2}$ for the sample size $\left[\widehat{\theta^-}, \widehat{\theta^+}\right] = [-0.377, 0.377]$	$p_\delta = 0.338$	Inconclusive
Unable	Cannot identify a pre-specified scientifically relevant indifference zone; any alternative no matter how close to the point null is meaningful.	Do not run SGPV test.	NA

Table 3.3: Results from COVID-19 ivermectin clinical trial dataset for different levels of collaborator uncertainty in SGPV analyses.

tween 0 and 14 days however the predicted result is only a portion of a day, 0.49 days difference. This measurement is worth statistical discussion and an indifference zone should be chosen carefully. Collaborators and statisticians should decide before looking at the data what the indifference zone is for mean time spent unwell. In Table we show what would happen if the collaborator has different levels of uncertainty for different hypotheses. The reported SGPVs in Table 3.3 are all either inconclusive or consistent with the null. Therefore, as long as the null zone is greater than a difference of 3 hours the inference outcome will not be that ivermectin improved time spent unwell. We are confident given the study design and discrete variable a collaborator would not identify an indifference zone smaller than 0.5 a day or 12 hours.

We have thoughtfully made these recommendations and would like for this conversation to continue in the literature. Shrinking the null zone due to collaborator uncertainty could be beneficial in brain imaging, other

COVID research, or with discrete variables modeled as continuous. Another area this could be beneficial is in new the techniques for SGPV variable selection proposed by Yi Zuo (Zuo et al., 2021, 2022). Collaborator uncertainty should be discussed and evaluated by collaborators and statisticians. Then this information should modify the analysis plan in an appropriate way. Results will then reflect the best version of the underlying truth. This concept will change the current framework of statistical collaboration.

CHAPTER 4

FDRestimation: Flexible False Discovery Rate Computation in R

4.1 Abstract

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

Keywords

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

4.2 Introduction

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

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

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

4.2.1 Simple Motivating Example

We begin with a simple example to fix ideas. Table 4.1 shows five unadjusted (raw) p-values for experimental features along with their corresponding Z-values. The third column lists the Benjamini-Hochberg adjusted p-values to be used for FDR control (Benjamini and Hochberg, 1995). Controlling the FDR at level γ amounts to selecting all of the adjusted p-values in column 3 that are below γ. Note here that the adjusted p-values are monotonically increasing, just like the raw p-values, but inflated.

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

Table 4.1: Example with 5 features using the Benjamini-Hochberg adjustment and assuming a two-sided normal distribution.

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

The fourth column presents FDR estimates for each feature. As we show later, there are several ways to invert the FDR control procedures to yield an estimate of the FDR. Our package performs this inversion for most popular methods. The FDRs here were obtained by inverting the Benjamini-Hochberg FDR control procedure, and so we will refer to them as the BH FDRs (Benjamini and Hochberg, 1995). In practice we find these estimates to be the most context useful when making scientific decisions about which findings to pursue.

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

Because FDRs are only estimates, and because there are a variety of estimation approaches, it helps to have a feature-specific benchmark for each FDR. The fifth column provides such a benchmark; it displays a well-known lower bound on the FDR assuming a Gaussian posterior and a null proportion of 50%. These assumptions are relatively benign for reasons we discuss later and represent a "best-case" scenario. This

benchmark shows two things: (1) the adjusted p-values are a poor substitute for the FDRs, and (2) the smoothness of the FDR estimation approach is important.

4.3 Methods

4.3.1 FDR Methods

4.3.1.1 p-value Based Approaches

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

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

$$p_{(i)} \leq \gamma \frac{i}{m} \text{ for } i \in \{1, 2, ..., m\} \tag{4.1}$$

This can be written compactly $k = \max\left[i : p_{(i)} \leq \gamma i/m\right]$. Then all features with $p_{(1)}, ..., p_{(k)}$ are deemed interesting at the FDR γ threshold and considered "findings". This is called a "step-up" procedure because not all of the rejected features will have unadjusted p-values that meet the above criterion. Only the largest of them must meet that criterion. Because this is a "step-up" procedure, the adjusted p-values will depend on the raw p-values from other features. The Benjamini-Hochberg adjusted p-value for the i^{th} feature is notated in this paper by $\tilde{p}_i$ and defined in Equation (4.2), where $:=$ means "is defined as".

$$\tilde{p}_{(i)} := \min_{j \geq i} \left(\frac{p_{(j)} m}{j} \right) \leq \gamma \tag{4.2}$$

These adjusted p-values are monotone increasing in raw p-value ranking, so one can directly compare $\tilde{p}_i$ to γ to see if a particular feature would be rejected as null for the FDR threshold γ. Importantly, the feature specific FDR estimates need not be monotone. To see this, re-arrange Equation (4.1) as follows in Equation (4.3).

$$FDR_i := \frac{p_i m}{\text{rank}(p_i)} \cdot \hat{\pi}_0 \tag{4.3}$$

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

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

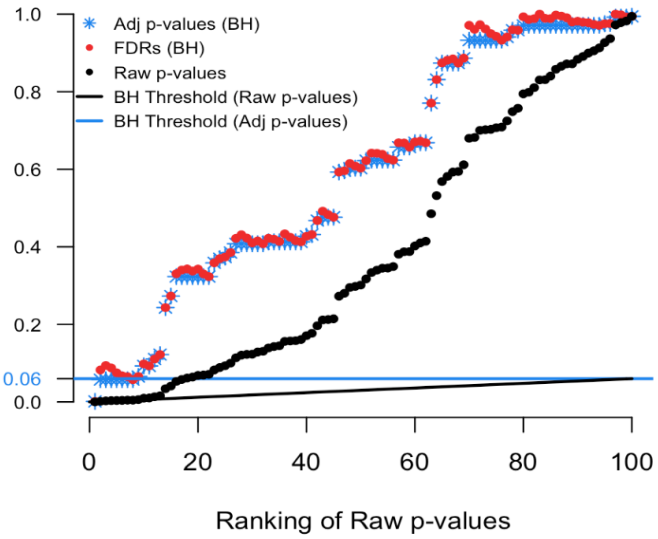

Figure 4.1: Simulated example of raw p-values and the threshold of interest.

To illustrate we simulated real data from 100 hypothesis tests and captured the 100 raw p-values. For context, 80 of these p-values were generated from a uniform distribution (and hence under the null) while the other 20 were generated from a skewed distribution representing the alternative. Results are computed using our `p.fdr` function, which we detail later. The raw p-values are displayed in Figure 4.1 as black points; Figure 4.2 shows only the 20 features with the smallest ranked raw p-values. The black sloped line is the BH rejection line from Equation (4.1). Also included in the plot are the BH adjusted p-values (blue stars), the BH

FDR threshold for interesting findings (blue horizontal line), and the BH FDR estimates (red points).

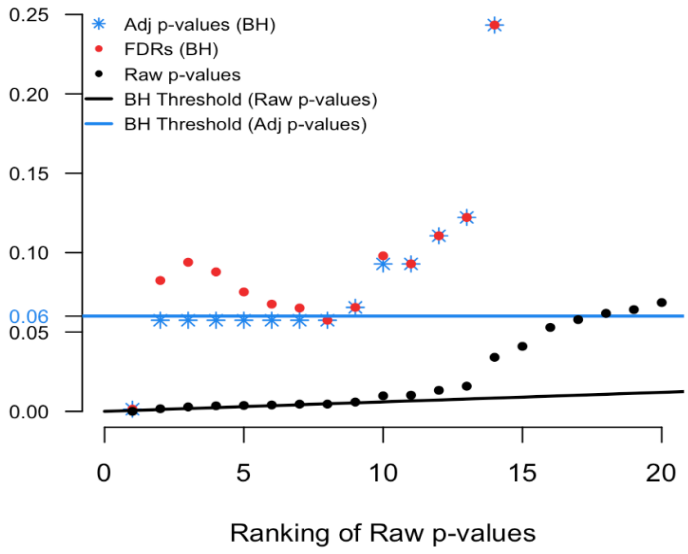

Figure 4.2: Magnified section of Figure 1.

In Figure 4.2 we see that exactly 8 of the adjusted p-values fall below our threshold of interest (blue line, set here to 0.06). Therefore, the BH FDR procedure that controls the group-wise FDR identifies the 8 smallest p-values as interesting findings. However, notice the non-monotonicity of the individual FDRs. Only the first and last of the 8 lowest FDRs are less than 0.06.

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

4.3.1.2 Z-value Based Approaches

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

We begin with the classic two-group model, which assumes each of the m features is either null (distribution known) or alternative (distribution unspecified), but that this status is unknown. As a group the combined data can be used to provide an estimate of the mixture distribution, where the mixing proportion (π_0) is also unknown. Let $f_0(z)$ be the probability density function of the z-values when they come from the true null

distribution and $f_1(z)$ be the probability density function of the z-values when they come from the alternative distribution. Then $F_0(\cdot)$ and $F_1(\cdot)$ denote the probability of rejection for any subset $\mathscr{Z}$ of the real line such that,

$$F_0(\mathscr{Z}) = \int_{\mathscr{Z}} f_0(z)dz \text{ and } F_1(\mathscr{Z}) = \int_{\mathscr{Z}} f_1(z)dz \tag{4.4}$$

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

$$F(\mathscr{Z}) = \pi_0 F_0(\mathscr{Z}) + \pi_1 F_1(\mathscr{Z}) \tag{4.5}$$

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

$$FDR(\mathscr{Z}) \coloneqq Pr\{null | z \in \mathscr{Z}\} = \frac{\pi_0 F_0(\mathscr{Z})}{F(\mathscr{Z})} \tag{4.6}$$

Substituting the natural empirical estimate of the mixture distribution $F(\mathscr{Z})$ results in empirical Bayes estimates the global FDR Equation (4.6) (Benjamini and Hochberg, 1995) (Efron, 2013). For example, the obvious empirical estimate of the mixing distribution function is the step function $\hat{F}(\mathscr{Z}_i) = \text{rank}(p_i)/m$. Notice that the right hand side of Equation (4.1) then looks like $\gamma \cdot \hat{F}(\mathscr{Z}_i)$ or γ times the step function. In some settings smoothing $\hat{F}(\mathscr{Z}_i)$ can be beneficial. Very often it is assumed $\pi_0 = 1$ and $F_0(\mathscr{Z}) = 1 - \Phi(\mathscr{Z})$ for one-sided tests. An advantage of estimating the FDR from the right hand side of Equation (4.6) is that one only needs to accurately estimate the mixture distribution function to get good estimates of the FDR and this does not require the independence of the z-values.

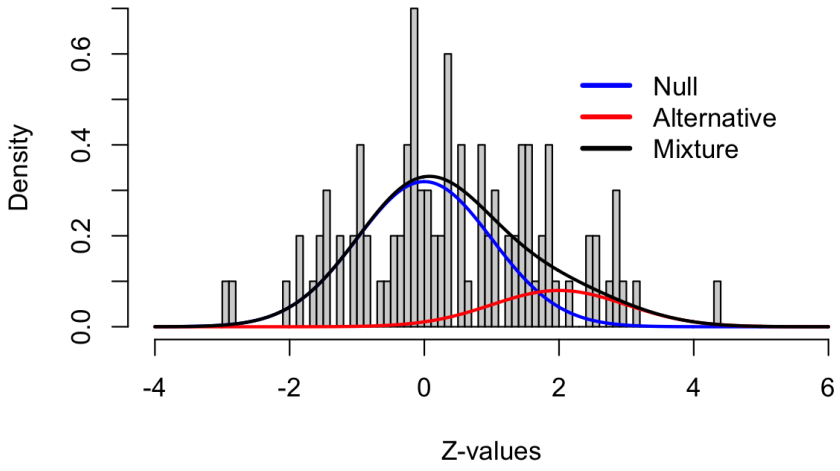

Figure 4.3: Density histogram of the simulated example.

Figure 4.3 and Figure 4.4 show the application of this framework using the same simulated data as in the last example (100 tests, 80 truly null). In the z-space, the null distribution is now the standard normal and the alternative distribution was set to $N(2,1)$ (of course this is unknown, in practice). Figure 4.3 shows these densities overlaid on a histogram of the raw data. The blue curve indicates the null density, the red curve indicates the alternative density, and the black curve is the mixture density with $\pi_0 = 0.8$. Clearly, the blue curve does not fit the histogram well, with a much shorter right tail than the histogram shows. So, assuming all 100 tests come from the null distribution does come with a penalty.

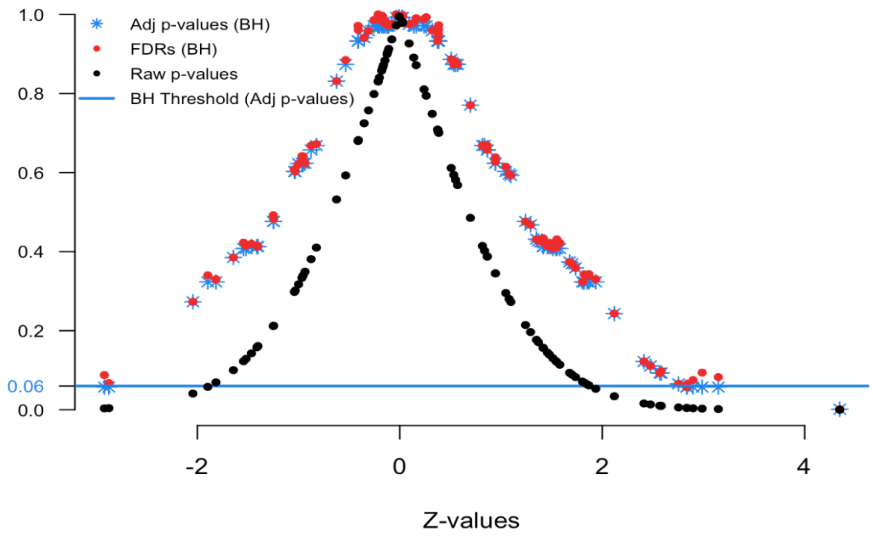

Figure 4.4: FDR simulated Z-values plot.

Figure 4.4 displays the relationship between the Z-values and various FDR quantities. The black dots show the raw p-values (y-axis) versus their Z-value (x-axis); the red dots show the estimated FDRs (y-axis) versus their Z-value (x-axis); and the blue stars show the BH adjusted p-values (y-axis) versus their Z-value (x-axis). This is the comparable plot to Figure 4.1, where the x-axis has been changed from p-value ranking to z-scale. The usefulness of this plot is that is shows what the desired FDR quantity is for a given Z-value. This provides context for our FDRs and adjusted p-values.

Here we see that Z-values greater than 2.85 and less than -2.5 have adjusted p-values less than 0.06 (blue threshold line, horizontal). This means in order to control the group-wise FDR, one would identify features with these Z-values as "interesting". Notice that the Z-value above 4 has a FDR less than 0.06. Also the Z-value of 2.9 has a FDR less than 0.06. In practice, we find that the display in Figure 4.1 is more intuitive for non-statisticians, but that Figure 4.4 provides some essential insight into the stability and smoothness of the estimation procedure.

4.3.1.3 Lower Bound on the FDR

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

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

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

$$P(null|x_1,...x_n) \geq \left(1 + \frac{L(\hat{\theta}_n)}{L(\theta_0)} \frac{\pi_1}{\pi_0}\right)^{-1} \approx \left(1 + \exp(z^2/2)\frac{\pi_1}{\pi_0}\right)^{-1} \tag{4.7}$$

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

For illustration, consider feature 4 in Table 4.1. Feature 4 has an observed p-value of 0.051, but has a univariate gaussian lower bound on the FDR of $0.13 = (1+\exp(1.951^2/2))^(-1)$. In this case the BH estimated FDR is 0.064, substantially below the lower bound. This discrepancy in estimates is due to differing underlying assumptions. In contrast, feature 2 has a p-value of 0.049 and FDR of 0.122, very close to its lower bound. Although feature 4 has nearly the same p-value as feature 2, its BH FDR is nearly half that of feature 2. The univariate gaussian lower bound is helpful for identifying when FDR estimates may be optimistic, as in the case above. Similarly, we see that the adjusted p-values can be much less than the lower bound, which is another reason why they should not be mistaken for FDR estimates.

4.3.2 Adjustment Methods

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

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

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

4.3.2.1 Benjamini-Yekutieli

Benjamini-Yekutieli (BY) is a step-up method for controlling the false discovery rate under arbitrary dependence (Benjamini and Yekutieli, 2001). For a pre-specified dependence structure, there exits an adjustment

function called $c(m)$ that is used to modify the Benjamini-Hochberg estimate of the FDR. For example, in the case of flexible positive dependence, the function $c(m) = \sum_{j=1}^{m} \frac{1}{j}$ is used. Then, the threshold criteria is to find the largest index i such that Equation (4.21) holds, which is a scaled version of the BH criterion given in Equation (4.1).

$$p_{(i)} \leq \gamma \frac{i}{m} \frac{1}{c(m)} \tag{4.8}$$

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

Mathematically, the adjusted p-values and estimated FDRs are

$$\tilde{p}_{(i)}^{BY} := \min_{j \geq i} \left(p_{(j)} \frac{m \cdot c(m)}{j} \right) \leq \gamma \tag{4.9}$$

$$FDR_i^{BY} := p_i \frac{m \cdot c(m)}{\text{rank}(p_i)} \cdot \hat{\pi}_0 \tag{4.10}$$

Comparing this form to the general formula for the FDR in Equation (4.6), we see that the BY correction amounts to changing the estimate of the mixture distribution $F(\mathscr{Z})$ from $[\text{rank}(p_i)/m]$ to $[\text{rank}(p_i)/(m \cdot c(m))]$ to account for dependence. Note that we have avoided using the ordered notation for False discovery rate estimates, say $FDR_{(i)}$, because although those estimates are dependent on ordered p-values the FDR estimates themselves do not have to be monotonic.

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

4.3.2.2 Bonferroni

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

$$\tilde{p}_i^{Bon} := p_i m \leq \gamma \tag{4.11}$$

$$FDR_i^{Bon} := p_i m \cdot \hat{\pi}_0 \tag{4.12}$$

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

4.3.2.3 Sidak

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

$$\tilde{p}_i^{Sid} := 1-(1-p_i)^m \leq \gamma \tag{4.13}$$

$$FDR_i^{Sid} := 1-(1-p_i)^m \cdot \hat{\pi}_0 \tag{4.14}$$

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

4.3.2.4 Holm

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

$$p_{(k)} \leq \frac{\gamma}{m+1-k} \tag{4.15}$$

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

$$\tilde{p}_{(i)}^{Holm} := \max_{j \leq i} \left(p_{(j)}(m+1-j) \right) \leq \gamma \tag{4.16}$$

$$FDR_i^{Holm} := p_i(m+1-\text{rank}(p_i)) \cdot \hat{\pi}_0 \tag{4.17}$$

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

4.3.2.5 Hochberg

The Hochberg method uses the same equation as the Holm method, Equation (4.15) (Hochberg, 1988). However for this method we use the step-up procedure. This means we would reject the null for those rankings $1,...,j$ such that j is the largest ranking where:

$$p_{(j)} \leq \frac{\gamma}{m+1-j} \tag{4.18}$$

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

$$\tilde{p}_{(i)}^{Hoch} := \min_{j \geq i} \left(p_{(j)}(m+1-j) \right) \leq \gamma \tag{4.19}$$

$$FDR_i^{Hoch} := p_i(m+1-\text{rank}(p_i)) \cdot \hat{\pi}_0 \tag{4.20}$$

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

4.3.3 Null Proportion (π_0) Estimation

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

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

4.3.3.1 Last Histogram Height

Under the null, a test statistic for a feature, say a Z-value, is standard normal. As such, the corresponding p-value has a uniform distribution over the unit interval. Therefore, if all the features were null, we would expect an empirical histogram of the observed p-values to be approximately flat. Moreover, we see that the distribution of non-null p-values tends to be shifted toward zero.

The "Last Histogram Height" method uses the bin height of p-values near 1 to estimate the true proportion of null features. We rely on the assumption that larger p-values are more likely to be come from null features.

Let bin heights be $H_1, H_2, ..., H_B$, where B is the total number of bins. When $B = m$ (m is the number of features) and all features are null, we would expect $H_i \approx 1$ for all $i = 1, ..., B$. The caveat is estimating bin height is sensitive to the choice of bin width. However, we have found that Scott's normal reference rule tends to work very well for this method (Scott, 1979).

When $\pi_0 < 1$, the empirical distribution of the p-values (as shown by the histogram) will not be uniform over the unit interval. Departure from the uniform becomes easier to detect as π_0 moves further from 1 because the histogram shape quickly deviates from a uniform appearance. An example is presented in Figure 4.5, which shows a histogram of raw p-values from our simulated example of $m = 1000$ features. The red horizontal line is drawn at the height of the last bin, H_B. In this approach H_B is our "null height" and $H_B \cdot B$ is an estimate of the total number of null features. We then divide that by the number of total features (m) to estimate the null proportion (see Equation (4.21)):

$$\hat{\pi}_0 = \frac{H_B B}{m} \tag{4.21}$$

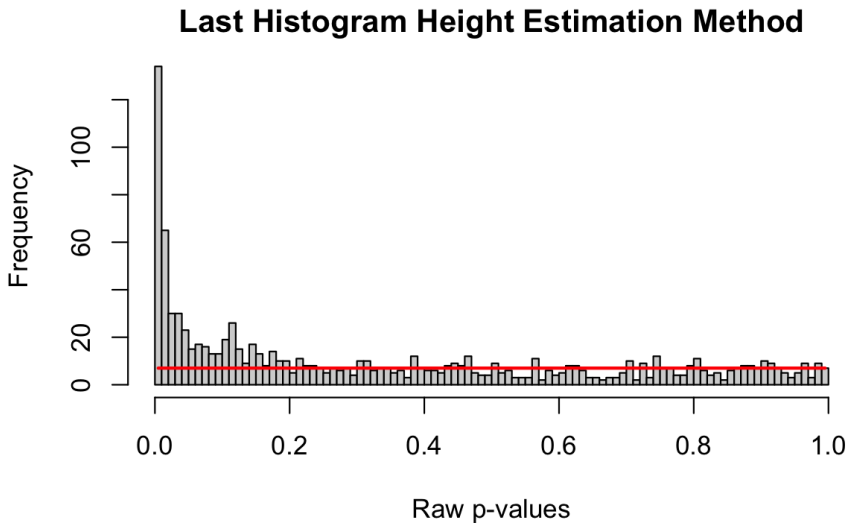

Figure 4.5: Simulated histogram of p-values with horizontal line at the last bin height.

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

small bin width. The "Last Histogram Height" algorithm is as follows:

Algorithm 1 Last Histogram Height Method

Resultado: Null proportion estimate

1. Plot a histogram of the raw p-values, $p_1, p_2, ...p_m$, with B number of bins, where $B < m$
    - The most consistent bin method is `scott`, according to our simulations
2. Store the histogram bin heights H_b for each bin $b = 1, 2, ..., B$
3. Call the height of last bin H_B the "null height"
4. Set the estimate of π_0 to be

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

4.3.3.2 Storey

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

Algorithm 2 Storey's Method

Resultado: Null proportion estimate

1. Let $p_{(1)}, p_{(2)}, ...p_{(m)}$ be the ordered p-values. This also denotes the ordering of the features in terms of their evidence against the null hypothesis.
2. For a range of λ, say $\lambda = 0, 0.05, 0.10, ..., 0.95$, and $i = 1, ..., m$, calculate

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

3. Let $\hat{f}$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ
4. Set the estimate of π_0 to be when $\lambda = 1$:

$$\hat{\pi}_0 = \hat{f}(1)$$

4.3.3.3 Comparison

Below in Figure 4.6 are three plots showing the range of behavior of the six methods for estimating the null proportion that are included in our R package. These plots show the arrogate behavior of each method for estimating π_0 over 1000 simulations where the methods are used on a set of 100 features. A standard normal distribution was used for null features and three different alternative distributions were examined for alternative features (three different plots). The x-axis represents the true π_0 used to generate data and ranges from 0 to 1. The y-axis represents the average estimate π_0 (over the 1,000) simulations) for each of the six methods. Figure 4.7 shows the corresponding mean squared error (MSE) for these simulations.

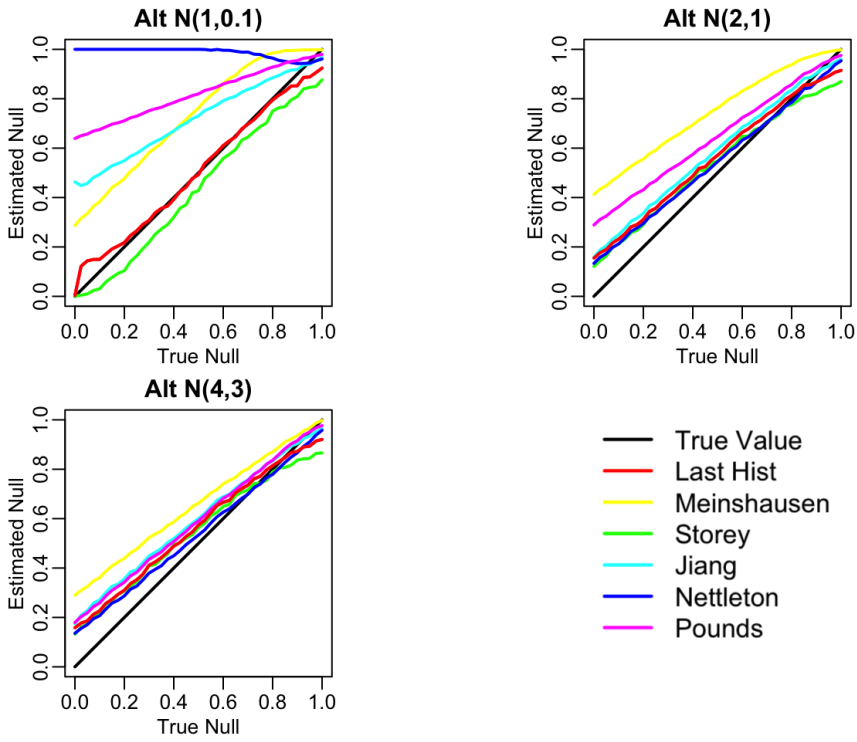

Figure 4.6: Comparison of null proportion estimation methods performance.

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

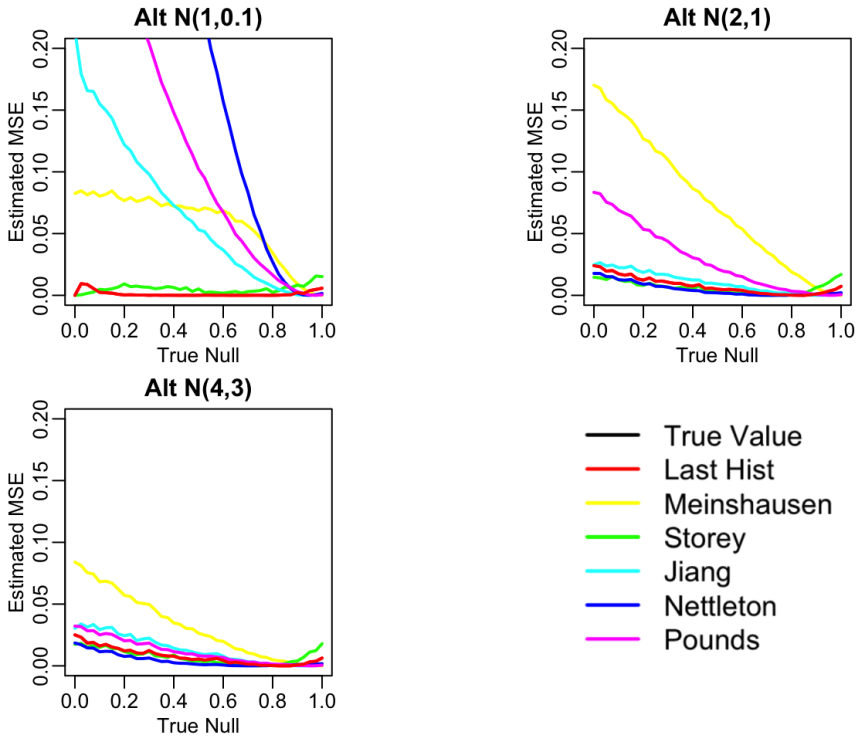

Figure 4.7: Comparison of null proportion estimation methods MSE.

4.3.4 Implementation

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

4.3.5 `p.fdr` Function

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

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

Table 4.2: Inputs to the `p.fdr` function taken directly from the R documentation (R Core Team, 2021).

This `FDRestimation::p.fdr` function allows for the following adjustment methods: Benjamini-Hochberg, Benjamini-Yekutieli (with both positive and negative correlation), Bonferroni, Holm, Hochberg, and Sidak (Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001); Bonferroni (1936); Holm

(1979); Hochberg (1988); Šidák (1967)). It also allows the user to specify the threshold for important findings, the assumed pi_0 value, the desired pi_0 estimation method, whether to sort the results, and whether to remove NAs in the imputed raw p-value vector count (`stats::p.adjust` actually counts NAs as viable features in its Bonferroni adjustment). Table 4.2 shows all of the inputs for this function and their descriptions.

The underlying methods for estimating the null proportion can be set by using the "estim.method" and "set.pi0" arguments. The default value of "set.pi0" is 1, meaning it assumes that all features are null features. Accordingly, this approach will yield conservative estimates of the FDR. Alternatively, and less conservatively, one can attempt to estimate the null proportion from the data. To do this, we recommend using "Last Histogram Height", as it was the simplest routine and one of the most accurate in our simulations (R Core Team, 2021).

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

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

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

$pi0
[1] 1

$threshold
[1] 0.05

$`Adjustment Method`
[1] "BH"

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

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

Figure 4.8: Example of output produced with p.fdr code.

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

```
set.seed(88888)


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

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

# First 5 p-values for Figure 7
p.fdr(p=sim.data.p[1:5], threshold=0.05, adjust.method="BH")
```

The function will return a list object of the `p.fdr` class. In Figure 4.8 we see this list object from the first five *p*-values for with the following components (R Core Team, 2021).

- **fdrs** A numeric vector of method adjusted FDRs.
- **Results Matrix** A numeric matrix of method adjusted FDRs, method adjusted *p*-values, and raw *p*-values.
- **Reject Vector** A vector containing Reject.H0 and/or FTR.H0 based off of the threshold value and hypothesis test on the adjusted *p*-values.
- **pi0** A numeric value for the pi0 value used in the computations.
- **threshold** A numeric value for the threshold value used in the hypothesis tests.
- **Adjustment Method** The string with the method name used in computation(needed for the plot.fdr function).

4.3.6 `get.pi0` Function

The `get.pi0` function is used to estimate the null proportion from the raw *p*-values. The user can choose one of six different methods included in our function: Last Histogram Height, Storey, Meinshausen, Jiang, Nettleton, and Pounds (Storey and Tibshirani (2003); Meinshausen et al. (2006); Jiang and Doerge (2008); Nettleton et al. (2006); Pounds and Morris (2003)). The user may also change the methods of determining the number of histogram breaks, which is an essential component for many of the methods implemented here. Table 4.3 shows function arguments and their descriptions.

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

Table 4.3: Inputs for the `get.pi0` function taken directly from the R documentation (R Core Team, 2021).

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

```
set.seed(88888)

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

[1] 0.8
```

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

[1] 0.85
```

4.3.7 `plot.p.fdr` Function

This `plot.p.fdr` function is used to plot the results of `p.fdr`. By default, the adjusted FDRs, adjusted *p*-values and raw *p*-values are plotted along with two threshold lines to help contextualize the points. Any combination of *p*-values and thresholds can be removed from the plot. The user can set the axis limits, the location of the legend, the title of the plot and the plotting symbols and colors. Table 4.4 shows all the function arguments and their descriptions.

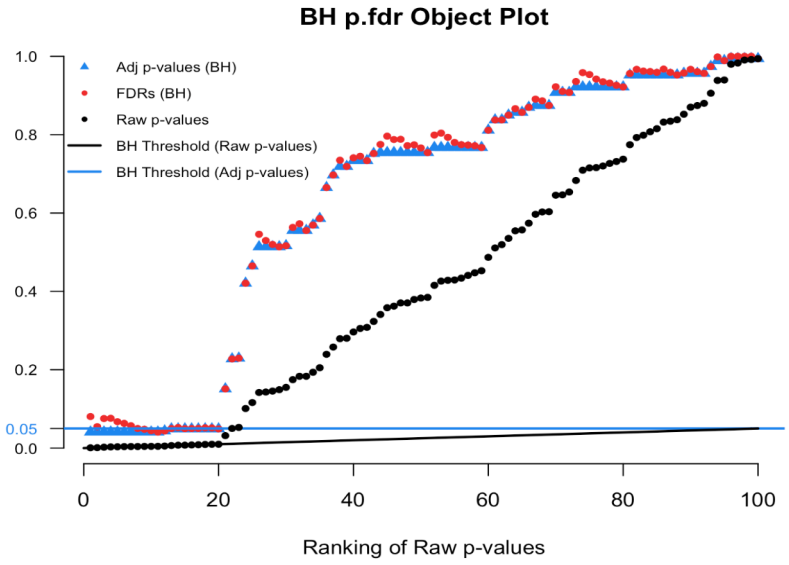

Figure 4.9: Benjamini-Hochberg p.fdr plot.

Here we see an example of the `plot.p.fdr` function in R. We used our simulated data `sim.data.p`, where the a true null proportion was 80%, for illustration. Figure 4.9 show the default plot, and Figure 4.10 zooms in on an interesting subset of findings.

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

Table 4.4: Inputs for the `plot.p.fdr` function taken directly from the R documentation (R Core Team, 2021).

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

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

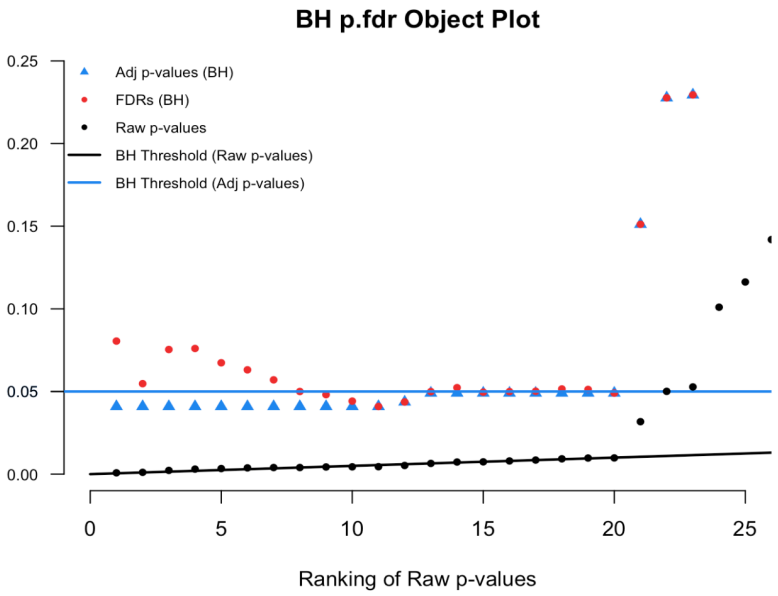

Figure 4.10: Magnified section of Figure 8.

4.3.8 Operation

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

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

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


# Load the package
library(FDRestimation)
```

4.4 Conclusions

We encourage the use of FDR methods and desire to illuminate the importance of contextualizing important findings. Our package provides useful and easy tool for those want to compute the false discovery rate, analogous to the role that `stats::p.adjust` plays for multiple comparison adjustments in everyday

practice. Importantly, we hope it is now clear that *p*-value adjustments are not interchangeable with FDRs. In addition, `FDRestimation` package clearly delineates between methods for FDR control and methods for FDR estimation, while still allowing the user to choose from many different inputs and assumptions for their data. The more flexibility the user has at their disposal with these methods, better interpretations and applications will result.

Data availability

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

Software availability

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

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

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

License: MIT + file LICENSE

CHAPTER 5

Conclusion

This dissertation aims to improve transparency and accessibility of statistical analyses, specifically for the methods of second-generation *p*-values (SGPVs) and false discovery rates (FDRs). First, we identify SGPV as the most flexible and easy-to-use method for establishing statistical equivalence with an interval null hypothesis when compared to the TOST test. Reasons for favoring SGPV include the benefit of a third inference outcome, ease of interpretation, clear statistical properties, and the amount overlap between intervals being used in the reported *p*-value. Second, we also propose a technique of shrinking the indifference zone over sample size in SGPV analyses to address collaborator uncertainty. In the case where a collaborator is uncertain in the hypothesis but can estimate a wide interval, shrinking the interval over sample size can beneficially balance the power and errors. This paper opens up a new discussion of how statisticians should discuss hypotheses with collaborators during the study planning phase. Third, we present an R package "FDRestimation" for flexible and transparent false discovery rate computation for classical *p*-values. It distinguishes between estimated FDRs and adjusted *p*-values for many different published adjusted methods.

The following points summarize the key contributions of this work. These contributions improve current statistical methodology and encourage transparent collaboration with other researchers.

- Our comparison clearly identifies SGPV as the most flexible and easy-to-use statistical method to establish statistical equivalence when compared to equivalence tests, like TOST.
- We present a solution to adjust for collaborator uncertainty in SGPV analyses and encourage statisticians to discuss hypothesis uncertainty with collaborators.
- We introduce a user-friendly R package to compute and distinguish between FDRs and adjusted *p*-values. This package can be used to account for multiple testing in high-dimensional data scenarios like genomics.

References

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52(6):708–721. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.200900299.

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

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319. Publisher: Institute of Mathematical Statistics.

Blume, J., Greevy, R., Welty, V., Smith, J., and Dupont, W. (2019). An Introduction to Second-Generation p -Values. *The American Statistician*, 73:157–167.

Blume, J. D., McGowan, L. D., Dupont, W. D., and Jr, R. A. G. (2018). Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLOS ONE*, 13(3):e0188299. Publisher: Public Library of Science.

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

Cohen, M. P. (2021). Why Not an Interval Null Hypothesis? *Journal of Data Science*, 17(2):383–390. Publisher: School of Statistics, Renmin University of China.

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

Ennis, D. M. and Ennis, J. M. (2010). Equivalence hypothesis testing. *Food Quality and Preference*, 21(3):253–256. Place: Netherlands Publisher: Elsevier Science.

Fisher, R. A. (1959). *Statistical methods and scientific inference.* Hafner, New York. OCLC: 1516472.

Fryar, C. D., Kruszon-Moran, D., Gu, Q., and Ogden, C. L. (2018). Mean body weight, height, waist circumference, and body mass index among adults: United states, 1999-2000 through 2015-2016. *Natl Health Stat Report*, (122):1–16.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36. Publisher: Radiological Society of North America.

Hauck, W. W. and Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, 12(1):83–91.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Hung, H. M. J., O'Neill, R. T., Bauer, P., and Kohne, K. (1997). The Behavior of the P-Value When the Alternative Hypothesis is True. *Biometrics*, 53(1):11–22. Publisher: [Wiley, International Biometric Society].

Interventions, A. C.-. T., Group, V. A.-. S., and Naggie, S. (2022). Ivermectin for treatment of mild-to-moderate covid-19 in the outpatient setting: A decentralized, placebo-controlled, randomized, platform clinical trial. *medRxiv*.

Jiang, H. and Doerge, R. (2008). Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer informatics*, 6:117693510800600001.

Kirkwood, T. B. L. and Westlake, W. J. (1981). Bioequivalence Testing – A Need to Rethink. *Biometrics*, 37(3):589–594. Publisher: [Wiley, International Biometric Society].

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280. Publisher: SAGE Publications Inc.

Lakens, D. and Caldwell, A. (2022). TOSTER: Two One-Sided Tests (TOST) Equivalence Testing.

Lakens, D. and Delacre, M. (2018). Equivalence Testing and the Second Generation P-Value. Technical report, PsyArXiv. type: article.

Mandallaz, D. and Mau, J. (1981). Comparison of Different Methods for Decision-Making in Bioequivalence Assessment. *Biometrics*, 37(2):213–222. Place: United States Publisher: Biometric Society.

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

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference*, 26:231–245.

Miller, R. G. (1981). Miscellaneous Techniques. In Miller, R. G., editor, *Simultaneous Statistical Inference*, pages 211–229. Springer, New York, NY.

Nettleton, D., Hwang, J. G., Caldo, R. A., and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p values. *Journal of agricultural, biological, and environmental statistics*, 11(3):337.

Neyman, J., Pearson, E. S., and Pearson, K. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337. Publisher: Royal Society.

Perlman, M. D. and Wu, L. (1999). The Emperor's new tests. *Statistical Science*, 14(4):355–369. Publisher: Institute of Mathematical Statistics.

Pounds, S. and 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.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robbins, H. (1970). Statistical Methods Related to the Law of the Iterated Logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409. Publisher: Institute of Mathematical Statistics.

Rogers, J. L., Howard, K. I., and Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3):553–565. Num Pages: 553-565 Publisher: American Psychological Association (US).

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(6):657–680.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Seaman, M. A. and Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, 3(4):403–411. Num Pages: 403-411 Place: Washington, US Publisher: American Psychological Association (US).

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

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

Tukey, J. W. (1953). "the problem of multiple comparisons.". *Mulitple Comparisons*.

Welty, V., Irlmeier, R., Stewart, T., Greevy, R., Jr, McGowan, L. D., and Blume, J. (2020). sgpv: Calculate Second-Generation p-Values and Associated Measures.

Zuo, Y., Stewart, T. G., and Blume, J. D. (2021). Variable Selection With Second-Generation P-Values. *The American Statistician*, 0(0):1–11. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2021.1946150.

Zuo, Y., Stewart, T. G., and Blume, J. D. (2022). ProSGPV: an R package for variable selection with second-generation p-values. Technical Report 11:58, F1000Research. Type: article.