NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

ONLINE LARGE-SCALE HYPOTHESIS TESTING WITH CORRUPTED DATA

by

Victor Benicio Ardilha da Silva Alves

June 2024

Thesis Advisor:	Roberto Szechtman
Co-Advisor:	Louis Chen
Second Reader:	Jefferson Huang

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2024	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE ONLINE LARGE-SCALE HYPOTHESIS TESTING WITH CORRUPTED DATA	**5. FUNDING NUMBERS**
6. AUTHOR(S) Victor Benicio Ardilha da Silva Alves	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT	**12b. DISTRIBUTION CODE**
Distribution Statement A. Approved for public release: Distribution is unlimited.	A

13. ABSTRACT (maximum 200 words)

Modern statistical inference involves processing extensive datasets, with multiple hypothesis testing being one methodology to draw conclusions on many features at once. Control of the false discovery rate (FDR) is essential. Target classification via satellite imagery and acoustic signal processing are examples in military applications where false detections can be costly for a Command and Control framework. These contexts also showcase another layer of complexity: the volume of data is often processed online, with decisions having to be made sequentially on evolving, incomplete datasets. This underscores the need for FDR control in an online environment. Current methods for online FDR control are successful in this regard; however, they are not designed with data error or, worse, data corruption in mind. This research will explore the level of robustness of the Levels Based On Recent Discovery (LORD) algorithm. The fundamental objective is to learn how to corrupt data and make it robust against such corruption efficiently. This work will draw insights from studying corruption-robust bandit algorithms and aim to advance the adversarial online multiple-hypothesis testing field.

14. SUBJECT TERMS false discovery rate, power, data corruption, cascading effect, LORD algorithm	**15. NUMBER OF PAGES** 81
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

ONLINE LARGE-SCALE HYPOTHESIS TESTING WITH CORRUPTED DATA

Victor Benicio Ardilha da Silva Alves
Capitão de Corveta, Brazilian Navy
BNS, Brazilian Navy Academy, 2011

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2024

Approved by: Roberto Szechtman
Advisor

Louis Chen
Co-Advisor

Jefferson Huang
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Modern statistical inference involves processing extensive datasets, with multiple hypothesis testing being one methodology to draw conclusions on many features at once. Control of the false discovery rate (FDR) is essential. Target classification via satellite imagery and acoustic signal processing are examples in military applications where false detections can be costly for a Command and Control framework. These contexts also showcase another layer of complexity: the volume of data is often processed online, with decisions having to be made sequentially on evolving, incomplete datasets. This underscores the need for FDR control in an online environment. Current methods for online FDR control are successful in this regard; however, they are not designed with data error or, worse, data corruption in mind. This research will explore the level of robustness of the Levels Based On Recent Discovery (LORD) algorithm. The fundamental objective is to learn how to corrupt data and make it robust against such corruption efficiently. This work will draw insights from studying corruption-robust bandit algorithms and aim to advance the adversarial online multiple-hypothesis testing field.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

ADDIS	Adaptive Algorithm that Discards Conservative Nulls
AJB	Brazilian Jurisdictional Waters
ANP	National Agency of Petroleum, Natural Gas and Biofuels
BAF	Brazilian Air Force
BH	Benjamini and Hochberg
BN	Brazilian Navy
CDF	Cumulative Density Function
FDP	False Discovery Proportion
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
GAI	Generalized Alpha-Investing
LORD	Levels Based On Recent Discovery
mFDR	Marginal False Discovery Rate
PDF	Probability Density Function
PEM	Strategic Plan of the Brazilian Navy
SAFFRON	Serial Estimate of the Alpha Fraction that is Futilely Rationed on True Null Hypotheses
SAR	Search and Rescue
SisGAAz	Blue Amazon Management System

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

This thesis examines the robustness of the Levels Based On Recent Discovery (LORD) algorithm when exposed to corrupted data, particularly within critical real-time processing environments like the Brazilian Navy's Blue Amazon Management System (SisGAAz). Our study reveals that maintaining the integrity of statistical testing is crucial, mainly where decision-making depends on the accuracy of data analysis conducted online.

Our research identifies and rigorously evaluates effective mitigation strategies against probabilistic data corruption scenarios. Key findings highlight the robust efficacy of "phantom" rejections and the strategic integration of the LORD algorithm with the online Benjamini and Hochberg (BH) algorithm, a variation adapted from the traditional offline BH method. These approaches, we assert, maintain testing power significantly, even under adversarial manipulations, instilling confidence in their effectiveness.

We propose a controlled adversarial setup involving two entities: "Blue," the defender who aims to make true discoveries, and "Red," the attacker focused on data corruption. Our analysis investigates several attack scenarios. The first is a singular anticipated attack that manipulates the first true discovery and traditionally triggers a cascade effect, countered by adjusting the decay rate of each test level to buffer against such disruptions. Additionally, we explore multiple p-value corruption scenarios where strategically placed "phantom" rejections can reclaim compromised testing power, although this strategy faces practical challenges due to the necessity of predicting attack probabilities. Lastly, indiscriminate attacks on any p-value show that integrating the LORD algorithm with the online BH algorithm is exceptionally effective, maintaining the algorithm's robustness even amidst widespread corruption.

The thesis concludes that while prevalent algorithms are adequate for handling FDR in trustworthy data scenarios, their effectiveness diminishes under adversarial data manipulation, a common issue in real-time data environments. Our findings suggest that enhancing algorithmic robustness against data corruption supports reliability in statistical testing and contributes to broader research and application in adversarial conditions. We propose new avenues for future investigation, such as exploring data corruption impacts on other exist-

ing algorithms and developing a “pure” algorithm. This new algorithm could offer a more robust alternative to the current mixed approach, providing a stronger defense against data manipulation.

Acknowledgments

I am deeply grateful to my wife, Priscilla, for constant support and encouragement throughout my thesis work. Her patience and understanding were vital in keeping me motivated every day.

I also want to express my love for my newborn child, Sophia. Her presence has brought immense joy and inspiration into my life.

Completing this research was truly a team effort. I am thankful for my advisor and co-advisor, whose guidance and support were crucial in accomplishing this task. I sincerely appreciate their dedication.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

Hypothesis testing is a cornerstone of empirical research, offering a systematic framework to validate theories and assumptions across diverse fields. This statistical tool has been pivotal in various domains, ranging from medical research, where it aids in determining the effectiveness of new treatments, to environmental science, where it is employed to assess the impact of human activities on ecosystems. It is also used in A/B testing, an approach used to compare two versions of a web page, application feature, or other product offerings to determine which one performs better in terms of specific metrics. In finance, it is instrumental in evaluating investment strategies and market trends. These applications underscore its versatility and critical role in evidence-based decision-making.

In the military context, particularly within the Brazilian Navy (BN), the significance of hypothesis testing takes on a more strategic dimension. The BN, entrusted with safeguarding Brazil's sovereignty and maritime interests, finds itself at the forefront of an era where data-driven decision-making is preeminent. The Blue Amazon Management System (SisGAAz) exemplifies this paradigm shift as a sophisticated surveillance mechanism designed to protect Brazil's vast maritime jurisdiction, known as the "Blue Amazon." SisGAAz is designed to leverage large-scale datasets and sophisticated statistical inference for immediate conclusions, especially in classifying targets. In such a system, controlling the False Discovery Rate (FDR) is critical, as errors in classification can lead to detrimental impacts on further (i.e., downstream) decision-making. While prevalent algorithms effectively manage FDR in trustworthy data scenarios, their effectiveness in environments susceptible to adversarial data manipulation remains unexamined.

The overarching goal of this thesis is to analyze and enhance the robustness of online multiple hypothesis testing methodologies to corruption. By investigating strategies for efficient data corruption targeted at the Levels Based On Recent Discovery (LORD) algorithm and developing countermeasures against different corruption scenarios, the study aims to strengthen the reliability and integrity of online multiple hypothesis testing.

1.1 Background

Brazil's sea and inland waterways, crucial for the nation's well-being, necessitate effective protection. In this regard, the Strategic Plan of the Brazilian Navy (PEM) is instrumental, steering both medium- and long-term strategic planning via Naval Objectives. Then, the Naval Strategic Actions are meticulously crafted by dissecting these objectives, delineating a clear execution strategy to fulfill the overarching mission of the BN.

According to the Third United Nations Convention on the Law of the Sea (UNCLOS III), Brazil has property rights and sovereignty in the Brazilian Jurisdictional Waters (AJB) up to 200 nautical miles. Beyond that, the nation also has the extension of the soil and subsoil of the submarine areas, defined by the limits of the continental Shelf. This area encompasses about 5.7 million square kilometers rich in natural resources, accounting for approximately 95% of Brazil's oil and 83% of the country's natural gas (Andrade and Franco 2018). Andrade et al. (2021), citing a 2002 report from the National Agency of Petroleum, Natural Gas and Biofuels (ANP), asserted that Brazil's reserves of these resources amounted to 9.81 billion barrels, and most of this, 8.87 billion barrels, originated from the AJB. In a related discussion, Husseini (2018) mentioned the discovery in 2006 of substantial oil deposits located beneath a salt layer approximately 2,000 meters thick, under a layer of sediment of similar thickness in the Santos Basin, about 300 kilometers southeast of Brazil's coast. More than two decades later, based on a new report released by the ANP, Smith (2023) highlighted that July's production from the pre-salt layers constituted 75% of Brazil's total oil output for that month. This substantial share underscores Brazil's capacity to emerge as the fourth-biggest oil producer worldwide.

Additionally, Andrade et al. (2021) claimed that data from the Brazilian Institute of Geography and Statistics reveals that a significant portion of Brazil's population, approximately 80%, resides within 200 kilometers of the coast. Consequently, according to these authors, this coastal proximity is a hub for economic activity, containing about 90% of the country's infrastructure and industrial production and roughly 80% of overall production. Furthermore, the oceans and river basins play a vital role as an intercommunicating element: 90% of the volume of this trade is made by sea (Rodrigues 2021).

The AJB are not merely conduits for transporting commodities; they represent an extensive reservoir of biodiversity and natural resources that are pivotal for the nation's advancement. The exigency for safeguarding and conserving these waters as a legacy for succeeding generations is of key importance (Rodrigues 2021).

1.1.1 Blue Amazon

The PEM explains the Blue Amazon concept, a term that the BN has spread to raise awareness among society and national institutions about the importance of the AJB's protection, a domain as vast as the Amazon rainforest. Figure 1.1 should not be perceived merely as an area encompassing the sea surface, waters overlying the seabed, and marine soil and subsoil within the Atlantic extension from the coast to the outer limit of the Brazilian continental Shelf. Rather, according to the Navy's plan document, it should be understood as a multifaceted concept embodying four distinct aspects:

1. Sovereignty – linked to the roles of the BN, which represents the authority of the state and oversees the use of force at sea.
2. Scientific – addresses the opportunities for research and technological advancement, the economic impact of using marine biodiversity, and the importance of maintaining knowledge about the maritime environment. Naval forces can use this knowledge to protect the interests of their respective nations.
3. Environmental – adopts a stance that goes beyond mere regulatory matters, considering that the unbroken expanse of oceanic areas and the movement of ocean currents enhance the risk of introducing and spreading non-native species and activities that endanger the marine ecosystem. This includes the need for mechanisms to monitor and tackle pollution, whether by accident or deliberate.
4. Economic – related to national development, based on the wealth of living and non-living resources in the AJB and the importance of maritime transportation for foreign trade.

Figure 1.1. Blue Amazon. Source: Gerhardinger et al. (2018).

Regarding the aspect of sovereignty, the BN undertakes strategic programs aligned with its institutional mission—the preparation and deployment of naval power as a component of national defense. Andrade et al. (2021) articulated that these initiatives are instrumental in overseeing and administrating the Blue Amazon.

Central to these projects is SisGAAz, a system primarily aimed at extensively monitoring and managing the BN's area of responsibility, enhancing the Navy's capability to respond to challenges, including threats, hostilities, illicit activities, emergencies, and ecological crises.

Consequently, Andrade et al. (2021) concluded that this system will bolster the situational awareness of national authorities in these zones, elevating their monitoring and regulatory capabilities and strengthening their surveillance and protection of these maritime areas.

1.1.2 Blue Amazon Management System

The discussion paper titled "Blue Amazon Management System (SisGAAz): Sovereignty, Surveillance, and Defense of the Brazilian Jurisdictional Waters" developed by researchers from the Institute for Applied Economic Research "aims to demonstrate the importance of developing and implementing SisGAAz to monitor the Blue Amazon. It also discusses the implications of its reformulation and the possible alternatives" (Andrade et al. 2021).

The Blue Amazon Management System (SisGAAz) project, initiated in 2009, as explained by the scholars, was designed to fulfill the need for effective monitoring, surveillance, and defense within the Blue Amazon. The study claimed that its objective is to establish a unified and cohesive monitoring system that leverages and integrates others, improving the application of resources that are already in place instead of creating something entirely new. The research also noted that it will enable data gathering, analysis, and generating supportive information for decision-making, ultimately facilitating informed decisions that will guide the deployment of available resources (protection). Figure 1.2 illustrates SisGAAz's conception.

According to the study, in terms of its integration with other platforms, SisGAAz will be interconnected with various systems both within and outside the BN, including the Military Command and Control System of the Ministry of Defense, which encompasses the Brazilian Army Integrated Border Monitoring System and the Brazilian Aerospace Defense System of the Brazilian Air Force (BAF). Additionally, the authors pointed out that SisGAAz will integrate with institutions outside the national defense realm, including those affiliated with ministries such as Finance, Transportation, Mines and Energy, Science and Technology, and Justice, as well as regulatory bodies and corporations.

Moreover, the paper stated further that the system will also receive data from various external sources, such as over-the-horizon radar, maritime patrol aircraft from the BAF, and unmanned aerial vehicles, and with systems from other nations and global entities, like the International Maritime Organization's Long Range Identification System (LRIT) and the Trans-Regional Maritime Network (T-RMN).

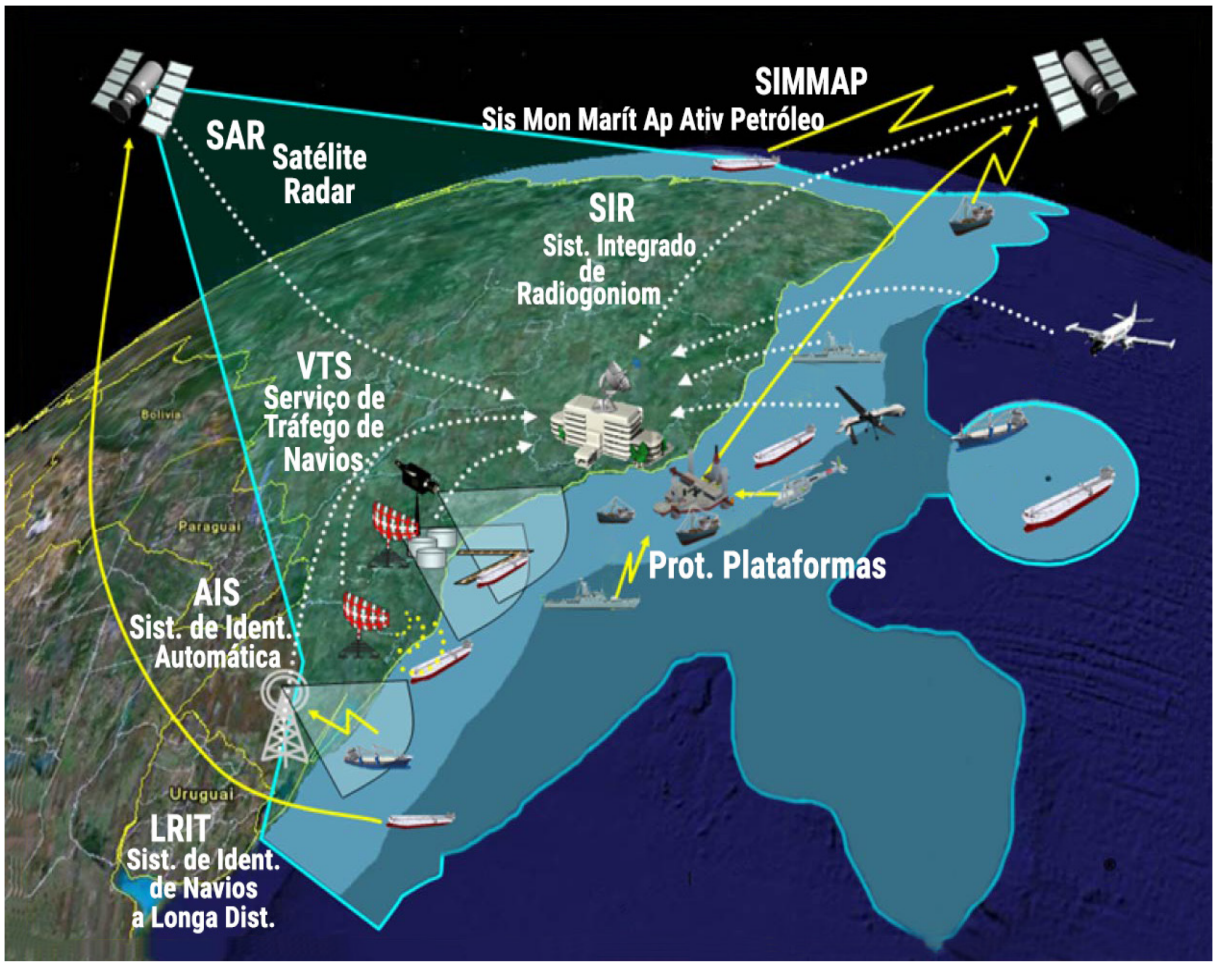

Figure 1.2. Operation and functioning of SisGAAz. Source: Andrade et al. (2021).

SisGAAz, tasked with the vigilant surveillance of the Blue Amazon, is instrumental in the continuous collection and analysis of comprehensive data to safeguard Brazil's maritime domain. It operates as an unceasing sentinel, meticulously sifting through current and historical data—unperturbed by the potential for data corruption—to categorize each vessel

as either suspect or non-suspect. This ongoing vigilance mirrors the principles of online hypothesis testing, whereby the system dynamically assesses new (online) information to make informed decisions.

1.2 Study Objective

This thesis investigates data corruption on online multiple hypothesis testing, along with mitigating strategies. While the literature on this topic is extensive, there remains a gap in developing tools or methodologies specifically designed for environments vulnerable to data corruption.

At the heart of this work is introducing a novel procedure tailored for scenarios where data corruption is likely. This approach aims to provide a robust framework for maintaining the integrity and reliability of hypothesis testing despite the challenges posed by data corruption.

1.3 Thesis Organization

This thesis is organized into five chapters: an introduction, a literature review, a comparison among main online hypothesis testing algorithms with the methodology section, different scenarios of online hypothesis testing with corrupted data and simulation results, and a conclusion.

In Chapter II, the literature review summarizes the concepts of hypothesis testing, reviews the main algorithms, and clearly distinguishes between an offline and an online setting.

Chapter III presents the LORD, Adaptive Algorithm that Discards Conservative Nulls (ADDIS), and Serial Estimate of the Alpha Fraction that is Futilely Rationed on True Null Hypotheses (SAFFRON) algorithms for online multiple hypothesis testing and the methodology for generating the data necessary for all simulations.

Chapter IV examines different data corruption scenarios to show its impact on power and FDR, and new procedures are proposed to mitigate it.

Chapter V summarizes the findings and suggests possible future research.

CHAPTER 2:
Literature Review

In this chapter, we comprehensively review the literature surrounding hypothesis testing, starting from the fundamental concepts of single hypothesis testing and extending to the more intricate and continually evolving domain of online multiple hypothesis testing. Through this detailed review, we will highlight these methods' significance, advantages, and limitations, offering a balanced and insightful perspective on this critical statistical tool.

2.1 Single Hypothesis Testing

In "Probability and Statistics for Engineering and the Sciences," Devore (2015) claimed that a statistical hypothesis, often referred to simply as a hypothesis, represents a statement or assertion regarding the value of a single parameter, of multiple parameters, or of the shape of an entire probability distribution. The book emphasized that, in hypothesis testing, there are typically two conflicting hypotheses to examine: the null hypothesis (H_0), the initially assumed claim, and the alternative hypothesis (H_a), which contradicts H_0. Using data from a sample, the author remarked that the prime objective is to determine which of these two hypotheses is true, pointing out that the null hypothesis will only be discarded in favor of the alternative hypothesis if the evidence from the sample strongly suggests that H_0 is incorrect. Consequently, hypothesis testing has two potential outcomes: rejecting H_0 or failing to reject H_0.

The statistician Fisher (1970) defined p-value as "the probability of the observed result, plus more extreme results if the null hypothesis were true" (p.66). This means that the p-value serves as a critical piece of information in hypothesis testing, quantifying the strength of evidence *against* the null hypothesis. Then, a p-value smaller than α (the test's significance level chosen by the analyst) suggests strong evidence against H_0, while a p-value greater than α suggests weaker evidence and the inability to reject H_0.

To illustrate this concept, Figure 2.1 shows a standard normal distribution's Probability Density Function (PDF) with p-value and α representing areas under the curve. For this example, the observed data x is used to decide whether H_0 should be rejected.

Then, since the p-value is less than α, we reject H_0.

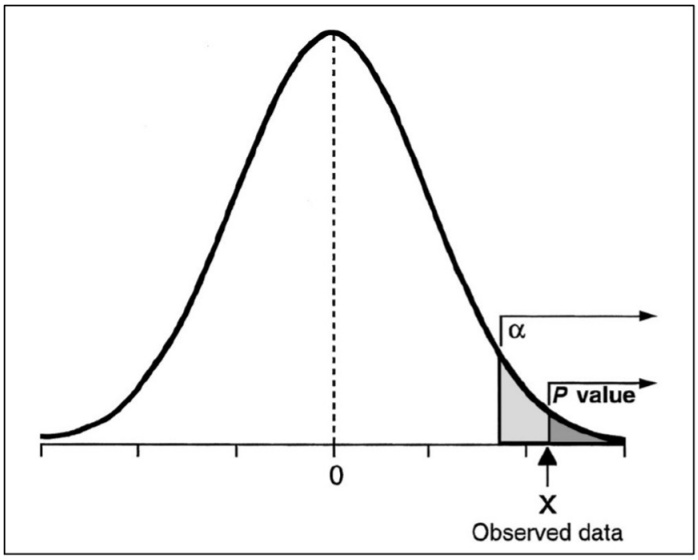

Figure 2.1. Probability of every possible outcome under the null hypothesis $H_0 = 0$ and the alternative $H_0 > 0$. Since p-value $\leq \alpha$, H_0 is rejected. Source: Goodman (1999).

Mathematically, Pr is the probability distribution of the observed data x under H_0, for any value of α between 0 and 1. Efron (2010) defined a rejection region R_α as

$$\Pr\{x \in R_\alpha\} = \alpha. \tag{2.1}$$

Moreover, the p-value $p(x)$ is also defined as the smallest α such that $x \in R_\alpha$:

$$p(x) = \inf_{\alpha}\{x \in R_\alpha\}. \tag{2.2}$$

Since the area under the curve for any PDF always equals 1, $p(x)$ conforms to a uniform distribution across the interval (0, 1):

$$p(x) \sim U(0, 1). \tag{2.3}$$

In this context, Austil et al. (2014) characterized the two potential mistakes that can be made: type I and type II errors. A type I error, also known as a false positive, happens when we reject the null hypothesis even though it is true. On the other hand, a type II error, or false negative, occurs when we fail to reject the null hypothesis when the alternative hypothesis is true. Table 2.1 summarizes the potential results.

Table 2.1. Possible outcomes for single hypothesis testing.

	Stated as True	Stated as False
True null hypothesis	Correct (1 - α)	Type I error (α)
False null hypothesis	Type II error (β)	Correct (1 - β)

Moreover, the authors emphasized that the test's significance level α is typically chosen to limit the probability of a type I error to a predetermined level, and the main objective is to maximize power (i.e., $1 - \beta$), while ensuring the probability of a type I error remains at the intended level.

2.2 Multiple Hypothesis Testing

In various fields where statistics are employed, military applications included, decisions are made by assessing many hypotheses. In these scenarios, as outlined by Austil et al. (2014), single hypothesis testing procedures are ineffective because the probability of committing at least one type I error significantly exceeds the nominal significance level employed for each test. The authors demonstrated that for N number of independent tests, with α as the threshold for each p-value, the probability of not committing any type I errors is $(1 - \alpha)^N$. Given that α falls between 0 and 1:

$$(1 - \alpha)^N < (1 - \alpha). \tag{2.4}$$

Hence, they deduced that when conducting multiple tests, the likelihood of avoiding any type I errors becomes significantly reduced compared to when only one test is performed. As a result, the chances of committing at least one type I error increase with the number of tests conducted. This situation highlights the increased complexity of controlling the rate of false positives while effectively managing the type I error rate in multiple testing scenarios.

In the literature, the most relevant type I error rates are the Family-Wise Error Rate (FWER) and the FDR. They will be described in the following subsections.

2.2.1 Family-Wise Error Rate

Austil et al. (2014) explained that the initial strategies developed to adjust for multiple hypotheses primarily focused on managing the FWER, defined as "the probability of committing at least one false rejection when all null hypotheses are true" (p.3). These early methods did not aim to keep the probability of a Type I error constant for each individual test; instead, they focused on maintaining the resultant FWER across all tests. Therefore, for V as the number of false rejections:

$$\text{FWER} = \text{Prob}\left(V \geq 1\right). \tag{2.5}$$

Nevertheless, they highlighted a significant drawback: controlling the FWER often results in overly conservative approaches, leading to tests with low power.

Bonferroni Inequality

Given the set of N multiple independent null hypotheses $H_1, \ldots, H_N$, the corresponding p-value p_i of each hypothesis, and the desired test significance level α, Austil et al. (2014) stated that the probability of at least one rejection is less than or equal to the sum of their marginal probabilities, and the suitable form of the inequality for $0 \leq \alpha \leq 1$ is:

$$\text{Prob}\left(\bigcup_{i=1}^{N}\left\{p_i \leq \frac{\alpha}{N}\right\}\right) \leq \alpha. \tag{2.6}$$

Bonferroni (1930) proposed the primary method based on this inequality and developed a straightforward correction to control the FWER. Menyhart et al. (2021) explained the two approaches to compute the adjusted p-values introduced by Bonferroni. The first approach divides α by the number of hypotheses to be tested N, and only p-values smaller than the adjusted α are considered statistically significant:

$$H_i \text{ is rejected if } p_i \leq \frac{\alpha}{N}. \tag{2.7}$$

As an alternative, in the second approach, the p-value of each test is multiplied by N ($\widetilde{p_i} = p_i N$). As a result, if the adjusted p-value is less than α, the null hypothesis is rejected:

$$H_i \text{ is rejected if } \widetilde{p_i} \leq \alpha. \tag{2.8}$$

Menyhart et al. (2021) argued that, though the Bonferroni adjustment is the most widely used procedure, it effectively controls the FWER at the desired level only when the quantity of statistical tests remains within several dozen to a few hundred.

Other researchers applied the concept proposed by Bonferroni. The Sidak procedure, introduced by Sidak (1971), controls the FWER more conservatively than the Bonferroni process does but still relies on the assumption of independence among individual tests. For independent hypotheses H_i:

$$H_i \text{ is rejected if } p_i \leq (1-\alpha)^{\frac{1}{N}}. \tag{2.9}$$

In addition, the equivalent adjusted p-value is defined as

$$\widetilde{p_i} = 1 - (1-p_i)^N. \tag{2.10}$$

Because $\frac{\alpha}{N} < 1 - (1-\alpha)^{\frac{1}{N}}$, this method exhibits slightly higher statistical power compared to Bonferroni's. However, as Abdi (2007) observed, the latter is more commonly used due to its simpler calculation method.

Furthermore, Holm (1979) developed an early instance of a step-down procedure, leading to a more potent and advanced strategy than Bonferroni's. Given the ordered p-values $p_{(1)} \leq \ldots \leq p_{(N)}$:

$$H_{(i)} \text{ is rejected if } p_{(j)} \leq \frac{\alpha}{N-j+1} \text{ for } j = 1, \ldots, i. \tag{2.11}$$

Simes Inequality

Simes (1986) proposed another method to control FWER. Given the ordered independent p-values $p_{(1)}, \ldots, p_{(N)}$, such as $p_{(1)} \leq \ldots \leq p_{(N)}$, for corresponding null hypotheses testing $H_{(1)}, \ldots, H_{(N)}$:

$$\text{Prob}\left(p_{(i)} \geq \frac{i\alpha}{N}\right) = 1 - \alpha. \tag{2.12}$$

Employing this inequality, Simes devised a straightforward rule for multiple testing:

$$H_{(i)} \text{ is rejected if } p_{(i)} \leq \frac{i\alpha}{N}. \tag{2.13}$$

Notably, as per Sarkar and Chang (1997), in the case of multivariate distributions displaying a type of positive dependence—a common occurrence in various multiple hypothesis testing situations—the Simes method adeptly manages and controls the probability of a type I error.

Two frequently used methods that also use the Simes inequality were developed by Hochberg (1988) and Hommel (1988). Hochberg's approach closely resembles Holm's proposed method, with the key distinction being its formulation as a step-up procedure. Moreover, as demonstrated in Hochberg's research, it has greater statistical power than Holm's. Once more, this analysis involves the use of ordered p-values :

$$\begin{aligned} &\text{Let } k \text{ be the largest } i \text{ for which } p_{(i)} \leq \frac{i\alpha}{N+1-\alpha}. \\ &\text{Reject all } H_{(i)} = 1, \ldots, k. \end{aligned} \tag{2.14}$$

Hommel, in his exploration, introduced an alternative algorithm that offers enhanced statistical power while necessitating only a marginally increased complexity in its implementation. Hommel's procedure is a variation, albeit a less widely adopted one:

$$\begin{aligned} &\text{Compute } j : \max\{i \in \{1, \ldots, N\}\} : p_{(N+k-1)} \leq \frac{k\alpha}{i} \text{ for } k = 1, \ldots, i\}. \\ &\text{If no maximum exists, reject all null hypotheses. Else, reject } \{H_i : p_i \leq \frac{\alpha}{j}\}. \end{aligned} \tag{2.15}$$

2.2.2 False Discovery Rate

Since it was introduced by Benjamini and Hochberg (1995), the concept of FDR has become a prominent focus in statistical research and remains the prevailing method applied, apparently attaining the "accepted methodology" status in scientific subject-matter journals (Efron 2010).

Following the explanation provided by Austin et al. (2014), the FDR is defined as "the expected proportion of rejected hypotheses that have been wrongly rejected" (p.5). The authors explained that if all hypotheses come from the null, the FDR and the FWER are equal. However, in a mixed model, with alternative hypotheses among the overall hypotheses to be tested, the FDR is smaller than or equal to the FWER. Moreover, they concluded, techniques that regulate the FWER will inherently regulate the FDR, and, since managing the FDR is less rigorous contrasted to controlling the FWER, FDR procedures have more statistical power.

In comparison, FWER methods focus on avoiding any type I errors, displaying caution even when some false positives might be acceptable, tending to be overly conservative as the number of tests increases. On the other hand, FDR methods, by considering the proportion of false positives among all rejected hypotheses, take a more permissive approach, with tolerance for some false positives, and are often more adaptable to large-scale studies.

Robertson et al. (2022) defined the False Discovery Proportion (FDP) up to time t, considering $R(t)$ as the number of rejected tests, and $V(t)$ as the number of falsely rejected hypotheses:

$$\mathrm{FDP}(t) = \frac{V(t)}{R(t) \vee 1}, \tag{2.16}$$

where $R(t) \vee 1 = \max(R(t), 1)$.

The FDR is the expectation of the FDP:

$$\mathrm{FDR}(t) = \mathbb{E}\{\mathrm{FDP}(t)\}. \tag{2.17}$$

Benjamini and Hochberg (BH) developed a method to maintain the FDR under a predetermined threshold, and, in line with Benjamini and Yekutieli (2001), the BH procedure is effective not just with independent tests but also with positive regression dependence on

those test statistics associated with the true null hypotheses. For Javanmard and Montanari (2018), this approach is advantageous in situations with a high number of true discoveries, particularly when numerous non-null hypotheses exist. The algorithm, instead of controlling the probability of a type I error at a set level for each test, controls the overall FDR at level α in (0,1):

$$\begin{aligned} &i_{\max} \text{ is the greatest index for which } p_{(i)} \leq \frac{i}{N}\alpha. \\ &\text{Reject all } H_{(i)} \text{ where: } i \leq i_{\max}. \end{aligned} \tag{2.18}$$

The BH procedure, enhanced with certain improvements, continues to be the leading approach in the field of multiple hypothesis testing (Javanmard and Montanari 2018).

2.3 Online Multiple Hypothesis Testing

Javanmard and Montanari (2018) asserted that standard FDR control methods, like the BH procedure, require the presence of all p-values under consideration before any discoveries are made. For them, this implies that decisions are made only after all the necessary data has been gathered. However, they argued that this approach is unfeasible in several applications better suited to an online hypothesis testing framework. The study defined online hypothesis testing as follows: "Hypotheses arrive sequentially in a stream. At each step, the analyst must decide whether to reject the current null hypothesis without having access to the number of hypotheses (potentially infinite) or the future p-values but solely based on the previous decisions" (Javanmard and Montanari 2018, p. 527)

More formally, the authors considered a sequence of hypotheses $H_1, \ldots, H_N$ arriving sequentially in a stream, as depicted in Figure 2.2, with p-values $p_1, \ldots, p_N$. The primary objective remains to keep the FDR under a predefined threshold α. A desired testing procedure offers, they proclaimed, a series of significance levels α_i with the following decision rule:

$$R_i = \begin{cases} 1, & \text{if } p_i \leq \alpha_i \text{ (reject } H_i), \\ 0, & \text{otherwise (accept } H_i). \end{cases} \tag{2.19}$$

Furthermore, each α_i depends on prior outcomes:

$$\alpha_i = \alpha_i(R_1, R_2, \ldots, R_{i-1}). \tag{2.20}$$

Offline multiple hypotheses testing

Online multiple hypotheses testing

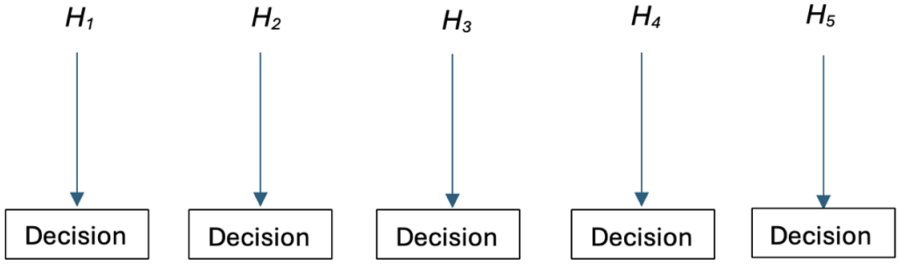

Figure 2.2. Offline and online FDR control. Decisions are made after all hypotheses have been available versus conclusions made sequentially for each incoming hypothesis online. Source: Jordan (2019).

The alpha-investing algorithm, first presented by Foster and Stine (2008), marked the beginning of online rate management techniques. According to Aharoni and Rosset (2014), the alpha-investing method focuses on controlling the marginal false discovery rate mFDR_η at level α for any given choice of η and α, a variant of the FDR. The mFDR_η is defined as

$$\text{mFDR}_\eta = \frac{\mathbb{E}\{\text{V}(t)\}}{\mathbb{E}\{\text{R}(t)\} + \eta}. \tag{2.21}$$

Based on this study, the approach diverges from the alpha-spending concept employed in Bonferroni-type corrections, where we start with an allowance for type I error, or the initial

α-wealth for a series of tests. At each time point t, a test is conducted at level α_t, reducing the α-wealth $W(t)$ by α_t, which means $W(t) = W(t-1) - \alpha_t$. Setting $W(0) = \alpha$ and ensuring $\forall j : W(j) \geq 0$ guarantee that $\sum_j \alpha_j \leq \alpha$. This approach ensures the control of the FWER at level α. However, as the authors pointed out, despite controlling the FWER, alpha-spending methods are often critiqued for their limited power and conservatism in multiple hypothesis testing. As an alternative to overcome this drawback, the alpha-investing rule earns a reward for each rejected null hypothesis. Aharoni and Rosset (2014) defined the wealth at any time point t as

$$W(t) = W(t-1) - (1 - R_t)\frac{\alpha_t}{1 - \alpha_t} + R_t\omega, \tag{2.22}$$

where $W(0) = \alpha\eta$.

If H_t is true, then $\frac{\alpha_t}{1-\alpha_t}$ is reduced from the wealth. If H_t is rejected, a reward ω is gained. By convention, ω is usually set to the maximal allowed value $\omega = \alpha$.

The research also extended the alpha-investing method to Generalized Alpha-Investing (GAI) algorithms. The potential function, previously known as alpha-wealth, operates as follows:

$$W(t) = W(t-1) - \phi_t + R_t\psi_t, \tag{2.23}$$

where $W(0) = \alpha\eta$.

Moreover, Aharoni and Rosset (2014) emphasized an important distinction: in the original alpha-investing, the quantity $\frac{\alpha_t}{1-\alpha_t}$ is deducted from the wealth only if the hypothesis H_t is not rejected. In contrast, in the GAI approach, ϕ_t is subtracted regardless of the test outcome.

Figure 2.3 summarizes this alpha-investing concept: when setting the initial FDR level, the algorithm is allocated a certain "initial wealth" W_0. At each time point t, the alpha-wealth $W(t)$ decreases by ϕ_t. If the hypothesis H_t is rejected ($R_t = 1$), then $W(t)$ is increased by ψ_t.

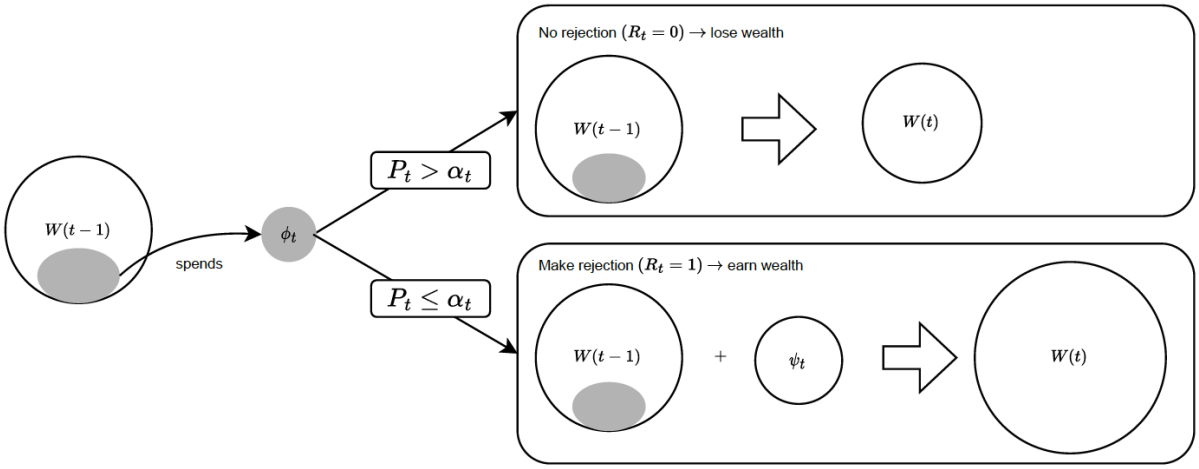

Figure 2.3. GAI representation showing how the wealth $W(t)$ changes depending on whether the hypothesis H_t is rejected. Source: Robertson et al. (2023).

Suppose that θ_t is the actual parameter value for test t, and $\mathcal{H}^0$ is the null hypothesis space, which includes all parameter values that would lead to the null hypothesis not being rejected. If $\theta_t \notin \mathcal{H}^0$, then it is in the alternative hypothesis space, and, according to Aharoni and Rosset (2014), the best power of the t-th test is defined as

$$\rho_t = \sup \text{Prob}_{\theta_t}(R_t = 1). \tag{2.24}$$

In simpler terms, this function finds the maximum probability that a test can achieve when it correctly rejects a null hypothesis across all possible alternative parameter values.

For the GAI method, as explained by the academics, any choice for the parameters α_t, ϕ_t, ψ_t is valid, as long as $W(t)$ does not become negative, meaning $\phi_t \leq W(t-1)$, and

$$\forall t : 0 \leq \psi_t \leq \min\left(\frac{\phi_t}{\rho_t} + \alpha, \frac{\phi_t}{\alpha_t} + \alpha - 1\right) \tag{2.25}$$

where ρ_t is the best power of the t-th test.

Javanmard and Montanari (2017) presented alternative versions of the GAI algorithms, which are designed to control the FDR, in contrast to the mFDR proposed by Foster and

Stine (2008). As described in Ramdas et al. (2017), the parameter B_0 and proved for monotone GAI rules and under independence, with $B_0 = \alpha - W_0$, the FDR is controlled. Now, for some user-defined B_0

$$\psi_t \leq \min\left\{\phi_t + B_0, \frac{\phi_t}{\alpha_t} + B_0 - 1\right\}. \tag{2.26}$$

Ramdas et al. (2017) also defined a class of improved GAI algorithms called GAI ++: the initial wealth W_0 is set to be $0 \leq w_0 \leq \alpha$ and the payout satisfies $\psi_t \leq \min\left\{\phi_t + b_t, \frac{\phi_t}{\alpha_t} + b_t - 1\right\}$, a modified version of Equation 2.26, where $b_t = \alpha - w_0 \mathbb{1}\{R(t-1) = 0\}$. As they demonstrated, any monotone GAI++ rule comes with the following guarantee:

Theorem 1 *If the null p-values (i.e., the p-values corresponding to the true null hypotheses) are independent of all other p-values, any monotone GAI++ rule satisfies the bound* $\mathbb{E}\left[\frac{V(t)+W(t)}{R(t)\vee 1}\right] \leq \alpha$ *for all* $t \geq 1$. *Since* $W(t) \geq 0$, *the FDR is controlled at level* α.

Finally, Javanmard and Montanari (2017) conceptualized the Levels Based On Recent Discovery (LORD) algorithm, an instance of GAI algorithms. Later enhanced by Ramdas et al. (2017), the so-called LORD++ (henceforth LORD) is widely considered one of the most advanced techniques in online multiple hypothesis testing.

This chapter highlighted the significance of managing error rates as the number of hypotheses tested grows. Initially, it explored the key strategies for handling the FWER and the FDR, with the latter being widely used. Finally, it presented the fundamental principles of online hypothesis testing, the focus of this thesis. A detailed discussion of the main online algorithms will follow in the next chapter.

CHAPTER 3:
Online Multiple Hypothesis Testing Algorithms

This chapter is dedicated to an in-depth explanation of the principal algorithms used in online hypothesis testing. It is our intent to meticulously delineate essential definitions and foundational concepts, thereby constructing a robust framework for comprehending these algorithms and the overarching methodology employed throughout this thesis.

3.1 LORD

We follow Ramdas et al. (2017) to explain the LORD algorithm. Given a sequence of p-values, the decisions (rejections or non-rejections) $R_1, \ldots, R_t$, where each R_i is an indicator of whether the i-th hypothesis is rejected. The decision at time t is adapted to the sequence of decisions until $t-1$ (meaning that it can depend on them); we store this information via $\mathcal{F}^t = \sigma(R_1, \ldots, R_{t-1})$. The same applies to the rejection thresholds $\alpha_t \in [0,1]$; they are adapted to the history up to $t-1$, which means $\alpha_t = f_t(R_1, \ldots, R_{t-1})$, where f_t is an arbitrary [0,1]-valued function of the first $t-1$ decisions.

If the hypothesis H_i is truly null, its corresponding p-value has a $U(0,1)$ distribution, so the p-value is unlikely to take on very small values. By definition, these p-values are *super-uniformly distributed*, meaning that

$$\text{Prob}\left\{p_t \leq \alpha_t \mid \mathcal{F}^{t-1}\right\} \leq \alpha_t, \text{ or equivalently, } \mathbb{E}\left[\frac{\mathbb{1}\{p_t \leq \alpha_t\}}{\alpha_t}\middle| \mathcal{F}^{t-1}\right] \leq 1. \tag{3.1}$$

The interpretation is that the probability of the t-th null p-value p_t being less than or equal to its corresponding threshold α_t is at most α_t, given the past information $\mathcal{F}^{t-1}$.

Ramdas et al. (2017) also defined, given any non-negative predictable sequence $\{\alpha_t\}$, the oracle FDP:

$$\text{FDP}^*(t) = \frac{\sum_{j \le t, j \in \mathcal{H}^0} \alpha_j}{R(t)}. \tag{3.2}$$

The interpretation is that the expected number of null (i.e., false) rejections up to time t is approximately the sum of all α_j, for $j \le t$ and H_j be a null hypothesis.

Since $\text{FDP}^*(t)$ cannot be calculated because the contents of $\mathcal{H}^0$ are unknown, a conservative estimate of the oracle FDP is

$$\widehat{\text{FDP}}_{\text{LORD}}(t) = \frac{\sum_{j=1}^{t} \alpha_j}{R(t)}. \tag{3.3}$$

The implication is that $\widehat{\text{FDP}}_{\text{LORD}}(t)$ overestimates the unknown FDP(t):

$$\widehat{\text{FDP}}_{\text{LORD}}(t) \ge \frac{\sum_{j \le t, j \in \mathcal{H}^0} \alpha_j}{R(t)} \approx \frac{\sum_{j \le t, j \in \mathcal{H}^0} \mathbb{1}\{p_j \le \alpha_j\}}{R(t)} = \text{FDP}(t). \tag{3.4}$$

The authors declared that a more straightforward approach to developing online FDR methods is to guarantee that $\sup_{t \in \mathbb{N}} \widehat{\text{FDP}}_{\text{LORD}}(t) \le \alpha$, eliminating the need for wealth, penalties, and rewards, as seen in the GAI procedure.

Based on these definitions, the following theorem is proved:

Theorem 2 *(a) If the null p-values are conditionally super-uniform, then the condition* $\widehat{FDP}_{LORD}(t) \le \alpha$, $\forall t \in \mathbb{N}$, *implies that* $mFDR(t) \le \alpha$, $\forall t \in \mathbb{N}$. *(b) If the null p-values are independent of each other and of the p-values corresponding to the non-null hypotheses, and* $\{\alpha_t\}$ *is chosen to be a monotone function of past rejections, then the condition* $\widehat{FDP}_{LORD}(t) \le \alpha$, $\forall t \in \mathbb{N}$, *implies that* $FDR(t) \le \alpha$, $\forall t \in \mathbb{N}$.

Leveraging this theorem, Ramdas et al. (2017) presented the LORD algorithm: given an infinite, non-increasing sequence of positive constants $\{\gamma_t\}_{t=1}^{\infty}$ that sums to one, and τ_j as the time of j-th rejection, the test level α_t is

$$\alpha_t = w_0\gamma_t + (\alpha - w_0)\gamma_{t-\tau_1}\mathbb{1}\{\tau_1 < t\} + \alpha \sum_{j:\tau_j<t,\tau_j\neq\tau_1} \gamma_{t-\tau_j}. \tag{3.5}$$

As explained by Robertson et al. (2022), the initial term $w_0\gamma_t$ represents the portion of the starting wealth w_0 allocated to the t-th test, while the subsequent terms are the gains from previous rejections before t that are used in round t: the reward for the first rejection is $(\alpha - w_0)$, and for subsequent rejections is α. Once these earnings are received, they are allocated to future rounds according to the same constants $\{\gamma_t\}$, shifted to start at the next instant. Ramdas et al. (2017) showed this rule ensures LORD always operates within its earned resources and maintains $\widehat{\text{FDP}}_{\text{LORD}}(t) \leq \alpha$. For them, default values are $w_0 = \frac{\alpha}{10}$ and $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, the former calculated in a Gaussian setting to maximize power.

3.2 SAFFRON

Ramdas et al. (2018) argued that the main limitation of $\widehat{\text{FDP}}_{\text{LORD}}$ arises when the (unknown) ground truth involves a substantial proportion of non-nulls. In such cases, $\widehat{\text{FDP}}_{\text{LORD}}$ becomes very conservative overestimation of FDP*, and consequently, of the true unknown FDP.

Bearing this limitation in mind and considering that non-nulls have a larger-than-uniform probability of smaller p-values, Ramdas et al. (2018) suggested a new estimator:

$$\widehat{\text{FDP}}_{\text{SAFFRON}(\lambda)}(t) \equiv \widehat{\text{FDP}}_{\lambda}(t) = \frac{\sum_{j\leq t} \alpha_j \frac{\mathbb{1}\{p_j>\lambda_j\}}{1-\lambda_j}}{R(t)}. \tag{3.6}$$

As the authors explained, $\{\lambda_j\}_{j=1}^{\infty}$ is a sequence of values in [0,1], adapted to the available information up to time $j-1$. To facilitate the analysis, λ_j is made constant for all j. The SAFFRON algorithm relies on the concept that the numerator of Equation 3.6 is a less conservative estimator for $\sum_{j\leq t, j\in\mathcal{H}^0} \alpha_j$ than $\sum_{j=1}^{t} \alpha_j$ used in LORD.

Ramdas et al. (2018) introduced a new indicator for candidacy $C_j = \mathbb{1}\{p_j \leq \lambda_j\}$, where p-

values with $C_j = 1$ are called candidates, and α_t is updated to $\alpha_t = f_t(R_1, \ldots, R_{t-1}, C_1, \ldots, C_{t-1})$. Furthermore, Equation 3.1 has a modified form:

$$\begin{aligned} &\text{Prob}\left\{p_t \leq \alpha_t \mid \mathcal{F}^{t-1}\right\} \leq \alpha_t, \text{ or equivalently,} \\ &\mathbb{E}\left[\frac{\mathbb{1}\{p_t > \alpha_t\}}{1-\alpha_t}\middle|\mathcal{F}^{t-1}\right] \geq 1 \geq \mathbb{E}\left[\frac{\mathbb{1}\{p_t \leq \alpha_t\}}{\alpha_t}\middle|\mathcal{F}^{t-1}\right]. \end{aligned} \tag{3.7}$$

The main result is somewhat different compared to LORD's Theorem 2:

Theorem 3 *(a) If the null p-values are conditionally super-uniform, then the condition* $\widehat{FDP}_\lambda(t) \leq \alpha$, $\forall t \in \mathbb{N}$, *implies that* $mFDR(t) \leq \alpha$, $\forall t \in \mathbb{N}$. *(b) If the null p-values are independent of each other and of the p-values corresponding to the non-null hypotheses, and* $\{\alpha_t\}$ *is chosen to be a monotone function of the vector* $R_1, \ldots, R_{t-1}, C_1, \ldots, C_{t-1}$, *then the condition* $\widehat{FDP}_\lambda(t) \leq \alpha$, $\forall t \in \mathbb{N}$, *implies that* $FDR(t) \leq \alpha$, $\forall t \in \mathbb{N}$.

Utilizing this theorem, they introduced the SAFFRON algorithm: given a desired FDR level α, the user should set a constant $\lambda \in (0, 1)$, an initial wealth $w_0 < (1-\lambda)\alpha$, and a positive non-increasing sequence $\{\gamma_t\}_{t=1}^{\infty}$ that sums to one. Considering the candidates after the j-th rejection C_{j+} as $C_{j+}(t) = \sum_{i=\tau_j+1}^{t-1} C_i$, where τ_j is the time of the j-th rejection, the test level α_t, $\forall t \geq 2$, is

$$\alpha_t = \min\{\lambda, \tilde{\alpha}_t\}, \tag{3.8}$$

where $\tilde{\alpha}_t = W_0\gamma_{t-C_{0+}} + ((1-\lambda)\alpha - W_0)\gamma_{t-\tau_1-C_{1+}} + \sum_{j\geq 2}(1-\lambda)\alpha\gamma_{t-\tau_j-C_{j+}}$.

For $t = 1$, $\alpha_1 = \min\{\lambda_1 W_0, \lambda\}$.

SAFFRON begins with an alpha-wealth of $(1-\lambda)w_0$, preserves wealth when testing candidate p-values, and increases wealth by $(1-\lambda)\alpha$ for every rejection after the first. Ramdas et al. (2018) advised as acceptable default option $\lambda = 0.5$ and $\gamma_t \propto t^{-1.6}$.

Lastly, the study remarked that with a significant presence of non-nulls and strong signals, SAFFRON has more power than LORD does.

3.3 ADDIS

Tian and Ramdas (2019) established that SAFFRON's increased power is evident only if the p-values are uniformly distributed under the null hypothesis. However, as these authors argued, in real-world scenarios where conservative null hypotheses are often encountered, SAFFRON may exhibit less power than LORD. They developed ADDIS to compensate for this power loss.

In hypothesis testing, it is always assumed that the p-value p is valid. In other words, if the null hypothesis is true, $\text{Prob}\{p \leq x\} \leq x$ for all $x \in (0, 1)$. Ideally, one expects $\text{Prob}\{p \leq x\} = x$, indicating a uniform distribution, but, in practice, we often encounter a stricter inequality. The p-value is called *conservative*, meaning that small p-values are less likely to occur under the null hypothesis than they would under a uniform distribution.

Formally, the authors present the definition of *uniformly conservative* null hypotheses:

$$\text{Prob}\left\{\frac{p}{c} \leq x \mid p \leq c\right\} \leq x \text{ for all } x, c \in (0, 1). \tag{3.9}$$

For instance, Zhao et al. (2019) showed that in a one-dimensional exponential family with parameter θ, if the actual parameter θ is strictly less than θ_0, the uniformly most powerful test for testing $H_0 : \theta \leq \theta_0$ against $H_a : \theta > \theta_0$ results in uniformly conservative nulls.

Adding complexity in comparison to the SAFFRON algorithm, besides indicators for candidacy $C_j = \mathbb{1}\{p_j \leq \lambda_j\}$ and for rejection $R_j = \mathbb{1}\{p_j \leq \alpha_j\}$, ADDIS introduces a new indicator $S_j = \mathbb{1}\{p_j \leq \eta_j\}$, where $S_j = 1$ indicates p-value p_j was selected (not discarded) for testing. The authors argue that α_t, λ_t and η_t are some function f_t of $\{R_{1:t-1}, C_{1:t-1}, S_{1:t-1}\} \to [0, 1]$.

Tian and Ramdas (2019) proposed a new estimator:

$$\widehat{\text{FDP}}_{\text{ADDIS}}(t) = \frac{\sum_{j \leq t} \alpha_j \frac{\mathbb{1}\{\lambda_j < p_j \leq \eta_j\}}{\eta_j - \lambda_j}}{R(t)} \equiv \frac{\sum_{j \leq t} \alpha_j \frac{\mathbb{1}\{p_j \leq \eta_j\}\mathbb{1}\{p_j/\eta_j > \theta_j\}}{\eta_j(1-\theta_j)}}{R(t)}, \tag{3.10}$$

where $\theta_j = \frac{\lambda_j}{\eta_j}$.

As these researchers made clear, given the users' defined sequences $\{\gamma_j\}_{j=1}^{\infty}$ and $\{\eta_j\}_{j=1}^{\infty}$,

such that $\gamma_j < \eta_j$ for all j, the numerator of Equation 3.10 is a preferable estimator for $\sum_{j \leq t, j \in \mathcal{H}^0} \alpha_j$ to what $\widehat{\text{FDP}}_{\text{SAFFRON}}(t)$ uses. Moreover, theorems 2 and 3 are customized to reflect ADDIS:

Theorem 4 *If the null p-values are uniformly conservative, and suppose we choose α_j, λ_j and η_j such that $\eta_j > \lambda_j \geq \alpha_j$ for each $j \in \mathbb{N}$, then: (a) the condition $\widehat{FDP}_{ADDIS}(t) \leq \alpha$, $\forall t \in \mathbb{N}$, implies that $mFDR(t) \leq \alpha$, $\forall t \in \mathbb{N}$. (b) If the null p-values are independent of each other and of the p-values corresponding to the non-null hypotheses, and α_t, γ_t and $1 - \eta_t$ are chosen to be monotonic functions of the past for all t, then the condition $\widehat{FDP}_{ADDIS}(t) \leq \alpha$, $\forall t \in \mathbb{N}$, implies that $FDR(t) \leq \alpha$, $\forall t \in \mathbb{N}$.*

In light of this, and for simplicity, with λ and η constants for all t, Tian and Ramdas (2019) introduced the ADDIS algorithm: given a desired FDR level α, the user should set a discarding threshold $\eta \in (0, 1)$, a candidate threshold $\lambda \in (0, \eta)$, an initial wealth $w_0 \leq \alpha$, and a positive non-increasing sequence $\{\gamma_t\}_{t=1}^{\infty}$ that sums to one.

Considering:

$$\begin{aligned} S^t &= \sum_{i<t} \mathbb{1}\{p_i \leq \eta\}, \\ C_{j+} &= \sum_{i=k_j+1}^{t-1} \mathbb{1}\{p_i \leq \lambda\}, \\ k_j &= \min\{i \in [t-1] : \sum_{k \leq i} \mathbb{1}\{p_k \leq \alpha_k\} \geq j\}, \\ k_j^* &= \sum_{i \leq k_j} \mathbb{1}\{p_i \leq \eta\}. \end{aligned} \tag{3.11}$$

The test level α_t is

$$\alpha_t = \min\{\lambda, \tilde{\alpha}_t\}, \tag{3.12}$$

where $\tilde{\alpha}_t = (\eta - \lambda)\left(W_0 \gamma_{S^t - C_{0+}} + (\alpha - W_0)\gamma_{S^t - k_1^* - C_{1+}} + \alpha \sum_{j \geq 2} \gamma_{S^t - k_j^* - C_{j+}}\right)$.

Selecting $\lambda = 0.25$, $\eta = 0.5$ and $\gamma_t \propto (t+1)^{-1.6}$as a practical default, the authors demonstrated numerically that ADDIS exhibits considerably more power in scenarios with numerous conservative nulls and seldom experiences a loss of power in settings devoid of conservative nulls.

3.4 Methodology

This thesis employs the open-source R package "online FDR," encompassing implementations of the LORD, SAFFRON, and ADDIS algorithms and nearly all subsequent advancements in online error rate control methods. Additionally, modifications to the source code of the LORD algorithm enabled further analysis under various data corruption scenarios.

Our study implements a straightforward experimental framework that tests Gaussian means across N hypotheses to evaluate the comparative efficacy among the algorithms, utilizing the default configurations recommended in the existing literature.

For all simulations conducted, null hypotheses $H_t : \mu_t = 0$ are tested against the alternative: $\mu_t > 0$, for t = 1, . . . , N. Consequently, we observe independent $Z_t \sim \mathcal{N}(\mu_t, 1)$ transformed into one-sided p-values $p_t = \Phi(-Z_t)$, where Φ denotes the standard Gaussian Cumulative Density Function (CDF). The values of μ_t are determined based on the mixture distribution:

$$\mu_t = \begin{cases} \mathcal{N}(0,1) & \text{with probability } \pi_0 = 1 - \pi_1 \\ \mathcal{N}(3,1) & \text{with probability } \pi_1. \end{cases} \tag{3.13}$$

In the composite model designated as G, the null hypothesis H_t stipulates that p_t is uniformly distributed within the interval [0,1]. Contrarily, the alternative hypothesis posits that p-values are derived from a distribution with the CDF represented by F. Consequently, the marginal distribution of these simulated p-values is expressed as $G(x) = \pi_0 x + \pi_1 F(x)$. As depicted in Figure 3.1, the histogram illustrates why online methodologies have more likelihood of rejecting non-null hypotheses, as they are characterized by lower p-values.

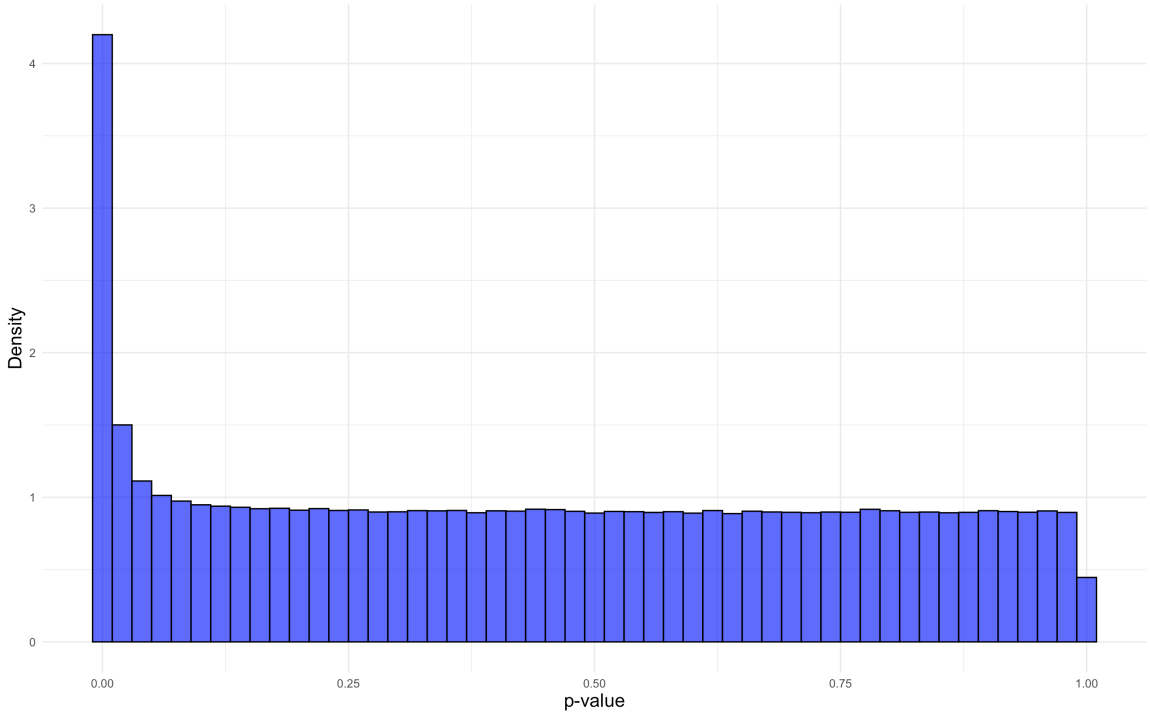

Figure 3.1. Histogram of the mixed model G. Results are based on $N = 100,000$.

3.5 Performance Comparison

Figure 3.2 compares the statistical power of LORD, SAFFRON, and ADDIS for the proportion of non-nulls π_1 varying from 0.1 to 0.9, $N = 1,000$, and $\alpha = 0.05$. For this matter, power is defined as (Robertson et al. 2023)

$$\text{Power}\ (N) = \mathbb{E}\left[\frac{\sum_{t\in\mathcal{H}^1} R_t}{\left(\sum_{t=1}^{N} \mathbb{1}\{t \in \mathcal{H}^1\}\right) \vee 1}\right], \tag{3.14}$$

where $\mathcal{H}^1$ denotes the set of indices corresponding to alternative hypotheses. Put into words, power is the expected value of the true discoveries divided by the total number of alternative hypotheses.

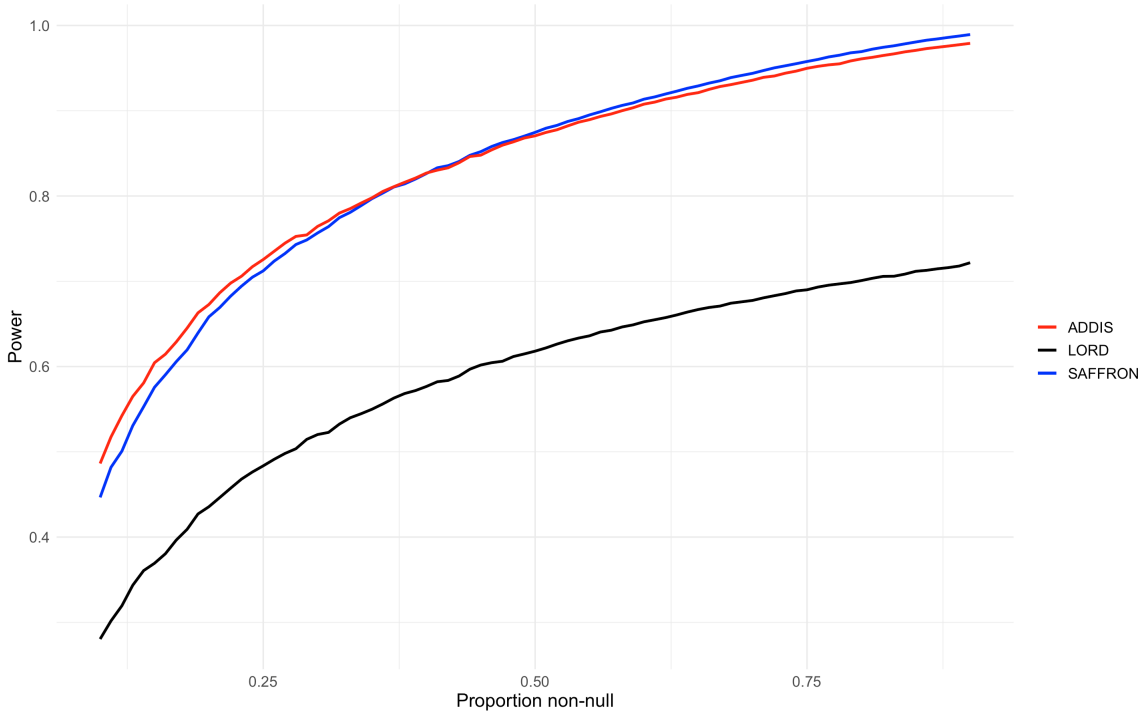

Figure 3.2. Power of LORD, SAFFRON, and ADDIS as the proportion of non-nulls varies. Results are based on 10^3 replications.

As anticipated, the power of LORD is lower than the power of both SAFFRON and ADDIS for all proportions of non-nulls π_1. Since this simulation deals with conditionally super-uniform null p-values, there is no clear advantage for ADDIS over SAFFRON. In fact, SAFFRON surpasses ADDIS in terms of power as the proportion of non-nulls increases ($\pi_1 > 0.40$).

Figure 3.3 shows the corresponding FDR against power for all the algorithms considered. As expected, all algorithms control the FDR below the nominal 0.05 level setting.

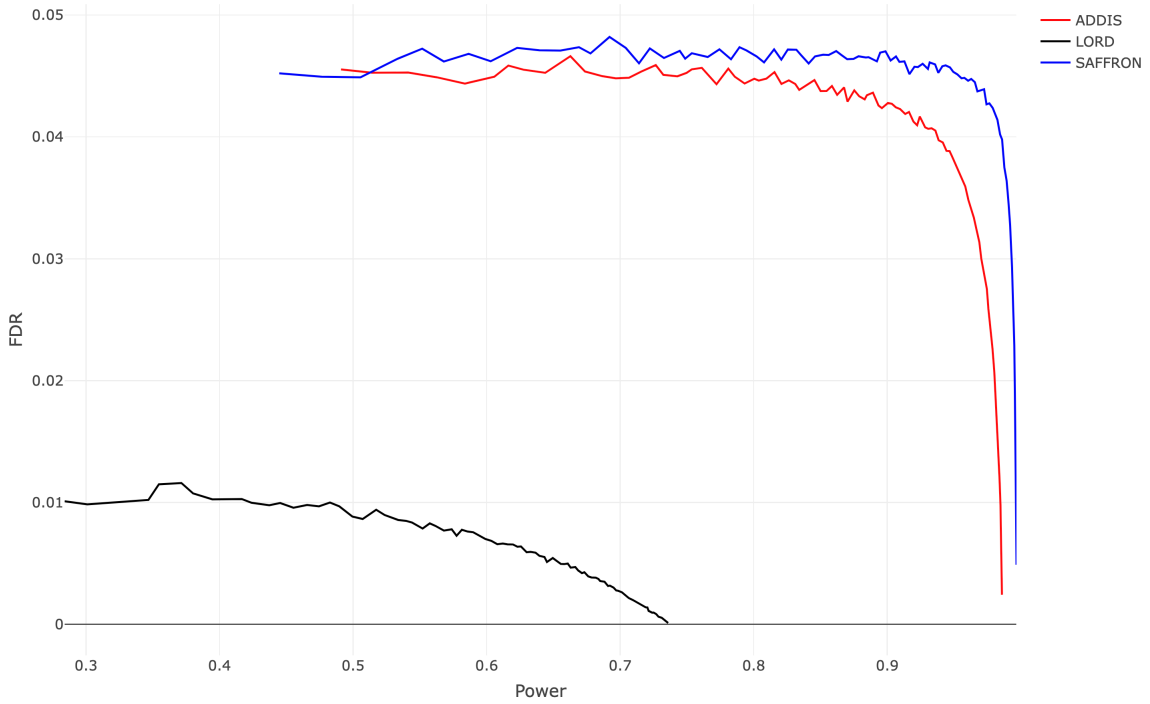

Figure 3.3. FDR and power of LORD, SAFFRON, and ADDIS as the proportion of non-nulls varies. Results are based on 10^3 replications.

As π_1 increases from 0.01 to 0.9, there is a noticeable increase in power alongside a reduction in the FDR. It becomes clear that the enhanced power exhibited by SAFFRON and ADDIS is accompanied by an elevated FDR. Furthermore, when comparing scenarios with identical π_1 and power levels, SAFFRON consistently exhibits a higher FDR than ADDIS does, except for very low values of π_1, but all FDR are lower than the threshold α specified by the user.

This simulation elucidates the distinctions among leading algorithms controlling the FDR under the presumption of valid and reliable p-values. Now, we will delve into an unexplored yet crucial area: the influence of data corruption on online hypothesis testing. The impact of such data on both power and FDR is uncharted territory in the existing literature, so this thesis seeks to bridge this gap by providing comprehensive answers to these pressing questions, thereby marking a novel and significant contribution to online hypothesis testing theory.

CHAPTER 4:
Online Multiple Hypothesis Testing with Corrupted Data

This chapter explores different data corruption scenarios and proposes an alternative procedure to complement the standard LORD algorithm, making it more robust to adversarial attacks.

4.1 Assumptions

The principal conclusion drawn from the preceding chapter is that LORD, ADDIS, and SAFFRON maintain the FDR beneath the threshold α for all periods. To simplify the analysis, in this chapter, we primarily focus on the LORD algorithm due to its clarity and methodological simplicity.

To rigorously evaluate data corruption in online multiple hypothesis testing, we propose a controlled adversarial setup featuring two entities: Blue, representing the side that performs the tests attempting to make true discoveries, and Red, who acts as the offensive agent by stealing discoveries. The model operates as follows:

1. In period t, Blue receives a single p-value, p_t, and must decide whether to accept or reject the hypothesis H_t, using only the information collected in rounds $1, \ldots, t-1$.
2. Red knows if H_t is true but does not know the p-values past p_t.
3. In case p_t is stolen, it is removed from the data stream without Blue noticing. This way, we simplify potentially more complicated scenarios, such as setting a new (corrupted) value of p_t.

4.2 Problem Formulation

The primary objective of Blue is to maximize power, ensuring that the FDR does not exceed the chosen threshold α over a time horizon of interest. Conversely, Red aims to min-max Blue's power, subject to an effort constraint.

In this setting, there are T periods. In periods $t = 1, \ldots, T$, Blue receives a single p-value and has to reject or fail to reject the hypothesis associated with p_t. This is done by comparing p_t with α_t, where α_t is determined by the LORD algorithm. If the p-value associated with the hypothesis H_t is greater than α_t, the hypothesis will not be rejected, but if $p_t \leq \alpha_t$ it will be rejected in favor of the alternative, and we get a so-called discovery (which could be a true or false one).

Before examining the impact of Red's attack, it is pertinent to revisit the LORD α_t equation, as delineated in Equation 3.5:

$$\alpha_t = w_0\gamma_t + (\alpha - w_0)\gamma_{t-\tau_1}\mathbb{1}\{\tau_1 < t\} + \alpha \sum_{j:\tau_j<t,\tau_j\neq\tau_1} \gamma_{t-\tau_j}.$$

Suppose the p-values are generated as those in section 3.4, $w_0 = \frac{\alpha}{10}$ and $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$:

1. For time $t = 1$, $\alpha_1 = w_0\gamma_1$. If $p_1 > \alpha_1$, there is no discovery, and no wealth is added to the initial budget.
2. For time $t = 2$, $\alpha_2 = w_0\gamma_2$. However, if $p_2 \leq \alpha_2$, a discovery is made and $(\alpha - w_0)\gamma_1$ is added to the budget. Thus, $\tau_1 = 2$.
3. For time $t = 3$, $\alpha_3 = w_0\gamma_3 + (\alpha - w_0)\gamma_1$. If $p_3 \leq \alpha_3$, then $\alpha\gamma_1$ is added to the budget.
4. For time $t = 4$, $\alpha_4 = w_0\gamma_4 + (\alpha - w_0)\gamma_2 + \alpha\gamma_1$. If $p_4 \leq \alpha_4$, then $\alpha\gamma_2$ is added to the budget.
5. For time $t = 5$, $\alpha_5 = w_0\gamma_5 + (\alpha - w_0)\gamma_3 + \alpha(\gamma_1 + \gamma_2)$. If $p_5 \leq \alpha_5$, then $\alpha\gamma_3$ is added to the budget.
6. This process continues until the last unit of time.

In this case, let us consider that the p-value p_4 at time $t = 4$ came from the alternative hypothesis and is stolen by Red. This results in a wealth amount equal to $\alpha\gamma_1$ being removed from α_5, $\alpha\gamma_2$ removed from α_6, and so on, for a total α removed from the subsequence α_{5+t}, for $t \geq 0$. This may induce a "cascade effect" in future values of α_t.

This "cascade effect" is significant because it influences the likelihood of future discoveries. Normally, if the conditions for discovery ($p_t \leq \alpha_t$) were met, additional wealth would be

added to the budget. However, due to Red's manipulation at $t = 4$, the subsequent α_t values are impacted. This means that potential discoveries that might have occurred under normal circumstances may no longer happen, as the altered α_t levels are now lower, making it harder to meet the discovery criteria. This illustrates how a single attack at a point in time can have lasting effects on the entire process, altering the trajectory outcomes.

4.3 Cascade Effect Formulation

This section estimates the expected number of lost discoveries until the next discovery, which is a lower bound for the expected number of discoveries lost.

Let $p_k^{(1)}$ be a random p-value from $\mathcal{H}_1$ at time k. As an example, in case the alternative distribution is $N(\mu_1, 1)$, we know that $p_k^{(1)} \sim 1 - \Phi(\mu_1 + Z)$, for $Z \sim N(0, 1)$. Likewise, $p_k^{(1)} \sim U(0, 1)$, if the alternative and null distributions coincide.

The starting wealth at time t, α_t, depends on rejections up to $t - 1$. We consider an attack taking place at time t. That is, $p_t \leq \alpha_t$, but the attacker prevents a rejection from taking place—it does not matter how this is done, whether by stealing the p-value or by corrupting it, the end effect is that there is no rejection in period t when there should have been one. Importantly, the decision-maker is unaware of this fact.

Let $\tilde{\alpha}_t = \alpha_t$ and $\tilde{\alpha}_{t+k}$ as the value of α_{t+k} if there were no rejections at $t, t+1, \ldots, t+k$. That is, the sequence of thresholds $\tilde{\alpha}_t$ is deterministic, conditional on the discoveries until $t - 1$.

Hence, the expected number of lost discoveries until the next discovery is

$$\underbrace{\pi_1 P(p_{t+1}^{(1)} \in (\tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1} + \alpha\gamma_1))}_{\text{P(missing an } \mathcal{H}_1 \text{ rejection in period } t+1 \text{ due to stolen discovery at } t)}$$
$$+ \sum_{k=2}^{\infty} \underbrace{\left(\prod_{j=1}^{k-1} \left(\pi_1 P(p_{t+j}^{(1)} > \tilde{\alpha}_{t+j}) + \pi_0 P(p_{t+j}^{(0)} > \tilde{\alpha}_{t+j}) \right) \right)}_{\text{prob of no rejections up to period } t+k-1} \underbrace{\pi_1 P(p_{t+k}^{(1)} \in (\tilde{\alpha}_{t+k}, \tilde{\alpha}_{t+k} + \alpha\gamma_k))}_{\text{P(missing an } \mathcal{H}_1 \text{ rejection in period } t+k \text{ due to stolen discovery at } t)}. \tag{4.1}$$

The reasoning of Expression 4.1 is that the expected number of true discoveries lost in period $t + k$ is the sum of (i) the probability of the p-value in $t + 1$ being from $\mathcal{H}_1$ and falling

in the range of values that would have triggered a rejection had an attack in period t not taken place, with the product of two terms, (ii) the probability of the p-values in $t+2, \ldots, t+k-1$ being below the rejection threshold, and (iii) a p_{t+k} being from $\mathcal{H}_1$ and falling in the range of values that would have triggered a rejection had an attack in period t not taken place.

Since α_k is decreasing as long as there are no discoveries, and the PDF of $p_t^{(1)}$ is non-increasing, the above expression can be lower bounded by

$$\sum_{k=1}^{\infty}(\pi_1 P(p^{(1)} > \tilde{\alpha}_{t+1}) + \pi_0 P(p^{(0)} > \tilde{\alpha}_{t+1}))^{k-1}\pi_1 P(p_{t+k}^{(1)} \in (\tilde{\alpha}_{t+k}, \tilde{\alpha}_{t+k} + \alpha\gamma_k)). \tag{4.2}$$

An even weaker lower bound is obtained by replacing $\tilde{\alpha}_{t+k}$ with $\tilde{\alpha}_{t+1}$ in Expression 4.2,

$$\sum_{k=1}^{\infty}(\pi_1 P(p^{(1)} > \tilde{\alpha}_{t+1}) + \pi_0 P(p^{(0)} > \tilde{\alpha}_{t+1}))^{k-1}\pi_1 P(p_{t+k}^{(1)} \in (\tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1} + \alpha\gamma_k)). \tag{4.3}$$

In case the alternative distribution is $N(\mu_1, 1)$, we get

$$P(p^{(1)} > \tilde{\alpha}_{t+1}) = P(1 - \Phi(\mu_1 + Z) > \tilde{\alpha}_{t+1}) = P(\Phi(\mu_1 + Z) < 1 - \tilde{\alpha}_{t+1})$$

$$= P(\mu_1 + Z < \Phi^{-1}(1 - \tilde{\alpha}_{t+1})) = \Phi(-\mu_1 + \Phi^{-1}(1 - \tilde{\alpha}_{t+1})).$$

Likewise,

$$P(p_{t+k}^{(1)} \in (\tilde{\alpha}_{t+1}, \tilde{\alpha}_{t+1} + \alpha\gamma_k)) = \Phi(\mu_1 - \Phi^{-1}(1 - \tilde{\alpha}_{t+1} - \alpha\gamma_k)) - \Phi(\mu_1 - \Phi^{-1}(1 - \tilde{\alpha}_{t+1})).$$

From here, we can compute Expression 4.2 numerically.

In contrast, when $p^{(1)} \sim U(0, 1)$ (meaning that the null and alternative distributions coincide), we get in Expression (4.2),

$$\alpha\pi_1 \sum_{k=1}^{\infty}(1 - \tilde{\alpha}_{t+1})^{k-1}\gamma_k. \tag{4.4}$$

It follows that the expected number of true discoveries lost approaches $\alpha\pi_1$ in (4.4), as $\tilde{\alpha}_{t+1} \to 0$.

Thusly motivated, we investigate two distinct scenarios of online hypothesis testing with corrupted data:

1. **Single attack**: Red's capacity to attack is limited by a single attack.
2. **Stochastic attacks**: Red attacks each alternative p-value with probability ζ.

The simulation of each scenario uses altered forms of the LORD algorithm, derived from the "onlineFDR" R package, to effectively incorporate Red's and Blue's strategies. The generation of p-values adhered to the process detailed in Section 3.4.

4.4 Single Attack

As previously discussed in this chapter, the corruption of a single alternative p-value may initiate a "cascade effect," where the stolen wealth imposes future reduced α_t values, leading consequently to fewer discoveries.

From the attacker's perspective, the earlier Red intervenes in the data stream directed towards Blue, the more promptly α_t will decrease, thereby suppressing a greater number of potential discoveries. Accordingly, this scenario examines the dynamics of power and FDR when Red attacks the first alternative p-value.

Red procedure for attacking the first alternative p-value

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, and sets $\tau_0 = 0$.
2. At each step t, Blue computes α_t according to Equation 3.5.
3. If $H_t \in \mathcal{H}^1$ and $p_t \leq \alpha_t$, Red steals p_t and the discovery is not allowed.
4. Go back to step 2 if $t < N$ or the attack has not taken place.

Figure 4.1 compares the statistical power of LORD without attacks (in black) and of LORD when only the first alternative p-value is attacked (in blue) for the proportion of non-nulls π_1 varying from 0.1 to 0.9, $N = 1,000$, and $\alpha = 0.05$.

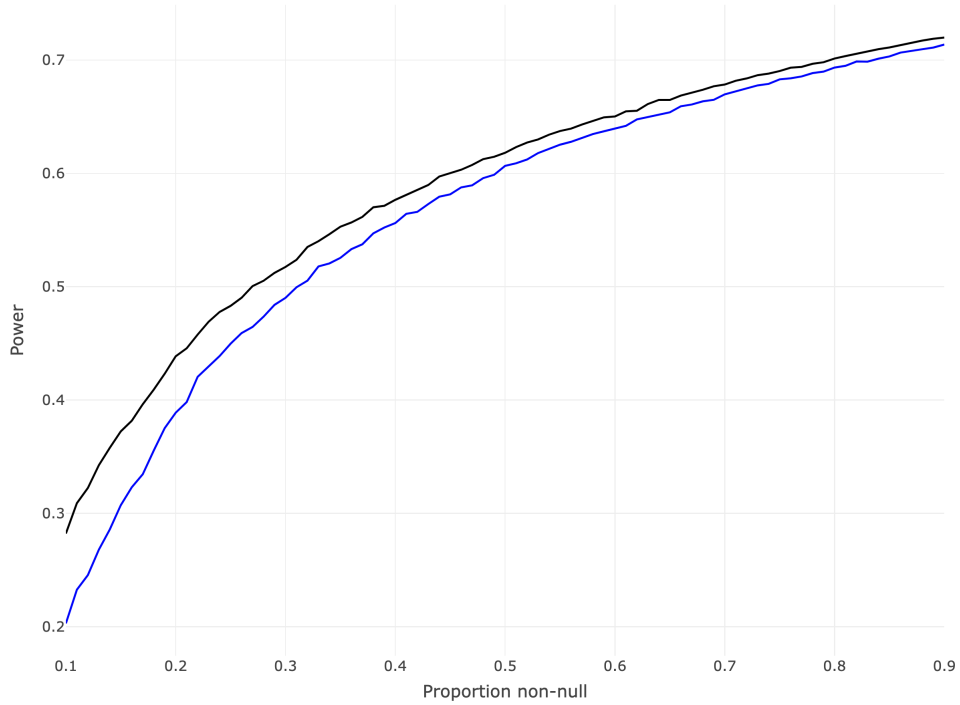

Figure 4.1. Power of LORD without attacks and of LORD with a single attack as the proportion of non-nulls varies. Results are based on 10^3 replications.

A single attack imposes an overall power decrease for different π_1. Notably, this decrement is more pronounced at lower π_1 values, while a single attack is largely inconsequential as π_1 gets bigger since there are many discoveries that remain to be made by Blue after Red's attack. Given the negative relationship between FDR and power, it is prudent to focus our investigation on lower π_1 values to ascertain the subsequent behavior of the FDR.

Figure 4.2 shows the corresponding FDR and power for $\pi_1 = 0.1$.

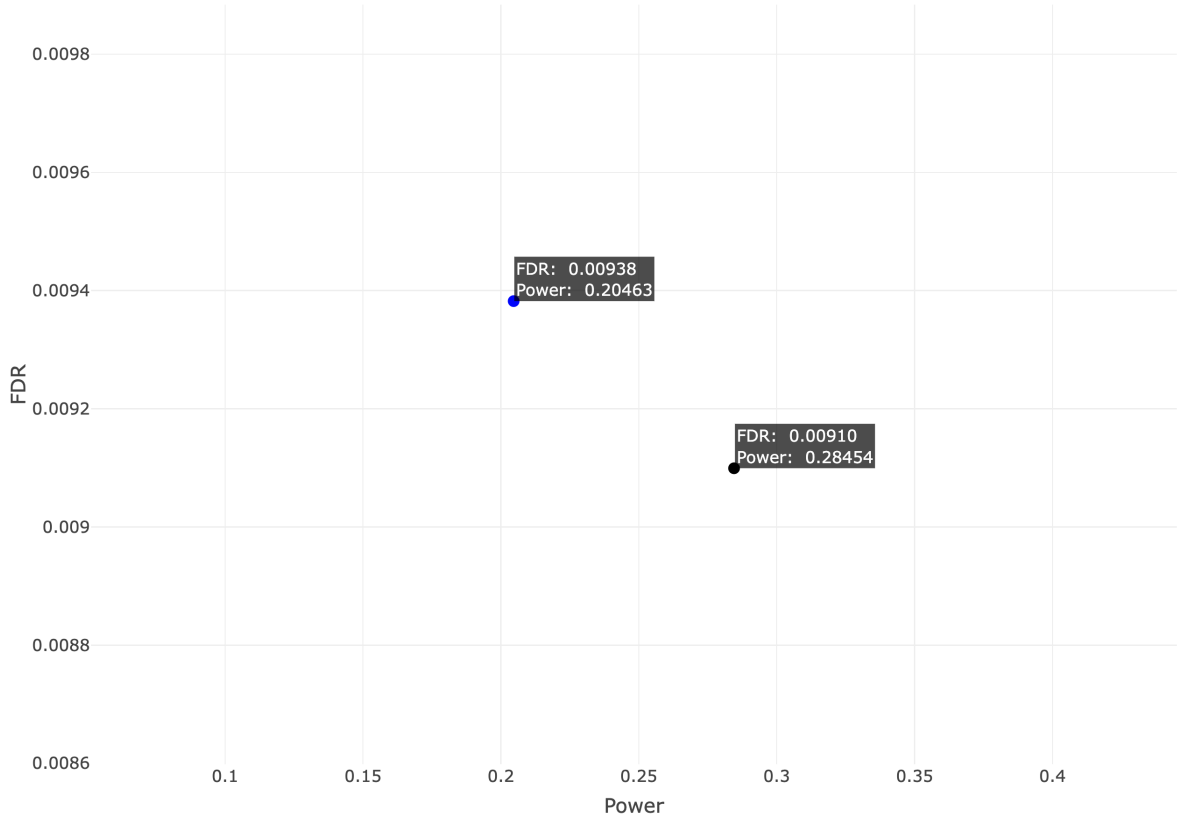

Figure 4.2. FDR and power of LORD without attacks and of LORD with a single attack for $\pi_1 = 0.1$. Results are based on 10^3 replications.

Stealing the first alternative p-value resulted in the average power dropping from around 0.28 to 0.20, marking a 28% decrease, while the FDR remained largely unaffected. The cascading effect resulted in missing seven extra alternative hypotheses with just a single steal, highlighting the effectiveness of this approach in undermining statistical power.

While Blue is certain of an imminent attack, the exact moment of its occurrence is undetermined. As a strategy, we propose reviewing the infinite, non-increasing sequence of positive constants $\{\gamma_t\}_{t=1}^{\infty}$ that sums to one, as the LORD algorithm does not impose any fixed formula for it.

The threshold α_t for each hypothesis H_t is a monotone decreasing function of past rejections, represented by the convolved sum of previous γ. This design implies that, as more hypotheses

are tested and potentially rejected over time, the threshold for deeming subsequent tests significant becomes progressively more restrictive. Consequently, as the testing process advances and the criterion for each test becomes more rigorous, the likelihood of achieving further discoveries diminishes. When Red prevents a discovery, the effect on the testing procedure is twofold. Firstly, the immediate outcome of such an attack is the failure to add a wealth $\alpha\gamma_1$. Secondly, the α_t value assigned to the ensuing tests becomes even more restrictive than without corruption.

Therefore, we propose as Blue's strategy to modify the original formula for the sequence $\{\gamma_t\}_{t=1}^{\infty}$ to reduce the rate of decay of each α_t until the first discovery, consequently increasing the probability of discoveries and after that go back to the default equation. Any function with a lower rate of decrease than $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$ can be applied.

Figure 4.3 displays the plot of the function $\gamma_t = C\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$ for $t = 1, \ldots, 1000$, when C has its default value of 0.0722 (black) and $C = 2$ (blue). As expected, for small values of t, the new function provides larger values, but as t increases, it converges toward the black curve. This behavior implies that the new α_t levels will be higher than using the default value of C, and they tend to take longer to decrease, leading to greater "wealth" until Red's attack.

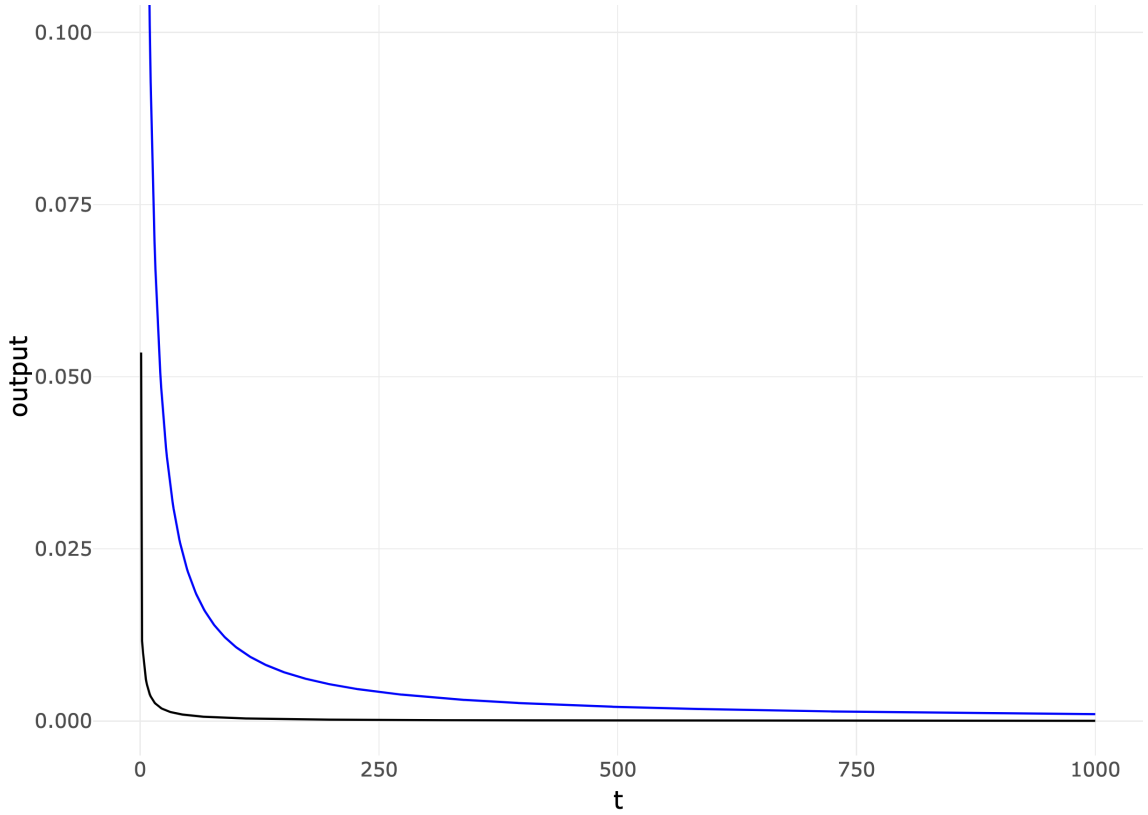

Figure 4.3. $\gamma_t = C\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$ for $t = 1, \ldots, 1000$, when $C = 0.0722$ and $C = 2$.

Blue procedure for defending against Red's attack at the first alternative p-value

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = C\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, and sets $\tau_0 = 0$, and $C = 2$.
2. At each step t, Blue computes α_t according to Equation 3.5.
3. If $H_t \in \mathcal{H}^1$ and $p_t \leq \alpha_t$, Red steals p_t and the discovery is not allowed.
4. Blue sets $C = 0.0722$.
5. Execute step 2 till $t = N$.

Figure 4.4 illustrates the statistical power of LORD without attacks, of LORD attacking only the first alternative p-value, and of LORD with the defender policy implemented for the proportion of non-nulls π_1 varying from 0.1 to 0.9, $N = 1,000$, and $\alpha = 0.05$.

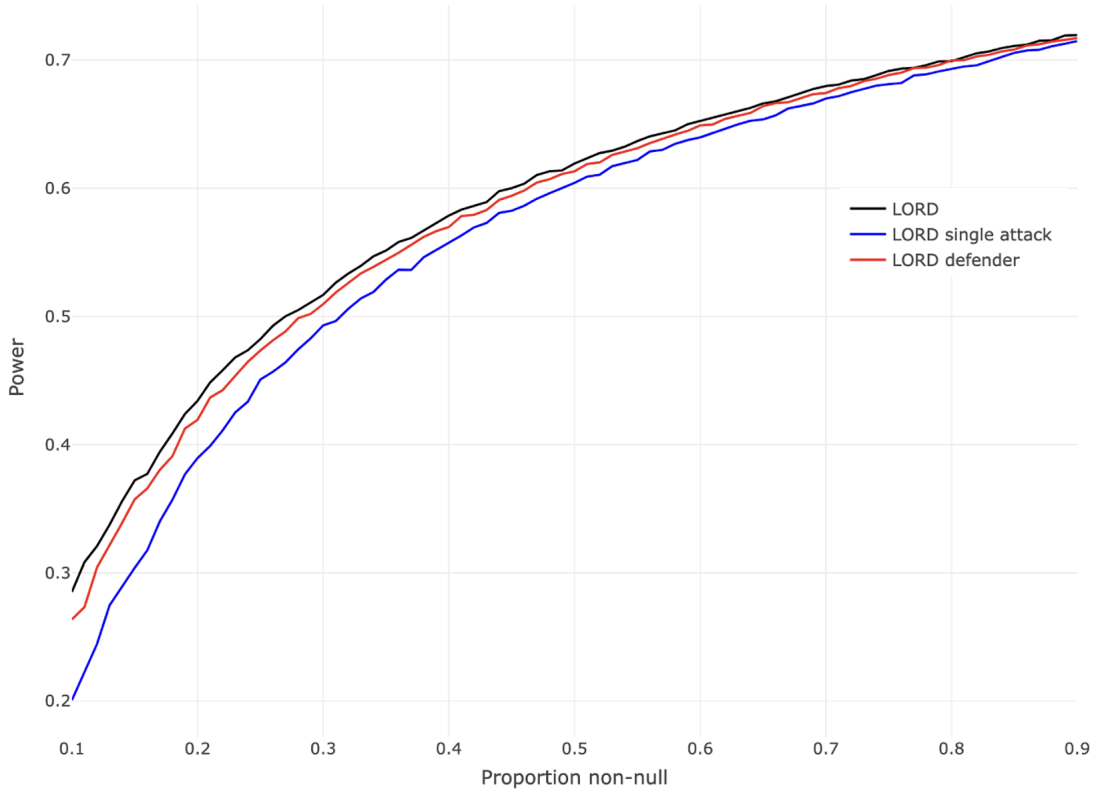

Figure 4.4. Power of LORD without attacks, of LORD with a single attack, and of LORD with the defender policy implemented for $\pi_1 = 0.1$. Results are based on 10^3 replications.

Implementing the previously mentioned defensive policy by Blue results in a less pronounced reduction in power compared to scenarios lacking data corruption for every value of π_1. Specifically at $\pi_1 = 0.1$, as illustrated in Figure 4.5, the average power diminishes from approximately 0.28 to 0.26 with the deployment of the defensive strategy, as opposed to 0.20 in the absence of any countermeasures, while the FDR remains virtually unaffected. This robustness is further exemplified by Blue's ability to recover six true discoveries out of eight lost (seven due to cascading). Without any defensive strategy, a single offensive maneuver by Red imposed seven additional discoveries, while the strategy actively limited the outcome to just one additional true discovery not being made.

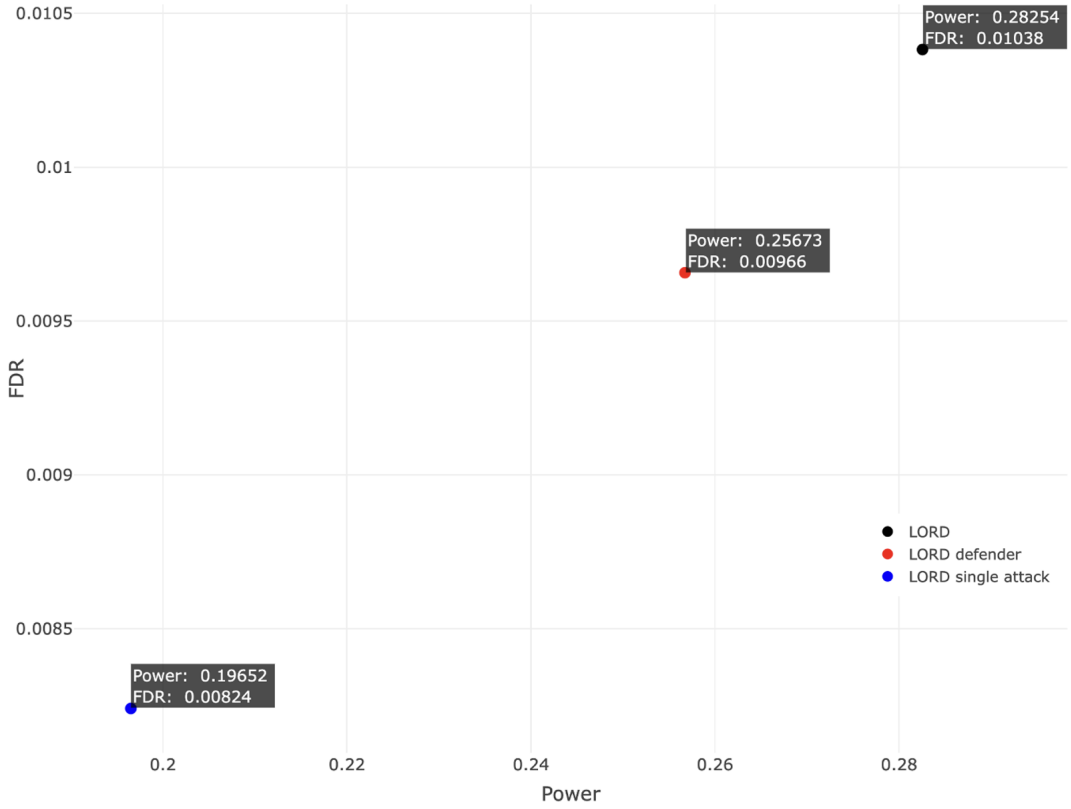

Figure 4.5. FDR and power of LORD without attacks, of LORD with a single attack, and of LORD with the defender policy implemented for $\pi_1 = 0.1$. Results are based on 10^3 replications.

In conclusion, by increasing the sequence $\{\gamma_t\}_{t=1}^{\infty}$ by a constant factor up to the first discovery, Blue can *protect* a large fraction of the discoveries that would otherwise be lost due to cascading from a single stolen discovery, with minimal increase in FDR.

4.5 Stochastic Attacks

In this scenario, we consider a setting where Red attacks only alternative p-values that would otherwise be rejected, with probability ζ. This setting may arise when Red has a great intelligence capability, allowing it to steal more than just one true discovery.

Red procedure for attacking alternative p-values with probability ζ

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$ and sets $\tau_0 = 0$.
2. At each step t, Blue computes α_t according to Equation 3.5.
3. If $H_t \in \mathcal{H}^1$ and $p_t \leq \alpha_t$, then Red steals the p-value with probability ζ.
4. Go back to step 2 till $t = N$.

Figure 4.6 compares the statistical power of LORD without attacks (in black) and of LORD with a probability $\zeta = 0.1$ of attacks (in blue) for the proportion of non-nulls π_1 varying from 0.1 to 0.9, $N = 1,000$, and $\alpha = 0.05$.

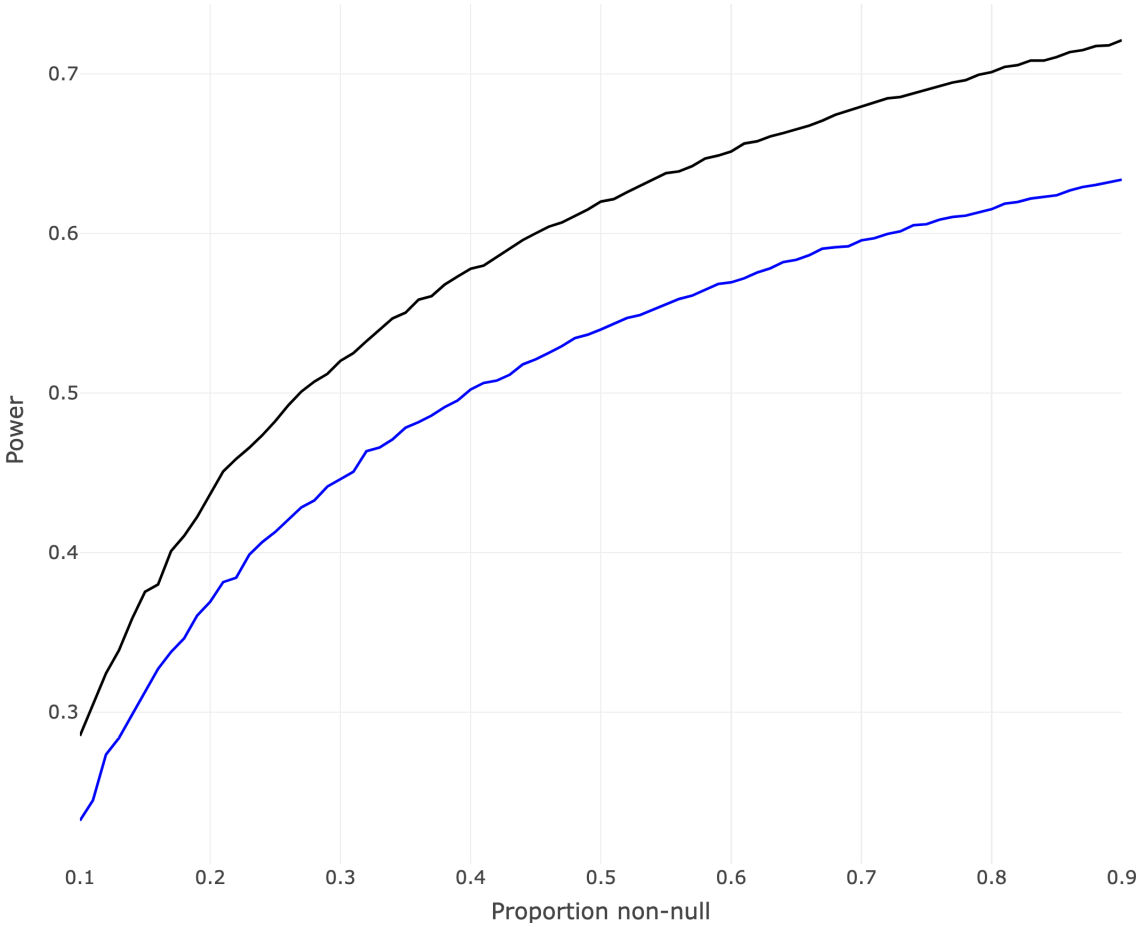

Figure 4.6. Power of LORD without attacks and of LORD with 10% of attack probability as the proportion of non-nulls varies. Results are based on 10^3 replications.

For a $\zeta = 0.1$ attack probability, the LORD algorithm shows a decreased power compared to its operation without any attacks for every level of non-null proportion π_1. The effect of a 10% attack is more pronounced as π_1 increases since there are more true discoveries to steal. Likewise, increasing the attack probability ζ leads to an even greater decrease in LORD's power.

The joint effect of FDR and power of increasing π_1 is shown in Figure 4.7.

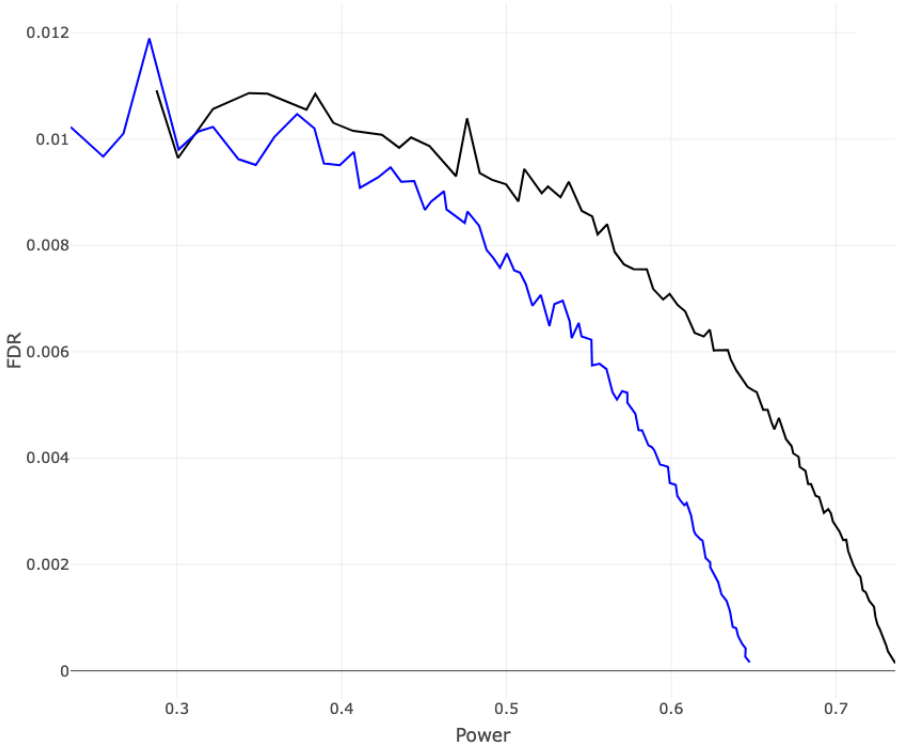

Figure 4.7. FDR and power of LORD without attacks and of LORD with 10% of attack probability as the proportion of non-nulls varies. Results are based on 10^3 replications.

Larger values of π_1 correspond to increased power for attack and no-attack cases, alongside a reduction in the FDR. The curve associated with an attack probability of $\zeta = 0.1$ lies below and to the left of the no-attack curve. Importantly, the FDR remains below the predefined $\alpha = 0.05$. In the particular instance of $\pi_1 = 0.1$, as depicted in Figure 4.8, there is a 21% reduction in the average power, from 0.29 to 0.23, exceeding 10%. This outcome merits

emphasis: even when subjected to an attack with a ζ probability, the power reduction exceeds the scale of ζ itself, as expected due to the cascade effect described in the last section.

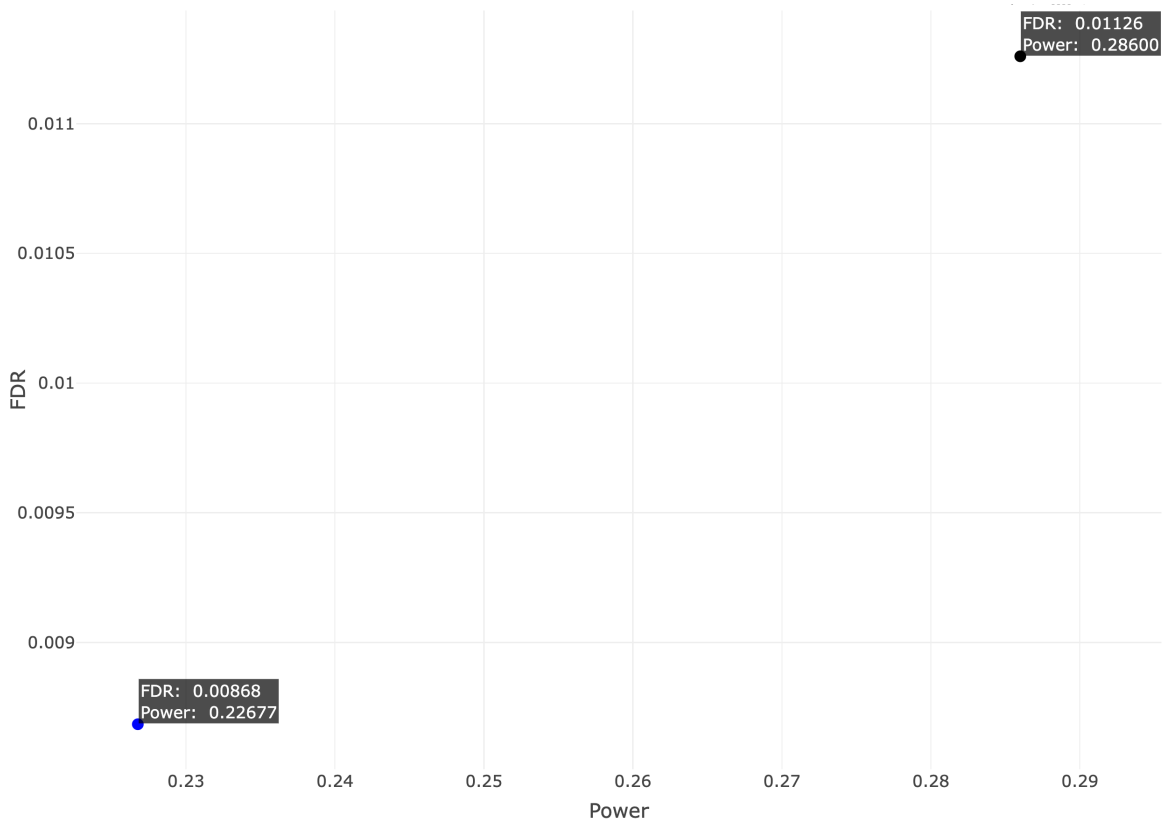

Figure 4.8. FDR and power of LORD without attacks and of LORD with 10% of attack probability for $\pi_1 = 0.1$. Results are based on 10^3 replications.

To counteract Red "usurping" discoveries from Blue, it becomes imperative to formulate mitigating strategies to prevent the cascade effect. In this context, we introduce the concept of "milestones." Blue strategically creates fathom discoveries at predetermined intervals $t = m$ for a certain attack probability, namely "milestones." This approach is designed to augment the overall count of discoveries, irrespective of their veracity, thus increasing the wealth function in the LORD algorithm, which instead results in a higher power.

Blue procedure for defending against Red attacking alternative p-values with probability ζ

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, and sets $\tau_0 = 0$ and milestone = m.
2. At each step t, Blue computes α_t according to Equation 3.5.
3. If $H_t \in \mathcal{H}^1$ and $p_t \leq \alpha_t$, then Red steals the p-value with probability ζ.
4. If $t \bmod m = 0$, a fathom discovery is created.
5. Go back to step 2 till $t = N$.

This defender policy can be mathematically expressed as:

$$\alpha_t = \left(w_0\gamma_t + (\alpha - w_0)\gamma_{t-\tau_1}\mathbb{1}\{\tau_1 < t\} + \alpha \sum_{\substack{j:\tau_j<t,\\ \tau_j\neq\tau_1}} \gamma_{t-\tau_j} \right) \mathbb{1}\{t \bmod m \neq 0\} + \alpha\mathbb{1}\{t \bmod m = 0\} \tag{4.5}$$

With the revised formulation of Equation 3.5, the condition when $t \bmod m = 0$ results in $\alpha_t = \alpha$. Hence, if $p_t \leq \alpha$, it leads to a discovery irrespective of the value initially determined by the original LORD algorithm. However, the challenge is to tune the m parameter to regain power and remain below the threshold α.

Figure 4.9 shows the corresponding FDR against power. The milestone m (in blue) values considered in this analysis include 1000, 500, 250, 200, 125, 100, 50, 40, 25, 20, 10, 5, 4, 2, 1.

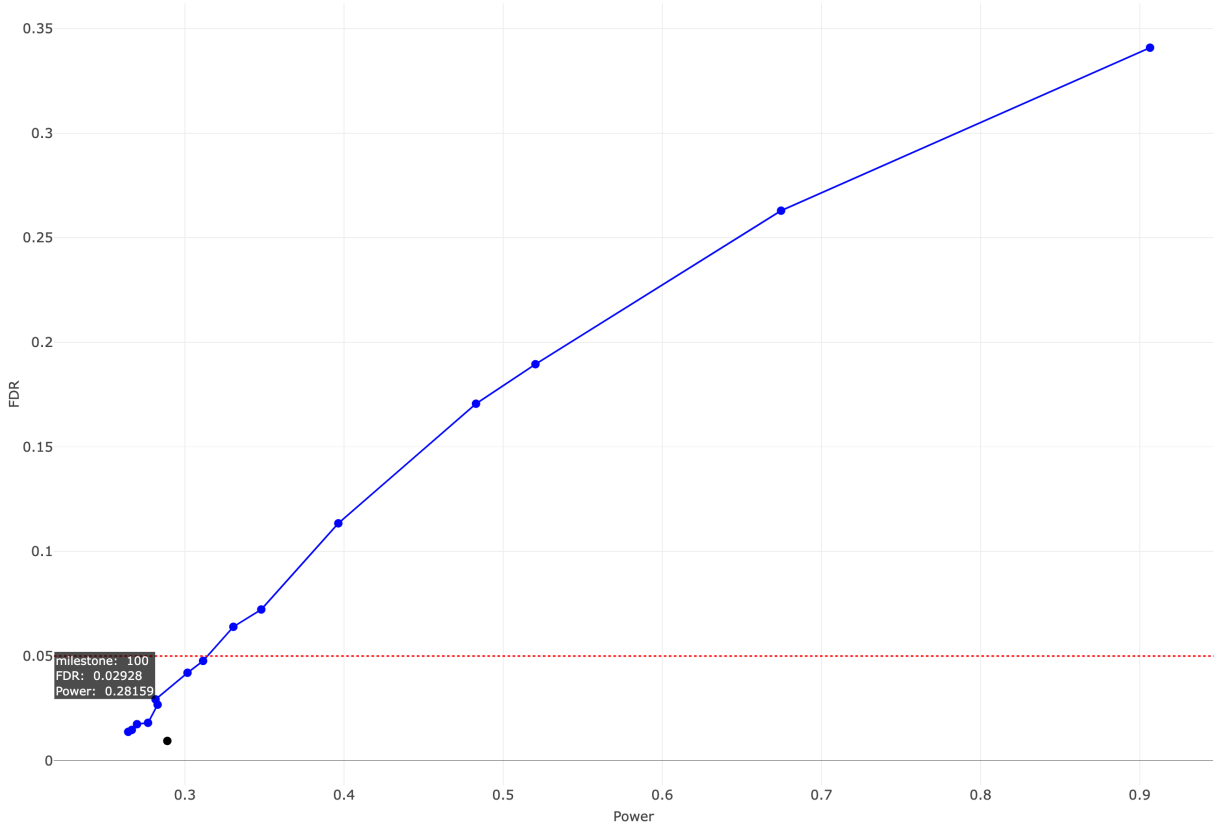

Figure 4.9. FDR and power as m decreases from 1,000 to 1 for $\pi_1 = 0.1$. Results are based on 10^3 replications.

Consider the extreme rightmost point where $m = 1$. In this case, $\alpha_t = \alpha$ for all t's, which results in an average power of 0.9 and a FDR of 0.34. At the other extreme, for $m = 1000$ the power drops to 0.26 with a FDR of 0.013. As evidenced by the plot, if the frequency of enforced discoveries is increased or m is decreased, both FDR and power show an upward trend. For example, when $m = 100$, all power can be recovered when we compare its value with the black dot representing the original LORD's power.

Figure 4.10 illustrates how the milestone should be adjusted according to ζ.

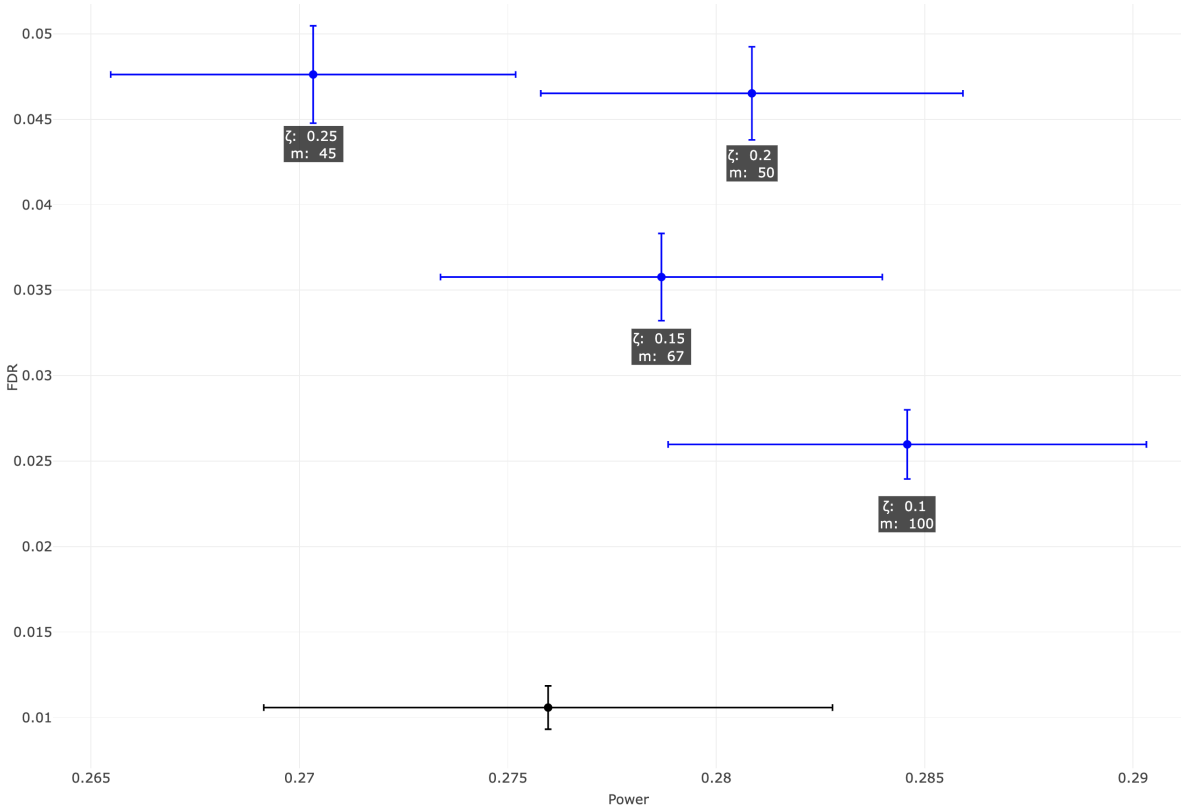

Figure 4.10. FDR and power with its corresponding confidence intervals for different ζ values as the milestone decreases. ζ values include 0.1, 0.15, 0.20, 0.25. Results are based on $2,000$ replications.

As the value of ζ increases, the milestone should decrease to encourage more discoveries, which leads to a higher FDR, allowing the recovery of all lost power. By comparing the power confidence intervals for Blue using the milestone (in blue) with LORD without any attacks (in black), it becomes clear that the power could be recovered.

A critical inference from this is that fully restoring power without breaching the established $\alpha = 0.05$ threshold is feasible using this milestone procedure but relies on the information about the probability of attacks ζ, which is impossible to know in real life, so a more robust procedure is needed.

4.6 Online BH Algorithm

Recall the setting where Red steals each discovery with probability ζ.

Red procedures for attacking any p-values with probability ζ

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, and sets $\tau_0 = 0$.
2. At each step t, Blue computes α_t according to Equation 3.5.
3. If $p_t \leq \alpha_t$ (meaning that the t term would be rejected), then with probability ζ, Red eliminates the p-value, and the discovery is not allowed. The next p-value in the sequence is fed to Blue.
4. Go back to step 2 till $t = N$.

Note that in Step 3, Red attacks regardless of whether the p-value is in H_0 or H_1. This would be the case of a *blind* attacker, who steals p-values smaller than α_t without considering the ground truth. In practice, only for alternative distributions without a strong signal (e.g., μ_1 close to zero) this type of blind attack would be impactful in relation to a not-blind attack (where only null p-values are stolen). When the signals are strong, most of the rejections are from the alternative distribution, so it does not matter whether the adversary—thanks to its intelligence capability—an discriminate between true and false discoveries.

As depicted in Figure 4.11, the impact of Red's attacks on power reduction remains consistent, irrespective of whether the attack is directed solely at p-values associated with alternative hypotheses or at all p-values, with the FDR consistently below α for all t's. This phenomenon is explicable because the probability of a p-value being from the alternative hypothesis conditional on the p-value being small is very close to 1.

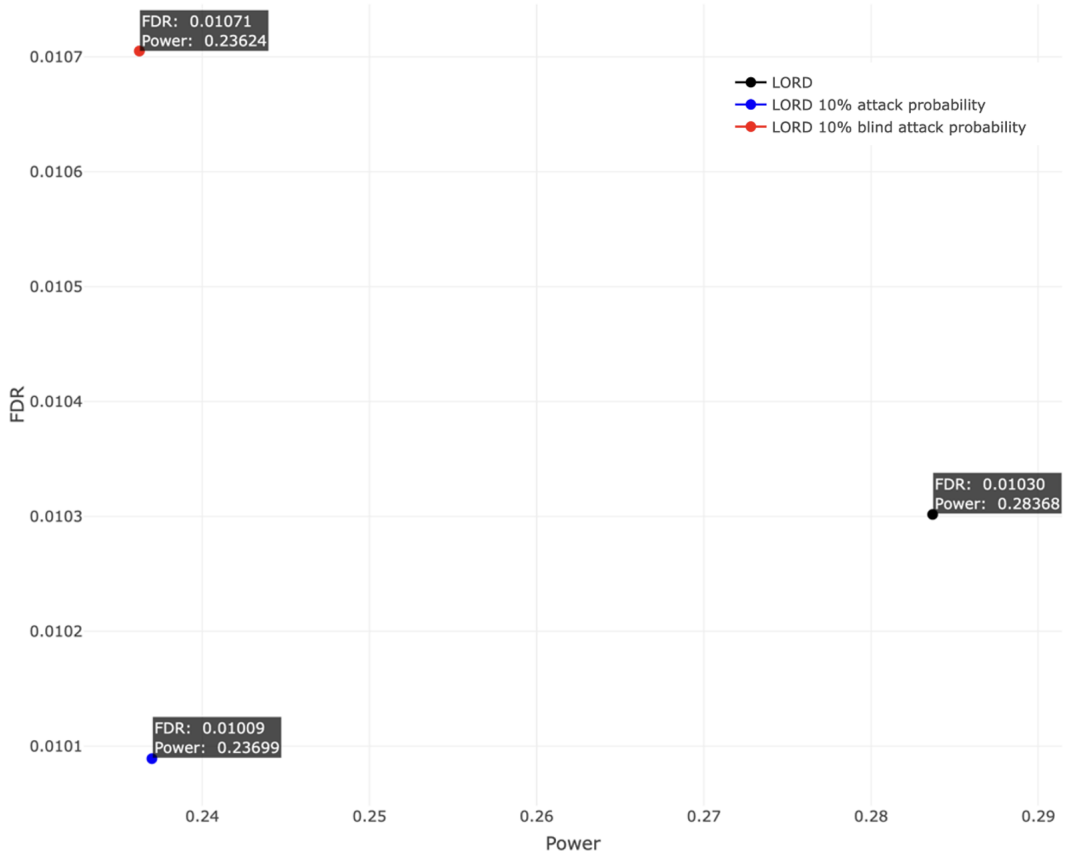

Figure 4.11. FDR and power of LORD without attacks, of LORD with 10% of attack probability at alternative p-values, and any p-values for $\pi_1 = 0.1$. Results are based on 10^3 replications.

To ameliorate the "cascade effect," we tested rejecting all p-values below some small threshold. Numerical testing indicated that the power was greatly increased while the FDR was kept below the guaranteed α. Thusly motivated, we devised a so-called online BH algorithm, which applies the BH procedure—traditionally employed in offline settings—in online fashion as the p-values roll in.

Following Section 3.4, consider the mixed model with the null mean $\mu_0 = 0$ and the alternative mean $\mu_1 > \mu_0$ as depicted in Figure 3.1, and the BH algorithm presented in Equation 2.18:

$$i_{\max} \text{ is the greatest index for which } p_{(i)} \leq \frac{i}{N}\alpha.$$
$$\text{Reject all } H_{(i)} \text{ where: } i \leq i_{\max}.$$

The idea is to perform the BH procedure at each period t. Therefore, Blue orders all p-values received till time t and calculates the corresponding position $i_{\max}$ of the current p-value p_t. Hence, a conservative dynamic threshold is:

$$\alpha_t = \frac{i_t}{t}\alpha, \tag{4.6}$$

where i_t is the position of p_t in the sorted vector $p_{(1)}, p_{(2)}, \ldots, p_{(t)}$.

If we consider a stream of p-values of length N, employing the LORD algorithm till time $t = N/2$ and the Online BH for the remaining sequence has demonstrated, through simulation, an enhanced power while adhering to the FDR control.

Blue procedures in a scenario with attacking probability of ζ

1. Blue initializes the LORD algorithm with $w_0 = \frac{\alpha}{10}$, $\gamma_t = 0.0722\frac{\log(t\sqrt{2})}{t\exp(\sqrt{\log t})}$, and sets $\tau_0 = 0$.
2. At each step $t \leq N/2$, Blue computes α_t according to the LORD algorithm.
3. For $t \geq N/2$, Blue computes α_t according to the BH algorithm.
4. If $p_t \leq \alpha_t$ (meaning that the t term would be rejected), then with probability ζ, Red eliminates the p-value, and the discovery is not allowed. The next p-value in the sequence is fed to Blue.
5. Go back to step 2 until $t = N$.

Figure 4.12 shows the results when we compare LORD's power with the aforementioned mixed procedure using the online BH algorithm for $N = 1,000$ and $\pi_1 = 0.1$.

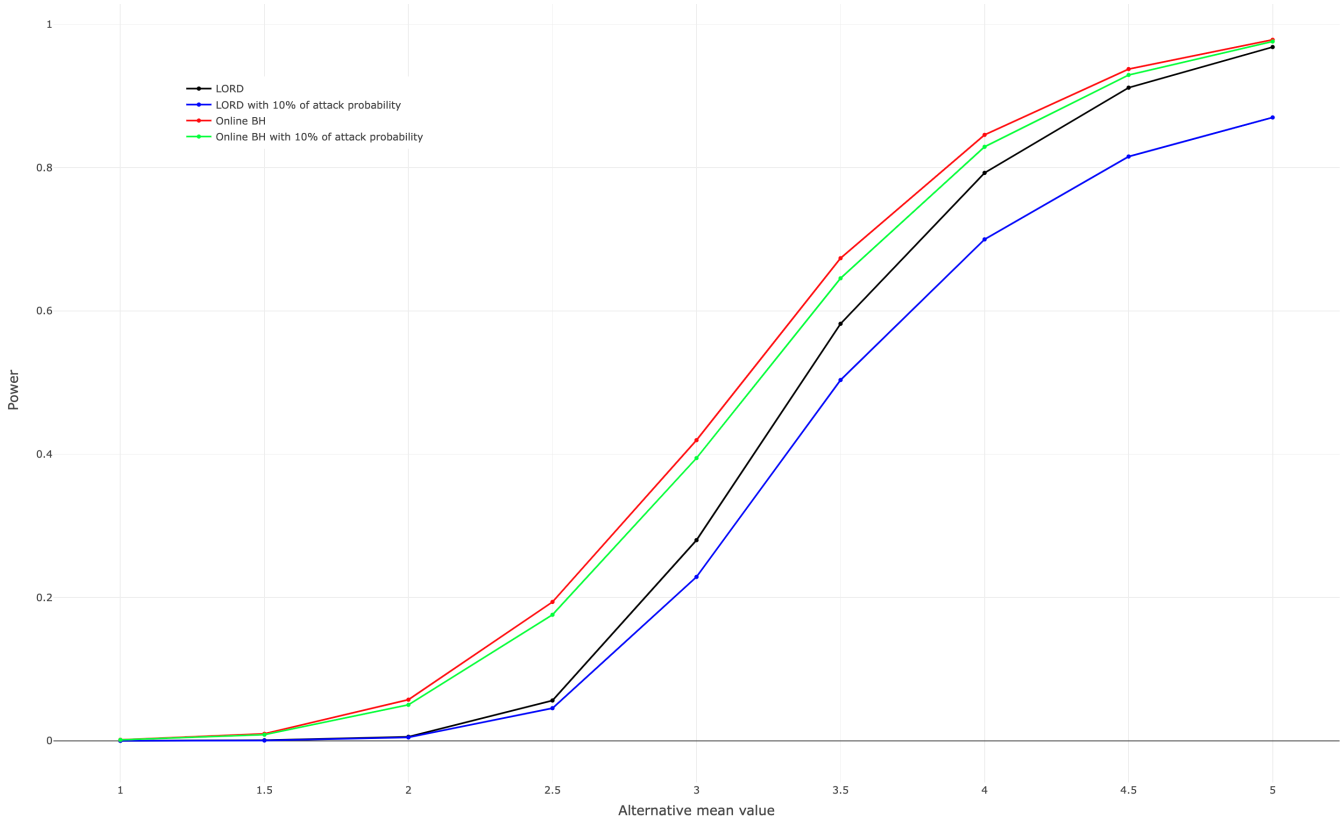

Figure 4.12. Power for $\mu_1 = 1, \ldots, 5$. Results are based on 2,000 replications.

It is apparent that augmenting the value of μ_1, so that the alternative signal is stronger, improves the power of the hypotheses tests. Indeed, both LORD and online BH procedures exhibit a monotonic increase in power. This trend underscores the intuitive principle that as the alternative hypothesis becomes more distinct from the null, the ability of these algorithms to identify true discoveries is enhanced, improving overall statistical power. Nevertheless, a comparative analysis between LORD (in black) and online BH (in red) reveals a substantial increase in power when Blue adopts online BH as a defender policy. Furthermore, when there is some probability of attacks $\zeta = 0.1$, the online BH (in green) shows more robustness compared to the LORD algorithm (in blue). This is evidenced by a less pronounced decrease in power, maintaining a stronger performance in the face of such adversarial conditions.

Even when we compare LORD (in black) absent of attacks with online BH (in green) with 10% of attack probability, it is clear that the latter has a better power performance for the tested values of μ_1.

Figure 4.13 illustrates the FDR behavior for this simulation:

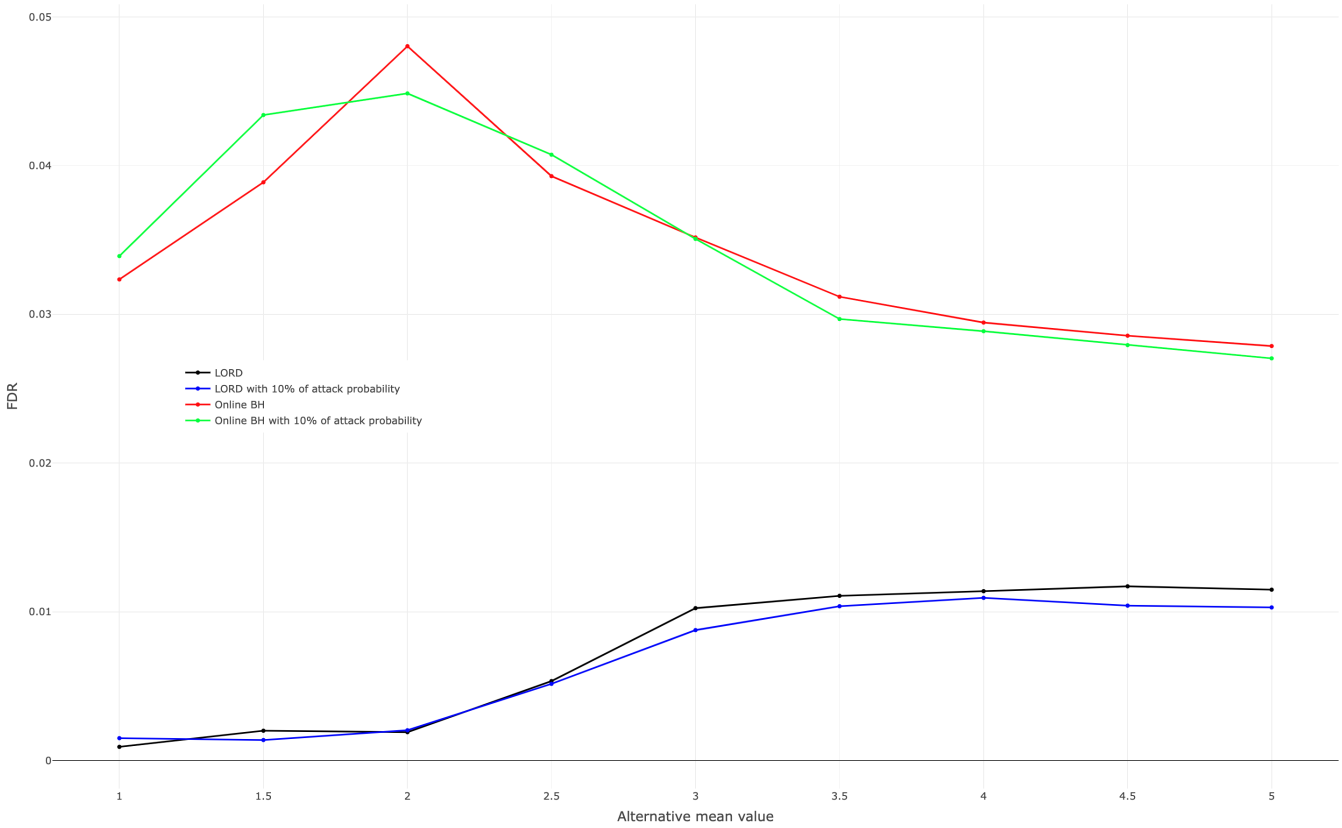

Figure 4.13. FDR for $\mu_1 = 1, \ldots, 5$. Results are based on 2,000 replications.

As μ_1 gets larger, the FDR marginally increases with the LORD algorithm with and without 10% attack probability; this is due to more rejections inducing more wealth, resulting in more false rejections. On the other hand, the simulation indicates that the FDR trajectory experiences an ascent only up to $\mu_1 = 2$ and a decrease beyond this point when using the online BH algorithm. Importantly, the FDR remains below 0.05 even with attacks for both algorithms.

Next, we investigate how the LORD and the online BH algorithms behave with greater probabilities of attacks ζ. Figure 4.14 depicts the power for $N = 1,000$, $\pi_1 = 0.1$, and different values of ζ.

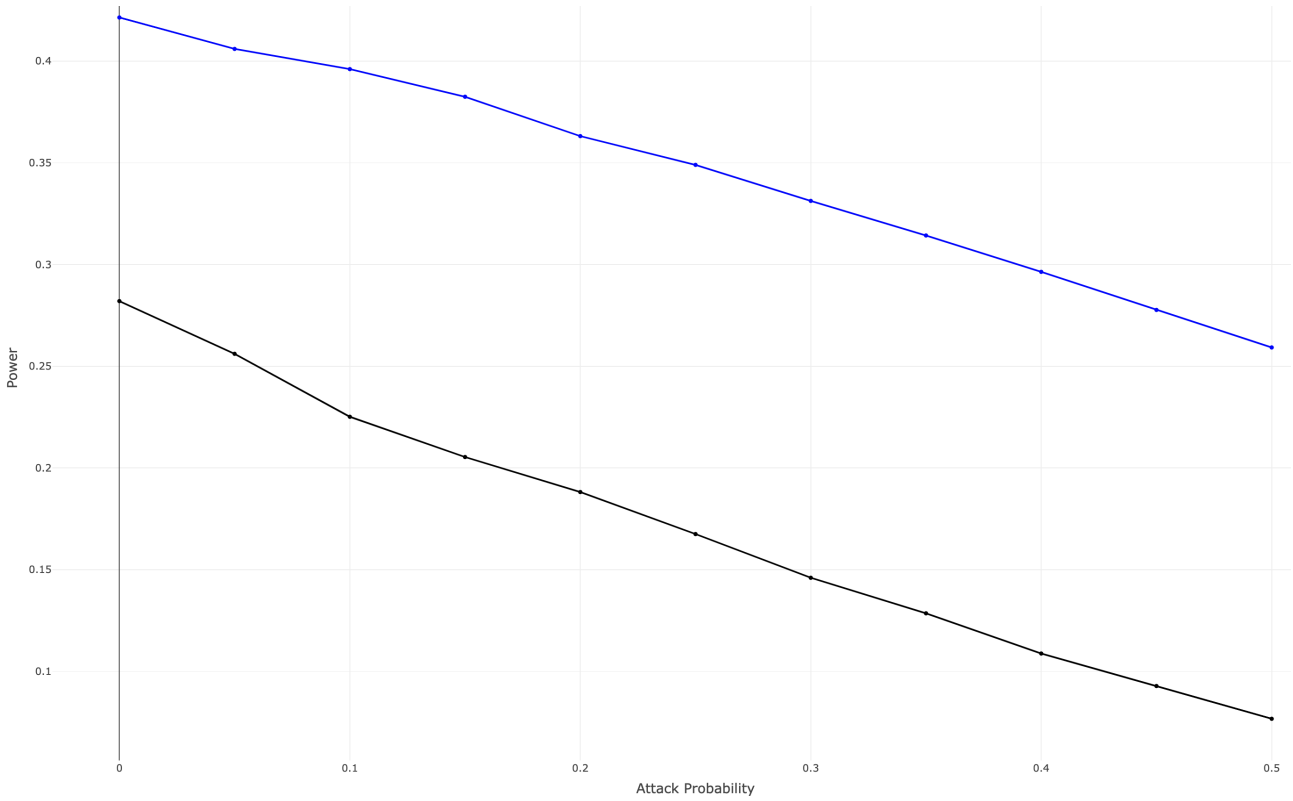

Figure 4.14. Power for $\zeta = 0.1, \ldots, 0.5$. Results are based on 2,000 replications.

As ζ increases, the power of both algorithms diminishes, as expected. Specifically, when ζ is set to 0.5, the power of the LORD algorithm (in black) falls below 0.1, indicating a reduced capacity for making true discoveries. In contrast, the online BH algorithm (in blue) demonstrates superior performance across all simulated ζ values, maintaining a noteworthy statistical power even at $\zeta = 0.5$. Within this context, the power of the online BH approximates that of the LORD algorithm when $\zeta = 0$, showcasing its robustness in highly contested environments.

Figure 4.15 proves that the FDR is always below 0.05 in this simulation, which is the FDR guarantee in our case.

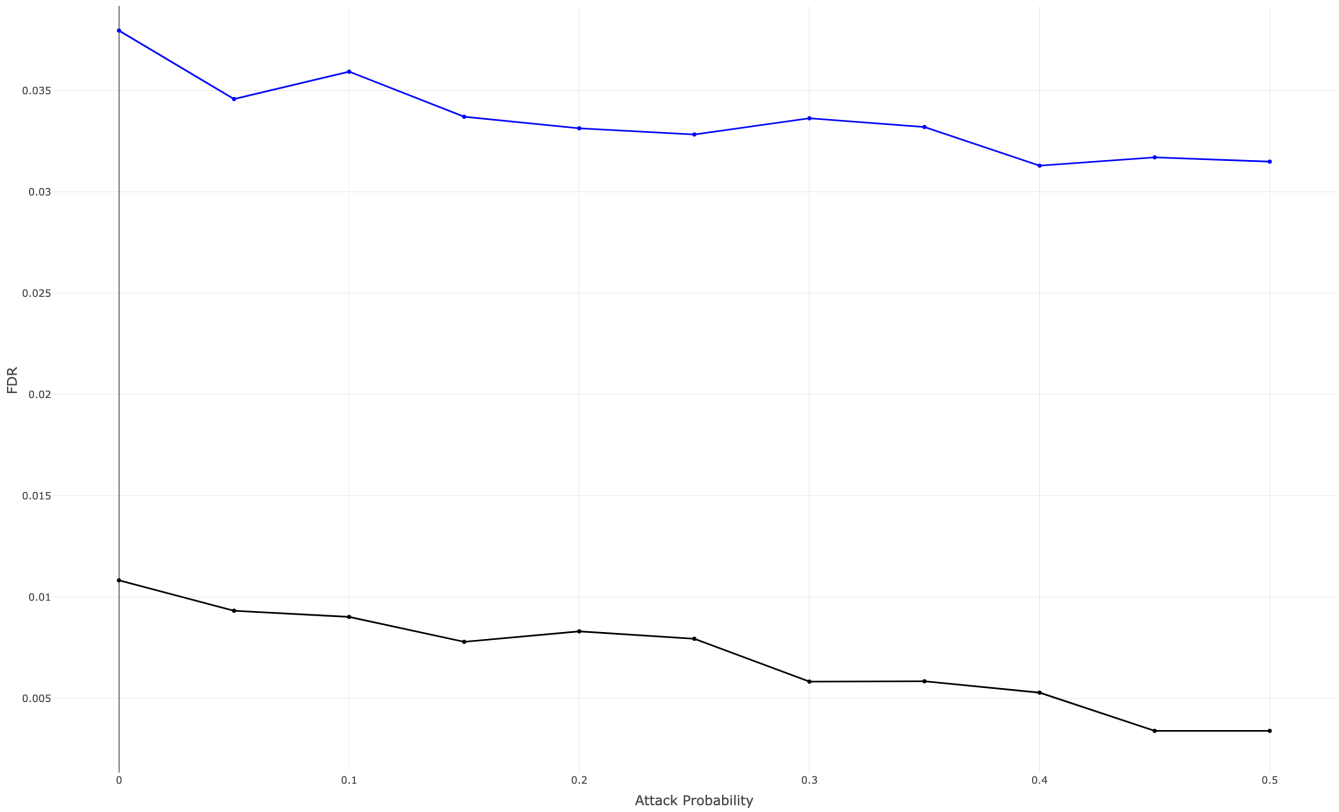

Figure 4.15. FDR for $\zeta = 0.1, \ldots, 0.5$. Results are based on 2,000 replications.

In summary, implementing the online BH algorithm, especially when integrated with the LORD algorithm, enables Blue to protect discoveries that the cascading effect might otherwise compromise, resulting in *increased* statistical power with only a slight rise in the FDR. That is, online BH, together with LORD, is more robust against corrupted data.

CHAPTER 5:
Conclusion

In this chapter, we present our concluding remarks by summarizing the findings of our study and providing recommendations for future research.

This thesis has successfully demonstrated the limitations of prevalent algorithms in maintaining the FDR when subjected to adversarial data manipulation, a common challenge in real-time data environments. Our evaluation reveals that while these algorithms perform adequately in trustworthy data scenarios, their efficacy diminishes significantly if subject to corruption. This finding underscores a crucial vulnerability in statistical testing, particularly in applications where data integrity is critical for accurate decision-making.

We have concentrated our efforts on enhancing the robustness of the LORD algorithm against adversarial conditions. Our research demonstrated that, when suitable for multiple attacks, the integrity and effectiveness of the algorithm could be preserved by inducing "phantom" rejections and integrating the LORD algorithm with the online BH algorithm. Furthermore, when only a single attack is allowed at the first true discovery, adjusting the algorithmic parameters to reduce the decay rate of each test level until this attack happens proved to mitigate the cascade effect effectively.

The practical implications of our research extend beyond the academic realm. In high-stakes environments like the Brazilian Navy's SisGAAz, where real-time data analysis is crucial for maritime surveillance and security, our adapted algorithm may ensure that the system can rely on the accuracy of its analyses.

In conclusion, our investigation highlights statistical algorithms' vulnerabilities to data corruption and introduces methodological advancements for safeguarding their reliability and effectiveness. These findings contribute to the ongoing discourse on algorithmic robustness, offering pathways for future research and application in uncertain adversarial environments.

5.1 Future Work

Despite the successes of recovering power and still being below the FDR threshold using the LORD algorithm as a baseline, there are opportunities for improvements.

Firstly, the same approach could be used to understand how ADDIS and SAFFRON algorithms handle corrupted data.

Secondly, the suggestion to increase the system's robustness against random attacks was to combine both LORD and online BH algorithms. However, the optimal approach would be to implement a "pure" algorithm that can handle data of varying lengths and mean values without using any other algorithm. Moreover, this thesis presented all findings through simulation rather than formal mathematical proof, an area that could also be further explored.

Lastly, additional simulations should be conducted to further enhance and verify the lower bound formulation for the expected number of lost discoveries until the next discovery, as discussed in the Cascade Formulation Effect section.

List of References

Abdi H (2007) Bonferroni and Sidak corrections for multiple comparisons. *Encycl. Meas. Stat* 3.

Aharoni E, Rosset S (2014) Generalized α-investing: Definitions, optimality results and application to public databases. *R. Stat. Soc. Ser. B. Stat. Methodol* 76(4):771–794, https://doi.org/10.1111/rssb.12048.

Andrade I, Franco LG (2018) A Amazônia Azul como fronteira marítima do Brasil: Importância estratégica e imperativos para a defesa nacional [Blue Amazon as Brazil's maritime frontier: Strategic importance and imperatives for national defense]. Pego B, Moura R, eds., *Fronteiras do Brasil: uma avaliação de política pública [Brazilian Borders: A public policy assessment].* (IPEA, Brasilia, Brazil), 151–178.

Andrade I, Rocha A, Franco LG (2021) Blue Amazon Management System (SisGAAz): Sovereignty, surveillance and defense of the Brazilian jurisdictional waters. *Discussion Paper* (12), https://doi.org/10.38116/dp261.

Austin SR, Dialsingh I, Altman N (2014) Multiple hypothesis testing: A review. *J. Indian Soc. Agric. Stat* 68(2):303–14.

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat* 29, https://doi.org/10.1214/aos/1013699998.

Devore JL (2016) *Probability and Statistics for Engineering and the Sciences*, https://www.academia.edu/49732665/Probability_and_Statistics_9E_2016_.

Efron B (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, 1st ed. (Cambridge University Press, Cambridge, UK).

Fisher RA (1992) Statistical methods for research workers. Kotz S, Johnson NL, eds., *Breakthroughs in Statistics* (Springer, New York), 66–70.

Foster DP, Stine RA (2008) α-investing: A procedure for sequential control of expected false discoveries. *R. Stat. Soc. Ser. B. Stat. Methodol* 70(2):429–444 (02), https://doi.org/10.1111/j.1467-9868.2007.00643.x.

Gerhardinger L, Gorris P, Gonçalves L, Herbst D, Nova DV, de Carvalho FG, Glaser M, Zondervan R, Glavovic B (2018) Healing Brazil's Blue Amazon: The role of knowledge networks in nurturing cross-scale transformations at the frontlines of ocean sustainability. *Front. Mar. Sci* 4:395, https://doi.org/10.3389/fmars.2017.00395.

Goodman SN (1999) Toward evidence-based medical statistics. 1: The p value fallacy. *Ann. Intern. Med* 130(12):995–1004.

Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75(4):800–802, https://doi.org/10.1093/biomet/75.4.800.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6(2):65–70.

Hommel G (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75(2):383–386, https://doi.org/10.1093/biomet/75.2.383.

Husseini T (2020) Tracing the history of exploration in the Brazilian pre-salt oil region. Accessed January 31, 2024, https://www.offshore-technology.com/features/pre-salt-oil-region-brazil/.

Javanmard A, Montanari A (2018) Online rules for control of false discovery rate and false discovery exceedance. *Ann. Statist* 46(2):526–554, https://doi.org/https://doi.org/10.48550/arXiv.1603.09000.

Jordan MI (2019) Optional material: Online false discovery rate control. Class notes for Data 102: Data, Inference, and Decisions, Fall 2019, UC Berkeley, Berkeley, CA, USA, https://data102.org/fa21/assets/notes/notes_online_FDR.pdf.

Menyhart O, Weltz B, Győrffy B (2021) Multipletesting.com: A tool for life science researchers for multiple hypothesis testing correction. *PLOS ONE* 16(6):e0245824, https://doi.org/10.1371/journal.pone.0245824.

Ramdas A, Yang F, Wainwright MJ, Jordan MI (2017) Online control of the false discovery rate with decaying memory. *Advances in Neural Information Processing Systems*, https://proceedings.neurips.cc/paper_files/paper/2017/file/7f018eb7b301a66658931cb8a93fd6e8-Paper.pdf.

Ramdas A, Zrnic T, Wainwright M, Jordan M (2018) SAFFRON: An adaptive algorithm for online control of the false discovery rate. *Proc. Mach. Learn. Res*, https://doi.org/https://doi.org/10.48550/arXiv.1802.09098.

Robertson DS, Wason JMS, Ramdas A (2023) Online Multiple Hypothesis Testing. *Statist. Sci* 38(4):557–575, https://doi.org/10.1214/23-STS901.

Rodrigues S (2021) Plano estratégico da Marinha PEM 2040 [Strategic plan of the Brazilian Navy]. *Revista da Escola de Guerra Naval (EGN) [Journal of the Brazilian Naval War College (EGN)]*, 13–30.

Sarkar SK, Chang CK (1997) The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Am. Stat. Assoc* 92(440):1601–1608, https://doi.org/https://doi.org/10.2307/2965431.

Smith M (2023) Brazil's pre-salt oil gains unprecedented global popularity. Yahoo!Finance. Accessed August 31, 2023, https://finance.yahoo.com/news/brazils-pre-salt-oil-gains-210000522.html.

Tian J, Ramdas A (2019) ADDIS: An adaptive discarding algorithm for online FDR control with conservative nulls. *Adv. Neural Inf. Process. Syst* 32, https://doi.org/https://doi.org/10.48550/arXiv.1905.11465.

Zhao Q, Small DS, Su W (2018) Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *J. Am. Stat. Assoc*, https://doi.org/https://doi.org/10.48550/arXiv.1703.09787.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Fort Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California