Statistical Perspectives on Algorithmic Fairness: Quantifying Group Fairness in Thresholding Decisions

Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37379283

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Statistical Perspectives on Algorithmic Fairness
Quantifying Group Fairness in Thresholding Decisions

A THESIS PRESENTED
BY
ANGELA Y. LI
TO
THE DEPARTMENT OF STATISTICS
AND
THE DEPARTMENT OF COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE JOINT DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECTS OF
STATISTICS AND COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2024
Abstract

Machine learning algorithms have become increasingly entrusted with consequential, high-impact decisions over the past few decades; however, numerous examples of the unfairness of their outcomes spawned the creation of the research field of algorithmic fairness over the past decade. Most of the work in algorithmic fairness has primarily focused on rigorously defining fairness as it relates to machine learning procedures and outcomes, and proposing a robust set of methods for correcting for unfairness; however, there still remains a gap in rigorously identifying and quantifying the extent of unfairness in a statistical sense. In this thesis, we provide novel derivations for the distributions of five of the most fundamental group fairness metrics—accuracy, acceptance rate, false positive rate, false negative rate, and positive predictive value—and the distributions of their differences across protected groups. These ultimately serve as the bases for the construction of confidence intervals, which provide a principled framework for rigorously assessing uncertainty and the extent of unfairness with respect to the disparity of group fairness quantities between protected groups. We hope these statistical tools contribute new perspectives and understanding to this highly multidisciplinary field of algorithmic fairness.
# Contents

1 Introduction

2 Background and Literature Review

2.1 Historical Context

2.1.1 Fairness

2.1.2 Mathematical Decision Rules

2.2 The Inception of Algorithmic Fairness

2.2.1 Automated Decision-Making

2.2.2 Sources of Unfairness in Machine Learning

2.3 Binary Classification Primer

2.3.1 Set-Up and Notation

2.3.2 Traditional Evaluation Metrics

2.4 Algorithmic Fairness Definitions

2.4.1 Group Fairness

2.4.2 Individual Fairness

2.4.3 Causal Fairness

2.4.4 Process Fairness

2.4.5 Table of Definitions

2.4.6 Tradeoffs Between Definitions

2.5 Fairness Interventions with Machine Learning Algorithms

2.5.1 Beyond Binary Classification

2.5.2 Evaluating Unfairness
4.6 Main Takeaways ......................................................... 94

5 Metric Difference Distributions and Applications ........................................... 95

5.1 Derivation of Difference Distributions ......................................................... 96
  5.1.1 Uniform Skill, Gaussian Noise ......................................................... 96
  5.1.2 Uniform Skill, General Noise ......................................................... 101

5.2 Empirical Trends of Difference Distributions .............................................. 103
  5.2.1 Difference Variation with Threshold $c$ ............................................. 105
  5.2.2 Difference Variation with Sample Size $n$ ........................................... 105
  5.2.3 Difference Variation with Noise Variance $\sigma^2$ .................................. 107
  5.2.4 Difference Variation with Underlying Distributions ............................. 108

5.3 Constructing Confidence Intervals to Quantify Unfairness ............................ 109
  5.3.1 Approximate 95% Confidence Intervals ............................................ 109
  5.3.2 Coverage Probability with General Noise .......................................... 112

5.4 Computing Fair Optimal Thresholds ......................................................... 114

6 Discussion ......................................................................................... 116

6.1 Our Contributions .................................................................................. 116

6.2 Future Directions ................................................................................. 118

References ............................................................................................. 128
# Listing of figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1</td>
<td>Confusion matrix for binary classification.</td>
<td>12</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Common binary classification metrics derived from the model’s confusion matrix.</td>
<td>15</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Table summarizing popular fairness definitions.</td>
<td>23</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Processing stages in a typical machine learning pipeline.</td>
<td>28</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Simulated FPR distribution vs. proposed approximate Binomial FPR distribution, for Unif(0,100) skill, $\mathcal{N}(0, 10^2)$ noise, threshold $c_a = 50$, and sample size $n_a = 100$.</td>
<td>51</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Simulated FNR distribution vs. proposed approximate Binomial FNR distribution, for Unif(0,100) skill, $\mathcal{N}(0, 10^2)$ noise, threshold $c_a = 50$, and sample size $n_a = 100$.</td>
<td>55</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Simulated PPV distribution vs. proposed approximate Binomial PPV distribution, for Unif(0,100) skill, $\mathcal{N}(0, 10^2)$ noise, threshold $c_a = 50$, and sample size $n_a = 100$.</td>
<td>59</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Simulated PPV distribution (white outline with light filling) vs. proposed approximate Binomial PPV distribution (solid outline with no filling), for multiple skill distributions, $\mathcal{N}(0, 10^2)$ noise, threshold $c_a = 50$, and sample size $n_a = 100$.</td>
<td>60</td>
</tr>
</tbody>
</table>
4.1.1 Mean and error bars (two standard deviations in width) for each of the five fairness metrics, across all integer thresholds $c \in [1, 99]$, with underlying $\text{Unif}(0, 100)$ and $\mathcal{N}(0, 10^2)$ noise.

4.1.2 Standard deviation of the five fairness metrics, across all integer thresholds $c \in [1, 99]$, with underlying $\text{Unif}(0, 100)$ and $\mathcal{N}(0, 10^2)$ noise.

4.1.3 Mean and standard deviation of each of the fairness metrics, across all integer thresholds $c \in [41, 59]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, 10^2)$ noise.

4.1.4 Accuracy mean across all integer thresholds $c \in [41, 59]$, with varying underlying skill distributions and $\mathcal{N}(0, 10^2)$ noise.

4.2.1 Mean of each of the five fairness metrics with error bars representing two standard deviations, across sample size $n \in [50, 500]$, for underlying $\text{Unif}(20, 60)$ skill and $\mathcal{N}(0, 10^2)$ noise.

4.2.2 Standard deviation of each of the five fairness metrics, across sample size $n \in [50, 500]$, for underlying $\text{Unif}(20, 60)$ skill and $\mathcal{N}(0, 10^2)$ noise.

4.3.1 Mean of each of the five fairness metrics with error bars representing two standard deviations, across noise $\sigma \in [1, 25]$, for underlying $\text{Unif}(20, 60)$ skill and $\mathcal{N}(0, \sigma^2)$ noise.

4.3.2 Mean of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, \sigma^2)$ noise.

4.3.3 Standard deviation of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, for underlying $\text{Unif}(20, 60)$ skill and $\mathcal{N}(0, \sigma^2)$ noise.

4.3.4 Standard deviation of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, \sigma^2)$ noise.

4.4.1 Rate of decay of mean of each of the five fairness metrics with error bars representing two standard deviations, across noise $\sigma \in [1, 25]$, for underlying $\text{Unif}(20, 60)$ skill and $\mathcal{N}(0, \sigma^2)$ noise.
4.4.2 Rate of decay of the mean of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, \sigma^2)$ noise. .......................... 89

4.4.3 Rate of decay of the standard deviation of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, for underlying Unif(20, 60) skill and $\mathcal{N}(0, \sigma^2)$ noise. .......................... 90

4.4.4 Standard deviation of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, \sigma^2)$ noise. .......................... 91

4.5.1 Mean and error bars (two standard deviations in width) for each of the five fairness metrics, across all integer thresholds $c \in [1, 99]$, with underlying Unif(0, 100), and zero-centered Gaussian noise (top) and Laplace noise (bottom) with standard deviation 10. .......................... 93

5.1.1 Simulated FPR distribution vs. proposed approximate Binomial difference distribution, for two groups each with Unif(0,100) skill, $\mathcal{N}(0, 10^2)$ noise, threshold 50, and sample size 100. .......................... 98

5.1.2 Simulated FPR distribution vs. proposed approximate Binomial difference distribution, for two groups each with Unif(0,100) skill, $\mathcal{N}(0, 10^2)$ noise, threshold 50, and sample size 100. .......................... 99

5.1.3 Simulated PPV distribution vs. proposed approximate Binomial difference distribution, for two groups each with Unif(0,100) skill, $\mathcal{N}(0, 10^2)$ noise, threshold 50, and sample size 100. .......................... 100

5.2.1 Probability density functions for the distribution of the difference of the five fairness metrics, across different noise standard deviations $\sigma$ 104

5.2.2 Mean and error bars (two standard deviations in width) for the distribution of the difference of each of the five fairness metrics, across all integer thresholds $c \in [41, 59]$, with underlying Unif(0, 60) skill for Group a and underlying Unif(40, 100) skill for Group b, and $\mathcal{N}(0, 10^2)$ noise and sample size 100 for both groups. .......................... 105
5.2.3 Mean and error bars (two standard deviations in width) for the distribution of the difference of each of the five fairness metrics, across different shared sample sizes \( n \in [50, 500] \), with underlying Unif\((0, 60)\) skill, \( \mathcal{N}(0, 10^2) \) noise, and threshold 50 for Group \( a \) and underlying Unif\((40, 100)\) skill, \( \mathcal{N}(0, 10^2) \) noise, and threshold 50 for Group \( b \).

5.2.4 Mean and error bars (two standard deviations in width) for the distribution of the difference of each of the five fairness metrics, across different noise standard deviations \( \sigma \in [1, 25] \) and underlying skill distributions, with threshold 50 and sample size 100 for both groups.

5.2.5 Mean and error bars (two standard deviations in width) for the distribution of the difference of each of the five fairness metrics, across different distances in the centers of the underlying skill distributions between the two groups, with \( \mathcal{N}(0, 10^2) \) noise, threshold 50 and sample size 100 for both groups.

5.3.1 Coverage probability of approximate 95% confidence intervals for the difference of metrics with different \( \), for intervals generated for Groups 1 and 2 both with Unif\((0,100)\) underlying skill distribution, \( \mathcal{N}(0, 10^2) \) noise, threshold 50, and sample size 100.
Acknowledgments

Thank you so much to everyone who has supported me during this senior thesis process, but more broadly all those who have played a role in different parts of my college journey.

To Professor Joe Blizstein, thank you for not only being a wonderful thesis advisor, but even more fundamentally, for playing such a crucial role in igniting and helping nurture my passion for statistics and teaching. Little did I know this class called Stat 110 that I enrolled in on a whim in the last few days before freshman fall, taught on Zoom in the midst of the pandemic, would so deeply pivot and shape my academic interests for the years to come. It has been an utmost privilege to learn from your deep expertise and insightful lecturing in four courses, serve as a teaching fellow (TF) for you seven times, and now work on this year-long exploration into algorithmic fairness with you. I truly cannot thank you enough for all your generous time and consistent support not only on this thesis, but also throughout my academic and personal growth over the past four years, and for always believing in me and guiding me to become a better student, teacher, and person.

To Professors Stephanie Gil and Milind Tambe, I will always be deeply grateful to have had the opportunity to work so closely with the two of you in CS182 last semester; thank you for entrusting me with the role of Head TF, and for sharing your knowledge and expertise in different areas with me. Finally, thank you so much for taking the time to read this thesis.
To Professors Ariel Procaccia, Lucas Janson, Subhabrata Sen, Pragya Sur, Sharad Goel, and Flavio Calmon, and Hans Gaebler, thank you for all the technical wisdom you have imparted on me in various courses, for inspiring me through your unwavering pursuits of academic excellence, and for generously taking the time to chat with me about this project.

Finally, to my family and friends, thank you for your steadfast love and presence in my life; none of this would be possible without your unwavering support.
1

Introduction

In today’s technology-driven age, as machine learning algorithms are increasingly entrusted as arbiters of decision-making, the pursuit—and frequent lack of achievement—of fairness by machines has emerged as a paramount concern among scholars and society alike. From hiring decisions, to credit approvals, to pretrial criminal release, to healthcare allocation, and countless other applications, algorithms have come to wield profound influence over our lives, not only automating trivial day-to-day decisions, but making critical decisions that have significant impact on individuals’ lives and futures. This growing reliance on algorithms and the consequential nature of their decisions thus elevate algorithmic fairness as a field beyond mere technical discourse, prompting reflection on fundamental questions of ethics, justice, and societal values.

The quest for fairness and justice is not merely a contemporary pursuit. Rooted in the age-old principles of distributive justice and moral reasoning, the
discourse surrounding algorithmic fairness traces its lineage thousands of years back to the philosophical musings of Plato, Aristotle, and Confucius, who grappled with questions of virtue, equality, and the common good; these themes continued to amplify through pivotal moments such as the Universal Declaration of Human Rights, the Civil Rights Act, suffrage movements, and labor rights campaigns, each marking a milestone in the ongoing struggle for equality and equity, fueled by an unwavering and innate human desire for and commitment to justice.

In the past decade, however, the proliferation of automated decision-making systems across various sectors during the 21st century has propelled algorithmic fairness to the forefront of scholarly inquiry and societal discourse, especially as a plethora of real-world examples reveal these algorithmic decisions being fraught with extreme unfairness. This recent emergence of algorithmic fairness as a pivotal field of computer science research in our digital age represents a convergence of ancient ideals with contemporary challenges—a synthesis of moral philosophy, social ethics, and technological innovation. Indeed, it is widely agreed upon within the research community today that while rooted in mathematical principles, algorithmic fairness ultimately transcends the boundaries of math, deeply intertwining computer science and statistics with disciplines such as ethics, philosophy, sociology, and law.

Most of the work in algorithmic fairness has focused on quantitatively defining fairness across different dimensions, and proposing interventions that correct for unfairness in different stages of the modeling process. A variety of fairness definitions across different categories—fairness for groups, for individuals, in causal senses, and many more—have been proposed, studied, and widely applied as definitions of fairness when it comes to machines, and many methods have been constructed to correct for unfairness in machine learning models. Algorithms have been found to be unfair for a variety of reasons, including biased data and models that amplify these biases captured in the data, and thus many methods have been proposed to eliminate unfairness. However, there exists relatively little work in actually rigorously identifying unfairness or quantify how
fair or unfair a set of model outcomes is in a statistical sense.

In this thesis, we thus seek to statistically quantify the extent of unfairness with respect to various fundamental group fairness definitions, specifically in the case of thresholding-based decision rules. We do so by deriving the distribution of five of the most foundational group fairness metrics—accuracy, acceptance rate, false positive rate, false negative rate, and positive predictive value—in the specific case where the decision is based on a noisy skill level, composed of a true, Uniformly-distributed underlying skill level and general continuous noise for simplicity. We analyze the trends of these metrics empirically across different theoretical input parameters, providing valuable insights into their behavior and variability.

A key aspect and contribution of our work lies in subsequently deriving and analyzing the distribution of the difference of each of these metrics across protected groups, as this encapsulates the essence of what group fairness seeks to analyze—the disparities between demographic groups. Leveraging these difference distributions, we then construct 95% approximate confidence intervals for the difference of each of these five distributions, which provides a principled framework for assessing the uncertainty inherent in fairness metrics’ estimations. Notably, the construction of these confidence intervals stand as a novel, robust methodology for quantifying and interpreting fairness disparities in thresholding-based algorithmic decision-making and better allow one to answer how fair or unfair a set of model outcomes are with respect to one of these five group fairness metrics.

While the field of algorithmic fairness is still plagued by ethical and philosophical questions that cannot be answered by math alone, we hope that our work provides tools for quantifying the extent of group unfairness and deepens statistical understanding of threshold-based group fairness. By offering insights into fairness metric distributions and using these for rigorous uncertainty quantification, we aspire to make meaningful statistical contributions to this collective effort at paving the way towards a future where algorithms serve as instruments of justice and equity, rather than perpetrators of inequality and
injustice.

The remainder of this thesis is organized as follows. In Chapter 2, we contextualize the broad issues surrounding fairness and justice, and review existing literature in the field of algorithmic fairness, including classes of fairness definitions and fairness interventions. Next, in Chapter 3, we introduce the threshold-based binary classification setting used throughout the rest of the thesis, and derive the distributions of five of the most fundamental binary classification metrics central to group fairness definitions. In Chapter 4, we empirically analyze these distributions to better understand their trends across different dimensions of the parameters of our setting at hand. In Chapter 5, we use the distributions derived in Chapter 3 to derive distributions of the difference of these metrics, which we then analyze and use to propose 95% confidence intervals for evaluating the extent of fairness in a given thresholding setting. Finally, in Chapter 6, we summarize our contributions and discuss future directions for this research.
Background and Literature Review

We begin by outlining a brief history of the field of fairness, widespread definitions and terminology central to algorithmic fairness, and current state-of-the-art techniques to address unfairness in machine learning algorithms.

2.1 Historical Context

2.1.1 Fairness

Injustice, inequality, and polarization — all inherently referring to a lack of fairness or justice, especially in the treatment of different groups of people — have been present societal themes for millennia. Social patterns of racism, slavery, patriarchy, genocide, classism, and more have been present in a number of societies around the world for thousands of years, and have inflicted generations.
of horrific dehumanization, pain, and suffering on affected people groups.

Even after the formal abolishment of slavery, these themes have persisted and made their way into formal regulations, for example in 1881 when Prudential announced that life insurance plan rates would explicitly take into account race and differ on the basis of it, upcharging Black individuals over their White counterparts [83]. Clearly discriminatory, these only further perpetuated cycles of deep-rooted injustice.

Starting from the mid-20th century, a significant period of unprecedented efforts to socially and legally rectify racial injustice began in the United States. Perhaps most notably, the Civil Rights Act of 1964 legally prohibited discrimination on the basis of race, sex, religion, and national origin [23], and the Fair Housing Act of 1968 was a similar attempt to alleviate injustice. Since then, there have also been many court rulings on the basis of fairness and equality of treatment, such as with respect to educational opportunity [32].

Especially into the 21st century, the presence and urgency of social movements, legislation, and scientific-based methods aimed at eliminating injustices against various people groups have only continued to surge, further calling attention to the issue of injustice in society and humans’ innate desire for fair treatment of individuals and differently-identifying groups. Next, we explore the history of using mathematical principles as a basis for regulating decisions.

2.1.2 Mathematical Decision Rules

The notion of formulating decisions or decision rules mathematically extends far into the past [67]. The usage of mathematical ideas to define notions of justice has been around since ancient history, dating as early as Aristotle discussing the usage of mathematics as the basis of justice rules [14].

This proliferated more in recent centuries after the fields of mathematics and statistics saw more revolutionary developments. Notably, eighteenth-century mathematicians Marquis de Condorcet and Pierre-Simon Laplace reasoned about using probability as a basis for judicial practices and certain areas of law [30]. In
the twentieth century, the influence of statisticians and eugenicists Francis Galton and Karl Pearson and their newly proposed statistical methods such as regression and correlation allowed for more detailed statistical analyses on an individual level, although the initial grounds on which these were first introduced likely are not free from prejudice, given their eugenics backgrounds, already hinting at some of the many challenging questions to come in future algorithmic fairness work [72].

In terms of formalizing decision rules for individuals mathematically, consumer credit bureaus began mass adopting statistical credit scoring in the 1960s [60]. At a similar time, in the 1970s, “actuarial fairness” was a response to the civil rights movement from the insurance industry regarding race-based insurance pricing, and was underlined by the idea that customers with the same risk should be charged the same rate, an intuitive idea of fairness that is still present in the field today [42].

2.2 THE INCEPTION OF ALGORITHMIC FAIRNESS

2.2.1 AUTOMATED DECISION-MAKING

The last decade especially has seen an explosion in popularity of algorithmic fairness, as machine learning has been increasingly used to automate decisions that are more and more impactful on peoples’ lives. Today, machines are used across a multitude of fields to make significantly consequential decisions [22], including but certainly not limited to lending [89], credit scoring [13], crime prediction [54], hiring [11], and childhood welfare systems [21].

A few large reasons for the rapid proliferation of machine learning algorithms over the past few decades has been the efficiency of these algorithms in significantly expediting repetitive processes, and often being more reliable and precise than humans [74]; there are predictions, for example, that 300 million jobs could be lost to artificial intelligence [53]. While these still largely hold, it has also become abundantly clear that while these models are extremely powerful
and useful, they are often also imperfect and unfair. Note that here, we generally refer to **fair machine learning** as the nonexistence of bias or prejudice against any individuals or groups, implicitly or explicitly, in the model and decision-making process.

Machine learning has certainly fallen short from this standard of fairness, as exhibited by a number of concerning examples in recent years. Perhaps most famously, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software was shown to inaccurately predict recidivism—one’s risk of recommitting a crime—to be higher for people of color [1] [20]; studies have even shown that COMPAS achieves no better results than the basic rational judgement of an average person [34], a troubling finding. Text data has been found to contain many implicit human biases [16], word2vec embeddings trained on Google News text data reproduced biases in the data [12], Google ad delivery was found to be biased towards blacks [77], STEM career ads were shown more often to men than women [59], all resulting from the usage of machine learning in these decision processes. These select examples, only a small sample of a variety of other issues [69], highlight some of the dimensions of unfairness exhibited by machine learning algorithms today, which we more formally explore in the next subsection.

At this point, we note that the fundamental question of whether or not it is even ***appropriate*** to use machine learning in a given setting, and other related topics such as how to appropriately correct for historical injustices and discrimination in data and to what extent machine learning should reflect human biases, are all open ethical debates which transcend many fields; our goal is not to answer these. Multiple scholars have provided their thoughts on the answers to such questions, including studying how philosophical fairness ideas relate to technical machine learning concepts [9], and these issues continue to be widely discussed and debated among the research community [2]. Instead of foraying into these philosophical questions, we instead focus on how statistics can better inform the field of algorithmic fairness, in settings where data has already been collected and machine learning is already the chosen tool.
2.2.2 Sources of Unfairness in Machine Learning

Having identified empirically that machine learning exhibits a large amount of unfairness, it becomes salient to examine where exactly this bias in machines originates from. This issue has been extensively studied in fairness literature [22] [65] [68] [51], as it is central to determining what interventions are possible and necessary to correct for such bias.

Formally, we isolate three main sources of bias in machine learning. The first centers around the idea that humans are inherently biased creatures, often rooted in historical and societal norms, and thus these natural human biases will also be reflected by the data.

The second source of bias is the data collection process itself, as there are a large number of ways in which bias can enter during the data creation and collection process. Specifically, over- or under-sampling from certain groups, sampling non-randomly, or not collecting enough data will all result in non-representative training sets for the model and potentially introduce additional human bias or prejudice into the model (if, for example, the under-sampled groups are historically marginalized or underprivileged). For example, a study on two very popular and widely-used image datasets for machine learning, ImageNet and Open Images, found that the datasets were misrepresentative and lacked diversity [76]. Additionally, bias in how features are chosen and measured is also prevalent, as are other sources of data collection bias [65].

Third, the models themselves can often enhance the bias in the data. Fundamentally, overfit models will over-learn and amplify trends in the data; mathematically, optimizing for loss functions, especially those that measure prediction error, will fit to majority data to optimize overall error, which inherently leads to larger errors and a worse fit for minority populations, which can certainly be argued to be unfair [22]. Modeling processes can also create positive feedback loops where bias in the model leads to data reflecting that bias, which when fed back into the model, continues amplifying the original bias in the
model itself, even in the absence of any of this bias in the original data [61].

From this, it is clear that there is a lot of existing human bias in data, and the data collection process often encodes more human biases. Models then learn these human biases from the data and can even amplify them, as they are inherently designed to learn the train data well, and thus naturally replicate existing biases in the data; there exist no mathematical or intuitive reasons to expect these models to naturally be able to remove any of the bias in the data [22].

Given the issues presented in this section, a wide range of fairness definitions compatible with machine learning algorithms have been identified and studied by the research community, which we now introduce.

2.3 Binary Classification Primer

A large majority of the work done in algorithmic fairness focuses on binary classification, which we will discuss further later in this literature review. Here, we also focus on the binary classification setting, and define common notation for different components of binary classification problems.

2.3.1 Set-Up and Notation

In binary classification, the goal is to predict the true binary class label of a particular observation, given some of its covariates. Typically, $X_i \in \mathbb{R}^d$ is taken to be a vector of observed non-protected covariates for a given individual with index $i$ (e.g., non-sensitive demographic information) where $d$ is the number of observed covariates, $A_i$ individual $i$’s discrete protected attribute (e.g., race or gender), $Y_i \in \{0, 1\}$ their true binary outcome for the decision of interest (e.g., whether they ended up recommitting a crime), and $\hat{Y}_i = f(X_i)$ a prediction for $Y_i$ outputted from the binary classification model, represented abstractly by $f$. Let the population size be $n$ here.

Protected (sometimes also called sensitive) features or attributes are typically determined case-by-case and can broadly refer to any characteristic of individuals that are considered sensitive as determined by the legal, ethical, and
contextual considerations of the setting at hand, which could be demographic features, attributes individuals were born into or cannot change about themselves, or attributes that should not affect one’s decision without being ethically considered discriminatory [71]; common examples include race, gender, age, socioeconomic status, disability status, sexual orientation, and religion. However, most of the time, protected features are characteristics of individuals that are subject to formal legal protection against discrimination, which encompass all of the aforementioned demographic attribute examples and more, via laws such as the Civil Rights Act of 1964, the Fair Housing Act, the Americans with Disabilities Act, and numerous others [2]. The same terminology is used in the term protected groups, which generally refer to a population split into groups based on a particular protected attribute (e.g., race).

As a simple motivating example consider a job application setting where \( X_i \in \mathbb{R} \) is candidate \( i \)'s score on a job screening examination, \( A_i \) is candidate \( i \)'s race, \( Y_i = 1 \) if candidate \( i \) is truly qualified for the job and 0 otherwise (which could be determined a variety of ways, e.g., some aggregate measure of ability or IQ), and \( \hat{Y}_i = 1 \) if candidate \( i \) is accepted for the job and zero otherwise. We will use this as a running example throughout this chapter to illustrate different notions and properties of fairness.

2.3.2 Traditional Evaluation Metrics

Most of the traditional metrics used to evaluate binary classification models can be derived from the model’s confusion matrix, which summarizes the number of observations that fall into each of the four categories of \((Y, \hat{Y})\) values; this is visually depicted in Figure 2.3.1. Using the confusion matrix, a variety of different evaluation metrics can be written for binary classification, many of which are widely used in literature and have become the fundamental metrics by which binary classification is typically evaluated.

We note that while the confusion matrix provides the foundation for defining these metrics, they are all sample quantities since they are computed on the data
in a particular setting; thus, we also provide definitions for the theoretical
versions of these quantities, which are also of central importance in literature [2]
and our later analysis.

**Definition 2.3.1 (Accuracy).** The **accuracy** of a model is the probability it
outputs a correct assignment. Theoretically, the accuracy is defined as

\[ \text{ACC} := P(\hat{Y} = Y). \]

For a given model, the (sample) accuracy is the proportion of people who were
correctly classified, and can be defined using confusion matrix quantities as

\[ \hat{\text{ACC}} := \frac{TP + TN}{TP + TN + FP + FN} \]

In our running example, the sample accuracy is the number of people who
received their “true” or “correct” acceptance with respect to the job, based on
their skill level.

**Definition 2.3.2 (Acceptance Rate).** The **acceptance rate** of a model is the
probability an observation is predicted to the positive class. Theoretically, the
acceptance rate is defined as

\[ AR := P(\hat{Y} = 1). \]

For a given model, the (sample) acceptance rate is the proportion of people who
were classified to the positive class, and can be defined using confusion matrix
quantities as
\[ \hat{AR} := \frac{TP + FP}{TP + TN + FP + FN} \]

In our example, the sample acceptance rate is the proportion of people who
were accepted for the job.

**Definition 2.3.3** (False Positive Rate). The **false positive rate (FPR)** of a
model is the probability a negative-class observation is predicted to the positive
class. Theoretically, the FPR is defined as
\[ FPR := \mathbb{P}(\hat{Y} = 1 \mid Y = 0). \]

For a given model, the (sample) FPR is the proportion of people in the negative
class who were predicted to the positive class, i.e.,
\[ \hat{FPR} := \frac{FP}{FP + TN}. \]

If \( FP + TN = 0 \), then let \( \hat{FPR} = 0 \).

In our running example, the sample FPR is the proportion of people who
should not have been accepted for the job but were accepted.

**Definition 2.3.4** (False Negative Rate). The **false negative rate (FNR)** of a
model is the probability a positive-class observation is predicted to the negative
class. Theoretically, the FNR is defined as
\[ FNR := \mathbb{P}(\hat{Y} = 0 \mid Y = 1). \]

For a given model, the (sample) FNR is the proportion of people in the positive
class who were predicted to the negative class, i.e.,
\[ \hat{FNR} := \frac{FN}{FN + TP}. \]

If \( FN + TP = 0 \), then let \( \hat{FNR} = 0 \).
In our example, the sample FNR is the proportion of people who should have been accepted for the job but were not.

**Definition 2.3.5 (Positive Predictive Value).** The *positive predictive value (PPV)* of a model is the probability a predicted-positive observation actually came from the positive class. Theoretically, the PPV is defined as

\[ PPV := \Pr(Y = 1 \mid \hat{Y} = 1). \]

For a given model, the (sample) PPV is the proportion of people predicted to the positive class who actually have positive labels, i.e.,

\[ \hat{PPV} := \frac{TP}{TP + FP}. \]

If \( TP + FP = 0 \), then let \( \hat{PPV} = 0 \).

In our running example, the sample PPV is the proportion of people who were accepted for the job who were truly qualified for the job.

As a comment on notation, we note that in the next chapter, we reintroduce all of the sample quantities with more rigorous mathematical notation for the specific setting we consider in all of our later work, and also explicitly define the quantities separately for different groups. For the rest of this chapter, we maintain these more loose intuitive definitions, and use a subscript notation when referring to a metric evaluated on one specific group (e.g., \( \hat{ACC}_a \) refers to the accuracy of the group with protected attribute \( a \) specifically, and is computed exactly the same as in 2.3.1, except with the analogous quantities for group \( a \) specifically).

We summarize these and other common confusion matrix-derived metrics [2] [80] in the Table 2.3.2 below.

Equipped with these basic binary classification evaluation metrics, we now introduce the different definitions used to evaluate algorithmic fairness.
Table 2.3.2: Common binary classification metrics derived from the model’s confusion matrix.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Shorthand</th>
<th>Theoretical</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>ACC</td>
<td>$\mathbb{P}(\hat{Y} = Y)$</td>
<td>$\frac{TP + TN}{TP + TN + FP + FN}$</td>
</tr>
<tr>
<td>Acceptance Rate</td>
<td>AR</td>
<td>$\mathbb{P}(\hat{Y} = 1)$</td>
<td>$\frac{TP + FN}{TP + FP}$</td>
</tr>
<tr>
<td>True Positive Rate</td>
<td>TPR</td>
<td>$\mathbb{P}(\hat{Y} = 1 \mid Y = 1)$</td>
<td>$\frac{TP + FN}{TP}$</td>
</tr>
<tr>
<td>True Negative Rate</td>
<td>TNR</td>
<td>$\mathbb{P}(\hat{Y} = 0 \mid Y = 0)$</td>
<td>$\frac{TN + FP}{TN}$</td>
</tr>
<tr>
<td>False Positive Rate</td>
<td>FPR</td>
<td>$\mathbb{P}(\hat{Y} = 0 \mid Y = 0)$</td>
<td>$\frac{TN + FP}{TN}$</td>
</tr>
<tr>
<td>False Negative Rate</td>
<td>FNR</td>
<td>$\mathbb{P}(\hat{Y} = 0 \mid Y = 1)$</td>
<td>$\frac{FP}{TP}$</td>
</tr>
<tr>
<td>Positive Predictive Value</td>
<td>PPV</td>
<td>$\mathbb{P}(Y = 1 \mid \hat{Y} = 1)$</td>
<td>$\frac{TP + FN}{TP}$</td>
</tr>
<tr>
<td>Negative Predictive Value</td>
<td>NPV</td>
<td>$\mathbb{P}(Y = 0 \mid \hat{Y} = 0)$</td>
<td>$\frac{TN + FP}{TN}$</td>
</tr>
<tr>
<td>False Discovery Rate</td>
<td>FDR</td>
<td>$\mathbb{P}(Y = 0 \mid \hat{Y} = 1)$</td>
<td>$\frac{TP + FN}{FP}$</td>
</tr>
<tr>
<td>False Omission Rate</td>
<td>FOR</td>
<td>$\mathbb{P}(Y = 1 \mid \hat{Y} = 0)$</td>
<td>$\frac{TN + FN}{FN}$</td>
</tr>
</tbody>
</table>

2.4 Algorithmic Fairness Definitions

Dozens of definitions have been proposed in the field of algorithmic fairness, all trying to capture different dimensions and notions of fairness. The well-known “21 Fairness Definitions and their Politics” lecture, which also served as some of the inspiration for this thesis, focuses on summarizing and drawing connections between these [66]. We introduce some of the most salient classes of definitions to contextualize the field, and more rigorously define the definitions that we will be using throughout the rest of this paper.

2.4.1 Group Fairness

Group fairness, also known as statistical fairness, is one of the main branches of algorithmic fairness; this class of definitions fixes a small number of demographic groups differentiated by some protected attribute and seeks rough parity of some statistical measure across this partition into groups [22]. Most of these measures seek to equalize some subset of the traditional binary classification evaluation metrics introduced in Section 2.3.2.

We formally define a few of the most foundational and widely-used group
fairness definitions here, which fundamentally depend on the five binary classification metrics we introduced in Section 2.3.2. Before doing so, however, we contextualize the significance of the metrics in our work specifically. Namely, these definitions largely focus on simple equality or inequality of a metric across protected groups when deeming a model is fair or unfair; exact numerical equality across all groups, however, is a very strong condition that is rarely achieved. Thus, in our work, which largely focuses on group fairness ideas, we seek to contribute new insights by analyzing the significance of differences of a given binary classification evaluation metric across protected groups (e.g., the difference in acceptance rates between two populations) across the entire spectrum of possible values, instead of just the event that acceptance rate equality is exactly achieved as traditional group fairness metrics require. So, while we do not exactly operate on the following group fairness metrics themselves, they serve as foundational principles for many of the ideas we introduce and consider later.

**Definition 2.4.1 (Accuracy Equality).** *Overall accuracy equality requires that the classifier’s accuracy is equal across all protected groups [6] [80]. With our notation, this requires the model to satisfy*

\[ AR_i = AR_j \]

*for all groups i, j.*

In our example, overall accuracy equality would be achieved if the model’s performance was equivalent across all of the protected groups, i.e., the racial subgroups within the population. To revisit our final point prior to Definition 2.4.1 above for clarity, accuracy equality requires all protected groups to have the same accuracy, which is a very strong condition; in our later work, we instead focus on quantifying the significance of the difference in accuracies between protected groups, which encompasses the case when accuracy equality is satisfied but also looks at cases where it is not exactly satisfied.
Definition 2.4.2 (Statistical Parity). **Statistical parity** (also known as **demographic parity**), originally formalized by Dwork et al. [35], asks that the acceptance rates as defined above are equalized across all protected groups, i.e.,

\[ ACC_i = ACC_j \]

for all groups \(i, j\). Demographic parity is sometimes also written as

\[ \hat{Y} \perp A \]

when considering the probabilities in terms of independence [25].

In our example, statistical parity would be achieved if all racial groups were accepted for the job at the same rate. Note that statistical parity essentially enforces that all groups have the same probability of obtaining favorable (and unfavorable) outcomes. Sometimes, analogous definitions are used where instead of requiring the acceptance rates to be exactly equalized, a relaxed constraint is considered which requires the difference between any two groups’ acceptance rates is within some small, often arbitrary deviation \(\epsilon\) of each other [2].

Definition 2.4.3 (Predictive Equality). **Predictive equality** (also known as **FPR balance**) requires that all protected groups have the same FPR, i.e.,

\[ \text{FPR}_i = \text{FPR}_j \]

for all groups \(i, j\) [20] [26]. In terms of independence relations, this is equivalent to

\[ \hat{Y} \perp A \mid Y = 0. \]

Definition 2.4.4 (Equal Opportunity). **Equal opportunity** (also known as **FNR balance**) requires that all protected groups have the same FNR, i.e.,

\[ \text{FNR}_i = \text{FNR}_j \]
for all groups \( i, j \) \[20\] \[41\]. In terms of independence relations, this is equivalent to

\[
\hat{Y} \perp A \mid Y = 1.
\]

**Definition 2.4.5** (Equalized Odds). Equalized odds\(^7\) (also known as disparate mistreatment and conditional procedure accuracy equality) requires that both predictive equality and equal opportunity are satisfied, i.e., both false positive and false negatives rates are equalized across all groups,

\[
FPR_i = FPR_j \quad \text{and} \quad FNR_i = FNR_j
\]

for all groups \( i, j \) \[41\] \[20\] \[57\] \[6\]. This is thus equivalent to

\[
\hat{Y} \perp A \mid Y.
\]

In our example, equalized odds requires that all racial groups have equal error rates (FPR and FNR) with respect to job acceptance.

**Definition 2.4.6** (Predictive Parity). Predictive parity requires that all protected groups have equal PPV, i.e.,

\[
PPV_i = PPV_j
\]

for all groups \( i, j \) \[20\]. In terms of independence, this can be written as

\[
Y \perp A \mid \hat{Y} = 1.
\]

In our example, predictive parity would require all racial groups’ predictive power for the positive class to be equivalent.

Here, we again emphasize the fact that traditional group fairness metrics rely on a very strong constraint, namely that a (or multiple) particular binary classification metric is equalized across all protected groups; this is an especially stringent constraint if there is a moderate or large number of groups, as with \( n \) protected groups, the number of pairs \( \binom{n}{2} \) grows quadratically in \( n \), since
Thus, instead of requiring these constraints necessarily be satisfied, we seek to statistically quantify the relative significance of values across the entire spectrum of feasible values for a particular metric or metric difference.

Beyond the six definitions above, a number of other popular group fairness definitions exist, such as **conditional parity** [73], which enforces equal prediction probabilities regardless of group membership for all individuals with the same covariates $X = x$, and **calibration within groups** [20], which enforces equally well-calibrated probability estimates with respect to a score function across all groups.

Group fairness definitions are popular for a variety of reasons, but also fall short in many desired qualities [22]. Their merit largely lies in being simple, intuitive measures that are easily computable, interpretable, and verifiable, and which require no additional assumptions on any of the data in order to evaluate.

Beyond scientists, policymakers, news reporters, and other scholars, among many others, often still turn to these group definitions to evaluate fairness with various ethical and philosophical reasons. For example, in Idaho, proposed legislation aimed to require equal error rates across groups [25]; Texas House Bill 588 (better known as the “Top 10 Percent Law”) guarantees that the top ten percent of the graduating class of all public and private Texas high schools receive automatic admission to all state-funded university, providing a baseline guarantee on the acceptance rates of schools from different socioeconomic areas of Texas [28]. A variety of other policies and principles put into place in other contexts have also focused on equalizing various of these group definitions.

As an example of some of the philosophical arguments for using group fairness metrics, when it comes to demographic parity, its merit lies in providing representation across groups in the outcomes of a model. Individuals’ qualifications may differ because of a variety of factors out of their control, such as historical injustices, and thus a common argument in favor of using demographic parity in appropriate settings is that because of historical prejudice, groups may differ in their objective qualification, but should be equally deserving of having access to an opportunity across groups [2]. Another argument for
demographic parity (instead of accuracy equality, for example) is that certain resources or allocations will benefit some groups—usually marginalized populations—more than others, and thus from a utility perspective, this is sound and justifiable, even if the actual model accuracy is slightly suboptimal.

As another example, many intuitive and ethical arguments exist in favor of equalized odds being a favorable choice of a fairness metric. The reason for this is that higher error rates are usually associated with historically marginalized groups, thus inflicting further damage on them, which is incredibly unfair, and thus forcing equalized error rates across groups alleviates this burden. When a group has disproportionately high error rates, this could also end up perpetuating cycles of inequality, which is fundamentally the antonym of fairness [2]. Concretely, mistakenly classifying someone (e.g., as not deserving of acceptance, or high risk for recidivism when they are actually low risk) could have lots of snowballing effects (e.g., denying pretrial release because they are incorrectly deemed high risk for recidivism), and thus unequal error rates across groups puts disproportionately large burden on groups with higher error rates, which again are usually the marginalized groups to begin with.

However, group fairness metrics have quite a few weaknesses. Importantly, they do not give meaningful guarantees to individuals by themselves, and rather only give average guarantees on groups as a whole, which is a relatively weak result; for example, statistical parity could be satisfied, but the classifier could be incredibly unfair against certain individuals, even if on average the acceptance rates across groups are equalized. Further, these definitions may only be satisfied for one set of subgroups, but not all partitions into subgroups which might be ethically meaningful in a given setting. Lastly, there are also key tradeoffs when satisfying group fairness metrics, which we discuss more later. Nonetheless, the popularity and widespread applicability of group fairness metrics remain, and we also aim to make some of this more robust in our work.
2.4.2 Individual Fairness

Individual fairness is centered around the concept introduced by Dwork et al. that “similar” individuals should be treated similarly [35]. Formally, if the distance between two individuals is small for some distance metric, then the distance between the distributions over outcomes of two individuals must also be small, where the distribution of outcomes for an individual with covariates $x$ is given by some function $M(x)$. In the case of our job application example, individual fairness would be satisfied if individuals with similar scores received similar acceptances (or rejections) from the model. A similar definition under the scope of individual fairness is that less qualified individuals should not be favored by the model over more qualified individuals [47].

Intuitively, these definitions seem very reasonable, but issues with their practicality and applicability quickly become evident. Of central concern is the question of what a reasonable distance metric is to measure the “distance” between individuals, which in itself, seems like a very complicated and controversial philosophical question, and does not have a clear answer [22]. The necessity of such a distance metric and corresponding assumptions in order to employ individual fairness definitions significantly limits their practicality.

2.4.3 Causal Fairness

Group fairness depends only on aggregate measures computable from the data itself; causality-based fairness, on the other hand, tries to model the actual causal structure between attributes in order to identify the causal effects of sensitive attributes on outputs [25]. Counterfactual fairness, for example, considered to be an individual-level decision, requires that if someone’s group membership were changed while holding all other features constant, then the probability they are predicted to the positive class should stay the same, with their modified group membership [58] [86]. In our running example, counterfactual fairness would be satisfied if every individual, in some alternate world, would receive the same outcome from the model if their race were counterfactually altered, holding their
screening exam score fixed. Multiple other related metrics also exist in the context of causal or counterfactual fairness [25].

While philosophically meaningful, counterfactual fairness is quite abstract and requires assumptions that are often unverifiable, and thus frequently impractical to apply in real-world modeling; for example, considering one’s counterfactuals if they also had their demographics modified starts to lean quite hypothetical, and requires assumptions that are challenging to verify. Nonetheless, a variety of meaningful work has been done in causality-based fairness [27] [44] [55].

2.4.4 Process Fairness

The main definition behind process fairness is “fairness through unawareness,” which says that as long as the protected attributes or group membership are not explicitly used in the decision-making process, then the model is fair [40]. In our example, as long as the decision process does not use race explicitly, then the model is considered to be fair with respect to this definition.

One of the first attempts at defining fairness formally in the context of machine learning, this definition has obvious shortcomings in the cases where other features are highly correlated with the protected attribute, as even if not explicit, this indirectly introduces the sensitive group membership information. Thus, while simple in nature, process fairness only really works in the rare case that there exists no correlation at all between the protected attribute and the rest of the features, which is obviously extremely unrealistic in most real-world contexts.

2.4.5 Table of Definitions

A summary of the definitions defined in the sections above are given in 2.4.1 below; more extensive versions of this can be found throughout literature as well [2] [80] [81]. In the table, “Equalized Quantity” refers to the quantity that must be equalized across all protected groups (i.e., across all values $a, a'$ of the protected attribute $A$), $y, \hat{y} \in \{0, 1\}$, score $S$, and individuals $i, j$, where present,
in order for the definition to be satisfied.

2.4.6 Tradeoffs Between Definitions

With such a wide range of fairness definitions, including those laid out above, a variety of other named definitions \cite{80} \cite{18}, and new definitions regularly introduced into the research community continuously, many philosophical questions remain regarding how to use them.

For example, regarding the question of whether there exists a “one size fits all” fairness definition or “best of both worlds” between group fairness and individual fairness, most scholars agree that the answer is no, and that the usage of these definitions is very application-dependent \cite{7}. The question of which definition is “best” or most appropriate to use in a given context is also relatively understudied, and one that certainly relies on both technical and ethical rationale.

This being said, a highly related central theme in algorithmic fairness that has been extensively studied is the idea that when actually choosing any definition over another, costly mathematical and moral tradeoffs almost always exist. We examine a few of these key tradeoffs below.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Class</th>
<th>Equalized Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy Equality</td>
<td>Group</td>
<td>$\mathbb{P}(Y = Y \mid A = a)$</td>
</tr>
<tr>
<td>Statistical Parity</td>
<td>Group</td>
<td>$\mathbb{P}(\hat{Y} = 1 \mid A = a)$</td>
</tr>
<tr>
<td>Predictive Equality</td>
<td>Group</td>
<td>$\mathbb{P}(\hat{Y} = 1 \mid Y = 0, A = a)$</td>
</tr>
<tr>
<td>Equalized Odds</td>
<td>Group</td>
<td>$\mathbb{P}(\hat{Y} = 0 \mid Y = 1, A = a)$</td>
</tr>
<tr>
<td>Predictive Parity</td>
<td>Group</td>
<td>$\mathbb{P}(\hat{Y} = \hat{y} \mid Y = y, A = a)$</td>
</tr>
<tr>
<td>Conditional Parity</td>
<td>Group</td>
<td>$\mathbb{P}(\hat{Y} = 1 \mid R = r, A = a)$</td>
</tr>
<tr>
<td>Calibration Within Groups</td>
<td>Group</td>
<td>$\mathbb{P}(Y = 1 \mid R = r, A = a) = r$</td>
</tr>
<tr>
<td>Lipschitz Property</td>
<td>Individual</td>
<td>$D(M(x_i, M(x_j)) \leq d(x_i, x_j)$</td>
</tr>
<tr>
<td>Counterfactual Fairness</td>
<td>Causal</td>
<td>$\mathbb{P}(\hat{Y}_{A=a'} = \hat{y} \mid A = a, x) = \mathbb{P}(\hat{Y} = \hat{y} \mid A = a, x)$</td>
</tr>
<tr>
<td>Fairness Through Unawareness</td>
<td>Process</td>
<td>$\hat{y} = F(x), \ A \notin X$</td>
</tr>
</tbody>
</table>

**Figure 2.4.1:** Table summarizing popular fairness definitions.
TRADEOFFS WITHIN GROUP FAIRNESS

A few critical tradeoffs exist within the class of group fairness definitions. One of the most influential results in the field is Kleinberg, Mullainathan, and Raghavan’s impossibility theorem which states that it is not possible to satisfy both equalized odds and protected parity across all protected groups except in trivial cases [57]; other works also deduce similar tradeoffs between group fairness metrics [20] [2].

Basic observations also give similar intuition on the permanent presence of trade-offs within group fairness definitions. Simply considering a case when the prevalence of the positive class is significantly different between groups, then an optimal predictor will likely have different error rates between groups; thus, enforcing equalized error rates will decrease model accuracy, as the model will perform less accurately than it could for at least some individuals, which is arguably unfair [22]. Then, depending on the setting, enforcing equalized odds at the cost of accuracy could be seen as more or less fair, but this simple case demonstrates a fundamental trade-off that exists between these constraints, all of which are desirable to achieve but cannot be simultaneously satisfied.

BETWEEN GROUP FAIRNESS AND INDIVIDUAL FAIRNESS

The question of bridging group fairness and individual fairness has been frequently asked in the field of algorithmic fairness. It has been shown that in very specified cases, individual fairness can guarantee demographic parity [35]; however, these conditions generally do not hold. There have been multiple attempts to find a middle ground between the two classes of definitions, for example expanding group fairness requirements to hold not only one particular demographic partition, but on exponentially large classes of protected groups [52], and using oracles to relax individual fairness assumptions [56]. Ultimately, neither of these are able to supply the mutual guarantees of both definitions, however.

Some argue that the seeming incompatibility between group and individual
fairness is more because of their rash application, and not fundamental incompatibility [9], though this remains a largely unanswered question.

**Philosophical Tradeoffs**

Earlier, we discussed some of the different motivations for different fairness metrics, and it is evident that the choice of metric largely differs based on the motivations of the party. For example, an individual might feel like they were treated unfairly if they were treated very differently to someone whose features match theirs exactly, but, from the perspective of a policymaker, the consideration might be more holistic. Thus, as previously mentioned, the question of when to use what metric could have some mathematical underpinnings, but certainly extends beyond just technical considerations to the ethical, philosophical, and legal fields.

Again, it is widely agreed that there is no one “correct” definition to satisfy all fairness desiderata, and that the (often still very subjective) choice of metric is highly setting-dependent and motivation-dependent [2]. Thus, we do not seek to answer whether a choice of fairness definition is correct or not, or which definitions are most reasonable to use in a given situation, as these questions go far beyond statistics itself and are much beyond the scope of this paper. Rather, we only aim to provide tools and insights to better understand the fairness of model with respect to a *a predetermined set of metrics*.

### 2.5 Fairness Interventions with Machine Learning Algorithms

#### 2.5.1 Beyond Binary Classification

As previously mentioned, the overwhelming majority of the work done in algorithmic fairness has focused on binary classification settings; however, this is of fundamental concern, as most problems of interest in the real world are much more complex than binary classification.
Even when considering the simplest regression models, the fairness metrics in a regression setting are not yet well-defined or common in the algorithmic fairness literature. Some definitions have been proposed for evaluating fairness in regression [6], and there has also been work in de-biasing data for regression [63] and enforcing fairness via constrained regression and decision trees [38], though the overall topic is still relatively under-studied.

Even more questions arise when considering even more complicated models, such as online learning, deep learning, and reinforcement learning. Some work has been done, for example in bandits [46] [62], reinforcement learning [45] [33], and ranking [85], but many open questions remain. For example, is exploration—a key component of reinforcement learning algorithms—unethical or unfair because it does not necessarily always choose exploitation, and thus does not optimize for all individuals even when it could [10]? Although we also focus on binary classification settings in this thesis, as more complicated models become increasingly used in impactful real-world settings, these questions will be crucial to ponder and answer.

2.5.2 Evaluating Unfairness

A critical step in eliminating unfairness in models is being able to rigorously identify whether a model is fair or unfair — and by how much. For the most part, designating models as “fair” or “unfair” is quite arbitrary or loose: for example, if equalized odds is not satisfied and the error rates appear to be quite different between groups, some may argue that the model is not fair. No widespread methods exist for general evaluation thus far [81], though there are some existing suggestions.

One simple proposed method in the case of group fairness is checking if the difference of outcome probabilities between groups is above some relatively arbitrary threshold $\epsilon$, i.e., if

$$|AR_i - AR_j| < \epsilon$$

(2.1)
or, using a relative rather than absolute difference,

\[
\frac{AR_i}{AR_j} \geq 1 - \xi,
\]

again for some relatively arbitrary \(\xi\), for all groups \(i, j\) \[19\]. In particular, choosing this threshold to be 0.8 correspond to the “four-fifths” principle from law, which states that in the context of employment, the selection rate for any protected group must be at least four-fifths the selection rate of the group with the highest selection rate \[24\]. While a relatively intuitive procedure, the four-fifths value determination seems quite arbitrary, and not necessarily appropriate in all scenarios, though it has sometimes been used as a threshold due to its legal grounding \[84\]. In part of our work, we aim to make the significance determination threshold \(\epsilon\) of these metric differences more statistically grounded.

There have been a few hypothesis testing frameworks proposed to more rigorously determine whether a model is statistically significantly fair or unfair with respect to group fairness metrics specifically. FairTest uses permutation tests and bootstrapping to check for associations \[79\], and a similarly general framework has been proposed for performing permutation tests for fairness using standard techniques and asymptotic variance estimates \[31\].

There also exists a hypothesis testing framework using optimal transport theory, where the null hypothesis of interest considers whether the true data-generating distribution falls in the entire space of distributions for which a desired fairness metric actually achieves fairness on. However, these are all either still relatively abstract or overly broad, so part of our work we discuss later aims to more specifically and concretely supply interval estimates and quantify estimation uncertainty, in order to make fairness evaluation even more statistically rigorous \[78\].

Finally, there exist some commercial evaluation tools, including Aequitas, which produces reports on the fairness of a model \[75\], and AI Fairness 360 (AIF360), which is a similar toolkit developed by IBM for industrial and collaborative use when analyzing fairness of models \[5\]. Neither of these dive too
deeply into statistical rigor, however.

In conclusion, rigorously identifying unfairness is still considered in the fairness literature to be a hard question [65], and thus part of our work aims to make some contributions towards increasing the statistical rigor of portions of the fairness evaluation process.

2.5.3 CORRECTING FOR UNFAIRNESS

Once unfairness in a model is identified, a natural next step is making the necessary modifications to make the model more fair. These can be categorized into three stages, namely methods that are applied to the data itself before any modeling is done, modifications to the model itself, and adjustments made to the model results after-the-fact; the standard fairness literature refers to these three stages as pre-processing, in-processing, and post-processing, respectively. Visually, this is represented in Figure 2.5.1 above.

We now take a closer look at some of the techniques within each stage.

PRE-PROCESSING

Modifications done in the pre-processing stage refer to any interventions that focus on de-biasing the data itself, before any sort of modeling is done. The main benefit of this set of methods is that they do not modify the actual model explicitly, thus not interfering with model training and remaining compatible with most existing models and optimizations today. However, modifying the original data significantly decreases the interpretability of the data in most cases,
and as a result could also have legal complications, given the existence of data protection legislation today [3].

A variety of data de-biasing techniques have been suggested in fairness literature. The most basic technique is the aforementioned “fairness through unawareness,” which simply excludes the explicit protected attributes from the data, although this is often ineffective as discussed earlier [40]. One of the earliest solutions focused on modifying the data itself proposed learning an unbiased model on biased data by de-biasing the dataset with the least intrusive modifications possible in order to build a Classification with No Discrimination (CND) [49]. Many later works have iterated on this, especially via fair representation learning, which focus on de-biasing the data such that the final features are independent of the protected attributes, while retaining as much of the information in the data as possible [87] [37] [64], as well as adversarial learning [36] [8] [88]. Other methods focus on learning data transformations to control discrimination while preserving utility [17] and modifying all non-protected features to ensure the protected attribute cannot be predicted from the non-protected features [37].

**IN-PROCESSING**

In-processing techniques focus on adjusting the model itself to achieve fairness. This is advantageous because it is the only processing stage during which we can actually optimize for fairness within the model itself, but its main downside is that it significantly interferes with model training, thus requiring that new modeling and optimization techniques be created in order to take into account the new fairness constraints or optimality conditions.

Many in-processing techniques thus focus on optimizing very specific types of models, for example eliminating discrimination in decision trees via a strategy of relabeling leaf nodes [50], constructing fair models using regularization [51], and modifying naive Bayes models to be independent with respect to a desired sensitive attribute [15]. There are also attempts in adding fairness constraints
into the classification learning problem itself [86], but this type of learning subject to group fairness constraints has been found to be computationally hard [82] [22], unless post-processing techniques are also used [41], or the model settings are very well-specified (for example, the constrained problem using a specific “disparate mistreatment” fairness definition is easy to solve [86]). Given this, part of our work discussed later will make contributions towards providing computationally tractable in-processing adjustments in thresholding models to optimize for certain fairness metrics.

**Post-processing**

Post-processing techniques adjust the model outcomes themselves after modeling in order to correct for unfairness. The upside of this is similar to that of pre-processing, in that it does not interfere with the model training process and thus makes it easily applicable. However, how to soundly modifying outputs is a tricky question in itself. Another common argument against post-processing is that post-processing usually uses group membership very explicitly in the process of optimizing or equalizing a certain metric by group, which counteracts the common desire that the decision process should be made independent of the protected attributes [3].

The most simple post-processing technique is satisfying equalized odds by directly flipping some of the model outputs [41], but this is arguably very crude and unfair towards those individuals. Other methods propose separate thresholds for different groups to maximize accuracy and minimize demographic parity [26] [48], or learning a different classifier for each group via decoupling and transfer learning [35].

In conclusion, we note that many interventions focus on correcting for unfairness; less techniques aim to rigorously evaluate the extent of unfairness, which we aim to provide some meaningful insights for.
3

Distributional Derivations of Fairness Metrics

3.1 Formal Data and Model Setup

For the remainder of this thesis, consider a population of $n$ individuals who differ based on some protected attribute (e.g., race, gender, etc.). For simplicity, we only consider scenarios where there are only two unique classes of this attribute represented in the population (e.g., men and women, cats and dogs, Republicans and Democrats, etc.); formally, the group membership $A$ is such that $A \in \{a, b\}$ here, for two values $a$ and $b$ of the protected attribute. Further, let there be $n_a$ people in the group with attribute $a$, $n_b$ people in the group with attribute $b$, and $n = n_a + n_b$. 
3.1.1 Feature Construction

Now, consider again a simplified scenario where there is only one scalar feature for each individual, aiming to measure their true skill level or qualification with respect to a specific discipline or setting. Let these be represented by

\[ X_{a,1}, \ldots, X_{a,n_a}, X_{b,1}, \ldots, X_{b,n_b} \]

where \( X_{a,i} \) is the observation for the \( i \)-th person of group \( a \), and \( X_{b,j} \) is the observation for the \( j \)-th person of group \( b \). In particular, let

\[
X_{a,i} = W_{a,i} + \epsilon_{a,i} \tag{3.1}
\]
\[
X_{b,j} = W_{b,j} + \epsilon_{b,j} \tag{3.2}
\]

where

\[
W_{a,1}, \ldots, W_{a,n_a} \overset{\text{i.i.d.}}{\sim} \text{Unif}(s_a, t_a)
\]
\[
W_{b,1}, \ldots, W_{b,n_b} \overset{\text{i.i.d.}}{\sim} \text{Unif}(s_b, t_b)
\]

represent the true, unobservable underlying skill level or qualification of each individual, with different skill distributions for the two groups, and

\[
\epsilon_{a,1}, \ldots, \epsilon_{a,n_a} \overset{\text{i.i.d.}}{\sim} f_{\epsilon_a}
\]
\[
\epsilon_{b,1}, \ldots, \epsilon_{b,n_b} \overset{\text{i.i.d.}}{\sim} f_{\epsilon_b}
\]

represent the aggregate bias or error from any possible parts of the modeling process as discussed in Section 2.2.2 and are meant to characterize the idea that real-world skill estimates are never perfect, where \( f_{\epsilon_a}, f_{\epsilon_b} \) are the PDFs of any two continuous univariate distributions on \( \mathbb{R} \). We also take all of the \( W_{a,i}, W_{b,j}, \epsilon_{a,k}, \epsilon_{b,l} \) to be pairwise independent for all \( i, k \in \{1, \ldots, n_1\}, j, l \in \{1, \ldots, n_2\} \). As a result, all of the \( X_{a,1}, \ldots, X_{a,n_a}, X_{b,1}, \ldots, X_{b,n_b} \) are all independent, and further identically distributed within their subgroups \( X_{a,1}, \ldots, X_{a,n_a} \) and \( X_{b,1}, \ldots, X_{b,n_b} \).

With this construction, the scalar features \( X_{a,1}, \ldots, X_{a,n_a}, X_{b,1}, \ldots, X_{b,n_b} \) thus represent noisy measurements or estimates of each individual’s true skill level in
a given setting.

3.1.2 Comments on Distribution Choices

Starting with the true skill values \( W_{a,1}, \ldots, W_{a,n_a}, W_{b,1}, \ldots, W_{b,n_b} \), the choice of the Uniform distribution for the true underlying skill distributions was largely for computational simplicity and practicality. However, while the Uniform distribution is not necessarily particularly common in raw data, Universality of the Uniform allows us to generate Uniformly-distributed data from any continuous data simply by plugging that data back into its own CDF [43]. Then, considering the fact that the probability integral transform (also known as Universality of the Uniform) essentially brings raw data to its percentile value, this Uniform assumption becomes reasonably well-founded in settings where representing one’s skill measure via their relative percentile within a population is a reasonable choice. In settings where, for example, the scale of skill level measurement is largely uninterpretable or changes over time (e.g., SAT scores), then one’s skill percentile becomes a much more useful measure, and is Uniformly distributed by the probability integral transform.

We place no constraints on the noise distribution except that it is continuous on \( \mathbb{R} \), again for computational soundness. Thus, note that \( f_{e_a} \) and \( f_{e_b} \) can (and should) be chosen to represent contextual information from real-world societal biases for the given setting. For example, if the noise distribution is chosen to be zero-centered and symmetric (e.g., \( \mathcal{N}(0, \sigma^2) \)), then this implies we believe there is no upwards or downwards bias from the various sources of modeling error, as the noisy skill estimates will still be centered on average at the true mean of the population’s skill level, and the symmetry makes all values of upwards and downwards error equally likely, not favoring either in particular. A choice of noise centered elsewhere or asymmetric, on the other hand, could be used to represent upwards or downwards bias or prejudice against a certain group; as mentioned in Section 2.2.2, this is not at all uncommon, for example with numerous real-world examples of undersampling, mismeasurement, and misrepresentation of the data.
for historically marginalized groups [4].

3.1.3 Threshold-Based Decision Rule

Now, we turn to the true and predicted outcomes corresponding to each individual in our data.

Here, we consider a simple threshold-based decision rule. Namely, we directly threshold based on one’s true skill and noisy skill estimates to obtain their true and predicted outcomes, respectively, and if one’s skill level is above a pre-determined threshold \( c \), then they are classified in the positive class, otherwise they are classified to the negative class.

Formally, we have that the true binary outcomes \( Y_{a,1}, \ldots, Y_{a,n_a} \), \( Y_{b,1}, \ldots, Y_{b,n_b} \in \{0, 1\} \) are

\[
Y_{a,i} = \mathbb{1}_{\{W_{a,i} \geq c_a\}} \quad (3.3)
\]

\[
Y_{b,j} = \mathbb{1}_{\{W_{b,j} \geq c_b\}} \quad (3.4)
\]

i.e., if a group \( a \) individual’s true skill level \( W \) is above the threshold \( c_a \), then their true qualification in this setting is to class 1, and class 0 otherwise, and analogously for group \( b \). We also assume \( c_a, c_b \) satisfy

\[
s_a, s_b \leq c_a, c_b \leq t_a, t_b,
\]

otherwise many results are trivial, since the true labels are identical across the entire group with \( c_a, c_b \notin [s_i, b_i] \). We separately notate the thresholds \( c_a \) and \( c_b \) for the two groups because depending on the setting, group-specific thresholds may be desired (e.g., when determining opportunity for two groups who significantly differ in socioeconomic status); in the case that group-specific thresholds are unnecessary and only a general threshold is desired, we can simply take a shared threshold \( c \) such that \( c = c_a = c_b \). In our example from Chapter 2, a threshold-based decision rule could be one that accepts any candidate as long as their screening exam score was above a certain value.
For the predictions \( \hat{Y}_{a,1}, \ldots, \hat{Y}_{a,n_a}, \hat{Y}_{b,1}, \ldots, \hat{Y}_{b,n_b} \in \{0, 1\} \), we use the individuals’ noisy skill estimates \( X \), since this is the data that is actually observed or collected. We have

\[
\hat{Y}_{a,i} = \mathbb{1}_{\{X_{a,i} \geq c_a\}} \quad (3.5)
\]

\[
\hat{Y}_{b,j} = \mathbb{1}_{\{X_{b,j} \geq c_b\}} \quad (3.6)
\]

so if a group \( a \) individual’s (noisy) measured skill level \( X_{a,i} \) is above the threshold \( c_a \), then they are classified to the positive class, and analogously for group \( b \).

Thresholding rules are widely used in many automated real-world decisions [41], such as determining loan decisions using credit scoring [29], patient diagnosis of diseases [70], spam email filtering [39], and more. Intuitively, many decisions are binary in nature, and require a line being drawn somewhere in terms of choosing between two classes, and thus thresholding is a natural solution. Many binary classification also inherently make predictions based on a linear combination of the input features being above or below a certain value (e.g., logistic regression, support vector machines). Thus, thresholding decisions are a central structure in the binary classification setting, and valuable to further analyze. Note that while we focus only on the case of one skill predictor here and one common thresholds, it is relatively straightforward to generalize to the cases with multiple thresholds, or multiple features of interest.

### 3.2 Fairness Metrics of Interest

#### 3.2.1 Chosen Fairness Definitions

In Section 2.3.2, we outlined a few of the evaluation metrics of interest in binary classification settings and their relation to group fairness definitions. Here, we formally (re)define the binary classification sample evaluation metrics of interest that we will be studying for the rest of the analyses in this paper, using the set-up and notation introduced in Section 3.1.1. We note in particular that these definitions are defined again as sample quantities, since they are computed on the
Definition 3.2.1 (Accuracy). The (sample) accuracies of groups a and b, call them $\hat{ACC}_a$ and $\hat{ACC}_b$ respectively, represent the proportion of predictions that were correct (i.e., $\hat{Y} = Y$). Formally,

$$\hat{ACC}_a = \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{1}\{\hat{Y}_{a,i} = Y_{a,i}\}$$ (3.7)

$$\hat{ACC}_b = \frac{1}{n_b} \sum_{i=1}^{n_b} \mathbb{1}\{\hat{Y}_{b,i} = Y_{b,i}\}$$ (3.8)

Note that we could have written an analogous expression for the overall accuracy of the model on all of the data points in aggregate as in Definition 2.3.1, but in group fairness literature, as discussed in Section 2.4, we are often interested in equalizing these group-specific measurements across all protected groups. Thus, since our focus is also on how these metrics differ by group, we specifically notate and compute them separately for each group.

Definition 3.2.2 (Acceptance Rate). The (sample) acceptance rates of groups a and b, call them $\hat{AR}_a$ and $\hat{AR}_b$ respectively, represent the proportion of predictions classified to the positive class (i.e., $\hat{Y} = 1$). Formally,

$$\hat{AR}_a = \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{1}\{\hat{Y}_{a,i} = 1\}$$ (3.9)

$$\hat{AR}_b = \frac{1}{n_b} \sum_{i=1}^{n_b} \mathbb{1}\{\hat{Y}_{b,i} = 1\}$$ (3.10)

Definition 3.2.3 (False Positive Rate). The (sample) false positive rates (FPR) of groups a and b, call them $\hat{FPR}_a$ and $\hat{FPR}_b$ respectively, represent the proportion of predictions classified to the negative class ($\hat{Y} = 0$) whose classification was inaccurate, i.e., those whose true classification should be to the
positive class ($Y = 1$). Formally,

\[
\bar{FPR}_a = \begin{cases} 
\frac{\sum_{i=1}^{n_a} \mathbb{1}_{\{\hat{Y}_{a,i}=1, Y_{a,i}=0\}}}{\sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=0\}}} & \sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=0\}} \neq 0 \\
0, & \sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=0\}} = 0
\end{cases} 
\] (3.11)

\[
\bar{FPR}_b = \begin{cases} 
\frac{\sum_{i=1}^{n_b} \mathbb{1}_{\{\hat{Y}_{b,i}=1, Y_{b,i}=0\}}}{\sum_{i=1}^{n_b} \mathbb{1}_{\{Y_{b,i}=0\}}} & \sum_{i=1}^{n_b} \mathbb{1}_{\{Y_{b,i}=0\}} \neq 0 \\
0, & \sum_{i=1}^{n_b} \mathbb{1}_{\{Y_{b,i}=0\}} = 0
\end{cases} 
\] (3.12)

Note that we can also think of $\bar{FPR}_a$ as the sample mean of the $\mathbb{1}_{\{\hat{Y}_{a,i}=1 \mid Y_{a,i}=0\}}$'s, though the notation above is slightly nicer.

**Definition 3.2.4 (False Negative Rate).** The (sample) false negative rates (FNR) of groups $a$ and $b$, call them $\bar{FNR}_a$ and $\bar{FNR}_b$ respectively, represent the proportion of predictions classified to the positive class ($\hat{Y} = 1$) whose classification was inaccurate, i.e., those whose true classification should be to the negative class ($Y = 0$). Formally,

\[
\bar{FNR}_a = \begin{cases} 
\frac{\sum_{i=1}^{n_a} \mathbb{1}_{\{\hat{Y}_{a,i}=0, Y_{a,i}=1\}}}{\sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=1\}}} & \sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=1\}} \neq 0 \\
0, & \sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=1\}} = 0
\end{cases} 
\] (3.14)

\[
\bar{FNR}_b = \begin{cases} 
\frac{\sum_{i=1}^{n_b} \mathbb{1}_{\{\hat{Y}_{b,i}=0, Y_{b,i}=1\}}}{\sum_{i=1}^{n_b} \mathbb{1}_{\{Y_{b,i}=1\}}} & \sum_{i=1}^{n_b} \mathbb{1}_{\{Y_{b,i}=1\}} \neq 0 \\
0, & \sum_{i=1}^{n_b} \mathbb{1}_{\{Y_{b,i}=1\}} = 0
\end{cases} 
\] (3.15)

Similarly, we can think of $\bar{FNR}_a$ as the sample mean of the $\mathbb{1}_{\{\hat{Y}_{a,i}=1 \mid Y_{a,i}=0\}}$'s.

**Definition 3.2.5 (Positive Predictive Value).** The (sample) positive predictive values (PPV) of groups $a$ and $b$, call them $\bar{PPV}_a$ and $\bar{PPV}_b$ respectively, represent the proportion of predictions classified to the positive class ($\hat{Y} = 1$) whose classification was inaccurate, i.e., those whose true classification should be
to the negative class \((Y = 0)\). Formally,

\[
\begin{align*}
\hat{PPV}_a &= \begin{cases} 
\frac{\sum_{i=1}^{n_a} I\{Y_{a,i} = 1, \hat{Y}_{a,i} = 1\}}{\sum_{i=1}^{n_a} I\{\hat{Y}_{a,i} = 1\}}, & \sum_{i=1}^{n_a} I\{\hat{Y}_{a,i} = 1\} \neq 0 \\
0, & \sum_{i=1}^{n_a} I\{\hat{Y}_{a,i} = 1\} = 0
\end{cases} \\
\hat{PPV}_b &= \begin{cases} 
\frac{\sum_{i=1}^{n_b} I\{Y_{b,i} = 1, \hat{Y}_{b,i} = 1\}}{\sum_{i=1}^{n_b} I\{\hat{Y}_{b,i} = 1\}}, & \sum_{i=1}^{n_b} I\{\hat{Y}_{b,i} = 1\} \neq 0 \\
0, & \sum_{i=1}^{n_b} I\{\hat{Y}_{b,i} = 1\} = 0
\end{cases}
\end{align*}
\]

Again, we can think of \(\hat{PPV}_a\) as the sample mean of the \(I\{\hat{Y}_{a,i} = 1 | Y_{a,i} = 0\}\)'s.

Note again that all five of the quantities above are simply standard binary classification evaluation metrics as discussed in Section 2.3.2, and the most fundamental metrics group fairness definitions seek to equalize, thus making them quantities of high significance in fairness work. Given this, we explore the specific importance of each of these metrics in more detail subsequently, before moving to deriving their distributions.

3.2.2 Mathematical and Philosophical Significance

As already discussed in Section 2.4.1, the five metrics above hold significant importance within the realm of group fairness, and are widely regarded as fundamental measures for evaluating different dimensions of fairness as it relates to protected groups in machine learning models.

The prominence of these definitions inherently stems from their intuitive interpretations and practical relevance in real-world decision-making contexts; each definition has its own moral significant with respect to some facet of equality. In particular, accuracy relates to how well individuals are being classified according to their true skill. Acceptance rates relate to the representation of different groups in the model outcome classes. FPR and FNR relate to how often the model makes different types of errors. PPV evaluates
predictive value of positive predictions. These intuitive interpretations of the
different metrics quickly reveal how any large discrepancy in a certain metric
between protected groups could be interpreted as unfair.

Beyond their mathematical formulations, these group fairness metrics carry
profound philosophical implications, reflecting societal values of fairness, justice,
and equity, and are frequently employed by policymakers, journalists, and various
other stakeholders when it comes to constructing regulatory policies surrounding
fairness. Thus, these metrics bear deep significance not only within the
algorithmic fairness research community, but also in the core values of fairness
from a variety of other academic and social disciplines.

3.3 Uniform Skill, Gaussian Noise

We first explore the setting where the measurement noise is Gaussian and
zero-centered, i.e., \( f_{\epsilon_a} = \mathcal{N}(0, \sigma_a^2) \) and \( f_{\epsilon_b} = \mathcal{N}(0, \sigma_b^2) \).

3.3.1 Aside: Key Integral

Before operating on the fairness metrics, we first derive the result of a key
integral used in all subsequent derivations in this section. Let \( \Phi \) and \( \varphi \) be the
CDF and PDF of the standard Gaussian distribution, respectively. Then, we
have the following lemma.

**Lemma 3.3.1** (Gaussian CDF Integral).

\[
\int_q^r \Phi \left( \frac{c - w}{\sigma} \right) \, dw = (r - c) \Phi \left( \frac{c - r}{\sigma} \right) - (q - c) \Phi \left( \frac{c - q}{\sigma} \right) \\
- \sigma \left[ \varphi \left( \frac{c - r}{\sigma} \right) - \varphi \left( \frac{c - q}{\sigma} \right) \right]
\]

**Proof.** Using integration by parts, we have

\[
\int_q^r \Phi \left( \frac{c - w}{\sigma} \right) \, dw
\]
\begin{align*}
&= w \Phi \left( \frac{c - w}{\sigma} \right) \bigg|_{q}^{r} - \int_{q}^{r} w \left( -\frac{1}{\sigma} \varphi \left( \frac{c - w}{\sigma} \right) \right) \, dw \\
&= \left[ r \Phi \left( \frac{c - r}{\sigma} \right) - q \Phi \left( \frac{c - q}{\sigma} \right) \right] - \int_{q}^{r} w \left( -\frac{1}{\sigma} \varphi \left( \frac{c - w}{\sigma} \right) \right) \, dw.
\end{align*}

Simplifying the latter term of the above expression, we have
\begin{align*}
\int_{q}^{r} w \left( -\frac{1}{\sigma} \varphi \left( \frac{c - w}{\sigma} \right) \right) \, dw \\
&= -\frac{1}{\sigma} \int_{q}^{r} w \left( \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(c - w)^2}{2\sigma^2} \right) \right) \, dw \\
&= -\frac{1}{\sigma} \int_{q}^{r} \left( \frac{w}{\sigma^2} - \frac{c}{\sigma^2} \right) \left( \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(c - w)^2}{2\sigma^2} \right) \right) \, dw \\
&= -\frac{1}{\sigma} \int_{q}^{r} \left( \frac{w - c}{\sigma} \right) \left( \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(c - w)^2}{2\sigma^2} \right) \right) \, dw \\
&= -\left[ \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(c - w)^2}{2\sigma^2} \right) \right]_{q}^{r} - \frac{c}{\sigma} \int_{q}^{r} \varphi \left( \frac{c - w}{\sigma} \right) \, dw \\
&= \sigma \varphi \left( \frac{c - r}{\sigma} \right) \bigg|_{q}^{r} + c \left[ \Phi \left( \frac{c - r}{\sigma} \right) - \Phi \left( \frac{c - q}{\sigma} \right) \right] \\
&= \sigma \left[ \varphi \left( \frac{c - r}{\sigma} \right) - \varphi \left( \frac{c - q}{\sigma} \right) \right] + c \left[ \Phi \left( \frac{c - r}{\sigma} \right) - \Phi \left( \frac{c - q}{\sigma} \right) \right]
\end{align*}

Then, returning to the original integral,
\begin{align*}
\int_{q}^{r} \Phi \left( \frac{c - w}{\sigma} \right) \, dw \\
&= \left[ r \Phi \left( \frac{c - r}{\sigma} \right) - q \Phi \left( \frac{c - q}{\sigma} \right) \right] - \int_{q}^{r} w \left( -\frac{1}{\sigma} \varphi \left( \frac{c - w}{\sigma} \right) \right) \, dw \\
&= \left[ r \Phi \left( \frac{c - r}{\sigma} \right) - q \Phi \left( \frac{c - q}{\sigma} \right) \right] - \sigma \left[ \varphi \left( \frac{c - r}{\sigma} \right) - \varphi \left( \frac{c - q}{\sigma} \right) \right] \\
&\quad - c \left[ \Phi \left( \frac{c - r}{\sigma} \right) - \Phi \left( \frac{c - q}{\sigma} \right) \right]
\end{align*}
\[(r - c)\Phi \left( \frac{c - r}{\sigma} \right) - (q - c)\Phi \left( \frac{c - q}{\sigma} \right) - \sigma \left[ \varphi \left( \frac{c - r}{\sigma} \right) - \varphi \left( \frac{c - q}{\sigma} \right) \right],\]

finishing our derivation. \qed

### 3.3.2 Derivation of Accuracy Distribution

#### Moment Calculations

Before deriving the distribution of \(\hat{\text{ACC}}_a\) and \(\hat{\text{ACC}}_b\), we first find their expected values and variances. We standardize by deriving all of the following distributions for group \(a\) only, but the resulting distributions for group \(b\) are exactly the same, with parameters \(s_b, t_b, \sigma_b,\) and \(c_b\) in place of \(s_a, t_a, \sigma_a,\) and \(c_a\) in the following results, and similarly in all subsequent subsections of this chapter as well.

**Lemma 3.3.2 (Expected Value of Accuracy).** The expected value of the accuracy of group \(a\), call it \(p_{\hat{\text{ACC}}_a}\), in this thresholding setting is

\[
p_{\hat{\text{ACC}}_a} := \mathbb{E}[\hat{\text{ACC}}_a]
= \frac{t_a - c_a}{t_a - s_a} - \frac{t_a - c_a}{t_a - s_a} \Phi \left( \frac{c_a - t_a}{\sigma_a} \right) + \frac{s_a - c_a}{t_a - s_a} \Phi \left( \frac{c_a - s_a}{\sigma_a} \right)
+ \frac{\sigma_a}{t_a - s_a} \left[ \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \right] - \frac{2\sigma_a}{(t_a - s_a)\sqrt{2\pi}}
\]

where \(c_a\) is the decision threshold for group \(a\), the true skill distribution is \(\text{Unif}(s_a, t_a)\), and the noise distribution is \(\mathcal{N}(0, \sigma_a^2)\).

**Proof.** By the identically distributed assumption and the fundamental bridge, we have

\[
p_{\hat{\text{ACC}}_a} := \mathbb{E}[\hat{\text{ACC}}_a]
= \mathbb{E} \left[ \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbbm{1}_{\{\hat{Y}_{a,i} = Y_{a,i}\}} \right]
= \mathbb{E}[\mathbbm{1}_{\{\hat{Y}_{a,i} = Y_{a,i}\}}]
\]

41
= \mathbb{P}(\hat{Y}_{a,i} = Y_{a,i})
= \mathbb{P}(\hat{Y}_{a,i} = 1, Y_{a,i} = 1) + \mathbb{P}(\hat{Y}_{a,i} = 0, Y_{a,i} = 0)
= \mathbb{P}(W_{a,i} \geq c_a, X_{a,i} \geq c_a) + \mathbb{P}(W_{a,i} < c_a, X_{a,i} < c_a)

For the first term above, we have

\begin{align*}
\mathbb{P}(W_{a,i} \geq c_a, X_{a,i} \geq c_a) \\
= \mathbb{P}(W_{a,i} \geq c_a, W_{a,i} + \epsilon_{a,i} \geq c_a) \\
= \int_{c_a}^{t_a} \mathbb{P}(W_{a,i} + \epsilon_{a,i} \geq c_a \mid W_{a,i} = w)f_{W_{a,i}}(w)dw \\
= \frac{1}{t_a - s_a} \int_{c_a}^{t_a} \mathbb{P}(W_{a,i} + \epsilon_{a,i} \geq c_a \mid W_{a,i} = w)dw \\
= \frac{1}{t_a - s_a} \int_{c_a}^{t_a} \mathbb{P}(\epsilon_{a,i} \geq c_a - w)dw \\
= \frac{1}{t_a - s_a} \int_{c_a}^{t_a} \left(1 - \Phi\left(\frac{c_a - w}{\sigma_a}\right)\right)dw \\
= \frac{1}{t_a - s_a} \int_{c_a}^{t_a} dw - \int_{c_a}^{t_a} \Phi\left(\frac{c_a - w}{\sigma_a}\right)dw \\
= \frac{t_a - c_a}{t_a - s_a} - \frac{1}{t_a - s_a} \left[(t_a - c_a)\Phi\left(\frac{c_a - t_a}{\sigma_a}\right) - \sigma_a \varphi\left(\frac{c_a - t_a}{\sigma_a}\right) + \frac{\sigma_a}{\sqrt{2\pi}}\right],
\end{align*}

plugging in our results from Lemma 3.3.1 and simplifying.
Similarly, for the second term, we have

\begin{align*}
\mathbb{P}(W_{a,i} < c_a, X_{a,i} < c_a) \\
= \int_{s_a}^{c_a} \mathbb{P}(W_{a,i} + \epsilon_{a,i} < c_a \mid W_{a,i} = w)f_{W_{a,i}}(w)dw \\
= \frac{1}{t_a - s_a} \int_{s_a}^{c_a} \Phi\left(\frac{c_a - w}{\sigma_a}\right)dw \\
= \frac{1}{t_a - s_a} \left[(c_a - s_a)\Phi\left(\frac{c_a - s_a}{\sigma_a}\right) - \frac{\sigma_a}{\sqrt{2\pi}} + \sigma_a \varphi\left(\frac{c_a - s_a}{\sigma_a}\right)\right],
\end{align*}
Then, combining,

\[ p_{\overline{\text{ACC}}_a} = \mathbb{P}(W_{a,i} \geq c_a, X_{a,i} \geq c_a) + \mathbb{P}(W_{a,i} < c_a, X_{a,i} < c_a) \]

\[ = \frac{t_a - c_a}{t_a - s_a} - \frac{t_a - c_a}{t_a - s_a} \Phi \left( \frac{c_a - t_a}{\sigma_a} \right) + \frac{s_a - c_a}{t_a - s_a} \Phi \left( \frac{c_a - s_a}{\sigma_a} \right) \]

\[ + \frac{\sigma_a}{t_a - s_a} \left[ \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \right] - \frac{2\sigma_a}{(t_a - s_a) \sqrt{2\pi}} \]

\[ = \frac{t_a - c_a}{t_a - s_a} - \frac{t_a - c_a}{t_a - s_a} \Phi \left( \frac{c_a - t_a}{\sigma_a} \right) + \frac{s_a - c_a}{t_a - s_a} \Phi \left( \frac{c_a - s_a}{\sigma_a} \right) \]

\[ + \frac{\sigma_a}{t_a - s_a} \left[ \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \right] - \frac{2\sigma_a}{(t_a - s_a) \sqrt{2\pi}} \]

\[ \square \]

**Lemma 3.3.3** (Variance of Accuracy). The variance of the accuracy of group \( a \) is

\[ \text{Var}(\overline{\text{ACC}}_a) = \frac{1}{n_a} p_{\overline{\text{ACC}}_a} (1 - p_{\overline{\text{ACC}}_a}) \]

with \( \overline{\text{ACC}}_a \) as defined in Lemma 3.3.2 above.

**Proof.** Using variance properties and the Bernoulli distribution variance since we know \( \mathbb{1}_{\{\hat{Y}_{a,i} = Y_{a,i}\}} \sim \text{Bern}(p_{\overline{\text{ACC}}_a}) \), we have

\[ \text{Var}(\overline{\text{ACC}}_a) = \text{Var} \left( \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{1}_{\{\hat{Y}_{a,i} = Y_{a,i}\}} \right) \]

\[ = \frac{1}{n_a^2} \sum_{i=1}^{n_a} \text{Var}(\mathbb{1}_{\{\hat{Y}_{a,i} = Y_{a,i}\}}) \]

\[ = \frac{1}{n_a^2} n_a \text{Var}(\mathbb{1}_{Y_{a,i} = Y_{a,i}}) \]

\[ = \frac{1}{n_a} \text{Var}(\mathbb{1}_{\{\hat{Y}_{a,i} = Y_{a,i}\}}) \]

\[ = \frac{1}{n_a} p_{\overline{\text{ACC}}_a} (1 - p_{\overline{\text{ACC}}_a}), \]

as desired. \( \square \)
**Exact, Approximate, and Asymptotic Distributions**

Having derived $p_{\hat{ACC}}$, we can now easily find the exact, approximate, and asymptotic distributions of $\hat{ACC}$.

**Lemma 3.3.4 (Exact Distribution of Accuracy).** The exact distribution of the accuracy for group $a$, $\hat{ACC}_a$, in this thresholding setting is

$$\hat{ACC}_a \sim \frac{1}{n_a} \text{Binomial}(n_a, p_{\hat{ACC}_a}),$$

with $p_{\hat{ACC}_a}$ as defined in Lemma 3.3.2, where $c_a$ is the decision threshold for group $a$, the true skill distribution is $\text{Unif}(s_a, t_a)$, and the noise distribution is $\mathcal{N}(0, \sigma_a^2)$.

**Proof.** We know $\sum_{i=1}^{n_a} 1_{\{Y_{a,i} = Y_{a,i}\}}$ follows a Binomial distribution with $n_a$ trials and success probability $p_{\hat{ACC}_a} = \mathbb{P}(\hat{Y}_{a,i} = Y_{a,i})$ by representation theory. Then, $\hat{ACC}_a = \frac{1}{n_a} \sum_{i=1}^{n_a} 1_{\{Y_{a,i} = Y_{a,i}\}}$ follows the scaled binomial distribution above. 

**Lemma 3.3.5 (Approximate Distribution of Accuracy).** The approximate distribution of the accuracy for group $a$ is

$$\hat{ACC}_a \sim \mathcal{N} \left( p_{\hat{ACC}_a},\frac{p_{\hat{ACC}_a}(1-p_{\hat{ACC}_a})}{n_a} \right),$$

for $n_a$ large.

**Proof.** For $n$ large, we have that the Normal distribution with mean $np$ and variance $np(1 - p)$ is an approximation for the Binomial($n, p$) distribution. Then, applying this to Lemma 3.3.4 and rescaling, we arrive at the above result.

**Lemma 3.3.6 (Asymptotic Distribution of Accuracy).** The asymptotic distribution of the accuracy for group $a$ is

$$\frac{\sqrt{n_a}(\hat{ACC}_a - p_{\hat{ACC}_a})}{\sqrt{p_{\hat{ACC}_a}(1-p_{\hat{ACC}_a})}} \overset{D}{\to} \mathcal{N}(0, 1).$$
The resulting approximate distribution is

\[ \widehat{ACC}_a \sim \mathcal{N}\left( p_{\widehat{ACC}_a}, \frac{p_{\widehat{ACC}_a}(1 - p_{\widehat{ACC}_a})}{n_a} \right), \]

for \( n_a \) large, which matches the result in Lemma 3.3.5.

Proof. This result follows directly from applying the Central Limit Theorem to \( \widehat{ACC}_a \), which is already a sample mean. \( \square \)

At this point, we note that distributions follow the exact same structure and justification for the four remaining metrics in Section 3.3 as well as the analogous derivations with general noise in the next part (Section 3.4), so we omit the identical derivational proofs, and focus only on derivations of new mathematical quantities (namely, the expected values).

3.3.3 DERIVATION OF ACCEPTANCE RATE DISTRIBUTION

MOMENT CALCULATIONS

Exactly analogously as above, we have

\[
p_{\widehat{AR}_a} := \mathbb{E}[\widehat{ACC}_a] = \mathbb{P}(\hat{Y}_{a,i} = 1)
\]

\[
\text{Var}(\widehat{AR}_a) = \frac{1}{n_a}p_{\widehat{AR}_a}(1 - p_{\widehat{AR}_a})
\]

Then, we compute \( p_{\widehat{AR}_a} \).

Lemma 3.3.7 (Expected Value of Acceptance Rate). The expected value of the acceptance rate of group \( a \), \( p_{\widehat{AR}_a} \), in this thresholding setting is

\[
p_{\widehat{AR}_a} := \mathbb{E}[\widehat{AR}_a]
= 1 - \frac{t_a - c_a}{t_a - s_a} \Phi\left( \frac{c_a - t_a}{\sigma_a} \right) - \frac{s_a - c_a}{t_a - s_a} \Phi\left( \frac{c_a - s_a}{\sigma_a} \right)
+ \frac{\sigma_a}{t_a - s_a} \left[ \varphi\left( \frac{c_a - t_a}{\sigma_a} \right) - \varphi\left( \frac{c_a - s_a}{\sigma_a} \right) \right]
\]
where \( c_a \) is the decision threshold for group \( a \), the true skill distribution is \( \text{Unif}(s_a, t_a) \), and the noise distribution is \( \mathcal{N}(0, \sigma_a^2) \).

**Proof.** We have

\[
p_{\overline{AR}_a} := \mathbb{E}[\overline{ACC}_a]
\]

\[
= \mathbb{E}[\mathbb{1}_{\{Y_{a,i} = 1\}} \cdot s]
\]

\[
= \mathbb{P}(\hat{Y}_{a,i} = 1)
\]

\[
= \mathbb{P}(X_{a,i} \geq c_a)
\]

\[
= \mathbb{P}(W_{a,i} + \epsilon_{a,i} \geq c_a)
\]

\[
= \int_{s_a}^{t_a} \mathbb{P}(W_{a,i} + \epsilon_{a,i} \geq c_a \mid W_{a,i} = w) f_{W_{a,i}}(w) \, dw
\]

\[
= \frac{1}{t_a - s_a} \int_{s_a}^{t_a} \left( 1 - \Phi \left( \frac{c_a - w}{\sigma_a} \right) \right) \, dw
\]

\[
= 1 - \frac{1}{t_a - s_a} \int_{s_a}^{t_a} \Phi \left( \frac{c_a - w}{\sigma_a} \right) \, dw
\]

\[
= 1 - \left[ \frac{t_a - c_a}{t_a - s_a} \Phi \left( \frac{c_a - t_a}{\sigma_a} \right) - \frac{s_a - c_a}{t_a - s_a} \Phi \left( \frac{c_a - s_a}{\sigma_a} \right) \right]
\]

\[
+ \frac{\sigma_a}{t_a - s_a} \left[ \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) - \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \right],
\]

applying Lemma 3.3.1 and simplifying.  

\[\square\]

**Lemma 3.3.8** (Variance of Acceptance Rate). The variance of the acceptance rate of group \( a \) is

\[
\text{Var}(\overline{AR}_a) = \frac{1}{n_a} p_{\overline{AR}_a} (1 - p_{\overline{AR}_a})
\]

with \( \overline{AR}_a \) as defined in Lemma 3.3.7 above.

**Proof.** Similar to the proof of Lemma 3.3.3, we have

\[
\text{Var}(\overline{AR}_a) = \text{Var} \left( \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i} = 1\}} \right)
\]
as desired. \hfill \qed

EXACT, APPROXIMATE, AND ASYMPTOTIC DISTRIBUTIONS

**Lemma 3.3.9** (Exact Distribution of Acceptance Rate). The exact distribution of the acceptance rate for group $a$ in this thresholding setting is

$$
\hat{AR}_a \sim \frac{1}{n_a} \text{Binomial}(n_a, p_{AR_a}),
$$

with $p_{AR_a}$ as defined in Lemma 3.3.7.

**Lemma 3.3.10** (Approximate Distribution of Acceptance Rate). The approximate distribution of the acceptance rate for group $a$ is

$$
\hat{AR}_a \sim \mathcal{N}\left(p_{AR_a}, \frac{p_{AR_a}(1 - p_{AR_a})}{n_a}\right),
$$

for $n_a$ large.

**Lemma 3.3.11** (Asymptotic Distribution of Acceptance Rate). The asymptotic distribution of the acceptance rate for group $a$ is

$$
\sqrt{n_a}(\hat{AR}_a - p_{AR_a}) \xrightarrow{d} \mathcal{N}(0, 1).
$$
The resulting approximate distribution is

$$
\tilde{AR}_a \sim \mathcal{N}(p_{\tilde{AR}_a}, \frac{p_{\tilde{AR}_a}(1 - p_{\tilde{AR}_a})}{n_a})
$$

for \( n_a \) large, which matches the result in Lemma 3.3.10.

### 3.3.4 Derivation of FPR Distribution

#### Moment Calculations

**Lemma 3.3.12** (Expected Value of FPR). The expected value of the FPR of group \( a \), \( p_{\text{FPR}_a} \), in this thresholding setting is

$$
p_{\text{FPR}_a} := \mathbb{E}[\tilde{FPR}_a] = 1 - \left[ \Phi \left( \frac{c_a - s_a}{\sigma_a} \right) - \frac{\sigma_a}{(c_a - s_a)\sqrt{2\pi}} + \frac{\sigma_a}{c_a - s_a} \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \right],
$$

where \( c_a \) is the decision threshold for group \( a \), the true skill distribution is \( \text{Unif}(s_a, t_a) \), and the noise distribution is \( \mathcal{N}(0, \sigma_a^2) \).

**Proof.** Recalling from Definition 3.2.3 that the false positive rate can also be thought of as the sample means of the \( \{\hat{Y}_{a,i} \mid Y_{a,i} = 0\} \)'s, we have

$$
p_{\text{FPR}_a} = \mathbb{P}(\hat{Y}_{a,i} = 1 \mid Y_{a,i} = 0) = \mathbb{P}(X_{a,i} \geq c_a \mid W_{a,i} < c_a) = \mathbb{P}(W_{a,i} + \epsilon_{a,i} \geq c_a \mid W_{a,i} < c_a) = \frac{\mathbb{P}(W_{a,i} + \epsilon_{a,i} \geq c_a, W_{a,i} < c_a)}{\mathbb{P}(W_{a,i} < c_a)} = \frac{\int_{s_a}^{c_a} \mathbb{P}(W_{a,i} + \epsilon_{a,i} \geq c_a \mid W_{a,i} = w) \frac{1}{t_a - s_a} dw}{\frac{c_a - s_a}{t_a - s_a}}
$$

$$
= 1 - \left[ \Phi \left( \frac{c_a - s_a}{\sigma_a} \right) - \frac{\sigma_a}{(c_a - s_a)\sqrt{2\pi}} + \frac{\sigma_a}{c_a - s_a} \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \right],
$$

48
applying Lemma 3.3.1 and simplifying.

For the variance, we note that because of the randomness of the denominator
\[ \sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i} = 0\}} \] of the definition of \( \hat{FPR}_a \) from Definition 3.2.3, the exact variance
derivation is not nearly as direct as that of Lemma 3.3.3 and Lemma 3.3.8 above.
For the same reason, the exact distribution of \( \hat{FPR}_a \) cannot be expressed in a
nice form. Thus, before proceeding with the variance, we first introduce a
Binomial approximation for \( \hat{FPR}_a \).

**Approximate Distribution**

**Lemma 3.3.13** (Approximate Distribution of FPR). An approximate
distribution of the FPR for group \( a \) is

\[ \hat{FPR}_a \sim \frac{1}{n_a p_{a,0}} \text{Binomial} \left( n_a, p_{a,0} p_{\hat{FPR}_a} \right), \]

for \( n_a \) large, with \( p_{a,0} := \mathbb{P}(Y_{a,i} = 0) = \frac{c_a - s_a}{t_a - s_a} \) the probability an individual’s true
skill is below the threshold \( c \) and \( p_{\hat{FPR}_a} \) as defined in Lemma 3.3.12.

Using the Normal approximation for the Binomial, we obtain another
approximate distribution for \( \hat{FPR}_a \),

\[ \hat{FPR}_a \sim \mathcal{N} \left( p_{\hat{FPR}_a}, \frac{p_{\hat{FPR}_a} (1 - p_{a,0} p_{\hat{FPR}_a})}{n_a p_{a,0}} \right), \]

again for \( n_a \) large.

**Proof.** Note preliminarily that from an intermediary step of the proof of Lemma
3.3.12 above,

\[ \mathbb{P}(\hat{Y}_{a,i} = 1, Y_{a,i} = 0) = \mathbb{P}(\hat{Y}_{a,i} = 1 | Y_{a,i} = 0) \cdot \mathbb{P}(Y_{a,i} = 0) \]

\[ = p_{\hat{FNR}_a} p_{a,0}, \]
for
\[ p_{a,0} := \mathbb{P}(Y_{a,i} = 0) = \mathbb{P}(W_{a,i} < c_a) = \frac{c_a - s_a}{t_a - s_a} \]
the probability of one’s true skill falling in the negative class.

Then, from Definition 3.2.3, we have
\[ \hat{FPR}_a = \frac{\sum_{i=1}^{n_a} \mathbb{1}\{Y_{a,i} = 1, Y_{a,i} = 0\} p_{a,0}}{\sum_{i=1}^{n_a} \mathbb{1}\{Y_{a,i} = 0\}}. \]

Then, applying the strong law of large numbers to the denominator, we have
\[ \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{1}\{Y_{a,i} = 0\} \xrightarrow{a.s.} \mathbb{P}(Y_{a,i} = 0) = p_{a,0} \]

Using the quotient form of Slutsky’s theorem to apply this result in the denominator of our \( \hat{FPR}_a \) expression above, we have
\[ \hat{FPR}_a \approx \frac{\frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{1}\{Y_{a,i} = 1, Y_{a,i} = 0\} p_{a,0}}{\mathbb{P}(Y_{a,i} = 0)} \sim \frac{1}{n_a p_{a,0}} \text{Binomial} \left( n_a, p_{a,0} \hat{FPR}_a \right) \]
for \( n_a \) large, as desired.

The mean and variance of the Binomial distribution above are
\[ \mu = \frac{1}{n_a p_{a,0}} \cdot n_a p_{a,0} \hat{FPR}_a \]
\[ = p_{\hat{FPR}_a} \]
\[ \sigma = \frac{1}{n_a^2 p_{a,0}^2} \cdot n_a p_{a,0} \hat{FPR}_a \left(1 - n_a p_{a,0} \hat{FPR}_a \right) \]
\[ = \frac{p_{\hat{FPR}_a} \left(1 - p_0 \hat{FPR}_a \right)}{n_a p_0}, \]
s respectively, so using the Normal approximation to the Binomial, the approximate Normal distribution is that given in Equation 3.3.13

We claim the math above is valid for a few reasons. Intuitively for \( n_a \) large, the sample mean in the denominator will approximately converge to its true mean in
value by the strong law of large numbers, but the numerator has an additional source of randomness from the $\hat{Y}_{a,i}$’s, hence why it is reasonable to leave it as is. Next, we see above that the mean of the approximate Binomial distribution derived above is equal to $p_{\hat{\text{ACC}}_a}$, which is desirable. Lastly, via simulation, our proposed approximate Binomial distribution from Equation 3.4.13 and a simulated distribution of the FPR across 10,000 trials are very similar, as in Figure 3.3.1 below.

![Figure 3.3.1: Simulated FPR distribution vs. proposed approximate Binomial FPR distribution, for Unif(0,100) skill, $\mathcal{N}(0,10^2)$ noise, threshold $c_a = 50$, and sample size $n_a = 100$.](image)

**Lemma 3.3.14 (Approximate Variance of FPR).** The approximate variance of $\hat{FPR}_a$ is

$$\text{Var}(\hat{FPR}_a) = \frac{p_{\hat{FPR}_a}(1 - p_{\hat{FPR}_a})}{n_a p_{a,0}}.$$
Proof. The derivation is given in the proof of 3.3.13.

3.3.5 Derivation of FNR Distribution

Moment Calculations

Lemma 3.3.15 (Expected Value of FNR). The expected value of the FNR of group $a$, $p_{\text{FNR}_a}$, in this thresholding setting is

$$p_{\text{FNR}_a} := \mathbb{E}[\hat{FNR}_a] = \Phi \left( \frac{c_a - t_a}{\sigma_a} \right) - \frac{\sigma_a}{t_a - c_a} \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \frac{\sigma_a}{(t_a - c_a)\sqrt{2\pi}},$$

where $c_a$ is the decision threshold for group $a$, the true skill distribution is $\text{Unif}(s_a, t_a)$, and the noise distribution is $\mathcal{N}(0, \sigma_a^2)$.

Proof. We have

$$p_{\text{FNR}_a} = \mathbb{P}(\hat{Y}_{a,i} = 0 \mid Y_{a,i} = 1)$$
$$= \mathbb{P}(X_{a,i} < c_a \mid W_{a,i} \geq c_a)$$
$$= \mathbb{P}(W_{a,i} + \epsilon_{a,i} < c_a \mid W_{a,i} \geq c_a)$$
$$= \mathbb{P}(W_{a,i} + \epsilon_{a,i} < c_a, W_{a,i} \geq c_a)$$
$$\frac{\mathbb{P}(W_{a,i} \geq c_a)}{\mathbb{P}(W_{a,i} \geq c_a)}$$
$$= \int_{c_a}^{t_a} \mathbb{P}(W_{a,i} + \epsilon_{a,i} < c_a \mid W_{a,i} = w) \frac{1}{t_a - s_a} dw$$
$$= \frac{1}{t_a - s_a} \int_{c_a}^{t_a} \Phi \left( \frac{c_a - w}{\sigma_a} \right) dw$$
$$= \Phi \left( \frac{c_a - t_a}{\sigma_a} \right) - \frac{\sigma_a}{t_a - c_a} \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \frac{\sigma_a}{(t_a - c_a)\sqrt{2\pi}},$$

applying Lemma 3.3.1 and simplifying.

Similar to $\hat{FPR}_a$, because of the randomness of the denominator of $\hat{FNR}_a$, we instead give an approximate Binomial distribution for $\hat{FNR}_a$ and its resulting
approximate variance.

**Approximate Distribution**

**Lemma 3.3.16** (Approximate Distribution of FNR). An approximate distribution of the FNR for group $a$ is

$$
\hat{\text{FNR}}_a \sim \frac{1}{n_a p_{a,1}} \text{Binomial} \left( n_a, p_{a,1} p_{\hat{\text{FNR}}_a} \right),
$$

for $n_a$ large, with $p_{a,1} := P(Y_{a,i} = 1) = \frac{t_a - c_a}{t_a - s_a}$ the probability an individual’s true skill is below the threshold $c_a$s and $p_{\hat{\text{FNR}}_a}$ as defined in Lemma 3.3.15.

Using the Normal approximation for the Binomial, we obtain another approximate distribution for $\hat{\text{FNR}}_a$,

$$
\hat{\text{FNR}}_a \sim \mathcal{N} \left( p_{\hat{\text{FNR}}_a}, \frac{p_{\hat{\text{FNR}}_a} (1 - p_{\hat{\text{FNR}}_a})}{n_a p_{a,1}} \right),
$$

again for $n_a$ large.

**Proof.** Note from the proof of Lemma 3.3.15 that

$$
P(\hat{Y}_{a,i} = 0, Y_{a,i} = 1) = P(\hat{Y}_{a,i} = 0 \mid Y_{a,i} = 1) \cdot P(Y_{a,i} = 1) = p_{\hat{\text{FNR}}_a} p_{a,1},
$$

for

$$
p_{a,1} := P(Y_{a,i} = 1) = P(W_{a,i} \geq c_a) = \frac{t_a - c_a}{t_a - s_a}
$$

the probability of one’s true skill falling in the positive class.

Then, from Definition 3.2.4, we have

$$
\hat{\text{FNR}}_a = \frac{\sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=0, Y_{a,i}=1\}}}{\sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=1\}}} = \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=0, Y_{a,i}=1\}} = \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{1}_{\{Y_{a,i}=1\}}.
$$

53
Then, applying the strong law of large numbers to the denominator, we have

\[
\frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{I}\{Y_{a,i} = 1\} \xrightarrow{a.s.} \mathbb{P}(Y_{a,i} = 1) = p_{a,1}
\]

Using the quotient form of Slutsky’s theorem to apply this result in the denominator of our \(FP_{Ra}\) expression above, we have

\[
\widehat{FNR}_a \approx \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{I}\{Y_{a,i} = 1, Y_{a,i} = 0\} \sim \frac{1}{n_a p_{a,0}} \text{Binomial} \left(n_a, p_{a,0} p_{FP_{Ra}}\right)
\]

for \(n_a\) large, as desired.

The mean and variance of the Binomial distribution above are

\[
\mu = \frac{1}{n_a p_{a,1}} \cdot n_a p_{a,1} p_{FP_{Ra}} = p_{FP_{Ra}}
\]

\[
\sigma = \frac{1}{n_a^2 p_{a,1}} \cdot n_a p_{a,1} p_{FP_{Ra}} (1 - n_a p_{a,1} p_{FP_{Ra}}) = \frac{p_{FP_{Ra}} (1 - p_{a,1} p_{FP_{Ra}})}{n_a p_{a,1}}
\]

respectively, so using the Normal approximation to the Binomial, the approximate Normal distribution is that given in Equation 3.3.16.

Our rationale for the validity of this Binomial approximation is exactly the same as that presented in the proof of Lemma 3.3.13. In terms of how our proposed approximate Binomial distribution from Equation 3.4.13 compares a simulated distribution of the FNR across 10,000 trials, they are again very similar as shown in Figure 3.3.2 below.
Figure 3.3.2: Simulated FNR distribution vs. proposed approximate Binomial FNR distribution, for Unif(0,100) skill, $N(0,10^2)$ noise, threshold $c_a = 50$, and sample size $n_a = 100$.

Lemma 3.3.17 (Approximate Variance of FNR). The approximate variance of $FNR_a$ is

$$\text{Var}(FNR_a) = \frac{p_{FNR_a}(1 - p_{FNR_a})}{n_a p_{a,1}}.$$ 

Proof. The derivation is given in the proof of 3.3.16.
3.3.6 Derivation of PPV Distribution

Moment Calculations

Lemma 3.3.18 (Expected Value of PPV). The expected value of the PPV of group \(a\), \(p_{\hat{PPV}_a}\), in this thresholding setting is

\[
p_{\hat{PPV}_a} := \mathbb{E}[\hat{PPV}_a]
\]

\[
= \frac{t_a - c_a}{t_a - s_a} - \frac{1}{t_a - s_a} \left[ \left( t_a - c_a \right) \Phi \left( \frac{c_a - t_a}{\sigma_a} \right) - \frac{\sigma_a}{\sqrt{2\pi}} \right] - \frac{c_a - t_a}{t_a - s_a} \Phi \left( \frac{c_a - s_a}{\sigma_a} \right) + \frac{\sigma_a}{\sqrt{2\pi}} - \frac{c_a - s_a}{\sigma_a} \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) - \frac{c_a - s_a}{\sigma_a} \varphi \left( \frac{c_a - s_a}{\sigma_a} \right),
\]

where \(c_a\) is the decision threshold for group \(a\), the true skill distribution is \(\text{Unif}(s_a, t_a)\), and the noise distribution is \(\mathcal{N}(0, \sigma_a^2)\).

Proof. We have

\[
p_{\hat{PPV}_a} = \mathbb{E}[\mathbb{I}\{Y_{a,i} = 1|\hat{Y}_{a,i} = 1\}]
\]

\[
= \mathbb{P}(Y_{a,i} = 1|\hat{Y}_{a,i} = 1)
\]

\[
= \frac{\mathbb{P}(Y_{a,i} = 1, \hat{Y}_{a,i} = 1)}{\mathbb{P}(\hat{Y}_{a,i} = 1)}
\]

\[
= \frac{t_a - c_a}{t_a - s_a} - \frac{1}{t_a - s_a} \left[ \left( t_a - c_a \right) \Phi \left( \frac{c_a - t_a}{\sigma_a} \right) - \frac{\sigma_a}{\sqrt{2\pi}} \right] - \frac{c_a - t_a}{t_a - s_a} \Phi \left( \frac{c_a - s_a}{\sigma_a} \right) + \frac{\sigma_a}{\sqrt{2\pi}} - \frac{c_a - s_a}{\sigma_a} \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) - \frac{c_a - s_a}{\sigma_a} \varphi \left( \frac{c_a - s_a}{\sigma_a} \right)
\]

by plugging in previous results, since we note the numerator was already computed in the proof of Lemma 3.3.2, and the denominator is exactly \(p_{\hat{AR}_a}\).

Similar to \(\hat{FPR}_a\) and \(\hat{FNR}_a\), because of the randomness of the denominator of \(\hat{PPV}_a\), we again give an approximate Binomial distribution for \(\hat{PPV}_a\) and its resulting approximate variance.
Approximate Distribution

Lemma 3.3.19 (Approximate Distribution of PPV). An approximate distribution of the PPV for group $a$ is

$$
\hat{PPV}_a \sim \frac{1}{n_a p_{\overline{AR}_a}} \text{Binomial}(n_a, p_{\overline{AR}_a} p_{\hat{PPV}_a}),
$$

for $n_a$ large, with $p_{\overline{AR}_a}$ as defined in Lemma 3.3.7 and $p_{\hat{PPV}_a}$ as defined in Lemma 3.3.18.

Using the Normal approximation for the Binomial, we obtain another approximate distribution for $\hat{PPV}_a$,

$$
\hat{PPV}_a \sim N\left(p_{\hat{PPV}_a}, \frac{p_{\hat{PPV}_a}(1 - p_{\overline{AR}_a} p_{\hat{PPV}_a})}{n_a p_{\overline{AR}_a}}\right),
$$

again for $n_a$ large.

Proof. Note from the proof of Lemma 3.3.15 that

$$
\mathbb{P}(Y_{a,i} = 1, \hat{Y}_{a,i} = 1) = \mathbb{P}(Y_{a,i} = 1 \mid \hat{Y}_{a,i} = 1) \cdot \mathbb{P}(\hat{Y}_{a,i} = 1) = p_{\hat{PPV}_a} p_{\overline{AR}_a}.
$$

Then, from Definition 3.2.5, we have

$$
\hat{PPV}_a = \frac{\sum_{i=1}^{n_a} \mathbb{I}_{\{Y_{a,i}=1, \hat{Y}_{a,i}=1\}}}{\sum_{i=1}^{n_a} \mathbb{I}_{\{\hat{Y}_{a,i}=1\}}} = \frac{\frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{I}_{\{Y_{a,i}=1, \hat{Y}_{a,i}=1\}}}{\frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{I}_{\{\hat{Y}_{a,i}=1\}}}.
$$

Then, applying the strong law of large numbers to the denominator, we have

$$
\frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{I}_{\{\hat{Y}_{a,i}=1\}} \xrightarrow{a.s.} \mathbb{P}(\hat{Y}_{a,i} = 1) = p_{\overline{AR}_a}.
$$

Using the quotient form of Slutsky’s theorem to apply this result in the
denominator of our $\hat{PPV}_a$ expression above, we have

$$\hat{PPV}_a \approx \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbb{I}_{\{Y_{a,i}=1, \ 1\hat{Y}_{a,i}=1\}} \sim \frac{1}{n_a p_{AR_a}} \text{Binomial} \left(n_a, p_{AR_a}, p_{\hat{PPV}_a}\right)$$

for $n_a$ large, as desired.

The mean and variance of the Binomial distribution above are

$$\mu = \frac{1}{n_a p_{AR_a}} \cdot n_a p_{AR_a} p_{\hat{PPV}_a}$$

$$= p_{\hat{PPV}_a}$$

$$\sigma = \frac{1}{n_a^2 p_{AR_a}^2 p_{\hat{PPV}_a}^2} \cdot n_a p_{AR_a} p_{\hat{PPV}_a} \left(1 - n_a p_{AR_a} p_{\hat{PPV}_a}\right)$$

$$= \frac{p_{\hat{PPV}_a} (1 - p_{AR_a} p_{\hat{PPV}_a})}{n_a p_{AR_a}},$$

respectively, so using the Normal approximation to the Binomial, the approximate Normal distribution is that given in Equation 3.3.19.

Our rationale for the validity of this Binomial approximation is exactly the same as that presented in the proof of Lemma 3.3.13. In terms of how our proposed approximate Binomial distribution from Equation 3.3.19 compares a simulated distribution of the PPV across 10,000 trials, there is quite a bit of difference this time, which is shown in Figure 3.3.3 below.
We investigate this a bit further by plotting the above graph across multiple different skill levels in Figure 3.3.4.

Both visually in Figure 3.3.4 and numerically, we observe that the simulated and theoretical means line up with each other, but the standard deviation of our estimated PPV distribution is \textit{roughly one and a half to two times larger} than the true simulated PPV variance for a given set of parameters. While we are not exactly sure why this is the case, especially given the accuracy of the FPR and FNR distributions above, we hypothesize that this may be due to greater correlation between the numerator and the denominator when the denominator is working with the predictions $\hat{Y}$ instead of the true outcomes $Y$, though we identify this issue as a clear next step for future work.

However, we note that while these distributions clearly have discrepancy in
Figure 3.3.4: Simulated PPV distribution (white outline with light filling) vs. proposed approximate Binomial PPV distribution (solid outline with no filling), for multiple skill distributions, $N(0,10^2)$ noise, threshold $c_a = 50$, and sample size $n_a = 100$.

terms of their spreads, we find empirically that the standard deviations are always an overestimate of the true standard deviation. While we would like to make these standard deviation estimates more precise in the future, we note that the validity of many of the empirical trends we analyze subsequently still hold, given that these approximate standard deviations appear to be a strict upper bound on the true standard deviations. Especially for the purposes of uncertainty quantification, the fact that our approximate standard deviations are an overestimate ensures that our uncertainty quantification is always more conservative (e.g., wider estimated margin of error), which does not bear significant negative consequences.
Lemma 3.3.20 (Approximate Variance of PPV). The approximate variance of $\widehat{PPV}_a$ is

$$\text{Var}(\widehat{PPV}_a) = \frac{p_{PPV_a}(1 - p_{\overline{AR}_a} p_{\overline{PPV}_a})}{n_a p_{\overline{AR}_a}}.$$  

Proof. The derivation is given in the proof of 3.3.19.  

3.4 Uniform Skill, General Noise

Now, we return to the general setting where $f_a$ and $f_b$ are any continuous distributions on $\mathbb{R}$. For clarity, we notate our fairness metrics with tildes instead of hats (i.e., the sample accuracy for group $a$ is now $\widetilde{ACC}_a$).

Note that for all of the distributional derivations, the results and corresponding proofs are exactly analogous to those in 3.3, so we omit fully writing out these proofs again in this section.

3.4.1 Aside: Key Integral

We derive the analogous integral as in Section 3.3.1 before proceeding to any of the distributional derivations.

Let $F_\epsilon$ and $f_\epsilon$ be the CDF and PDF of $\epsilon$, a continuous random variable on $\mathbb{R}$. Then, we have the following lemma.

Lemma 3.4.1 (General CDF Integral).

$$\int_{c-q}^{c-r} F_\epsilon(c - w)dw = (r - c)F_\epsilon(c - r) - (q - c)F_\epsilon(c - q)$$

$$- (F_\epsilon(c - q) - F_\epsilon(c - r))E[\epsilon_{(c-r, c-q)}],$$

where $E[\epsilon_{(c-r, c-q)}]$ represents the expected value of $\epsilon$ when truncated to the integral $(c-r, c-q)$.

Proof. Again using integration by parts, we have

$$\int_{c-q}^{c-r} F_\epsilon(c - w)dw$$
\[ w F_e (c - w) \left|_q^r \right. - \int_q^r w (-f_e (c - w)) \, dw \]
\[ = [r F_e (c - r) - q F_e (c - q)] - \int_{c-q}^{c-r} (c - u) f_e (u) \, du \]
\[ = [r F_e (c - r) - q F_e (c - q)] + \int_{c-r}^{c-q} (c - u) f_e (u) \, du \]
\[ = [r F_e (c - r) - q F_e (c - q)] + \int_{c-r}^{c-q} c f_e (u) \, du - \int_{c-r}^{c-q} u f_e (u) \, du \]
\[ = [r F_e (c - r) - q F_e (c - q)] + \left[ c (F_e (c - q) - F_e (c - r)) - \int_{c-r}^{c-q} u f_e (u) \, du \right] \]
\[ = (r - c) F_e (c - r) - (q - c) F_e (c - q) - \int_{c-r}^{c-q} u f_e (u) \, du \]
\[ = (r - c) F_e (c - r) - (q - c) F_e (c - q) \]
\[ - (F_e (c - q) - F_e (c - r)) \int_{c-r}^{c-q} \frac{f_e (u)}{F_e (c - q) - F_e (c - r)} \, du \]
\[ = (r - c) F_e (c - r) - (q - c) F_e (c - q) \]
\[ - (F_e (c - q) - F_e (c - r)) \mathbb{E}[\epsilon (c - r, c - q)]. \]

3.4.2 Derivation of Accuracy Distribution

Moment Calculations

**Lemma 3.4.2** (Expected Value of Accuracy). The expected value of the accuracy of group \( a \), \( p_{\text{ACC}_a} \), in this thresholding setting is

\[
\begin{align*}
p_{\text{ACC}_a} & := \mathbb{E}[\text{ACC}_a] \\
& = \frac{t_a - c_a}{t_a - s_a} - \left( \frac{t_a - c_a}{t_a - s_a} F_{\epsilon_a} (c_a - t_a) + \frac{s_a - c_a}{t_a - s_a} F_{\epsilon_a} (c_a - s_a) \right) \\
& \quad + \frac{1}{t_a - s_a} \left( (F_{\epsilon_a} (0) - F_{\epsilon_a} (c_a - t_a)) \mathbb{E}[\epsilon_{a, (c_a - t_a, 0)}] \\
& \quad - (F_{\epsilon_a} (c_a - s_a) - F_{\epsilon_a} (0)) \mathbb{E}[\epsilon_{a, (0, c_a - s_a)}] \right) \\
\end{align*}
\]
where \( c_a \) is the decision threshold for group \( a \), the true skill distribution is 
\( \text{Unif}(s_a, t_a) \), and the noise distribution has PDF \( f_{\epsilon_a} \).

**Proof.** The proof follows exactly analogous to that of Lemma 3.3.2, except using the result of Lemma 3.4.1 instead of Lemma 3.3.1.

---

**Exact, Approximate, and Asymptotic Distributions**

**Lemma 3.4.3** (Exact Distribution of Accuracy). *The exact distribution of the accuracy for group \( a \) in this thresholding setting with general noise is

\[
\hat{\text{ACC}}_a \sim \frac{1}{n_a} \text{Binomial}(n_a, p_{\hat{\text{ACC}}_a}),
\]

with \( p_{\hat{\text{ACC}}_a} \) as defined in Lemma 3.4.2.*

**Lemma 3.4.4** (Approximate Distribution of Accuracy). *The approximate distribution of the accuracy for group \( a \) is

\[
\hat{\text{ACC}}_a \sim \mathcal{N}(p_{\hat{\text{ACC}}_a}, \frac{p_{\hat{\text{ACC}}_a}(1 - p_{\hat{\text{ACC}}_a})}{n_a}),
\]

for \( n_a \) large.*

**Lemma 3.4.5** (Asymptotic Distribution of Accuracy). *The asymptotic distribution of the accuracy for group \( a \) is

\[
\sqrt{n_a}(\hat{\text{ACC}}_a - p_{\hat{\text{ACC}}_a}) \xrightarrow{D} \mathcal{N}(0, 1).
\]

The resulting approximate distribution is

\[
\hat{\text{ACC}}_a \sim \mathcal{N}(p_{\hat{\text{ACC}}_a}, \frac{p_{\hat{\text{ACC}}_a}(1 - p_{\hat{\text{ACC}}_a})}{n_a}),
\]

which matches the result in Lemma 3.4.4.*
3.4.3 Derivation of Acceptance Rate Distribution

Moment Calculations

Lemma 3.4.6 (Expected Value of Acceptance Rate). The expected value of the acceptance rate of group $a$, $p_{\bar{AR}_a}$, in this thresholding setting is

$$p_{\bar{AR}_a} := \mathbb{E}[\bar{AR}_a]$$

$$= 1 - \left[\frac{t_a - c_a}{t_a - s_a} F_{\epsilon_a}(c_a - t_a) - \frac{s_a - c_a}{t_a - s_a} F_{\epsilon_a}(c_a - s_a)\right]$$

$$+ \frac{1}{t_a - s_a} (F_{\epsilon_a}(c_a - s_a) - F_{\epsilon_a}(c_a - t_a)) \mathbb{E}[\epsilon_a(c_a - t_a, c_a - s_a)]$$

where $c_a$ is the decision threshold for group $a$, the true skill distribution is $\text{Unif}(s_a, t_a)$, and the noise distribution has PDF $f_{\epsilon_a}$.

Proof. Similarly, the proof follows exactly analogous to that of Lemma 3.3.7, except using the result of Lemma 3.4.1 instead of Lemma 3.3.1.

Exact, Approximate, and Asymptotic Distributions

Lemma 3.4.7 (Exact Distribution of Acceptance Rate). The exact distribution of the acceptance rate for group $a$ in this thresholding setting with general noise is

$$\tilde{AR}_a \sim \frac{1}{n_a} \text{Binomial}(n_a, p_{\bar{AR}_a})$$

with $p_{\bar{AR}_a}$ as defined in Lemma 3.4.6.

Lemma 3.4.8 (Approximate Distribution of Acceptance Rate). The approximate distribution of the acceptance rate for group $a$ is

$$\tilde{AR}_a \sim \mathcal{N}\left(p_{\bar{AR}_a}, \frac{p_{\bar{AR}_a}(1 - p_{\bar{AR}_a})}{n_a}\right)$$

for $n_a$ large.
**Lemma 3.4.9** (Asymptotic Distribution of Acceptance Rate). *The asymptotic distribution of the acceptance rate for group \( a \) is*

\[
\frac{\sqrt{n_a}(\hat{AR}_a - p_{\hat{AR}_a})}{\sqrt{p_{\hat{AR}_a}(1 - p_{\hat{AR}_a})}} \xrightarrow{D} N(0, 1).
\]

*The resulting approximate distribution is*

\[
\hat{AR}_a \sim N\left(p_{\hat{AR}_a}, \frac{p_{\hat{AR}_a}(1 - p_{\hat{AR}_a})}{n_a}\right),
\]

*which matches the result in Lemma 3.4.8.*

### 3.4.4 Derivation of FPR Distribution

**Moment Calculations**

**Lemma 3.4.10** (Expected Value of FPR). *The expected value of the FPR of group \( a \), \( p_{\overline{FPR}_a} \), in this thresholding setting is*

\[
p_{\overline{FPR}_a} := \mathbb{E}[\overline{FPR}_a]
= 1 - F_{\epsilon_a}(c_a - s_a) + \frac{1}{c_a - s_a} (F_{\epsilon_a}(c_a - s_a) - F_{\epsilon_a}(0)) \mathbb{E}[\epsilon_a(0,c_a-s_a)]
\]

*where \( c_a \) is the decision threshold for group \( a \), the true skill distribution is Unif\((s_a, t_a)\), and the noise distribution has PDF \( f_{\epsilon_a} \).*

*Proof.* Similarly, the proof follows exactly analogous to that of Lemma 3.3.12, except using the result of Lemma 3.4.1 instead of Lemma 3.3.1. \( \square \)
**Approximate Distribution**

**Lemma 3.4.11** (Approximate Distribution of FPR). An approximate distribution of the FPR for group $a$ is

$$\widetilde{FPR}_a \sim \frac{1}{n_a p_{a,0}} \text{Binomial}(n_a, p_{a,0} p_{\widetilde{FPR}_a}),$$

for $n_a$ large, with $p_{a,0} := \mathbb{P}(Y_{a,i} = 0) = \frac{c_a - s_a}{t_a - s_a}$ the probability an individual’s true skill is below the group $a$ threshold $c_a$ and $p_{\widetilde{FPR}_a}$ as defined in Lemma 3.4.10.

Using the Normal approximation for the Binomial, we obtain another approximate distribution for $\widetilde{FPR}_a$,

$$\widetilde{FPR}_a \sim \mathcal{N} \left( p_{\widetilde{FPR}_a}, \frac{p_{\widetilde{FPR}_a} (1 - p_{a,0} p_{\widetilde{FPR}_a})}{n_a p_{a,0}} \right),$$

again for $n_a$ large.

Thus, it is clear that the approximate variance of $\widetilde{FPR}_a$ is

$$\text{Var}(\widetilde{FPR}_a) = \frac{p_{\widetilde{FPR}_a} (1 - p_0 p_{\widetilde{FPR}_a})}{n_a p_0}.$$  

**3.4.5 Derivation of FNR Distribution**

**Moment Calculations**

**Lemma 3.4.12** (Expected Value of FNR). The expected value of the FNR of group $a$, $p_{\widetilde{FNR}_a}$, in this thresholding setting is

$$p_{\widetilde{FNR}_a} := \mathbb{E}[\widetilde{FPR}_a]$$

$$= F_{\epsilon_a}(c_a - t_a) - \frac{1}{c_a - t_a} (F_{\epsilon_a}(0) - F_{\epsilon_a}(c_a - t_a)) \mathbb{E}[\epsilon_{a,(c_a - t_a,0)}]$$

where $c_a$ is the decision threshold for group $a$, the true skill distribution is $\text{Unif}(s_a, t_a)$, and the noise distribution has PDF $f_{\epsilon_a}$. 


**Proof.** Similarly, the proof follows exactly analogous to that of Lemma 3.3.15, except using the result of Lemma 3.4.1 instead of Lemma 3.3.1.

**Approximate Distribution**

**Lemma 3.4.13 (Approximate Distribution of FNR).** An approximate distribution of the FNR for group \(a\) is

\[
\tilde{FNR}_a \sim \frac{1}{n_ap_{a,1}} \text{Binomial} \left( n_a, p_{a,1}p_{\tilde{FNR}_a} \right),
\]

for \(n_a\) large, with \(p_{a,1} := \mathbb{P}(Y_{a,i} = 1) = \frac{t_a - c_a}{t_a - s_a}\) the probability an individual’s true skill is below the threshold \(c_a\) and \(p_{\tilde{FNR}_a}\) as defined in Lemma 3.4.12.

Using the Normal approximation for the Binomial, we obtain another approximate distribution for \(\tilde{FNR}_a\),

\[
\tilde{FNR}_a \sim \mathcal{N} \left( p_{\tilde{FNR}_a}, \frac{p_{\tilde{FNR}_a}(1 - p_{a,1}p_{\tilde{FNR}_a})}{n_ap_{a,1}} \right),
\]

again for \(n_a\) large.

Thus, it is clear that the approximate variance of \(\tilde{FNR}_a\) is

\[
\text{Var}(\tilde{FNR}_a) = \frac{p_{\tilde{FNR}_a}(1 - p_{a,1}p_{\tilde{FNR}_a})}{n_ap_{a,1}}.
\]

**3.4.6 Derivation of PPV Distribution**

**Moment Calculations**

**Lemma 3.4.14 (Expected Value of PPV).** The expected value of the PPV of group \(a\), \(p_{\tilde{PPV}_a}\), in this thresholding setting is

\[
p_{\tilde{PPV}_a} := \mathbb{E}[\tilde{PPV}_a]
\]

67
where \( c_a \) is the decision threshold for group \( a \), the true skill distribution is \( \text{Unif}(s_a, t_a) \), and the noise distribution has PDF \( f_{\epsilon_a} \).

**Proof.** Similarly, the proof follows exactly analogous to that of Lemma 3.3.18, except using the result of Lemma 3.4.1 instead of Lemma 3.3.1. \( \square \)

### Approximate Distribution

**Lemma 3.4.15** (Approximate Distribution of PPV). An approximate distribution of the PPV for group \( a \) is

\[
\hat{PPV}_a \sim \frac{1}{n_a p_{\overline{AR}_a}} \text{Binomial}\left(n_a, p_{\overline{AR}_a}, p_{\overline{PPV}_a}\right),
\]

for \( n_a \) large, with \( p_{\overline{AR}_a} \) as defined in Lemma 3.4.6 and \( p_{\overline{PPV}_a} \) as defined in Lemma 3.4.14.

Using the Normal approximation for the Binomial, we obtain another approximate distribution for \( \hat{PPV}_a \),

\[
\hat{PPV}_a \sim \mathcal{N}\left(p_{\overline{PPV}_a}, \frac{p_{\overline{PPV}_a} (1 - p_{\overline{AR}_a} p_{\overline{PPV}_a})}{n_a p_{\overline{AR}_a}}\right),
\]

again for \( n_a \) large.

Thus, it is clear that the approximate variance of \( \hat{PPV}_a \) is

\[
\text{Var}(\hat{PPV}_a) = \frac{p_{\overline{PPV}_a} (1 - p_{\overline{AR}_a} p_{\overline{PPV}_a})}{n_a p_{\overline{AR}_a}}.
\]
Having derived the distributions of our five fairness metrics of interest, we now present insights and observations drawn directly from these distributional quantities.

Beforehand, we note that we focus on the Gaussian results from Section 3.3 for all of this chapter until the last subsection (Section 4.5) where we explore some more general trends. We also emphasize in particular that the variances used throughout this chapter are the *exact* variances for accuracy and acceptance rate from Lemmas 3.3.4 and 3.3.9, and the *approximate* variances for FPR, FNR, and PPV from Lemmas 3.3.14, 3.3.17, and 3.3.20 (and their analogs for the general noise case). As discussed in the proof of Lemma 3.3.19, while our proposed approximate PPV variance is an overestimate of the true simulated variance, it still acts as a heuristic and upper bound of the true variance and thus allows for insightful observations nonetheless.
We also note that for much of this chapter, our empirical analyses focus on the case where \( s_a, s_b, t_a, t_b \in [0, 100] \). While this decision was largely for simplicity and maintaining consistency throughout the empirical analyses presented in this thesis, it also has a deeper significance, namely in being a range appropriate for percentile values. Recall that in Section 3.1.2 we mentioned that any skill measurements can be converted to skill *percentiles* by the probability integral transform. Then, \( s_i \) and \( t_i \) can be chosen to represent the skill level of a particular protected group within the full range of percentiles in the overall population, which must be values between 0 and 100. Thus, this choice provides increased interpretability of some of the empirical plots we produce below. Note that we also typically take the sample size to be \( n = 100 \) as this is sufficiently large to justify our approximate variances, the threshold \( c = 50 \) for simplicity as it represents the median percentile value across the entire population, and \( \sigma = 10 \), unless otherwise specified in a given figure.

### 4.1 Metric Variation with Threshold \( c \)

We begin by focusing on the case with Gaussian noise again, because of its nice properties and well-defined nature. We first analyze empirically how the metrics, as defined by their mean and variance which we derived in Section 3.3, evolve as a function of the threshold \( c \).

Figure 4.1.1 below shows a simple example of how the theoretical means in particular of the five metrics vary with the threshold \( c \), fixing the underlying skill distribution to be \( \text{Unif}(0, 100) \) and the noise to be \( \mathcal{N}(0, 10^2) \) with \( n = 100 \) people.
Figure 4.1.1: Mean and error bars (two standard deviations in width) for each of the five fairness metrics, across all integer thresholds $c \in [1, 99]$, with underlying Unif(0, 100) and $\mathcal{N}(0, 10^2)$ noise.

Namely, we notice that acceptance rate and PPV decrease monotonically with $c$, the accuracy remains roughly consistent, which makes sense since we are not altering the amount of noise at all, and the FPR and FNR behave symmetrically about the Uniform distribution center of 50, which also makes sense intuitively.

In Figure 4.1.2 below, we plot the actual magnitudes of the standard deviations of the metrics over different thresholds $c$. We note in particular that the acceptance rate and accuracy standard deviations appear to be maximized at $c = 50$ and are symmetric about this point; further, the standard deviation of the acceptance rate is strictly greater than that of the accuracy across all $c$, which reveals fundamental differences in certainty or uncertainty about these different
measurements. It is also visually evident from the graph that the standard deviations of the FPR and FNR are symmetric about $c = 50$, which makes sense. The PPV has no apparent trends, except that it tracks much of the curve of the FNR, except at the extreme threshold values.

![Metric standard deviations for Unif(0,100) skill, N(0,10^2) noise](image)

**Figure 4.1.2:** Standard deviation of the five fairness metrics, across all integer thresholds $c \in [1, 99]$, with underlying Unif(0, 100) and $\mathcal{N}(0, 10^2)$ noise.

We extend a version of the above plots for the means and standard deviations separately, while also introducing variability in the underlying noise distribution, in Figure 4.1.3.
Figure 4.1.3: Mean and standard deviation of each of the fairness metrics, across all integer thresholds $c \in [41, 59]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, 10^2)$ noise.
We notice that for Uniform skill distributions with higher-valued supports, the FPR mean is strictly lower and FNR mean strictly higher than their counterparts with lower-valued supports. The PPV mean trend is roughly the same as that of FNR. For the accuracy and acceptance rate means, there seems to be some sort of symmetry depending on the distance of the Uniform center from relative decision thresholds; Figure 4.1.4 below illustrates this more clearly for the accuracy means.

![Figure 4.1.4: Accuracy mean across all integer thresholds $c \in [41, 59]$, with varying underlying skill distributions and $\mathcal{N}(0, 10^2)$ noise.](image)

From Figure 4.1.4, we notice an interesting trend; namely, the more extreme the distribution relative to the decision threshold (which is 50 here), the uniformly higher expected accuracy across all thresholds $c$, when compared to all other underlying distributions which are biased in the same direction with
respect to the decision threshold. Concretely, in our plot, the expected accuracy for the group with Unif(40,100) underlying skill is uniformly higher than all other groups for which the decision threshold $c = 50$ lies in the lower half of the skill distribution’s support (e.g., Unif(35,95), Unif(30,90), Unif(25,85), and Unif(20,80) in the graph above); the same is true for the group with Unif(0,60) underlying skill relative to the corresponding set of distributions for which $c = 50$ lies in the upper half of the support.

This observation has some interesting philosophical implications. Namely, it suggests that any group with a relatively extreme skill distribution with respect to the decision threshold should have, on average, uniformly higher accuracy, compared to any group with a less extreme skill distribution on the same side of the decision threshold. Then, this implies, for example, that the group with the highest skill should have uniformly higher model accuracy than a group with relatively high skill. In a case where skill level is directly correlated with, say, socioeconomic status, then these theoretical results would suggest the model should always do better on the highest income subpopulation regardless of the decision threshold, relative to other higher-than-average income subpopulations.

We see two main ways these conclusions could be interpreted. In one sense, this could imply that it should not be considered unfair if more extreme groups have higher observed accuracy relative to less extreme groups on the same side of the decision threshold, since in expectation, this is what is expected to happen theoretically. However, on another hand, others might use this to argue that thresholding models are not appropriate for this reason that a discrepancy fundamentally exists in accuracy expected values between groups with different underlying skill distributions.

4.2 Metric Variation with Sample Size $n$

We also examine how the metrics vary with the sample size $n$. Note that we choose relatively large values of $n$ ($n > 50$) to analyze here, so that our approximate distribution assumptions are reasonable.
Figure 4.2.1 below plots the mean of the fairness metrics with errors bars representing two standard deviations, for different values of \( n \); Figure 4.2.2 does the same for the standard deviations only. We note that the standard deviations decrease monotonically as the sample size \( n \) increases, as expected, and their relative magnitudes are smaller for accuracy, roughly similar for acceptance rate, FPR, and FNR, and much larger for PPV, in the Unif\((20,60)\) and \( \mathcal{N}(0,10^2) \) case.

![Mean with two standard deviation bars for Unif(0,100) skill and \( \mathcal{N}(0,10^2) \) noise, threshold 50](image)

**Figure 4.2.1:** Mean of each of the five fairness metrics with error bars representing two standard deviations, across sample size \( n \in [50,500] \), for underlying Unif\((20,60)\) skill and \( \mathcal{N}(0,10^2) \) noise.
Figure 4.2.2: Standard deviation of each of the five fairness metrics, across sample size $n \in [50, 500]$, for underlying $\text{Unif}(20, 60)$ skill and $\mathcal{N}(0, 10^2)$ noise.

4.3 Metric Variation with Noise Variance $\sigma^2$

Now, we consider how the distributions of these metrics evolve with the amount of noise $\sigma$. First, we visualize the evolution of these metrics in Figure 4.3.1, which plots the mean and standard deviation of the distribution of each of the five metrics, as a function of the noise standard deviation $\sigma$ for one specific underlying distribution.
From the plot, it is already evident that the different metrics grow and decay at different rates across $\sigma$; we examine this idea further in the next section.

Similar to Figure 4.1.3, we plot the mean for a variety of underlying skill distributions, this time as a function of the noise standard deviation, for each of the five metrics separately. The top graphs looks at a wider range of underlying distributions, while the bottom graphs look at a more constrained set of distributions which particularly enunciate the shape of these mean curves as they vary with $\sigma$. 

**Figure 4.3.1**: Mean of each of the five fairness metrics with error bars representing two standard deviations, across noise $\sigma \in [1, 25]$, for underlying Unif(20, 60) skill and $\mathcal{N}(0, \sigma^2)$ noise.
Figure 4.3.2: Mean of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, \sigma^2)$ noise.
We note one key high-level trend for each metric. Namely, when the underlying skill distribution is farther from the decision threshold, the mean theoretical accuracy is strictly higher across all $\sigma$ compared to the average accuracies when the underlying skill distribution is centered closer to the decision threshold. Further, for underlying skill distributions with higher-valued supports, the mean theoretical acceptance rate, FPR, and PPV are strictly higher across all $\sigma$, but strictly lower across all $\sigma$ for FNR, when compared with underlying skill distributions with lower-valued supports.

Perhaps more consequential, we also observe that the metrics all seem to decay faster with the noise when the underlying skill distributions are narrower in width, when considering the pairwise plots in the upper and lower grids in Figure 4.3.2. Then, this suggests that groups with narrower skill distributions may experience disproportionately faster decay across multiple metrics as the noise increases compared to groups with wider skill distributions, which is concerning when considering the fact that groups with narrower skill distributions could reasonably correspond to smaller or historically marginalized populations.

We repeat a similar process for the standard deviations instead of the means. Figure 4.3.3 gives the standard deviation values corresponding to the same parameters as Figure 4.3.1, and Figure 4.3.4 does the same for Figure 4.3.2.
Figure 4.3.3: Standard deviation of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, for underlying $\text{Unif}(20, 60)$ skill and $\mathcal{N}(0, \sigma^2)$ noise.
Figure 4.3.4: Standard deviation of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, \sigma^2)$ noise.
In particular, when the underlying skill distribution is farther from the decision threshold, the standard deviation of the accuracy is strictly lower, indicating less uncertainty. For underlying skill distributions with higher-valued supports, the theoretical standard deviation for acceptance rates, FNR, and PPV are all strictly lower than that of corresponding skill distributions with lower-valued supports, and strictly higher for FPR.

4.4 Metric Rate of Decay with $\sigma$

We now look at how the metrics decay with $\sigma$, to identify whether certain metrics are strictly better in terms of robustness to the amount of noise present, and under what conditions this happens.

To do this, we wish to derive the partial derivatives of our metric means and standard deviations with respect to $\sigma$. Then, first, we begin with two quick asides during which we derive the partial derivative of the quantity from Lemma 3.3.1 with respect to $\sigma$, since this will be a useful result when deriving the actual metric derivatives later on.

Lemma 4.4.1 (Gaussian PDF Derivative).

$$\frac{\partial}{\partial \sigma} \varphi \left( \frac{c - r}{\sigma} \right) = \varphi \left( \frac{c - r}{\sigma} \right) \cdot \frac{(c - r)^2}{\sigma^3}$$

Proof.

$$\frac{\partial}{\partial \sigma} \varphi \left( \frac{c - r}{\sigma} \right) = \frac{\partial}{\partial \sigma} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(c - r)^2}{2\sigma^2} \right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(c - r)^2}{2\sigma^2} \right) \cdot \left( -\frac{2(c - r)^2}{2\sigma^3} \cdot (-2) \right)$$

$$= \varphi \left( \frac{c - r}{\sigma} \right) \cdot \frac{(c - r)^2}{\sigma^3}$$

$\square$
Lemma 4.4.2 (Derivative of Gaussian CDF Integral).

\[
\frac{\partial}{\partial \sigma} \int_q^r \Phi \left( \frac{c - w}{\sigma} \right) \, dw = -\varphi \left( \frac{c - r}{\sigma} \right) + \varphi \left( \frac{c - q}{\sigma} \right)
\]

Proof.

\[
\frac{\partial}{\partial \sigma} \int_q^r \Phi \left( \frac{c - w}{\sigma} \right) \, dw \\
= \frac{\partial}{\partial \sigma} \left[ (r - c)\Phi \left( \frac{c - r}{\sigma} \right) - (q - c)\Phi \left( \frac{c - q}{\sigma} \right) - \sigma \left[ \varphi \left( \frac{c - r}{\sigma} \right) - \varphi \left( \frac{c - q}{\sigma} \right) \right] \right] \\
= (r - c)\varphi \left( \frac{c - r}{\sigma} \right) \cdot \frac{r - c}{\sigma^2} - (q - c)\varphi \left( \frac{c - q}{\sigma} \right) \cdot \frac{q - c}{\sigma^2} \\
\quad - \left( \varphi \left( \frac{c - r}{\sigma} \right) + \sigma \varphi \left( \frac{c - r}{\sigma} \right) \cdot \frac{(c - r)^2}{\sigma^3} \right) \\
\quad + \left( \varphi \left( \frac{c - q}{\sigma} \right) + \sigma \varphi \left( \frac{c - r}{\sigma} \right) \cdot \frac{(c - q)^2}{\sigma^3} \right) \\
= -\varphi \left( \frac{c - r}{\sigma} \right) + \varphi \left( \frac{c - q}{\sigma} \right)
\]

\[
\Box
\]

Before proceeding to deriving the rates of decay of the mean of each of our five metrics, we introduce one final supplementary lemma which helps simplify the derivative calculations.

Lemma 4.4.3 (Metric Expected Values Rewritten). We can more succinctly express the means of our five metrics of interest using the integrals

\[
A = \int_{c_a}^{t_a} \Phi \left( \frac{c_a - w}{\sigma_a} \right) \, dw \\
B = \int_{c_a}^{t_a} \Phi \left( \frac{c_a - w}{\sigma_a} \right) \, dw
\]

as follows:

\[
p_{\text{ACC}_a} = \frac{t_a - c_a}{t_a - s_a} + \frac{A - B}{t_a - s_a} \tag{4.1}
\]

\[
p_{\text{AR}_a} = 1 - \frac{A + B}{t_a - s_a} \tag{4.2}
\]

\[
p_{\text{FPR}_a} = 1 - \frac{A}{c_a - s_a} \tag{4.3}
\]

84
\[
\frac{p_{\text{FPF}}}{\text{a}} = \frac{t_a - c_a}{t_a - s_a} - \frac{B}{t_a - s_a} \left( \frac{t_a - c_a - B}{t_a - s_a} \right) \tag{4.4}
\]

\[
\frac{p_{\text{FPV}}}{\text{a}} = \frac{t_a - c_a}{t_a - s_a} - \frac{B}{t_a - s_a} \left( \frac{t_a - c_a - B}{t_a - s_a} \right) \tag{4.5}
\]

Proof. These directly follow from intermediate steps during the proofs of Lemmas 3.3.2, 3.3.7, 3.3.12, 3.3.15, and 3.3.18.

Finally, equipped with Lemmas 4.4.1, 4.4.2, and 4.4.3, we give expressions for the partial derivatives of our metric means, with respect to \( \sigma \).

**Theorem 4.4.1** (Derivative of Metric Means). The partial derivatives of the means of our five sample metrics are as follows:

\[
\frac{\partial}{\partial \sigma_a} p_{\text{ACC}} = \frac{1}{t_a - s_a} \left( \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \right) - \frac{2}{(t_a - s_a)\sqrt{2\pi}} \tag{4.6}
\]

\[
\frac{\partial}{\partial \sigma_a} p_{\text{AR}} = \frac{1}{t_a - s_a} \left( \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) - \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \right) \tag{4.7}
\]

\[
\frac{\partial}{\partial \sigma_a} p_{\text{FPF}} = \frac{1}{(c_a - s_a)\sqrt{2\pi}} - \frac{1}{c_a - s_a} \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) \tag{4.8}
\]

\[
\frac{\partial}{\partial \sigma_a} p_{\text{FNR}} = \frac{1}{(t_a - c_a)\sqrt{2\pi}} - \frac{1}{t_a - c_a} \varphi \left( \frac{c_a - t_a}{\sigma_a} \right) \tag{4.9}
\]

\[
\frac{\partial}{\partial \sigma_a} p_{\text{FPV}} = -\frac{\varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \frac{1}{\sqrt{2\pi}} p_{\text{AR}}}{t_a - s_a} - \frac{\varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \frac{1}{\sqrt{2\pi}} p_{\text{AR}}}{t_a - s_a} \tag{4.10}
\]

Proof. These follow immediately when applying the results from Lemma 4.4.2 to the metric expressions from Lemma 4.4.3. More specifically, starting from the expressions in Lemma 4.4.3, we have

\[
\frac{\partial}{\partial \sigma_a} p_{\text{ACC}} = \frac{\partial}{\partial \sigma_a} A - \frac{\partial}{\partial \sigma_a} B
\]

\[
\frac{\partial}{\partial \sigma_a} p_{\text{AR}} = \frac{\partial}{\partial \sigma_a} A + \frac{\partial}{\partial \sigma_a} B
\]
\[
\frac{\partial}{\partial \sigma_a} p_{\text{FPRA}_a} = -\frac{\partial}{\partial \sigma_a} \frac{A}{c_A - s_A} \\
\frac{\partial}{\partial \sigma_a} p_{\text{FNR}_a} = \frac{\partial}{\partial \sigma_a} \frac{B}{t_a - c_a} \\
\frac{\partial}{\partial \sigma_a} p_{\text{PPV}_a} = \frac{-\frac{\partial}{\partial \sigma_a} B}{t_a - s_a} p_{\text{AR}_a} - \left( \frac{t_a - c_a}{t_a - s_a} - \frac{B}{t_a - s_a} \right) \left( \frac{\partial}{\partial \sigma_a} p_{\text{AR}_a} \right)
\]

using basic derivative rules. Then, from Lemma 4.4.2, we have

\[
\frac{\partial}{\partial \sigma_a} A = -\varphi(0) + \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) = \varphi \left( \frac{c_a - s_a}{\sigma_a} \right) - \frac{1}{\sqrt{2\pi}} \\
\frac{\partial}{\partial \sigma_a} B = -\varphi \left( \frac{c_a - t_a}{\sigma_a} \right) + \varphi(0) = \frac{1}{\sqrt{2\pi}} - \varphi \left( \frac{c_a - t_a}{\sigma_a} \right).
\]

Plugging these into the derivative expressions above give us our desired results.

\[\square\]

**Theorem 4.4.2 (Derivative of Metric Standard Deviations).** The partial derivatives of the variances of our five sample metrics are as follows:

\[
\frac{\partial}{\partial \sigma_a} \text{Var}(\hat{\text{ACC}}_a) = \frac{1}{n_a} \left( 1 - 2 \cdot p_{\hat{\text{ACC}}_a} \cdot \frac{\partial}{\partial \sigma_a} p_{\hat{\text{ACC}}_a} \right) 
\]

\[\text{(4.11)}\]

\[
\frac{\partial}{\partial \sigma_a} \text{Var}(\hat{\text{AR}}_a) = \frac{1}{n_a} \left( 1 - 2 \cdot p_{\hat{\text{AR}}_a} \cdot \frac{\partial}{\partial \sigma_a} p_{\hat{\text{AR}}_a} \right) 
\]

\[\text{(4.12)}\]

\[
\frac{\partial}{\partial \sigma_a} \text{Var}(\hat{\text{FPR}}_a) = \frac{1}{n_a p_0} \left( 1 - 2p_0 \cdot p_{\hat{\text{FPR}}_a} \cdot \frac{\partial}{\partial \sigma_a} p_{\hat{\text{FPR}}_a} \right) 
\]

\[\text{(4.13)}\]

\[
\frac{\partial}{\partial \sigma_a} \text{Var}(\hat{\text{FNR}}_a) = \frac{1}{n_a p_1} \left( 1 - 2p_1 \cdot p_{\hat{\text{FNR}}_a} \cdot \frac{\partial}{\partial \sigma_a} p_{\hat{\text{FNR}}_a} \right) 
\]

\[\text{(4.14)}\]

\[
\frac{\partial}{\partial \sigma_a} \text{Var}(\hat{\text{PPV}}_a) = \frac{\left( \frac{\partial}{\partial \sigma_a} p_{\hat{\text{PPV}}_a} \right) \cdot np_{\hat{\text{AR}}_a} - np_{\hat{\text{PPV}}_a} \cdot \left( \frac{\partial}{\partial \sigma_a} p_{\hat{\text{AR}}_a} \right)}{(np_{\hat{\text{AR}}_a})^2} \\
- \frac{2p_{\hat{\text{PPV}}_a} \cdot \frac{\partial}{\partial \sigma_a} p_{\hat{\text{PPV}}_a}}{n} 
\]

\[\text{(4.15)}\]
\textit{Proof.} From before, we found that
\[ \text{Var}(\hat{ACC}_a) = \frac{p_{\hat{ACC}_a}(1 - p_{\hat{ACC}_a})}{n_a}. \]

Then, its partial derivative with respect to \( \sigma_a \) is
\[ \frac{\partial}{\partial \sigma_a} \text{Var}(\hat{ACC}_a) = \frac{\partial}{\partial \sigma_a} \frac{p_{\hat{ACC}_a}(1 - p_{\hat{ACC}_a})}{n_a} = \frac{1}{n_a} \left( 1 - 2 \cdot p_{\hat{ACC}_a} \cdot \frac{\partial}{\partial \sigma_a} p_{\hat{ACC}_a} \right). \]

An analogous derivation follows for \( \text{Var}(\hat{AR}_a) \) due to their similar structure. For FPR, we have
\[ \frac{\partial}{\partial \sigma_a} \text{Var}(\hat{FPR}_a) = \frac{\partial}{\partial \sigma_a} \frac{p_{\hat{FPR}_a}(1 - p_{\hat{FPR}_a})}{n_p_{\hat{FPR}_a}} = \frac{1}{n_p_{\hat{FPR}_a}} \left( 1 - 2p_0 \cdot p_{\hat{FPR}_a} \cdot \frac{\partial}{\partial \sigma_a} p_{\hat{FPR}_a} \right), \]
and similarly for FPR.

Finally, for PPV, we have
\[ \frac{\partial}{\partial \sigma_a} \text{Var}(\hat{PPV}_a) = \frac{\partial}{\partial \sigma_a} \frac{p_{\hat{PPV}_a}(1 - p_{\hat{AR}_a}p_{\hat{PPV}_a})}{n_p_{\hat{AR}_a}} = \frac{2p_{\hat{PPV}_a} \cdot \frac{\partial}{\partial \sigma_a} p_{\hat{PPV}_a}}{n}, \]
by the quotient rule and previously derived quantities. \( \square \)

Using these above derivations, we move to visualizing the rate of decay of our metrics, with a set of analogous plots to Section 4.3. First, Figure 4.4.1 gives the derivative of the mean of each of the five metrics, as a function of the noise.
standard deviation $\sigma$.

![Derivative of metrics for Unif(20,60) skill, N(0,\sigma^2) noise, threshold 50](image)

**Figure 4.4.1:** Rate of decay of mean of each of the five fairness metrics with error bars representing two standard deviations, across noise $\sigma \in [1, 25]$, for underlying Unif(20,60) skill and $\mathcal{N}(0, \sigma^2)$ noise.

In Figure 4.4.2 below, we plot the rate of decay of the mean for the corresponding plots from Figure 4.3.2. The top graphs again look at a wider range of underlying distributions, and the bottom graphs look at a more constrained set of distributions.
Figure 4.4.2: Rate of decay of the mean of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, with varying underlying skill distributions of equal width and $\mathcal{N}(0, \sigma^2)$ noise.
The relative shapes give intuition as to how the metrics grow or decay with respect to sigma across different underlying skill distributions. In particular, the above plots also confirm our findings in Section 4.3, as the magnitude of derivatives are strictly higher for the lower set of plots—which correspond to the set of distributions narrower in width—rigorously confirming that these metrics decay faster for narrower distributions, the implications of which we have already briefly touched on.

We repeat the same process now for the rate of decay of the standard deviations. Figure 4.4.3 again gives the standard deviation values corresponding to the same parameters as Figure 4.4.1, and Figure 4.4.4 does the same for Figure 4.4.2.

**Figure 4.4.3:** Rate of decay of the standard deviation of each of the five fairness metrics, across noise $\sigma \in [1, 25]$, for underlying Unif$(20, 60)$ skill and $\mathcal{N}(0, \sigma^2)$ noise.
Figure 4.4.4: Standard deviation of each of the five fairness metrics, across noise \( \sigma \in [1, 25] \), with varying underlying skill distributions of equal width and \( \mathcal{N}(0, \sigma^2) \) noise.
We observe based on the y-axis magnitudes that the standard deviations also decay faster with $\sigma$ for narrower underlying skill distributions, similar to the mean.

4.5 Expanding to General Noise

We briefly also aim to glean insights from more general noise distributions.

An aspect of particular interest to us is the robustness of these distributional results; namely, we would like the insights we draw in this section and the next to extend beyond just Gaussian noise to more general noise distributions.

Then, in Figure 4.5.1 below, we show a similar plot to Figure 4.1.1, except we plot it for both Gaussian (top) and Laplace (bottom) noise, with the same mean and variance. We note that the trends are relatively similar for most of the metrics, just with small differences in terms of the rate of decay of the metrics, as well as the standard deviations, particularly for PPV.

We dive more into this idea of robustness later when we use these distributions to construct interval estimates of various fairness-related quantities, but for now, the similarity in these plots suggests that our closed-form results assuming Gaussian noise may be relatively generalizable to general noise distributions.
Figure 4.5.1: Mean and error bars (two standard deviations in width) for each of the five fairness metrics, across all integer thresholds $c \in [1, 99]$, with underlying $\text{Unif}(0,100)$, and zero-centered Gaussian noise (top) and Laplace noise (bottom) with standard deviation 10.
4.6 **Main Takeaways**

Here, we summarize a few of the most consequential observations from the figures and math presented in this chapter.

First, we found in Section 4.1 that skill distributions relatively more extreme with respect to the decision threshold had uniformly higher expected accuracy across all decision thresholds, in comparison to less extreme skill distributions. This result could be interpreted in a number of ways, either to justify accuracy inequality between different classes in a given thresholding setting, or to fundamentally contest the usage of threshold-based models when equality of accuracy is particularly desirable.

Second, in Section 4.3, we found that groups with underlying skill distributions of narrower support had uniformly faster decay across the mean and standard deviation of all five metrics of interest, as a function of the noise standard deviation $\sigma$. This becomes particularly concerning when considering the fact that data collected from marginalized populations may reflect this relatively narrow range of skill values simply due to small sample sizes or low variability, as this observation suggests that these groups could face disproportionately high metric decay as the noise increases.

Last, we observed in Section 4.4 that basic trends with respect to the five metric means and standard deviations for Gaussian noise and Laplace noise of identical mean and standard deviation were relatively similar. This starts to hint at the relative robustness of our closed-form expressions derived from Gaussian noise, but we explore this further later.

In the next chapter, we begin using these distributions to explore more complex quantities which are closer aligned to traditional notions of fairness, such as the *difference* in a particular fairness metrics between groups $a$ and $b$. 

94
Recall from Section 2.4.1 that when it comes to group fairness, the desired condition for a certain fairness definition to be satisfied is that some binary classification evaluation metric, for example one of those introduced in Section 2.3, achieves equal value across all protected subgroups. However, as discussed in Section 2.4.1, perfect numerical equality is a very strong condition that is not often achieved unless in trivial classification settings.

Thus, instead of specifically analyzing whether exact equality is achieved or not with respect to a certain metric across protected groups, we instead focus on better understanding the significance of the discrepancy of a particular metric’s value between groups, largely on the basis of the metric distributions we derived in Sections 3.3 and 3.4. For example, in a given thresholding setting, what is the
difference in acceptance rates *expected* to be between two protected groups, and where do we expect most (e.g., 95%) of the values of the acceptance rate difference to fall?

5.1 Derivation of Difference Distributions

Since we wish to better quantify the significance of a difference in a certain metrics between groups, and how fair or unfair these discrepancies are, we begin by deriving the distributions of the difference of metrics using the distributions we derived in 3.3, starting with the case of zero-centered Gaussian noise.

5.1.1 Uniform Skill, Gaussian Noise

**Lemma 5.1.1 (Distributions of Accuracy Difference).** The exact distribution of the difference of accuracies for groups $a$ and $b$, $\widehat{ACC}_a - \widehat{ACC}_b$, in this thresholding setting is

$$\widehat{ACC}_a - \widehat{ACC}_b \sim \frac{1}{n_a} \text{Bin} \left( n_a, p_{\widehat{ACC}_a} \right) - \frac{1}{n_b} \text{Bin} \left( n_b, p_{\widehat{ACC}_b} \right),$$

with $p_{\widehat{ACC}_a}$ as defined in Lemma 3.3.2 and $p_{\widehat{ACC}_b}$ the analogous quantity for group $b$, where $c_a, c_b$ are the decision thresholds, the true skill distributions are $\text{Unif}(s_a, t_a)$ and $\text{Unif}(s_b, t_b)$, and the noise distributions are $\mathcal{N}(0, \sigma_a^2)$ and $\mathcal{N}(0, \sigma_b^2)$, for groups $a$ and $b$ respectively. Note that the two Binomial distributions are independent by independence of the groups.

An approximate distribution of the difference of accuracies is

$$\widehat{ACC}_a - \widehat{ACC}_b \sim \mathcal{N} \left( p_{\widehat{ACC}_a} - p_{\widehat{ACC}_b}, \frac{p_{\widehat{ACC}_a} (1 - p_{\widehat{ACC}_a})}{n_a} + \frac{p_{\widehat{ACC}_b} (1 - p_{\widehat{ACC}_b})}{n_b} \right),$$

for $n_a$ and $n_b$ large.

**Proof.** Using the result from Lemma 3.3.4 and the fact that the two groups are independent, the result immediately follows for the exact distribution. For the
approximate distribution, the same is true, except using the result from Lemma 3.3.5 and Normal properties to simplify the distribution as written above.

Lemma 5.1.2 (Distributions of Acceptance Rate Difference). The exact distribution of the difference of acceptance rates for groups $a$ and $b$, $\widehat{AR}_a - \widehat{AR}_b$, in this thresholding setting is

$$\widehat{AR}_a - \widehat{AR}_b \sim \frac{1}{n_a} \text{Binomial}(n_a, p_{\widehat{AR}_a}) - \frac{1}{n_b} \text{Binomial}(n_b, p_{\widehat{AR}_CC_b}),$$

with $p_{\widehat{AR}_a}$ as defined in Lemma 3.3.7 and $p_{\widehat{AR}_b}$ the analogous quantity for group $b$, where $c_a, c_b$ are the decision thresholds, the true skill distributions are $\text{Unif}(s_a, t_a)$ and $\text{Unif}(s_b, t_b)$, and the noise distributions are $\mathcal{N}(0, \sigma_a^2)$ and $\mathcal{N}(0, \sigma_b^2)$, for groups $a$ and $b$ respectively.

An approximate distribution of the difference of acceptance rates is

$$\widehat{AR}_a - \widehat{AR}_b \sim \mathcal{N}(p_{\widehat{AR}_a} - p_{\widehat{AR}_b}, \frac{p_{\widehat{AR}_a}(1 - p_{\widehat{AR}_a})}{n_a} + \frac{p_{\widehat{AR}_b}(1 - p_{\widehat{AR}_b})}{n_b}),$$

for $n_a$ and $n_b$ large.

Lemma 5.1.3 (Distribution of FPR Difference). An approximate distribution of the difference of FPRs is

$$\widehat{FPR}_a - \widehat{FPR}_b \sim \frac{1}{n_a p_{a,0}} \text{Binomial}(n_a, p_{a,0} p_{\widehat{FPR}_a}) - \frac{1}{n_b p_{b,0}} \text{Binomial}(n_b, p_{b,0} p_{\widehat{FPR}_b}),$$

for $n_a$ and $n_b$ large, with $p_{\widehat{FPR}_a}$ as defined in Lemma 3.3.12 and $p_{\widehat{FPR}_b}$ the analogous quantity for group $b$, and $p_{a,0} := \mathbb{P}(Y_{a,i} = 0), p_{b,0} := \mathbb{P}(Y_{b,i} = 0)$, where $c_a, c_b$ are the decision thresholds, the true skill distributions are $\text{Unif}(s_a, t_a)$ and $\text{Unif}(s_b, t_b)$, and the noise distributions are $\mathcal{N}(0, \sigma_a^2)$ and $\mathcal{N}(0, \sigma_b^2)$, for groups $a$ and $b$ respectively.

Another approximate distribution of the difference of FPRs is

$$\widehat{FPR}_a - \widehat{FPR}_b \sim \mathcal{N}(p_{\widehat{FPR}_a} - p_{\widehat{FPR}_b}, \frac{p_{\widehat{FPR}_a}(1 - p_{a,0} p_{\widehat{FPR}_a})}{n_a p_{a,0}} + \frac{p_{\widehat{FPR}_b}(1 - p_{b,0} p_{\widehat{FPR}_b})}{n_b p_{b,0}}).$$

97
Proof. This follows immediately from independence and the proof of Lemma 3.3.13.

Similar to the single metric case, we confirm that this approximate distribution for the difference of FPRs matches simulated data in below.

![Figure 5.1.1: Simulated FPR distribution vs. proposed approximate Binomial difference distribution, for two groups each with Unif(0,100) skill, \( \mathcal{N}(0,10^2) \) noise, threshold 50, and sample size 100.](image)

**Lemma 5.1.4** (Distribution of FNR Difference). An approximate distribution of the difference of FNRs is

\[
\widehat{\text{FNR}}_a - \widehat{\text{FNR}}_b \sim \frac{1}{n_a p_{a,1}} \text{Binomial} \left( n_a, p_{a,1} p_{\text{FNR}_a} \right) - \frac{1}{n_b p_{b,1}} \text{Binomial} \left( n_b, p_{b,1} p_{\text{FNR}_b} \right),
\]

for \( n_a \) and \( n_b \) large, with \( p_{\text{FNR}_a} \) as defined in Lemma 3.3.15 and \( p_{\text{FNR}_b} \) the analogous quantity for group \( b \), and \( p_{a,1} := \mathbb{P}(Y_{a,i} = 1) \), \( p_{b,1} := \mathbb{P}(Y_{b,i} = 1) \), where \( c_a, c_b \) are the decision thresholds, the true skill distributions are Unif\( (s_a, t_a) \) and Unif\( (s_b, t_b) \).
Unif($s_b, t_b$), and the noise distributions are $\mathcal{N}(0, \sigma^2_a)$ and $\mathcal{N}(0, \sigma^2_b)$, for groups $a$ and $b$ respectively.

Another approximate distribution of the difference of FN Rs is

$$\widehat{FN R}_a - \widehat{FN R}_a \sim \mathcal{N}\left(p_{\widehat{FN R}_a} - p_{\widehat{FN R}_b}, \frac{p_{\widehat{FN R}_a}(1 - p_{\cdot, 1}p_{\widehat{FN R}_a})}{n_a p_{\cdot, 1}} + \frac{p_{\widehat{FN R}_b}(1 - p_{\cdot, 1}p_{\widehat{FN R}_b})}{n_b p_{\cdot, 1}}\right)$$

Again, we confirm for $n_a, n_b$ large that the proposed approximate distribution is reasonable via simulation, in below.

![Theoretical vs. Simulated Distribution of Difference of FNR for Group $a$ with Unif(0,100) skill, $\mathcal{N}(0,10^2)$ noise, threshold 50 and Group $b$ with Unif(0,100) skill, $\mathcal{N}(0,10^2)$ noise, threshold 50](image)

**Figure 5.1.2:** Simulated FPR distribution vs. proposed approximate Binomial difference distribution, for two groups each with Unif(0,100) skill, $\mathcal{N}(0,10^2)$ noise, threshold 50, and sample size 100.

**Lemma 5.1.5** (Distribution of PPV Difference). An approximate distribution of the difference of PPVs is

$$\widehat{PPV}_a - \widehat{PPV}_b \sim \frac{1}{n_a p_{\cdot, 1} p_{\widehat{PPV}_a}} \text{Binomial}\left(n_a, p_{\cdot, 1} p_{\widehat{PPV}_a}\right) - \frac{1}{n_b p_{\cdot, 1} p_{\widehat{PPV}_b}} \text{Binomial}\left(n_b, p_{\cdot, 1} p_{\widehat{PPV}_b}\right),$$

99
for \( n_a \) and \( n_b \) large, with \( p_{F\text{NR}_a} \) as defined in Lemma 3.3.18 and \( p_{F\text{NR}_b} \) the analogous quantity for group \( b \), where \( c_a, c_b \) are the decision thresholds, the true skill distributions are \( \text{Unif}(s_a, t_a) \) and \( \text{Unif}(s_b, t_b) \), and the noise distributions are \( \mathcal{N}(0, \sigma_a^2) \) and \( \mathcal{N}(0, \sigma_b^2) \), for groups \( a \) and \( b \) respectively.

Another approximate distribution of the difference of PPVs is

\[
\hat{\text{PPV}}_a - \hat{\text{PPV}}_b \sim \mathcal{N}
\left(
\left(p_{\hat{\text{PPV}}_a} - p_{\hat{\text{PPV}}_b}\right)
\frac{\left(1 - p_{\text{AR}_a} p_{\hat{\text{PPV}}_b}\right)}{n_a p_{\text{AR}_a}}
\right)
\]

\[
\left(p_{\hat{\text{PPV}}_b}(1 - p_{\text{AR}_b} p_{\hat{\text{PPV}}_b})\right)
\]

We again plot the proposed versus simulated distributions in Figure 5.1.3 below.

![Theoretical vs. Simulated Distribution of Difference of PPV for Group a with Unif(0,100) skill, N(0,10) noise, threshold 50 and Group b with Unif(0,100) skill, N(0,10) noise, threshold 50](image)

**Figure 5.1.3:** Simulated PPV distribution vs. proposed approximate Binomial difference distribution, for two groups each with Unif(0,100) skill, \( \mathcal{N}(0,10^2) \) noise, threshold 50, and sample size 100.

As in Lemma 3.3.19, we note that the variance of the difference of PPV is again an overestimate of the simulated variance, as expected. We maintain
improving this approximate variance as an immediate next step of our work, and
again emphasize the fact that our estimate, which we strongly believe is
consistently an overestimate based on extensive simulations, therefore still is able
to provide (overly conservative) uncertainty quantification.

5.1.2 Uniform Skill, General Noise

Lemma 5.1.6 (Distributions of Accuracy Difference). The exact distribution of
the difference of accuracies for groups a and b, \( \widetilde{ACC}_a - \widetilde{ACC}_b \), in this
thresholding setting is

\[
\widetilde{ACC}_a - \widetilde{ACC}_b \sim \frac{1}{n_a} \text{Bin}(n_a, p_{\widetilde{ACC}_a}) - \frac{1}{n_b} \text{Bin}(n_b, p_{\widetilde{ACC}_b}),
\]

with \( p_{\widetilde{ACC}_a} \) as defined in Lemma 3.4.2 and \( p_{\widetilde{ACC}_b} \) the analogous quantity for
group b, where \( c_a, c_b \) are the decision thresholds, the true skill distributions are
\( \text{Unif}(s_a, t_a) \) and \( \text{Unif}(s_b, t_b) \), and the noise distributions are \( f_{\epsilon_a} \) and \( f_{\epsilon_b} \), for groups
a and b respectively.

An approximate distribution of the difference of accuracies is

\[
\widetilde{ACC}_a - \widetilde{ACC}_b \sim N\left( p_{\widetilde{ACC}_a} - p_{\widetilde{ACC}_b}, \frac{p_{\widetilde{ACC}_a}(1 - p_{\widetilde{ACC}_a})}{n_a} + \frac{p_{\widetilde{ACC}_b}(1 - p_{\widetilde{ACC}_b})}{n_b} \right),
\]

for \( n_a \) and \( n_b \) large.

Lemma 5.1.7 (Distributions of Acceptance Rate Difference). The exact
distribution of the difference of acceptance rates for groups a and b, \( \widetilde{AR}_a - \widetilde{AR}_b \),
in this thresholding setting is

\[
\widetilde{AR}_a - \widetilde{AR}_b \sim \frac{1}{n_a} \text{Bin}(n_a, p_{\widetilde{AR}_a}) - \frac{1}{n_b} \text{Bin}(n_b, p_{\widetilde{ARCC}_b}),
\]

with \( p_{\widetilde{AR}_a} \) as defined in Lemma 3.4.6 and \( p_{\widetilde{AR}_b} \) the analogous quantity for group b,
where \( c_a, c_b \) are the decision thresholds, the true skill distributions are \( \text{Unif}(s_a, t_a) \)
and \( \text{Unif}(s_b, t_b) \), and the noise distributions are \( f_{e_a} \) and \( f_{e_b} \), for groups a and b.
respectively.

An approximate distribution of the difference of acceptance rates is

$$\tilde{AR}_a - \tilde{AR}_b \sim N \left( p_{\tilde{AR}_a} - p_{\tilde{AR}_b}, \frac{p_{\tilde{AR}_a} (1 - p_{\tilde{AR}_a})}{n_a} + \frac{p_{\tilde{AR}_b} (1 - p_{\tilde{AR}_b})}{n_b} \right),$$

for $n_a$ and $n_b$ large.

**Lemma 5.1.8** (Distribution of FPR Difference). An approximate distribution of the difference of FPRs is

$$\tilde{FPR}_a - \tilde{FPR}_b \sim \frac{1}{n_a p_{a,0}} \text{Binomial} \left( n_a, p_{a,0} \tilde{FPR}_a \right) - \frac{1}{n_b p_{b,0}} \text{Binomial} \left( n_b, p_{b,0} \tilde{FPR}_b \right),$$

for $n_a$ and $n_b$ large, with $p_{\tilde{FPR}_a}$ as defined in Lemma 3.4.10 and $p_{\tilde{FPR}_b}$ the analogous quantity for group $b$, and $p_{a,0} := \mathbb{P}(Y_{a,i} = 0)$, $p_{b,0} := \mathbb{P}(Y_{b,i} = 0)$, where $c_a, c_b$ are the decision thresholds, the true skill distributions are $\text{Unif}(s_a, t_a)$ and $\text{Unif}(s_b, t_b)$, and the noise distributions are $f_{\epsilon_a}$ and $f_{\epsilon_b}$, for groups $a$ and $b$ respectively.

Another approximate distribution of the difference of FPRs is

$$\tilde{FPR}_a - \tilde{FPR}_a \sim N \left( p_{\tilde{FPR}_a} - p_{\tilde{FPR}_b}, \frac{p_{\tilde{FPR}_a} (1 - p_{\tilde{FPR}_a})}{n_a p_{a,0}} + \frac{p_{\tilde{FPR}_b} (1 - p_{\tilde{FPR}_b})}{n_b p_{b,0}} \right).$$

**Lemma 5.1.9** (Distribution of FNR Difference). An approximate distribution of the difference of FNRS is

$$\tilde{FNR}_a - \tilde{FNR}_b \sim \frac{1}{n_a p_{a,1}} \text{Binomial} \left( n_a, p_{a,1} \tilde{FNR}_a \right) - \frac{1}{n_b p_{b,1}} \text{Binomial} \left( n_b, p_{b,1} \tilde{FNR}_b \right),$$

for $n_a$ and $n_b$ large, with $p_{\tilde{FNR}_a}$ as defined in Lemma 3.4.12 and $p_{\tilde{FNR}_b}$ the analogous quantity for group $b$, and $p_{a,1} := \mathbb{P}(Y_{a,i} = 1)$, $p_{b,1} := \mathbb{P}(Y_{b,i} = 1)$, where $c_a, c_b$ are the decision thresholds, the true skill distributions are $\text{Unif}(s_a, t_a)$ and $\text{Unif}(s_b, t_b)$, and the noise distributions are $f_{\epsilon_a}$ and $f_{\epsilon_b}$, for groups $a$ and $b$ respectively.
Another approximate distribution of the difference of FNRs is

\[
\widetilde{F_{\text{NR}a}} - \widetilde{F_{\text{NR}b}} \sim \mathcal{N}

\left(p_{\widetilde{F_{\text{NR}a}}} - p_{\widetilde{F_{\text{NR}b}}}, \frac{p_{\widetilde{F_{\text{NR}a}}}(1 - p_{\widetilde{F_{\text{NR}a}}})}{n_a p_a, 1} + \frac{p_{\widetilde{F_{\text{NR}b}}}(1 - p_{\widetilde{F_{\text{NR}b}}})}{n_b p_b, 1}\right)
\]

**Lemma 5.1.10** (Distribution of PPV Difference). An approximate distribution of the difference of PPVs is

\[
\widetilde{P_{\text{PPV}a}} - \widetilde{P_{\text{PPV}b}} \sim \frac{1}{n_a p_{\text{AR}a}} \text{Binomial}(n_a, p_{\text{AR}a} p_{\text{PPV}a}) - \frac{1}{n_b p_{\text{AR}b}} \text{Binomial}(n_b, p_{\text{AR}b} p_{\text{PPV}b}),
\]

for \(n_a\) and \(n_b\) large, with \(p_{\widetilde{F_{\text{NR}a}}}\) as defined in Lemma 3.4.14 and \(p_{\widetilde{F_{\text{NR}b}}}\) the analogous quantity for group b, where \(c_a, c_b\) are the decision thresholds, the true skill distributions are Unif\((s_a, t_a)\) and Unif\((s_b, t_b)\), and the noise distributions are \(f_{\epsilon_a}\) and \(f_{\epsilon_b}\), for groups a and b respectively.

Another approximate distribution of the difference of FNRs is

\[
\widetilde{P_{\text{PPV}a}} - \widetilde{P_{\text{PPV}b}} \sim \mathcal{N}

\left(p_{\widetilde{P_{\text{PPV}a}}} - p_{\widetilde{P_{\text{PPV}b}}}, \frac{p_{\widetilde{P_{\text{PPV}a}}}(1 - p_{\text{AR}a} p_{\text{PPV}a})}{n_a p_{\text{AR}a}} + \frac{p_{\widetilde{P_{\text{PPV}b}}}(1 - p_{\text{AR}b} p_{\text{PPV}b})}{n_b p_{\text{AR}b}}\right)
\]

### 5.2 Empirical Trends of Difference Distributions

Equipped with these distributions, we first take the time to better understand the basic behavior of some of these distributions, especially how they vary across different dimensions of parameters (sample size, threshold, noise, etc.). We note in particular that we focus on the zero-centered Gaussian noise case for these results, and again use *exact* variances for accuracy and acceptance rate and *approximate* variances for FPR, FNR, and PPV.

We first visualize the shapes of some of these distributions in Figure 5.2.1 below. In particular, we observe the differing degradation of metric means across different shared noise standard deviations \(\sigma\) for two different sets of underlying distributions (identical between the two groups in the top set of graphs and differing in the bottom set of graphs).
Figure 5.2.1: Probability density functions for the distribution of the difference of the five fairness metrics, across different noise standard deviations $\sigma$
5.2.1 Difference Variation with Threshold \( c \)

Figure 5.2.2 below shows the plots of the means of the difference of each of the metrics as the threshold \( c \) changes, for two groups with different underlying skill distributions. For most of the metrics, the change in value is relatively small over the range of thresholds, except for PPV, which degrades quickly.

![Figure 5.2.2: Mean and error bars (two standard deviations in width) for the distribution of the difference of each of the five fairness metrics, across all integer thresholds \( c \in [41, 59] \), with underlying Unif(0, 60) skill for Group a and underlying Unif(40, 100) skill for Group b, and \( \mathcal{N}(0, 10^2) \) noise and sample size 100 for both groups.](image)

5.2.2 Difference Variation with Sample Size \( n \)

We also look at how the distributions vary with the sample size \( n \) in Figure 5.2.3 below.
Figure 5.2.3: Mean and error bars (two standard deviations in width) for the distribution of the difference of each of the five fairness metrics, across different shared sample sizes $n \in [50, 500]$, with underlying $\text{Unif}(0, 60)$ skill, $\mathcal{N}(0, 10^2)$ noise, and threshold 50 for Group $a$ and underlying $\text{Unif}(40, 100)$ skill, $\mathcal{N}(0, 10^2)$ noise, and threshold 50 for Group $b$. The 95% CI for distribution of difference of metrics across different shared $n$ for Group $a$ with $\text{Unif}(0, 60)$ skill, $\mathcal{N}(0, 10^2)$ noise, and $c_1 = 50$, Group $b$ with $\text{Unif}(40, 100)$ skill, $\mathcal{N}(0, 10^2)$ noise, $c_2 = 50$. 
Again, we notice that all of the standard deviations decrease as \( n \) increases, as expected, and the exact standard deviations are much smaller in magnitude than the approximate standard deviations.

5.2.3 Difference Variation with Noise Variance \( \sigma^2 \)

Figure 5.2.4 below shows how the distributions vary with the noise standard deviation \( \sigma \), for a few different underlying skill distribution combinations. In particular, we note that as \( \sigma \) increases, the expected difference in acceptance rates stays the same or decreases, but increases for FPR, FNR, and PPV.

**Figure 5.2.4:** Mean and error bars (two standard deviations in width) for the distribution of the difference of each of the five fairness metrics, across different noise standard deviations \( \sigma \in [1, 25] \) and underlying skill distributions, with threshold 50 and sample size 100 for both groups.
5.2.4 Difference Variation with Underlying Distributions

Lastly, Figure 5.2.5 visualizes the evolution of the metrics as a function of the difference between true skill distribution centers for the two groups, as the true skill distribution of Group \( a \) increases from Uni(0,60) to Uni(40, 100), and the opposite direction for Group \( b \).

\[ \text{Figure 5.2.5: Mean and error bars (two standard deviations in width) for the distribution of the difference of each of the five fairness metrics, across different distances in the centers of the underlying skill distributions between the two groups, with } \mathcal{N}(0, 10^2) \text{ noise, threshold 50 and sample size 100 for both groups.} \]

We observe that all of the quantities seem to be symmetric about zero in terms of their relative magnitudes, which makes sense since the groups should be symmetric in terms of their set-up here. We also notice that for FPR, FNR, and
PPV, the variance estimate is smaller when the distance between distribution centers is smaller.

Perhaps most importantly, however, we notice that for acceptance rate, FPR, FNR, and PPV, the magnitude of the difference in means for any of those metrics grows monotonically with the distance between the underlying skill distributions of the two populations. Combined with the math, these results confirm that in a thresholding setting, one should expect a discrepancy, on average, in any of these metrics between groups.

Given that there exists discrepancy on average, a natural follow-up question is what range of discrepancies are within some reasonable or probable range for a given set of populations. We aim to formally answer this in the following section.

5.3 Constructing Confidence Intervals to Quantify Unfairness

Using these difference distributions, we now propose a set of confidence intervals which capture an expected set of values for the difference in each of our five fairness metrics of interest.

5.3.1 Approximate 95% Confidence Intervals

Theorem 5.3.1 (Confidence Intervals for Difference of Distributions).

Approximate 95% confidence intervals for the difference of each of the five metrics are as follows:

Accuracy difference $\hat{\text{ACC}}_a - \hat{\text{ACC}}_b$:

$$p_{\hat{\text{ACC}}_a} - p_{\hat{\text{ACC}}_b} \pm 1.96 \cdot \sqrt{\frac{p_{\hat{\text{ACC}}_a}(1-p_{\hat{\text{ACC}}_a})}{n_a} + \frac{p_{\hat{\text{ACC}}_b}(1-p_{\hat{\text{ACC}}_b})}{n_b}}$$
Acceptance rate difference $\widehat{AR}_a - \widehat{AR}_b$:

$$p_{\widehat{AR}_a} - p_{\widehat{AR}_b} \pm 1.96 \sqrt{\frac{p_{\widehat{AR}_a}(1 - p_{\widehat{AR}_a})}{n_a} + \frac{p_{\widehat{AR}_b}(1 - p_{\widehat{AR}_b})}{n_b}}$$

FPR Difference $\widehat{FPR}_a - \widehat{FPR}_b$:

$$p_{\widehat{FPR}_a} - p_{\widehat{FPR}_b} \pm 1.96 \sqrt{\frac{p_{\widehat{FPR}_a}(1 - p_{\widehat{FPR}_a})}{n_ap_{a,0}} + \frac{p_{\widehat{FPR}_b}(1 - p_{\widehat{FPR}_b})}{n_bp_{b,0}}}$$

FNR Difference $\widehat{FNR}_a - \widehat{FNR}_b$:

$$p_{\widehat{FNR}_a} - p_{\widehat{FNR}_b} \pm 1.96 \sqrt{\frac{p_{\widehat{FNR}_a}(1 - p_{\widehat{FNR}_a})}{n_ap_{a,1}} + \frac{p_{\widehat{FNR}_b}(1 - p_{\widehat{FNR}_b})}{n_bp_{b,1}}}$$

PPV Difference $\widehat{PPV}_a - \widehat{PPV}_b$:

$$p_{\widehat{PPV}_a} - p_{\widehat{PPV}_b} \pm 1.96 \sqrt{\frac{p_{\widehat{PPV}_a}(1 - p_{\widehat{AR}_a}p_{\widehat{PPV}_a})}{n_ap_{\widehat{AR}_a}} + \frac{p_{\widehat{PPV}_b}(1 - p_{\widehat{AR}_b}p_{\widehat{PPV}_b})}{n_bp_{\widehat{AR}_b}}}$$

Proof. The approximate 95% coverage of all of these intervals follows directly from the Normal approximation to the Binomial results from the lemmas in Section 5.1.1, using the fact that the 0.025 and 0.975 standard Gaussian quantiles are −1.96 and 1.96, respectively.

These confidence intervals are a novel derivation and can be directly used for identifying unfairness, as discussed in Section 2.5.2, for a given thresholding setting in a statistically rigorous manner. We provide a straightforward example below.

**Example 5.3.1.** Consider the case where Group a has underlying distribution Uniform(0,60), Group b has underlying distribution Uniform (40,100), and both groups have $\mathcal{N}(0,10^2)$ noise, sample size 100, and threshold 50. Then, using
Theorem 5.3.1 above, 95% approximate confidence intervals for the difference in each of the five metrics between the two groups are as follows:

- Accuracy difference: [-0.0898 0.0898]
- Acceptance rate difference: [-0.7455 -0.5323]
- FPR difference: [-0.5048 0.0332]
- FNR difference: [-0.0332 0.5048]
- PPV difference: [-0.6637 0.0555]

Consider in particular the confidence interval for the difference in acceptance rates between Groups a and b, [-0.7455 -0.5323]. We note that this interval is relatively tight, and most importantly, does not capture the value 0, which represents no difference in acceptance rates between Groups a and b. Here, this then says that equal acceptance rates (i.e., demographic parity) are not within the likely set of values achieved, and thus it is not reasonable to expect a model with such parameters to be able to achieve equal acceptance rates.

Lastly, say a team claims their model follows the above parameters and achieves a discrepancy in acceptance rates of -0.2. Then, purely in terms of distributional results, this confidence interval (or hypothesis test derived from the interval) would reject -0.2 as a reasonable value for the difference in acceptance rates. In some sense, we can argue that their model is unfair with respect to acceptance rates, where “unfairness” here is characterized by their model not falling within an expected set given the model parameters.

The above example demonstrates the utility of these confidence intervals in rigorously identifying a range of values of high confidence, indicating whether equalized values of a metric across protected groups is reasonably achievable or not, and serving as an evaluation framework for determining whether a model is operating fairly or as expected with respect to a given fairness metric.
Next, we confirm that for data from the corresponding underlying skill and noise distributions, these intervals empirically achieve roughly 95% coverage, in order to justify the validity of these proposed intervals.

5.3.2 Coverage Probability with General Noise

Here, we empirically explore the coverage probability of our intervals generated using 5.3.1. In particular, in addition to exploring the coverage probability of our intervals for new data generated from those same sets of distributions and parameters, we also explore the robustness of our results by checking the coverage probability of our Gaussian noise intervals on new data with non-Gaussian noise.

Note that a simple distribution fact which we implicitly use going forth is that a Laplace random variable with location \( \mu \) and scale \( b \), its mean and variance are \( \mu \) and \( 2b^2 \) respectively, so for a \( N(0, \sigma^2) \) random variable, the Laplace distribution with equivalent mean and variance is Laplace(0, \( \sigma / \sqrt{2} \)). Similarly, for a Unif(\(-a, a\)) random variable, the mean is 0 and variance is \( a^2 / 3 \), so for a \( N(0, \sigma^2) \) random variable, the Uniform distribution with equivalent mean and variance is Unif(\(-\sigma \sqrt{3}, \sigma \sqrt{3}\))

For the data generated with the same noise as the intervals (\( N(0, 10^2) \)), we notice that the coverage probabilities are all very close to 95%, except that of PPV, which is 100%. This confirms the approximate 95% coverage of our intervals, except for PPV, where we hypothesize from other empirical observations that our proposed approximate variance is an overestimate of the true variance, and thus our interval is wider than it needs to be to achieve 95% coverage in the case of PPV.

When we change the noise to Laplace noise or Uniform noise, the coverage probabilities of the Gaussian noise intervals are all very close to 95% still; this suggests that our Gaussian intervals may actually be quite robust to the actual distribution underlying noise within our dataset (as long as the mean and variance are properly calibrated), which is a highly desirable outcome.

Thus, these intervals achieve very close to 95% coverage, are robust to the
<table>
<thead>
<tr>
<th>Noise</th>
<th>Metric</th>
<th>Gauss CI Start</th>
<th>Gauss CI End</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{N}(0, 10^2)$</td>
<td>ACC</td>
<td>-0.0751</td>
<td>0.0751</td>
<td>94.88%</td>
</tr>
<tr>
<td>$\mathcal{N}(0, 10^2)$</td>
<td>AR</td>
<td>-0.1386</td>
<td>0.1386</td>
<td>94.21%</td>
</tr>
<tr>
<td>$\mathcal{N}(0, 10^2)$</td>
<td>FPR</td>
<td>-0.1085</td>
<td>0.1085</td>
<td>95.28%</td>
</tr>
<tr>
<td>$\mathcal{N}(0, 10^2)$</td>
<td>FNR</td>
<td>-0.1085</td>
<td>0.1085</td>
<td>95.03%</td>
</tr>
<tr>
<td>$\mathcal{N}(0, 10^2)$</td>
<td>PPV</td>
<td>-0.2763</td>
<td>0.2763</td>
<td>100.0%</td>
</tr>
<tr>
<td>$\text{Lap}(0, 10/\sqrt{2})$</td>
<td>ACC</td>
<td>-0.0751</td>
<td>0.0751</td>
<td>93.84%</td>
</tr>
<tr>
<td>$\text{Lap}(0, 10/\sqrt{2})$</td>
<td>AR</td>
<td>-0.1386</td>
<td>0.1386</td>
<td>94.79%</td>
</tr>
<tr>
<td>$\text{Lap}(0, 10/\sqrt{2})$</td>
<td>FPR</td>
<td>-0.1085</td>
<td>0.1085</td>
<td>97.63%</td>
</tr>
<tr>
<td>$\text{Lap}(0, 10/\sqrt{2})$</td>
<td>FNR</td>
<td>-0.1085</td>
<td>0.1085</td>
<td>97.61%</td>
</tr>
<tr>
<td>$\text{Lap}(0, 10/\sqrt{2})$</td>
<td>PPV</td>
<td>-0.2763</td>
<td>0.2763</td>
<td>99.77%</td>
</tr>
<tr>
<td>$\text{Unif}(-10\sqrt{3}, 10\sqrt{3})$</td>
<td>ACC</td>
<td>-0.0751</td>
<td>0.0751</td>
<td>92.47%</td>
</tr>
<tr>
<td>$\text{Unif}(-10\sqrt{3}, 10\sqrt{3})$</td>
<td>AR</td>
<td>-0.1386</td>
<td>0.1386</td>
<td>94.76%</td>
</tr>
<tr>
<td>$\text{Unif}(-10\sqrt{3}, 10\sqrt{3})$</td>
<td>FPR</td>
<td>-0.1085</td>
<td>0.1085</td>
<td>96.07%</td>
</tr>
<tr>
<td>$\text{Unif}(-10\sqrt{3}, 10\sqrt{3})$</td>
<td>FNR</td>
<td>-0.1085</td>
<td>0.1085</td>
<td>95.92%</td>
</tr>
<tr>
<td>$\text{Unif}(-10\sqrt{3}, 10\sqrt{3})$</td>
<td>PPV</td>
<td>-0.2763</td>
<td>0.2763</td>
<td>99.54%</td>
</tr>
</tbody>
</table>

**Figure 5.3.1**: Coverage probability of approximate 95% confidence intervals for the difference of metrics with different , for intervals generated for Groups 1 and 2 both with $\text{Unif}(0,100)$ underlying skill distribution, $\mathcal{N}(0, 10^2)$ noise, threshold 50, and sample size 100.
noise distribution, and can be used to identify unfairness with respect to a certain metric by quantifying the set of reasonable values.

5.4 Computing Fair Optimal Thresholds

The final use case of our difference distributions we explore in this thesis is deriving thresholds that achieve, in expectation, a desired metric value, using our closed-form expected value expressions for each of the five metrics.

Specifically, because we have exact expressions for the expected value of the accuracy, acceptance rate, FPR, FNR, and PPV in a given thresholding setting, then if a policymaker or scientist has a desired metric value they would like their model to achieve in expectation (e.g., expected acceptance rate of 20% across all groups), we can directly solve for the roots of an existing equation to derive the thresholds that would guarantee this expected metric value. We give a simple example to demonstrate this idea.

Example 5.4.1. Consider the case again where Group $a$ has underlying distribution Uniform(0,60), Group $b$ has underlying distribution Uniform (40,100), and both groups have $\mathcal{N}(0,10^2)$ noise, sample size 100, and threshold 50.

Say a policymaker wishes for this model to have expected 20% acceptance rate for both groups. Then, given our existing expressions for the expected acceptance rate, $p_{AR}$, from Lemma 3.3.7, we can simply solve the equations

\[
p_{AR_a} = 0.2
\]
\[
p_{AR_b} = 0.2
\]

to derive the thresholds $c_a$ and $c_b$ that guarantee expected acceptance rates of 20% for both groups. Here, we find

\[
c_a = 48.64, \quad c_b = 88.64,
\]
which we empirically confirm indeed give expected acceptance rates for both groups of 20% using the equation from Lemma 3.3.7.

We first note that solving for such thresholds does not require anything beyond what we have already derived in this thesis, so long as a sufficiently accurate root solver exists, such as \texttt{scipy.optimize.root}, which is what we use in the example above; a root solver is needed here because of the challenges of solving for \( c \) in closed form starting from any of the expected value expressions. Thus, this can be almost directly applied and utilized based only on the theoretical results from Chapter 3 of this thesis.

This is a very powerful and useful application of our theoretical results in practice, as it can be used to correctly calibrate a model via the model’s threshold to satisfy the model creator’s intended output (in expectation), with respect to a certain fairness metric, as long as there exist reasonable estimates on the underlying skill distribution parameters and noise standard deviation.

There are a number of situations in which providing such thresholds would be of significant utility. For example, in our initial running example from Chapter 2, a company hiring employees may want to set acceptance rates equal to a certain value across all racial groups, and derive group-specific thresholds such that these desired acceptance rates are achieved in expectation. There also exist a multitude of more involved constrained optimization scenarios, such as maximizing the expected accuracy under the condition that the group-specific error rates are below a certain chosen value, which can also be directly solved for using our theoretical expressions from Chapter 3. While we only briefly skim the surface of methodologies for this particular application, all of the necessary quantities have already been derived in this thesis and can be used directly to compute the desired thresholds as in Example 5.4.1 above; we leave the application-specific framework formalization as a future step for our work.
We conclude this thesis with a brief discussion on our work, and future extensions.

6.1 Our Contributions

Most of the existing work on algorithmic fairness focuses on formally characterizing different ethical and philosophical dimensions of fairness mathematically via a wide range of definitions, or proposing interventions in the pre-processing, in-processing, and post-processing stages that correct for model unfairness and systemic injustice. These have largely been contributed by computer scientists, and while the field of algorithmic fairness is inherently grounded in probabilistic methods, very few statistical tools have been proposed to rigorously evaluate the extent of unfairness within any subfield of algorithmic
In this thesis, we thus formalized the distributions of five of the most fundamental binary classification metrics of interest in group fairness, in the setting of threshold-based decision rules where individual’s true skill levels follow a Uniform distribution with general continuous noise. Using these distributions, we first studied some of their basic properties and trends along different dimensions of the distributional parameters, including underlying skill level, amount of noise, sample size, and threshold cutoff, identifying a few key trends where certain fairness metrics are, in expectation, disproportionately consequential for certain groups over others.

We subsequently derived the distributions of the difference of these metrics between protected groups, as group fairness definitions tend to seek exact equality (or zero discrepancy) of these metrics across protected groups, which is a very strong and unlikely condition to achieve in real-world settings. Thus, after studying similar properties and trends of these difference distributions as they varied across different dimensions of parameters, we aimed to use these distributions to propose a new framework for fairness evaluation.

By using pivotal quantities derived directly from these difference distributions assuming Gaussian noise, we constructed approximate 95% confidence intervals to quantify, in any given thresholding setting, a reasonable range of values for a given fairness metric difference between groups; in testing, although approximate, these intervals achieved coverage probability within a few tenths of a percent from 95%. We also find that in practice, these confidence intervals are very robust to the underlying noise distribution (as long as the mean and variance are properly calibrated), achieving close to 95% coverage probability empirically, and thus making them generalizable tools for fairness evaluation.

Finally, we also demonstrated how these difference distributions can be used to derive the threshold that should be used by the model to achieve a certain desired (expected) fairness metric value, which is a very powerful and desirable tool for real-world modeling.
6.2 Future Directions

While these distributional characterizations of group fairness metrics are relatively novel and understudied, there are a number of valuable extensions for future research that build on top of these analyses.

First, we would like to continue deriving exact distributions for FPR, FNR, and PPV, to then construct exact 95% intervals for all metrics. This is especially true for PPV, whose approximate variance is currently an overestimate of the true variance. While FPR, FNR, and PPV appear to have relatively complicated distributions which do not follow any “nice” distribution due to their ratio structure and highly correlated and random nature of both the numerator and denominator, more formally quantifying their density functions will allow for the construction of more precise intervals, and subsequently increased sophistication of fairness evaluation in this thresholding setting.

Next, we hope to find methods to relax the constraint that the underlying skill distribution is Uniform. Although we present compelling reasoning for the generalizability of the Uniform distribution by noting that general skill values can be converted to skill percentiles which follow a Uniform distribution by the well-known probability integral transform, there are certainly settings where the raw skill is a more desirable choice of feature, and thus deriving similar results under general skill level distributions would be insightful. We would also like to investigate this for estimated distributions, noting that it may be highly unrealistic to perfectly know the true skill and noise distributions in real-world settings, and also explore methods to estimate the true skill and noise distributions from noisy skill observations alone.

Third, we would like to develop more formal hypothesis testing and thresholding frameworks that take into account the above relaxations, in the form of an open source Python library, to introduce these tools into general algorithmic fairness evaluation procedures.

Last, but certainly not least, we would like to explore a variety of more complicated settings with respect to this problem, for example settings with
non-zero-centered bias, or a large number of protected groups.

Ultimately, countless questions remain unanswered, both in this thesis and the broader field of algorithmic fairness as a whole. Yet, within these pages, we hope to have uncovered some novel glimpses of insight and contributed valuable dimensions to the discourse on algorithmic fairness from a statistical standpoint. We certainly all long for a distant world where fairness is not merely an aspiration but an undeniable reality for all; until that day dawns, however, let us continue to navigate these multidisciplinary terrains, steadfastly persevering in our pursuit of a future where every algorithmic decision radianty reflects virtues of equity and justice.
References


122


Jack Kelly. Goldman Sachs predicts 300 million jobs will be lost or degraded by artificial intelligence. Forbes, 5 2023.


Kristina Lerman and Tad Hogg. Leveraging position bias to improve peer recommendation. PLoS ONE, 9, 6 2014.


