My amazing title

Dasha Asienga

April DD, 20YY

Submitted to the Department of
Mathematics and Statistics
of Amherst College in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with honors.

Advisors:
Professor Katharine Correia
Your Other Advisor

Abstract

The abstract should be a short summary of your thesis work. A paragraph is usually sufficient here.

Acknowledgments

Use this space to thank those who have helped you in the thesis process (professors, staff, friends, family, etc.). If you had special funding to conduct your thesis work, that should be acknowledged here as well.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

The public and private sector are increasingly turning to data-driven methods to automate and to guide simple and complex decision-making. However, this trend raises an important question of bias. There is a lot of misinterpretation when it comes to the collection of data in many application areas, and there is a major concern for data-driven methods to further introduce and perpetuate discriminatory practices, or to otherwise be unfair because of the social and historical processes that operate to the disadvantage of certain groups.

For example, within healthcare, using mortality or readmission rates to measure hospital performance penalizes hospitals serving poor or non-White populations as those inherently have higher mortality and readmission rates due to confounding societal factors. Outside healthcare, credit-scoring algorithms predict outcomes based on income, which disadvantages low-income groups further perpetuating economic immobility. Policing algorithms result in increased scrutiny of Black neighborhoods because of the bias against Black people that is already present in the U.S. policing system, and hiring algorithms, which predict employment decisions, are affected by historical race and gender biases.

Yet, these algorithms are regarded as ground truth and free of human limitations because they are based on mathematics, statistics, and computer science – otherwise regarded as objective disciplines. In theory, this should lead to greater fairness. However, left unregulated, these mathematical models privilege majority groups and discriminate against minority groups because they often learn from inherently biased data. If the data used to train models contains bias, then the resulting algorithms will learn the bias and reflect it into their predictions. In some cases, this can be detrimental.

While there are widely-accepted, though sometimes disputed, societal notions of fairness, one key question emerges: are there any established statistical notions of fairness and bias? Is it possible to mathematically and statistically define algorithmic bias and unfairness, thereby paving a way for addressing the challenges they pose? This thesis paper aims to explore and answer precisely this question.

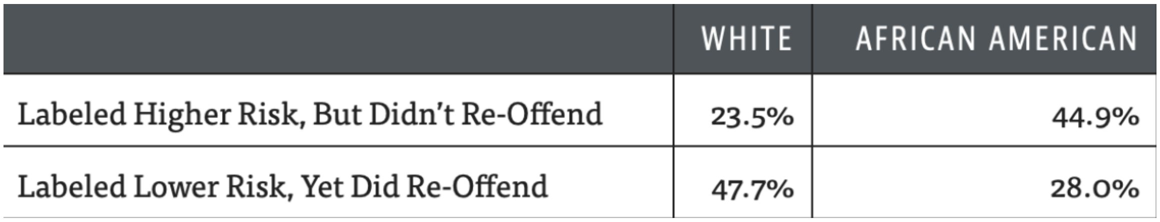
1.1 Algorithmic Bias

There are multiple different types and sources of bias in the realm of statistics. In particular, algorithmic bias arises when an algorithm's decisions are skewed towards a particular group of people, either positively or negatively (Mehrabi, 2021). The danger with biased algorithmic outcomes is that they can generate a feedback loop. Take, for example, a hiring algorithm that discriminates against female applicants for a specific job. In the long run, this can perpetuate, and even amplify, existing gender biases by further widening the gender-based class imbalance.

One such key example of algorithmic bias often cited in literature is regarding the broad use of the COMPAS – or the Correctional Offender Management Profiling for Alternative Sanctions – tool to predict a defendant's risk of recidivism – committing another crime – within two years. COMPAS is more likely to have higher false positive rates for African-American offenders than Caucasian offenders (Mehrabi, 2021). Across the country, scores of similar assessments are given to judges, which injects bias into courts (Angwin, 2016).

COMPAS is based on data from 7000 people arrested in Broward County, Florida in 2013 and 2014 (Angwin, 2016). The response variable, recidivism, was encoded based on who was charged with new crimes over the next two years. Analyses on the predictive efficacy of the COMPAS algorithm found that the algorithm was 61% accurate for a full range of crimes, including misdemeanors, and only 20% of people forecasted to commit violent crimes actually went on to do so. While this is better than a coin flip, there exists scope for enhancing the predictive performance, especially for a decision as critical as whether or not to grant a defendant bail or parole.

What's more concerning, however, is that when the effects of race, age, and gender are isolated, a statistical analysis showed that Black defendants were still 77% more likely to be predicted at higher risk of committing a future violent crime and 45% more likely to be predicted of committing a future crime of any kind, highlighting the role that proxies of race play into the predictions (Angwin, 2016). The table below highlights the performance discrepancy across race:



	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure 1.1: Prediction Fails Differently for Black v White Defendants

Although the tool had 61% accuracy, Black defendants are almost twice as likely to be labeled as higher risk without re-offending than White defendants. It makes the opposite mistake among White defendants. The reason for this is that classification models are trained to minimize average error, which fits majority populations (Chouldechova, 2020). For example, different factors lead to different SAT scores between majority and minority populations. It would, thus, seem fair that the relationship between SAT and college admissions be calibrated differently for the each demographic group. However, if a group-blind classifier is trained, it cannot simultaneously fit both groups optimally and will fit the majority population as it's more important to overall error.

COMPAS is just one such algorithm. In order to set fairness constraints that reduce, or even correct for, algorithmic bias, one must first define fairness mathematically, statistically, and quantifiably.

1.2 Statistical Definitions of Fairness

Statistical notions of fairness can be defined at a group level or an individual level. Group notions fix a few demographic groups and assess the parity of some statistical measures across all the groups (Chouldechova, 2020). Note that group measures, on their own, do not guarantee fairness to individuals or structured subgroups within protected demographic groups, but rather, give guarantees to "average" numbers of protected groups. These notions are the focus of this thesis paper.

Individual notions, on the other hand, are assessed on specific pairs of individuals rather than averaged across groups (Chouldechova, 2020). In other words, similar individuals should be treated similarly along some defined similarity or inverse distance metrics. Counter-factual fairness, for example, relies on the intuition that a decision is fair towards an individual if it's the same in both the real world and a counter-factual world where the individual belongs to a different demographic group (Mehrabi, 2020). This can be impractical, rely on strong assumptions about the data, and approach the realm of causality (Chouldechova, 2020). Moreover, there is a gap in literature with regard to individual notions of fairness.

Ultimately, group notions and individual notions are not in conflict per se. Instead, they are on the same spectrum of how much dependence are allowed between predictions and the sensitive attribute (Castelnovo, 2022). Subgroup fairness is an alternative notions that intends to obtain the best properties of both, for example, by picking a group fairness constraint and assessing whether it holds over a large collection of subgroups (Mehrabi, 2021). Group and individual fairness notions can be defined in both classification settings and regression settings, although most of the literature focuses on fairness within classification.

1.2.1 Group Fairness in Regression Settings

Fair regression is the quantitative notion of fairness of real-valued targets (Agarwal, 2019). Statistical parity refers to minimizing the expected loss function and mean squared errors (MSE) such that the probability that each predicted $\hat{Y}$ is above a certain threshold for each sensitive attribute is the same as the probability over the entire data set, given some margin:

$$\min_{f \in F} E[l(Y, f(X))] \text{ such that } \forall a \in A, z \in [0,1]:$$

$$|P[f(X) \geq z | A = a] - P[f(X) \geq z] \leq \epsilon_a|.$$

This is akin to the classification setting where it may be desirable to have the probability of being in the positive class is the same for each group as across the entire data set. A similar notion, known as bounded loss, requires that the MSE for each group is below some pre-specified level (Agarwal, 2019):

$$\min_{f \in F} E[l(Y, f(X))] \text{ such that } \forall a \in A:$$

$$E[l(Y, f(X)) | A = a] \leq c_a.$$

1.2.2 Group Fairness in Classification Settings

Group notions of fairness, at the core, refer to treating different groups equally. They aim to remedy or prevent disparate impact, which is a setting where there is unintended disproportionate adverse impact on a particular group (Chouldechova, 2016). There are three broad notions of observational group fairness: independence, separation, and sufficiency (Castelnovo, 2022).

Independence

Separation

Sufficiency