UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TESTING FAIRNESS IN INSURANCE

DISSERTATION
PRESENTED
AS A PARTIAL REQUIREMENT
OF THE MASTER'S DEGREE IN MATHEMATICS

BY
RAWANDA MATAR

MARCH 2022

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TESTER L'ÉQUITÉ EN ASSURANCE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAITRISE EN MATHÉMATIQUES

PAR
RAWANDA MATAR

MARS 2022

ACKNOWLEDGEMENTS

Before I start my report, I would like to express my sincere appreciation for a group of people, without them, I wouldn't have been able to finish my journey. First of all, a special thanks to Université du Québec à Montréal (UQAM) and all the staff and proffessor.

I am highly grateful to professor Arthur Charpentier for giving me this opportunity and for providing me with all the resources and for his dedication with me despite his busy schedule, to finish my thesis. I must confess that I am very grateful for the selfless help that he gave me.

And certainly, I would like to thank professor Emiliano A. Valdez for his time and dedication with me.

Also, I would like to thank Mitac's also for there help.

I deeply thank the correctors on my thesis, who gave time and effort to read and evaluate my thesis professor Emiliano A. Valdez and professor Yang Lu.

In the end, my thanks and appreciation go to my parents and my friends, for their non-stop moral support.

DEDICACE

This work is dedicated to each person who has been a victim of discrimination.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i

DEDICACE ii

LIST OF FIGURES vii

LIST OF TABLES viii

RÉSUMÉ ix

ABSTRACT xi

INTRODUCTION 1

Chapter 1 **MOTIVATION** 2

1.1 Legal context 2

1.2 Bias is in algorithms 3

1.2.1 Evidence of biased algorithms 4

1.2.2 Evidence of unfairness in insurance data . . . 4

1.3 Types and causes of discrimination 7

1.3.1 Types of discrimination 7

1.3.2 Causes of machine learning bias 7

Chapter 2 **NOTIONS OF FAIRNESS** 9

2.1 Background 9

2.1.1 Definition 9

2.1.2 Mathematical Notation 11

2.1.3 The Notions 12

2.2 Testing Fairness Notion on the German Credit Dataset 19

2.2.1 Fairness through unawareness 19

2.2.2 Group fairness 20

2.2.3 Predictive parity 22

2.2.4 Calibration 26

Chapter 3 **THE NOTION OF CAUSALITY DISCRIMINATION** 28

3.1 Causal graphs 28

3.2 Connecting causal fairness to observational models 30

3.2.1 Strategy used to translate from causal model to an observational model 31

3.3 d-separation tests 31

3.3.1 Testing the German Data set. 32

3.4 Structural equation model 34

3.4.1 Testing path models using maximum likelihood. 34

3.4.2 Modeling non-normally distributed variables . 37

3.4.3 Simulation 38

3.5 Using Path specific effect to quantify causality . . . 40

3.5.1 Causal inference 40

3.5.2 Causal discrimination 44

3.5.3 Causal discrimination in an Unidentifiable Situation 53

3.5.4 Experimentation 59

CONCLUSION 61

ANNEXE CODING TO TEST 62

BIBLIOGRAPHY 75

LIST OF FIGURES

2.1 Fairness through unawareness 20

2.2 Group fairness Boostrap 21

2.3 Predictive parity Boostrap 22

2.4 Predictive Equality Boostrap 23

2.5 Equality of Opportunity Boostrap 24

2.6 NPV Boostrap 25

2.7 Overall Accuracy Equality Boostrap 25

2.8 Treatment Equality Boostrap 26

2.9 Calibration 27

3.1 Directed graph showing the causal relation between 6 vertices 28

3.2 Acyclique Graph 33

3.3 Acyclique Graph based on German data set 34

3.4 Acyclique Graph based on German data set 35

3.5 Model to simulate 38

3.6 R test Output 40

3.7 Tetrad test Output 41

3.8 Tetrad test Output with knowlege 41

3.9 Causal graphs of a Markovian model 42

3.10 Causal graphs of a semi-Markovian model 42

LIST OF TABLES

2.1 Tabular definition of TP, FP, FN and TN 10
2.2 Confusion matrix 17
3.1 d-separation statements based on Figure 3.2 33
3.2 d-separation statements based on figure 3.2 36

RÉSUMÉ

L'une des normes sociales les plus importantes de l'ère moderne est l'équité. Notre comportement est déterminé par notre expérience personnelle, notre contexte culturel et notre conscience historique, et ce comportement est généralement injuste.

Afin de combattre les décisions injustes basées sur la race, le sexe, l'âge, etc., de nombreuses législations ont été mises en place dans le monde entier.

D'autre part, le processus de prise de décision dans le monde actuel a été confié à des algorithmes d'apprentissage automatique, et est appliqué automatiquement dans de nombreux domaines tels que l'admission au crédit et les primes d'assurance. L'objectif principal de ces algorithmes a toujours été la précision, comment minimiser notre erreur et créer un programme qui a la meilleure performance. Et c'est là qu'intervient l'injustice envers les groupes défavorisés, puisque dans ces algorithmes nous essayons de prédire le comportement humain et donc de prédire un comportement injuste.

Dans cette optique, notre objectif est de pouvoir quantifier l'équité. Tout d'abord, comme une notion associative, qui a été proposée par les chercheurs dans la littérature, certaines de ces notions sont la parité statistique, l'impact disparate, l'égalité des chances. Ensuite, comme une notion causale. En fait, nous allons réunir plusieurs méthodes pour détecter la relation causale entre notre attribut sensible et notre décision.

Mots clés : Race, Sexe, Discrimination, Équité, Causalité, d-séparation, Effet spécifique du chemin, Modèle d'équation structurelle, Discrimination directe, Discrimination indirecte.

ABSTRACT

One of the most important social norms in the modern era is fairness. Our behavior is determined by our personal experience, our cultural context, and our historical awareness, and that behavior is usually unfair.

In order to combat unfair decisions based on race, gender, age, etc., many legislations have been implemented around the world. On the other hand, the process of decision making in the current world has been handed over to machine learning algorithms and is applied automatically in multiple areas such as loan admission and insurance premium determination. The main goal of these algorithms has always been accuracy, how to minimize our error, and create a program that has the best performance. And this is where the unfairness towards disadvantaged groups comes in, since in these algorithms we are trying to predict human behavior and therefore predict unfair behavior.

With this in mind, our objective is to be able to quantify fairness. First, as an association-based notion, which has been proposed by researchers in the literature, some of these notions are statistical parity, disparate impact, and equality of opportunity. Then we will be exploring unfairness as a causal notion.In fact, we will bring together several methods to detect the causal relationship between our sensitive attribute and our decision.

Key Words: Race, Sex, Discrimination, Fairness, Causality, d-separation, Path specific effect, Structural equation model, Direct discrimination, Indirect discrimination.

INTRODUCTION

Machine learning tools have recently become one of the key players in decision making. And they are designed to make that decision based on human behaviors that may contain discrimination against certain groups like women or certain races.

Therefore, regulations have been put in place to combat this discriminatory behavior. It is thus urgent to eliminate this discrimination.

First we define what discrimination is and how it manifests itself in our decision-making process, and then we quantify discrimination using different concepts.

With this in mind, this report focuses on the different notions of fairness and how we can test them on a data set.

Next, we will review at the notion of causal fairness, how to quantify it, and how to test the causal relationship between our protection groups and the decision we want to make, whether direct or indirect.

CHAPTER 1
MOTIVATION

1.1 Legal context

Unequal opportunity has been a concern in many areas such as employment rates and credit scoring. Multiple pieces of legislation have been put in place to define and ensure non-discrimination against minorities and disadvantaged groups. Some of these laws include the UK Equality of Act [Feast and Hand(2015)],which requires non-discriminatory policy decisions by Ministers of the Crown and others, and is an update and consolidation of the Disability Discrimination Act, [Merry and Edwards(2002)], the Sex Discrimination Act [Act(1976)], and the Race Relations Act [Murphy(1976)]. In the United States, the Equal Pay Act of 1963 [Elisburg(1978)] prevents wage disparities based on sex; the Equal Credit Opportunity Act [Smith(1977)] prohibits discrimination in any aspect of a credit transaction; and the Civil Rights Act [Legislation(1968)] prohibits discrimination against disadvantaged groups in the workplace. In Canada, numerous laws have been put in place to ensure the right to fair and non-discriminatory treatment and prohibit discrimination against minorities and disadvantaged groups: the Canadian Human Rights Act [Legislation(1985b)], the Employment Equity Act [Legislation(1995)], and Canada Labour Code [Legislation(1985a)]. In the European Union, the legislation in place to combat discrimination is the Racial Equality Directive [Legislation(2000)],

the Employment Equality Framework Directive of 2000 [European Union Legislation(2
and the Equal Treatment Directive of 2006 [European Union Legislation(2006)].

One of the major roles of legislation has been to define protected characteristics. Under the UK Equality Act [Feast and Hand(2015)], the protected characteristics are:

- age;
- disability;
- gender reassignment;
- marriage and civil partnership;
- pregnancy and maternity;
- race (including color, nationality, ethnic or national origins);
- religion or belief;
- sex;
- sexual orientation.

1.2 Bias is in algorithms

It may seem that turning your decision-making over to an AI system will prevent you from discriminating, but that's not the case. In fact, machine learning algorithms are models designed by humans to conclude and predict based on an analysis of data related to a question (for example, who would get a loan?), and the way they work is simple: it's a process in which data is fed to a model, which mathematically processes it and creates the desired outcome. As new data is fed to the algorithm, the process loops around to improve accuracy. Thus, the algorithms are influenced by human behavior, its errors and discrimination, starting with the collection of data through its analysis and finally the decision.[Rovatsos et al.(2019)Rovatsos, Mittelstadt, and Koene]

1.2.1 Evidence of biased algorithms

Multiple algorithms have been shown to be biased, among the most notable:

- There was evidence that Google's ads showed fewer high-paying jobs for women[Datta et al.(2014)Datta, Tschantz, and Datta].
- Amazon Prime's same-day delivery offers were provided by algorithms that reinforced racial bias by not offering the same-day delivery option in predominantly minority American neighborhoods[Ingold and Soper
- U.S. insurance companies provide quotes based more on credit score than driving record. For example, a customer with a low credit rating and a good driving record pays more than a customer with a high credit rating and a poor driving record[O'neil(2016)].

These systems were not set up to discriminate explicitly; in fact, ignoring the sensitive attribute is not enough to achieve non-discriminatory systems.

1.2.2 Evidence of unfairness in insurance data

1.2.2.1 Racial Discrimination

As Wolff 2006 [Wolff(2006)] reminds us, in 1896, Frederick L. Hoffman, an actuary with Prudential Life Insurance, published a report demonstrating, with statistics, that a black American was uninsurable (see Hoffman 1896 [Hoffman(1896)]). Du Bois 1896 [Du Bois(1896)] noted ironically that the death rate of blacks in the United States was only slightly higher (but comparable) to that of white citizens in Munich, Germany, at the same time. More importantly, the main criticism is that it aggregated all sorts of data, preventing a finer analysis of other causes of (possible) excess mortality (this is also the argument made by O'Neil 2016 [O'neil(2016)]). At that time, in the United States, several states were passing anti-discrimination laws, prohibiting the charging of different premiums based on racial information. For example, as Wiggins 2013 [Wiggins(2013)] points out, in the summer of 1884, the Massachusetts state legislature passed the Act to Prevent Discrimination by Life Insurance Companies Against People of Color. This law prevented life insurers operating in the state from making any distinction or discrimination between white persons and colored persons wholly or partially of African descent, as to the premiums or rates charged for policies upon the lives of such persons. The law also required insurers to pay full benefits to African-American policyholders. It is on the basis of these laws that the argument of uninsurability was made: insuring blacks at the same rate as white would be statistically inequitable, argued Hoffman 1896 [Hoffman(1896)], and not insuring blacks was the only way to comply with the law (see also Heen 2009 [Heen(2009)]). As Bouk 2015 [Bellhouse(2016)] recounts "Industrial insurers operated a high-volume business; so to simplify sales they charged the same nickel to everyone. The home office then calculated benefits according to actuarially defensible discriminations, by age initially and then by race. In November 1881, Metropolitan decided to mimic Prudential, allowing policies to be sold to African Americans once again, but with the understanding that black policyholders' survivors only received two-thirds of the standard benefit".

1.2.2.2 Gender Discrimination

The 2004 European Goods and Services Directive, Council of the European Union 2004 [Conseil de l'Union Européenne (2004)()], aimed to reduce gender gaps in access to all goods and services, discussed for example by Thiery and Van Schoubroeck 2006 [Thiery and Van Schoubroeck(2006)]. A special derogation in Article 5, paragraph 2, allowing insurers to set gender-based prices for men and women. Indeed, "Member States may decide (...) to allow proportionate differences in premiums and benefits for individuals where the use of sex is a determining factor in the assessment of risk, on the basis of relevant and accurate actuarial and statistical data". In other words, this clause allowed for an exception for companies, provided that they provide actuarial and statistical data that establishes that gender is an objective factor in assessing risk. The European Court of Justice struck down this legal exception in 2011, in a ruling discussed at length by Schmeiser et al. 2014 [Schmeiser et al.(2014)Schmeiser, Störmer, and Wagner] or Rebert and Van Hoyweghen 2015 [Rebert and Van Hoyweghen(2015)], for example. This regulation, which generated a lot of comment in Europe in 2007 and then in 2011, had also raised many questions in the United States, several decades earlier, such as this discussion in the late 1970s, with Martin 1977[Martin(1977)], Hedges 1977 [Hedges(1977)] and Myers 1977[Myers(1977)]. For example, in City of Los Angeles, Department of Water and Power v. Manhart, the Supreme Court considered a pension system in which female employees made higher contributions than males for the same monthly benefit because of longer life expectancy. The majority ultimately determined that the plan violated Title VII of the Civil Rights Act of 1964 because it assumed that individuals would conform to the broader trends associated with their gender. Such discrimination, the court suggested, is troubling from a civil rights rights perspective because it does not treat individuals as individuals, as opposed to merely members of the mere members of the groups to which they belong. These laws were motivated, in part, that employment decisions are generally individual: a specific person is hired, fired, or demoted, based on his or her past or expected contribution to the to the employer's mission. In contrast, stereotypes about individuals based on group characteristics are generally more tolerated in fields such as insurance, where individualized decision making does not make sense.

1.3 Types and causes of discrimination

1.3.1 Types of discrimination

Discrimination in the literature has been divided into two types, disparate treatment and disparate impact. [Dwork et al.(2012)Dwork, Hardt, Pitassi, Reingold,

- Disparate treatment is the most straightforward and intuitive form of discrimination, when an individual is being treated in a different way based on his or her sensitive attributes. To address disparate treatment, a used approach is reverse Tokenism, that involves rejecting individuals from a minority while also rejecting a qualified member of the majority.
- For disparate impact, sensitive attributes are not considered in the decision-making process, but minorities are treated differently from majorities due to the correlation between sensitive attributes and other attributes that are not protected under the legislation.

1.3.2 Causes of machine learning bias

As mentioned earlier, discrimination can occur at different stages of the decision-making process. In fact, biases in machine learning can be caused by:

- Data collection: selected training data may be biased or sometimes incorrect, which may be related to an incorrect distribution of ground truth. Not all ground truths are objective, some are based on human decisions that may be biased, and poor data collection may result in poor or no representation of some groups.

- Selecting the Features: the choice of attributes can be a source of bias. One problem is disparate impact or redlining, where the predicted outcome is determined by a surrogate variable, i.e., a variable that correlates with a sensitive attribute. Another problem will be incomplete information, i.e., not including useful information due to confidentiality or difficulty of access, which can have a negative impact on minorities.

- Wrong Assumptions: two assumptions severely affecting minorities are usually made: the data is reflective of the population as a whole and the data is reflective of the future.

- Masking: intentionally applying the practices listed above to mask the incorporation of bias in the dataset can result in biased machine learning predictors.

CHAPTER 2
NOTIONS OF FAIRNESS

One question that always arises is how to define fairness. For this reason, in this chapter we will bring together several versions of the notion of fairness and then apply them to a data set.

2.1 Background

2.1.1 Definition

Before we start presenting the different notions, we will establish some definitions.

2.1.1.1 Machine Learning Definitions

- Training data and samples: collection of data used to train our algorithm to make a decision.
- Classifier: a predictor that assigns a class to each individual, binary when $y \in \{0, 1\}$.
- Positive and negative class: binary classification where the positive is relative to the favorable result, and the negative to the unfavorable result.
- Prediction (Pred.) and Ground Truth (G.T.): differentiate between

what the individual belongs to (Ground Truth) and what the algorithm predicts (Prediction).

- True and False Positives:
  - True Positives (TP): individual that are correctly classified in the positive class.
  - False Positives (FP): the prediction of an individual is different then the ground truth with the ground truth being the negative class.
- True and False Negatives: this definition is analogous to the previous definition:
  - True Negatives (TN): prediction and ground truth are the same with ground truth being the negative class.
  - False Negatives (FN): the prediction is negative class but the ground truth is positive class.

	G.T: Positive	G.T: Negative
Pred.: Positive	TP	FP
Pred.: Negative	FN	TN

Table 2.1: Tabular definition of TP, FP, FN and TN

2.1.1.2 Fairness Definitions

- Protected attribute: an attribute of an individual that should not affect decision making, such as gender and race.

- Privileged and non-privileged group, by considering a binary protected attribute, our population is divided into two groups, one of which will be discriminated against and is called the non-privileged group and the other the privileged group.
- Favorable and Unfavorable outcome refer respectively to the prediction of belonging to a positive and a negative class.
- Qualified and Unqualified Individuals refer to the ground truth which is a positive class and a negative class respectively.

2.1.2 Mathematical Notation

First, we need to define the database in which our work will take place.

Consider a finite data set of n individuals D in which each individual is defined as a triple (X, Y, Z)

- X the attributes used to predict.
- Y the outcome we want to predict
- Z the protected attribute that is binary, $Z \in \{0, 1\}$ could be included in X which means is used to predict.

$$Z = \begin{cases} 1 & \text{for privileged group} \\ 0 & \text{for unprivileged group} \end{cases}$$

$$Y = \begin{cases} 1 & \text{for favourable group} \\ 0 & \text{for unfavourable group} \end{cases}$$

Then we will set how the classification is defined.

- Assumption:
  - Binary classifiers
  - Talk about a given fixed classifier.
- Prediction:
  - $h : X \to [0, 1]$, with $S = h(X)$ score, corresponding to the predicted probability of an individual to belong to the positive class.
  - For a given threshold σ, Y is predicted to belong to the positive class, if $h(X) > \sigma$.
    $\hat{Y}$: the final prediction based on $\sigma, \hat{Y} = 1$ if $h(X) > \sigma$
  - Probability that the favourable outcome will be predicted for individuals from the privileged group, $P(\hat{Y} = 1|Z = 1) = P_1(\hat{Y} = 1)$
  - Probability that a positively classified individual from the unprivileged group is actually unqualified. $P(Y = 0|Z = 0, \hat{Y} = 1)$

2.1.3 The Notions

In order to achieve fairness in an algorithm, the first step is to define a fairness metric. According to Verma and Rubin [Verma and Rubin(2018)], we will list the notions of fairness, define them, and quantify them. The different notions of fairness that will be defined are:

- unawareness;
- group fairness;

- predictive parity;
- calibration; and
- individual fairness.

2.1.3.1 Fairness through unawareness

Fairness through unawareness is one of the basic notions of fairness; it is simply achieved when the sensitive attribute is not used in the classification problem. In fact, fairness by unawareness will be achieved when desperate impact is avoided.

Formally:

$$P(\hat{Y} = y|X = x) = P(\hat{Y} = y|X = x, Z = z) \tag{2.1}$$

Knowing that other attributes can substitute for a sensible attribute, this notion is insufficient to achieve fairness.

2.1.3.2 Group fairness

Group fairness, also known as statistical parity, is achieved when an individual in a non-privileged group has the same probability of a favorable outcome as the privileged group.

Formally, statistical parity is achieved if:

$$P(\hat{Y} = 1|Z = 1) = P_1(\hat{Y} = 1) = P_0(\hat{Y} = 1) = P(\hat{Y} = 1|Z = 0) \tag{2.2}$$

A more relaxed practical form of equation 2.2 is the notion of **ϵ-fairness**:

- Let $\epsilon > 0$, a parity based fairness notion is ϵ-fair if

$$|P_1(\hat{Y}=1) - P_0(\hat{Y}=1)| < \epsilon \tag{2.3}$$

And a more expanded notion of group fairness is taking into account legitimate factors $L \subset X$, and require equal decision despite the protected attribute, is **Conditional Statistical Parity**

- Formally defined as:

$$P_0(\hat{Y}=1|L=l) = P_1(\hat{Y}=1|L=l) \tag{2.4}$$

Finally the notion of group can be normalized and defined as **Normalised Difference** δ

$$\delta = \frac{P_1(\hat{Y}=1) - P_0(\hat{Y}=1)}{d_{max}} \tag{2.5}$$

where $d_{max} = \min\left\{\frac{P(\hat{Y}=1)}{P(Z=1)}, \frac{P(\hat{Y}=0)}{P(Z=0)}\right\}$

And

$$\delta = \begin{cases} 0, & \text{indicating complete fairness} \\ 1, & \text{indicating maximum discrimination} \end{cases}$$

2.1.3.3 Predictive parity

Since statistical parity only evaluates prediction, we introduce the notion of Predictive parity which takes into account the ground truth of the sample. To reach this notion, the precision must be equal in all the protection groups.

In this context precision is the positive predictive value $PPV = \dfrac{TP}{(TP+FP)}$

that is required to be equal for all demographic groups.

Formally, predictive parity is achieved if

$$P_0(Y = 1|\hat{Y} = 1) = P_1(Y = 1|\hat{Y} = 1) \tag{2.6}$$

Instead of comparing PPVs, there is a list of values that can be compared:

1. **Predictive Equality**: where, if satisfied, the probability of classifying an unqualified individual in the favorable outcome is similar in the privileged and non-privileged groups, i.e., the false positive rate $FPR = \dfrac{FP}{(TN + FP)}$ is similar in both groups.

   Formally:

   $$P_0(\hat{Y} = 1|Y = 1) = P_1(\hat{Y} = 1|Y = 1) \tag{2.7}$$

2. **Equality of Opportunity**: where, if satisfied, the probability of classifying a qualified individual in the unfavorable outcome is similar in the privileged and non-privileged groups, i.e. the false negative rate $FNR = \frac{FN}{(FN+TP)}$ is similar in both groups.

   Formally:

   $$P_0(\hat{Y} = 0|Y = 1) = P_1(\hat{Y} = 0|Y = 1) \tag{2.8}$$

3. **Equality of odds:** where, if satisfied, predictive equality and equal opportunity are met. Formally:

   $$P_1(\hat{Y} = 1|Y = i) = P_0(\hat{Y} = 1|Y = i),\ i \in |0, 1\} \tag{2.9}$$

4. **Conditional Use Accuracy Equality:** where, if satisfied, both predictive parity ($PPV = \dfrac{TP}{TP + FP}$) and equal negative prediction values ($NPV = \dfrac{TN}{FN + TN}$) are required across both groups

Formally:

$$P_1(Y = \hat{Y}|Y = i) = P_0(Y = \hat{Y}|Y = i), i \in 0, 1 \tag{2.10}$$

In summary, when the POSITIVE TRUE RATE and NEGATIVE TRUE RATE are similar in the two groups. Noting that NPV and PPV are not required to be equal.

5. **Equal overall accuracy**: where, if satisfied, the prediction accuracy is equal between the preferred and non-preferred groups, focusing on both True Positives and True Negatives.

   Formally:

$$P_1(Y = \hat{Y}) = P_0(Y = \hat{Y}) \tag{2.11}$$

6. **Treatment Equality:** where, if satisfied, the ratio of false positives to false negatives is equal for the preferred and non-preferred groups:

$$\frac{FN_1}{FP_1} = \frac{FN_0}{FP_0} \iff \frac{FP_1}{FN_1} = \frac{FP_0}{FN_0} \tag{2.12}$$

In fact, if the false negatives are higher than the false positives for the privileged group, there are more unqualified people who score favorably than qualified people who score unfavorably. And when the non-preferred group gets an equal ratio, we have misclassification discrimination.

2.1.3.4 Calibration

Calibration is a notion of fairness accompanied by a score S defined by the predicted probability of an individual with attributes X to be classified in

	Actual-Positive	Actual–Negative

Predicted-Positive	**True-Positive (TP)**
	PPV=$\frac{TP}{(TP+FP)}$
	TPR=$\frac{TP}{(TP+FN)}$
		False-Positive (FP)
		FDR=$\frac{FP}{(TP+FP)}$
		FPR=$\frac{FP}{(FP+TN)}$

Predicted-Negative	**False-Negative (FN)**
	FPV=$\frac{FN}{(TN+FN)}$
	FNR=$\frac{FN}{(TP+FN)}$
		True-Negative (TN)
		NPV=$\frac{TN}{(TN+FN)}$
		TNR=$\frac{TN}{(TN+FP)}$

Table 2.2: Confusion matrix

the favorable outcome.

Define S as:

$$S = P(\hat{Y} = 1|X) \tag{2.13}$$

If

$$P_1(Y = 1|S = s) = P_0(Y = 1|S = s), \forall s \in [0, 1] \tag{2.14}$$

Then the classifier is calibrated.

We also define extended notions of calibration:

- **Well-calibration**: this notion requires:

$$P_1(Y = 1|S = s) = s = P_0(Y = 1|S = s), \forall s \in [0, 1] \tag{2.15}$$

  The purpose of this concept is to ensure that the probability of awarding the favorable outcome and the percentage of qualified individuals are approximately equal.

- **Balance for negative class:** this notion requires an average of scores among all unqualified individuals, equal between privileged and non-privileged groups.

  A formal way to define the realization of this notion of fairness.

$$E_1(S|Y = 0) = E_0(S|Y = 0) \tag{2.16}$$

- **Balance for positive class**: this notion is similar to the notion of equilibrium for the negative class, and is satisfied if the average of the qualified individuals is equal between the privileged and non-privileged groups.

  Formaly:

$$E_1(S|Y = 1) = E_0(S|Y = 1) \tag{2.17}$$

2.1.3.5 Individual Fairness

To counter the previous notions, the notion of individual fairness is based on metrics above the individuals themselves, formulating a (D, d)-Lipschitz property, the classifier is said to fulfil individual fairness if:

$$D(h(x_i), h(x_j)) \leq d(x_i, x_j) \quad \forall x_i, x_j \tag{2.18}$$

With:

- D: a distance metric over the room of possible classifications
- d: a distance metric over individuals
- x_i, x_j denote individual

2.2 Testing Fairness Notion on the German Credit Data-set

In this section we will be testing the different fairness Notion on German Credit Data-set this data-set is mostly used in fairness literature it has information about 1000 individuals and has 21 attributes describing each individual.

This data-set is divided into:

- X: Attributes used for predictions regarding the data sample (19 attributes): *Status of existing checking account,Duration of Credit (in months),Credit history,Purpose of Loan,Credit amount,Savings account/bonds, Savings account/bonds, Present employment since, Installment rate (%), Guarantors,Present residence since, Most valuable available asset, Age in years, Concurrent Credits, Type of housing, Number of existing credit at this bank, Job, No of dependent, Telephone, Foreign Worker.*

- Y: Corresponding ground-truth of the sample: *Loan quality (Binary).*

- Z: Binary protected attribute: *Personal status and sex* note that we will transform this attribute into a binary based on sex only.

The classification used to test over our data-set is *binomial* using *logit.*

2.2.1 Fairness through unawareness

According to 2.1 two regression will be used.

- The first regression will be over the attributes X and Z:

- The second regression will be over the attributes X:

To be able to compare between the two component of equation 2.1 we will be using the Boostrap method i.e. sample the data base multiple times and then create multiple models so we are able to compare the estimated amount and plot their difference so we can conclude if the difference is centered around zero or not.

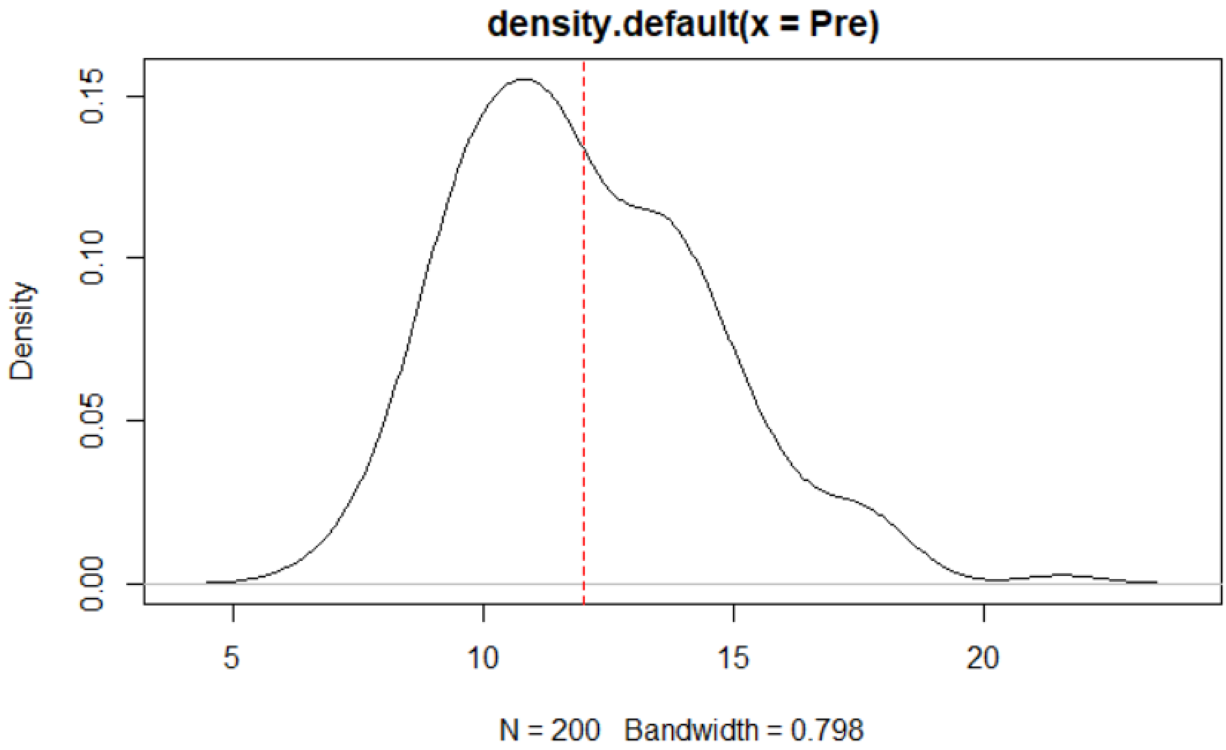


Figure 2.1: Fairness through unawareness

The figure 2.1 shows the density of the difference between the prediction using the two regressions is not centered around null which indicate that the gender is playing a big role in our estimation.

2.2.2 Group fairness

According to 2.2:
We calculate the number of privileged and nonprivelegier group of individuals.

Then we calculate the number of non-privileged individuals with a favorable result and the number of privileged individuals with a favorable result we consider the result coming from the regression over all the attributes. And compare the probability for an individual to be assigned the favourable outcome for each group.

We will use the Boostrap method to compare two component of 2.2.

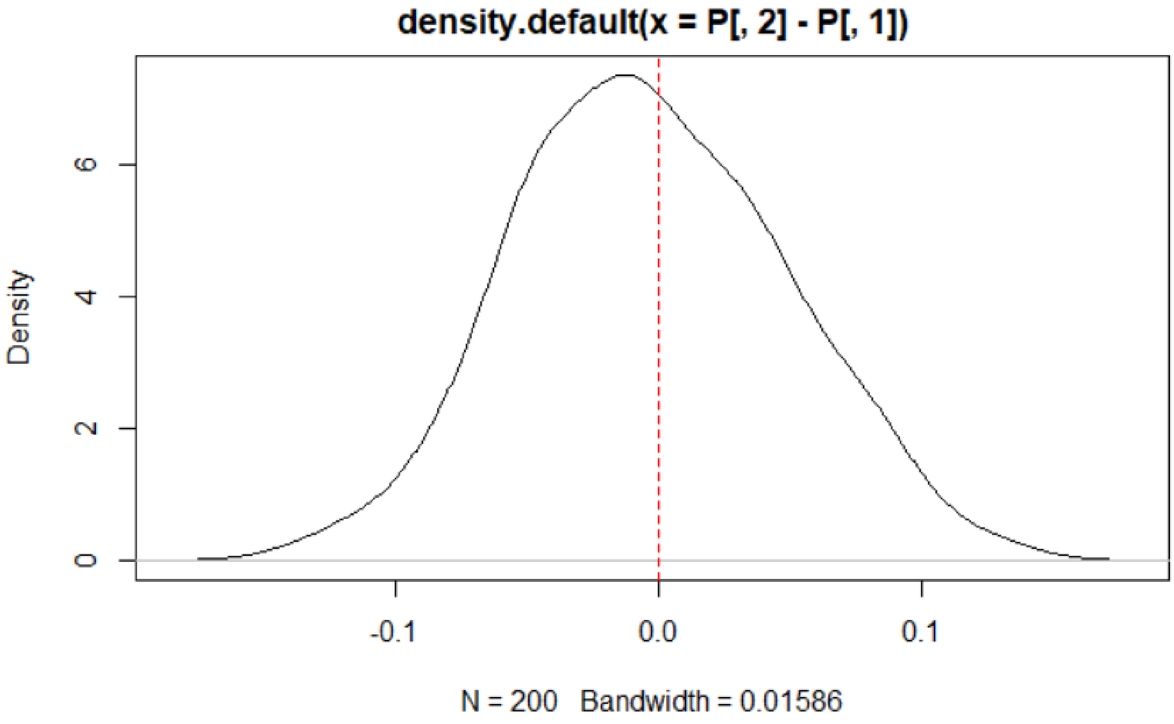


Figure 2.2: Group fairness Boostrap

The figure 2.2 shows the density of the difference between the two probability that is not centered around zero indicating an **unfaire** situation.

2.2.2.1 Epsilon group fairness

According to 2.3 with epsilon 0.05 the decision is **fair**.

2.2.3 Predictive parity

According to 2.6 we have to compare the PPV for both groups:
To be able to compare PPV for both group we will be using the Boostrap method.

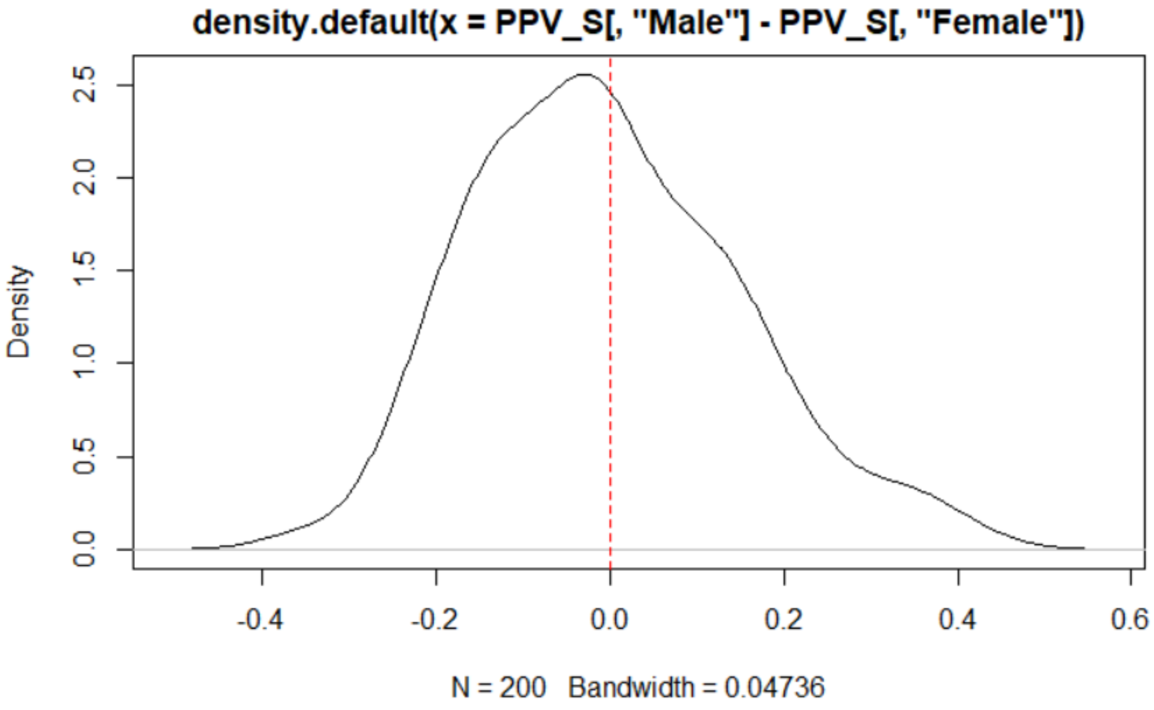


Figure 2.3: Predictive parity Boostrap

The figure 2.3 shows the density of the difference between the two PPV that is not centered around zero indicating an **unfaire** situation i.e. more females were accurately predicted to be part in the positive group then men.

1. **Predictive Equality**: According to 2.7 we have to compare the FPR for both groups.

   To be able to do the comparison of FPR for both group we will be using the Boostrap method i.e sampling multiple time the data to be able to plot the density of the difference in FPRs so we can deduct if the data is fair according to the predictive parity based on the mean. The figure 2.4 shows the density of the difference between the two FPR that is not centered around zero indicating an **unfaire** situation i.e. more male were wrongly predicted to be part in the negative group then women.

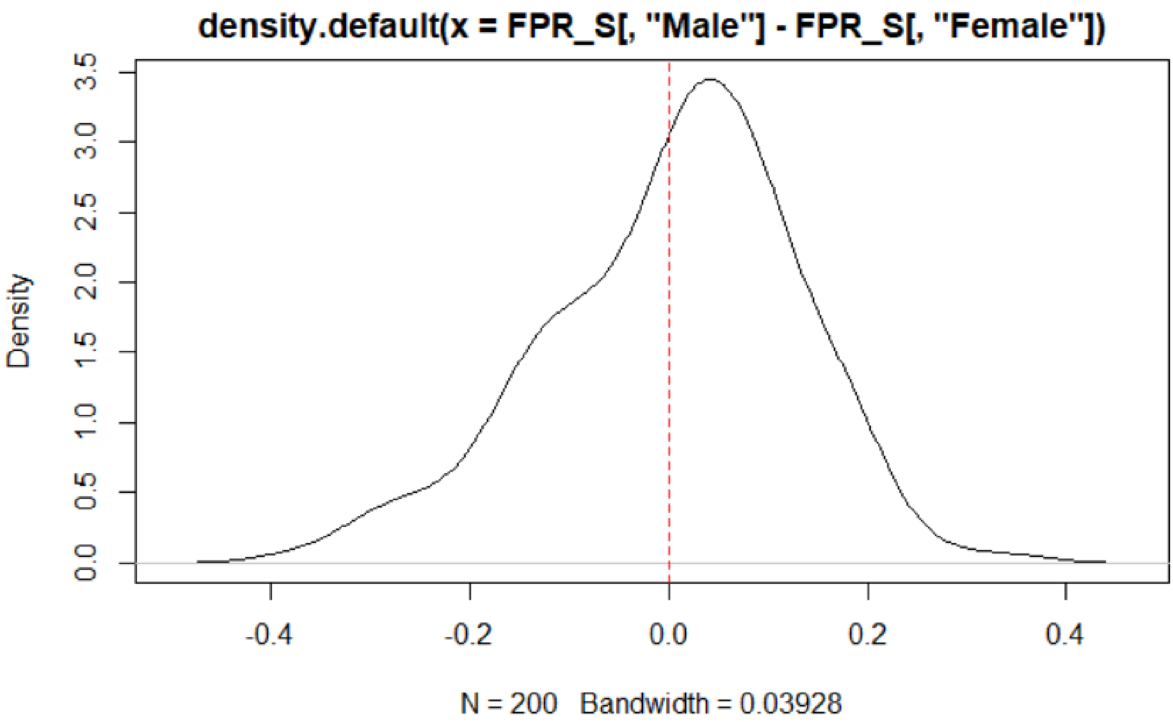


Figure 2.4: Predictive Equality Boostrap

2. **Equality of Opportunity** According to 2.8 we have to compare the FNR for both groups:
   To be able to compare FNR for both group we will be using the Boostrap method.

   The figure 2.5 shows the density of the difference between the two FNR is centered around zero indicating a **faire** situation.

3. **Equality of odds**: According to 2.9 we have to compare the FPR and FNR for both groups.

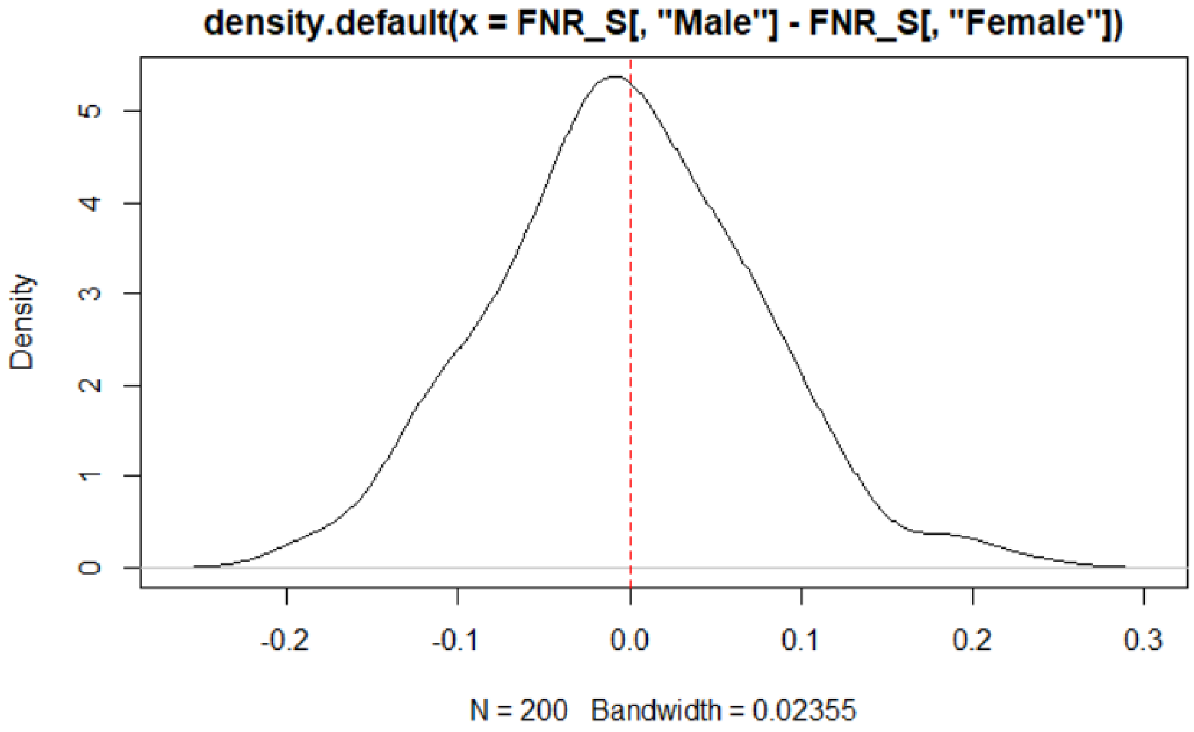


Figure 2.5: Equality of Opportunity Boostrap

The figure 2.4 shows the density of the difference between the two FPR is not centered around zero indicating an **unfaire** situation.

4. **Conditional Use Accuracy Equality:**According to 2.10 we have to compare the PPV and NPV for both groups:
   Bootstrapping:

   Since using predictive parity we saw unfaireness than we can deduct **unfairness** too.

5. **Overall Accuracy Equality:** According to 2.11 we have to compare the $P_1(Y = \hat{Y})$ and $P_0(Y = \hat{Y})$
   Bootstrapping:

   The figure 2.7 shows the density of the difference between the two probability that is not centered around zero indicating an **unfair** situation.

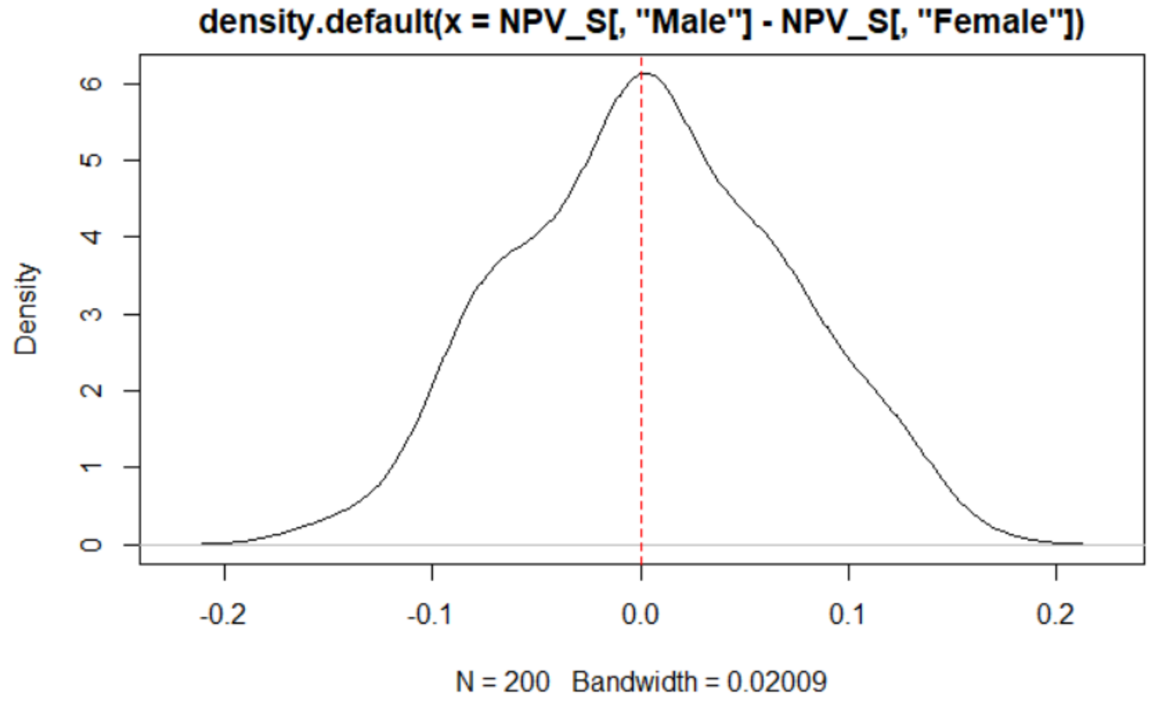


Figure 2.6: NPV Boostrap

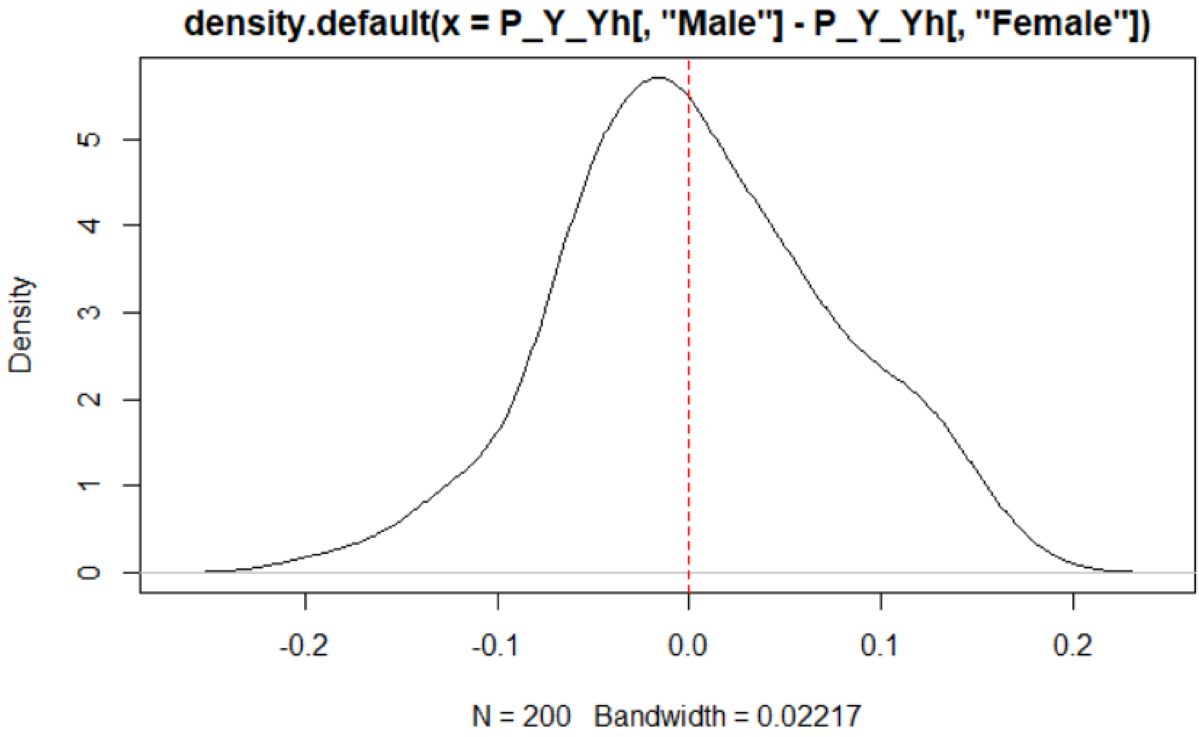


Figure 2.7: Overall Accuracy Equality Boostrap

6. **Treatment Equality**:

   According to 2.12 we have to test equal ratio of false positives and false negatives.

   Bootstrapping

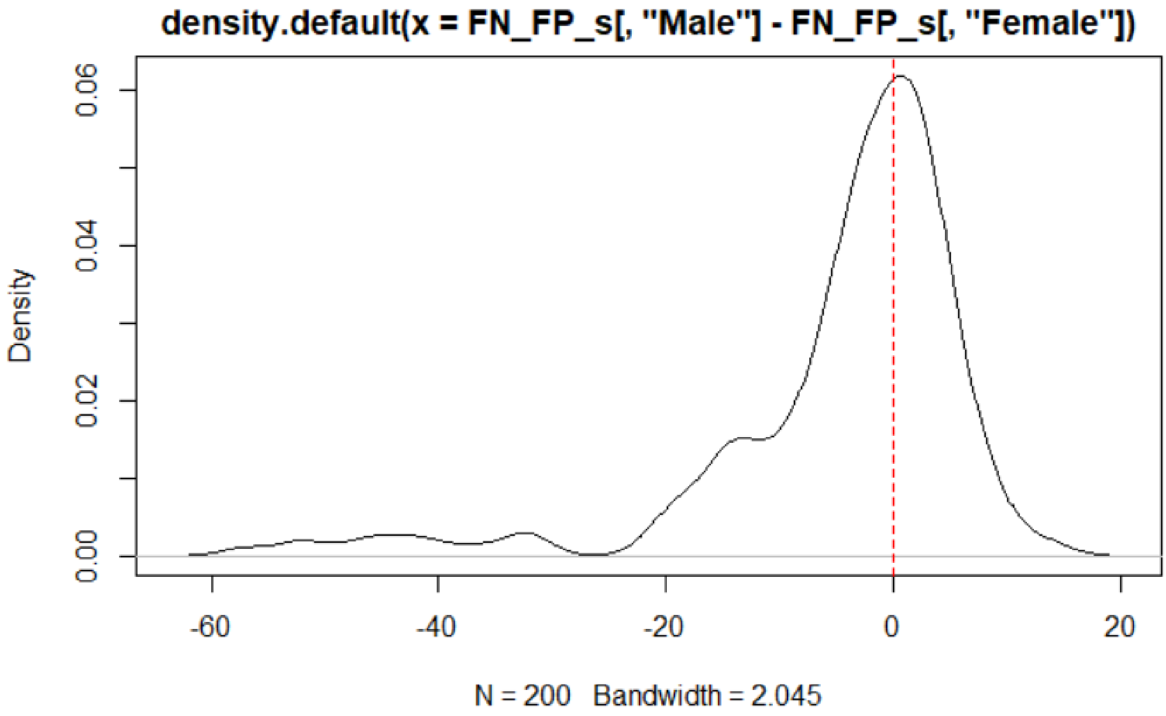


Figure 2.8: Treatment Equality Boostrap

The figure 2.8 shows the density of the difference between the two ratio that is centered around zero indicating an **fair** situation i.e. female have a higher prediction accuracy than male.

2.2.4 Calibration

According to 2.14 we have to compare $P_1(Y = 1|S = s) = P_0(Y = 1|S = s), \forall s \in [0,1]$ Figure 2.9 compare the probability between men and women.

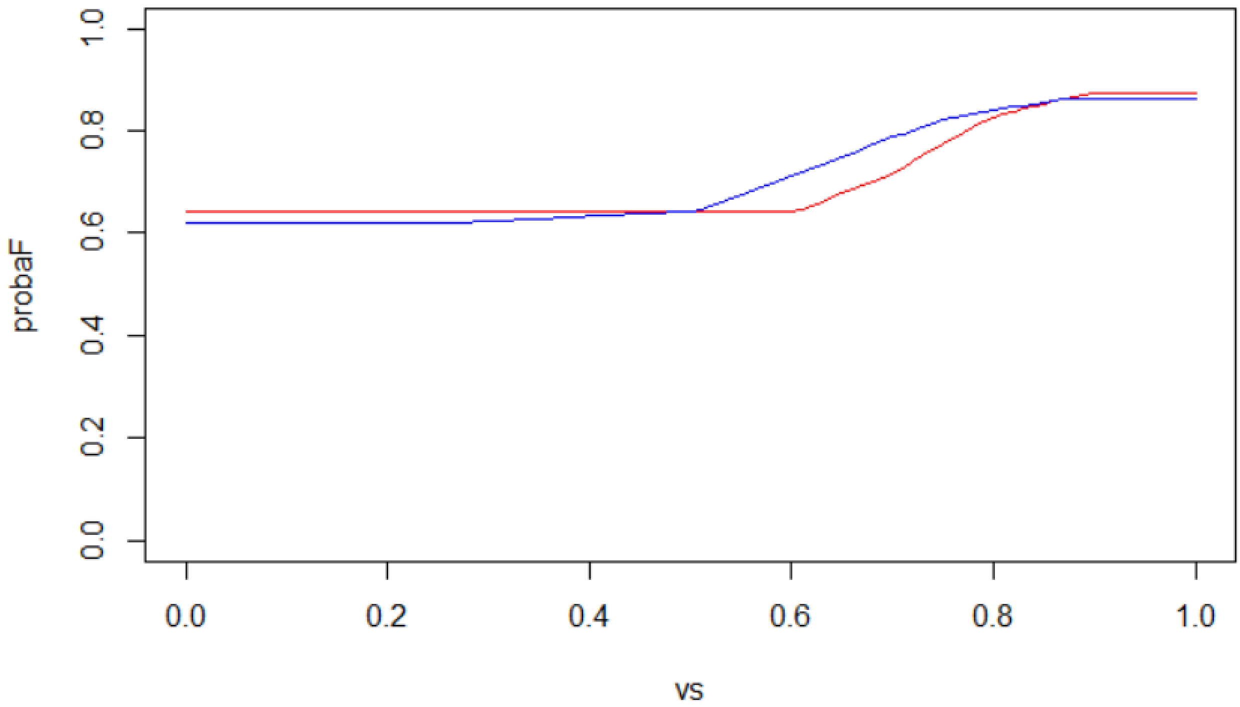


Figure 2.9: Calibration

CHAPTER 3
THE NOTION OF CAUSALITY DISCRIMINATION

One notion of fairness at the individual level is the notion of causality.

3.1 Causal graphs

In order to understand causality, we need to present some key definitions.

Definition 1 ***Directed graph****:*

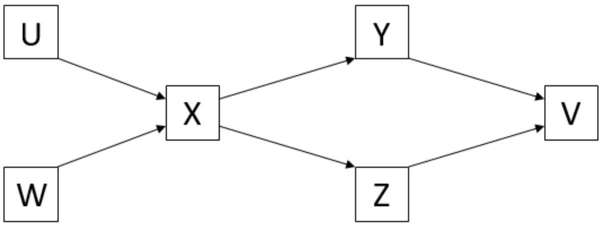


Figure 3.1: Directed graph showing the causal relation between 6 vertices

1. *The arrow symbol → will be used to indicate a causal relationship.*

2. *U and W are causally independent (i.e. if U changes W is not affected and vice versa)*

3. *X, Y, Z and V are causally dependent on U and W (i.e. a change in U and W will cause a change in X, Y, Z and V but the change in X, Y, Z and V does not affect U and W):*

- *X is directly caused by U and W (i.e. changes in U and W will affect X independently of Y, Z or V).*
- *Y, Z and V are **indirectly caused** by U and W (i.e. a change in U or W will cause a change in Y and Z only by causing changes in X.*
- *X is a common direct cause of Y and Z.*
- *X is an indirect cause of V through Y and Z.*
- *Y and Z are a direct cause of V.*

Definition 2 ***Type of causal paths, and type of vertix***

Path:

1. ***Directed Path** between two vertices, exists if one or more ordered sequences of vertices that must be crossed in one direction exist between these two vertices, if this condition is not satisfied, then these vertices are causally dependent an example of a directed path is the path between U and V ($U \rightarrow X \rightarrow Y \rightarrow V$)*
2. ***Undirected path** between two vertices, exists, if one or more ordered sequences of vertices must be crossed, regardless of the direction, an example of undirected path is the path between U and W ($U \rightarrow X \leftarrow W$)*

Vertix:

1. ***Collider vertix** is a vertix that has two arrows pointing in a path example V in ($Y \rightarrow V \leftarrow Z$), it is important to note that a collider is defined in terms of the path.*

2. ***Unshielded collider vertices*** *is a set of three vertices* $(X \rightarrow Y \leftarrow Z)$ *along a path such that* Y *is a collider and, additionally, there is no edge between* X *and* Z.

Definition 3 ***Causal model***

$\mathcal{M} =< U, V >$ *is a causal model with:*

- ***U****: exogenous variables determined by factors outside the model.*
- ***V****: endogenous variables determined by variables in* $U \cup \mathbf{V}$.
  This causal models are associated with a directed graph.

3.2 Connecting causal fairness to observational models

Definition 4 ***Causal conditioning*** *A vertix in a causal path can either be a non-collider in a path called an active variable, or a collider called inactive.*
Conditioning a vertix means changing its status from active to inactive and vice versa.

Definition 5 ***d-separation*** *Our goal in this definition is to explain the "independence" of vertices or groups of vertices in a causal graph while conditioning on a set of vertices Q; this property is called direct separation (d-separation).*
For any undirected path between X *and* Y *:*

Step 1: ***If*** *any non-colliding vertix in this path are in* Q. ***Then*** *the path is blocked and there is no causal influence between* X *and* Y.

Step 2: ***If*** $\exists$ *a collider that is not in* Q *and don't have a causal descendent in* Q. ***Then*** *the path is blocked and there is no causal influence between* X *and* Y.

If *every indirect path between X and Y is blocked.* ***Then*** *X and Y are d-separated.*

Theorem 1 *According to [Pearl(1988)] for any directed acyclic Graph if vertices are d-seperated then the relation in the joint distribution of the random variable associated is independence.*

Corollary 1 *If one statistical independency in the data disagree with what d-separation of the causal graph predict, then the causal model is assumed to be wrong.*

3.2.1 Strategy used to translate from causal model to an observational model

Step 1: Express a causal hypothesis using a directed graph.

Step 2: Translate, using d-separation, from the directed graph into mathematical language i.e. probability theory

Step 3: Establish the type of independence relationship that must occur in the resulting joint probability distribution.

Consequence of d-separation: two causally independent variables will be correlated if we condition on one of their common children, which can be a misleading result if we interpret these results as giving information about causal relationships.

3.3 d-separation tests

Due to the relation between the causal conditional independence and the probabilistic independence, through d-separation, an intuitive way to be able to test a causal model:

Step 1: List all the d-separation statements that are implied by the causal model

Step 2: Test each of the d-separation using an appropriate test of conditional independence.

This test has a couple of setbacks:

- Large number of d-separation statements.
- The necessity of a method to combine all the tests of independence into one test.
- The d-separation statements are not completely independent usually.

In able to adjust this setbacks a method to find the minimum set of this d-separation also known as basis set was suggested by [Pearl(1988)]:

Step 1: List all the variables in the causal model that are not linked by an arrow.

Step 2: List all causal parents of each vertex in the pair.

To be able to illustrate how to find this d-separation we gonna list them for the directed graph in figure 3.2
All the d-separation in table **??** predict the conditional independence, according to the variable types.

3.3.1 Testing the German Data set

The Chi-square test indicates if the clusters in a population are mutually dependent or not. It is important to note, however, that showing a sta-

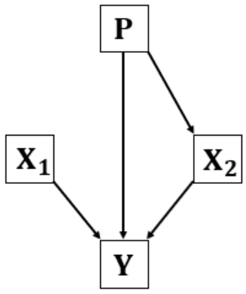


Figure 3.2: Acyclique Graph

Not Linked Variable	Parent variables of either non-adjacent variables	d-separation statement
P *and* X_1	None	$P \perp\!\!\!\perp X_1$
X_1 *and* X_2	P	$X_1 \perp\!\!\!\perp X_2 \vert P$

Table 3.1: d-separation statements based on Figure 3.2

tistical association using chi-square analysis may not necessarily indicate a cause-and-effect link between two clusters.

However we will use the Chi-square to test the d-separation statement for the causal graph based on the German Data set used in the previous sections.

The variables illustrated in this graph are the gender the credit Risk and the Property.

- First d-separation statement: Credit Risk $\perp\!\!\!\perp CreditRisk$
  $Hypotheses$:
  H$_0$: *The Credit score and the gender are independent*
  H_1: *The Credit score and the gender are not independent*
  We have no evidence to suggest that Credit score is related to gender

$$(\chi^2(df = 1) = 0.013\ p > .05)$$

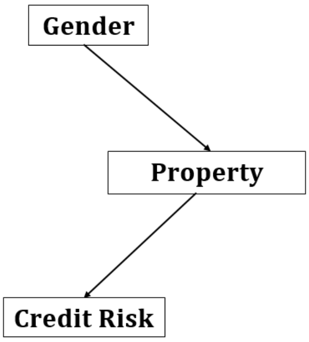


Figure 3.3: Acyclique Graph based on German data set

.

- Second d-separation statement: Credit Risk $\perp\!\!\!\perp$ $CreditRisk|Property$
  $Hypotheses$:
  H_0: The Credit score and the gender are independent under the condition Property
  H_1: The Credit score and the gender are not independent under the condition Property
  We have no evidence to suggest that Credit score is related to gender under the condition of property.

End finally we will test the d-separation statement based on the 3.4 graph of the German data-sets and the result is in ??.

3.4 Structural equation model

3.4.1 Testing path models using maximum likelihood

Step 1:

In order to test, we will need to translate the hypothesized causal system into a path diagram, defining the different types of variables in a structural equation model (SEM):

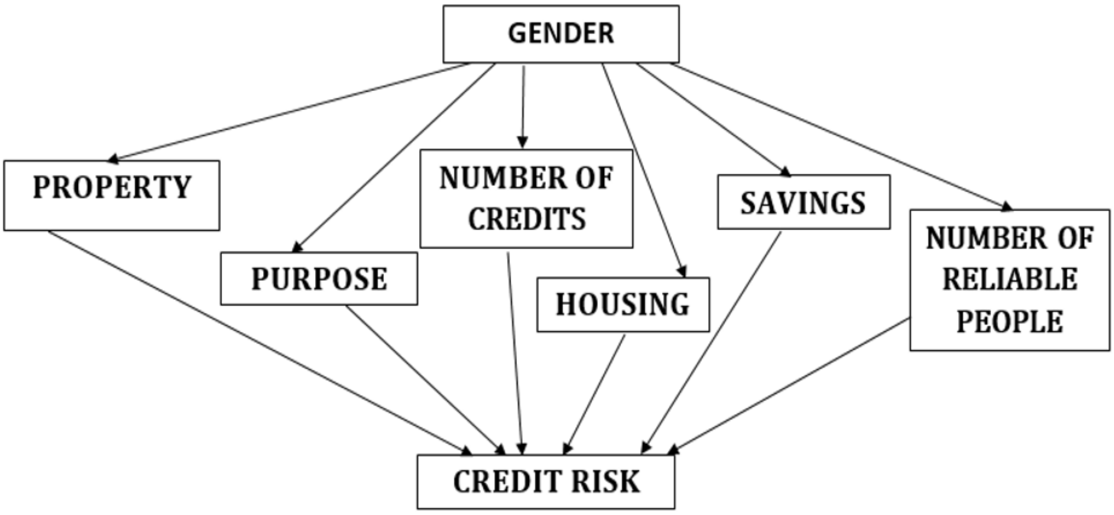


Figure 3.4: Acyclique Graph based on German data set

- Directly observed and measured variables are labelled *manifest* variables in a SEM.

- Non-measured variables that are assumed to have a causal role in the model, designated *latent* as variables in a SEM circled in the graph.

- The residual *error* variables.

And also the type of arrows:

- Straight arrow indicating a cause and effect relationship.

- A double-sided arrow indicating an unknown relationship.

Step 2:

Transform the causal model that we want to test into stuctural equations, assuming that the variables follow a multivariate normal distribution and the relationship between them is additive.

This transformation is not perfect since we are turning a directed relation

d-separation statement	Result
$Gender \perp\!\!\!\perp CreditRisk \mid Property$	No evidence of relation
$Gender \perp\!\!\!\perp CreditRisk \mid Purpose$	No evidence of relation
$Gender \perp\!\!\!\perp CreditRisk \mid Numberofcredits$	No evidence of relation
$Gender \perp\!\!\!\perp CreditRisk \mid Savings$	No evidence of relation
$Gender \perp\!\!\!\perp CreditRisk \mid Housing$	No evidence of relation
$Gender \perp\!\!\!\perp CreditRisk \mid NumberofCredit$	No evidence of relation

Table 3.2: d-separation statements based on figure 3.2

into an equivalence relation in the sence will mask the orientation of the causal relation.

Step 3:

Using covariance algebra, calculate the variance and covariance between variables.
If two vertices are not d-separated, then the corresponding random variables are not independently distributed, i.e. the covariance is zero.
Here we actually compare the causal model to a shadow, where the shadow is fixed but the numerical values are free and can be estimated.

Step 4:

Minimize the gap between observed and predicted variances in order to estimate the parameters. When using MLE ("maximum likelihood parameters"), the general idea behind finding the best free parameter value is to choose it in such a way that the covariates are as close as possible to the measured covariance of the actual data.

Step 5:

Assuming that the observed covariate is equal to the predicted covariate everywhere except for the sample variables, find P, probability of observing the minimum measured difference.

Null hypothesis: No difference between the observed and predicted covariance. Under this hypothesis, the following statistic (the maximum likelihood chi-square statistic) will follow a chi-square distribution as follows:

$$(N-1)F_{ML} \longrightarrow \chi^2_{[v(v+1)/2]-(p+q)}$$

with

- v: number of variables

- q: number of free variances of exogenous variables

- p: the number of free path coefficients in the model

Step 6

The model will be considered false if the probability is sufficiently low (i.e. less than 0.05), and vice versa the model is rejected if the probability is sufficiently high (i.e. greater than 0.05), after which lead the conclusion of the consistency of the data with the process.

- The steps prior to this one are more mathematically complicated but easily automated.

- The last step requires interpretation.

3.4.2 Modeling non-normally distributed variables

Knowing that the maximum likelihood chi-square statistic is a method that assumes multi-normality of the variables, but since we are working on a classification model based on a sensitive attribute that is a discrete

variable, we need to find a solution to this problem.
The most common solution used in most SEM algorithms is a statistic based on least squares generalization, elliptic estimators as well as distribution-free estimators.

3.4.3 Simulation

In this simulation we will create a causal graph that is based on credit scoring and is shown in Figure 3.5:

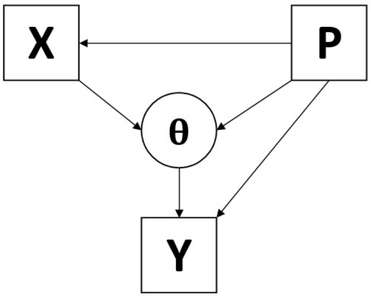


Figure 3.5: Model to simulate

- P: Sensitive attribute (example gender)
- X: endogenous attribute (example income)
- θ: latent non-measured variable
- Y: decision binary attribute (credit score)

where we will consider P as a binomial distribution X is based of on P with a normal error, and θ is a linear combination of X and P that can not be measured as for Y will be affected by θ.
Based on the causal graph in Figure 3.5:

$P = B(1, p)$

$$
\begin{aligned}
X &= a_1 P + N(0, \sigma_1) \\
\theta &= a_2 X + a_3 P \\
Y &= \frac{e^{\theta}}{(1 + e^{\theta})} + \frac{e^{P}}{(1 + e^{P})}
\end{aligned}
$$

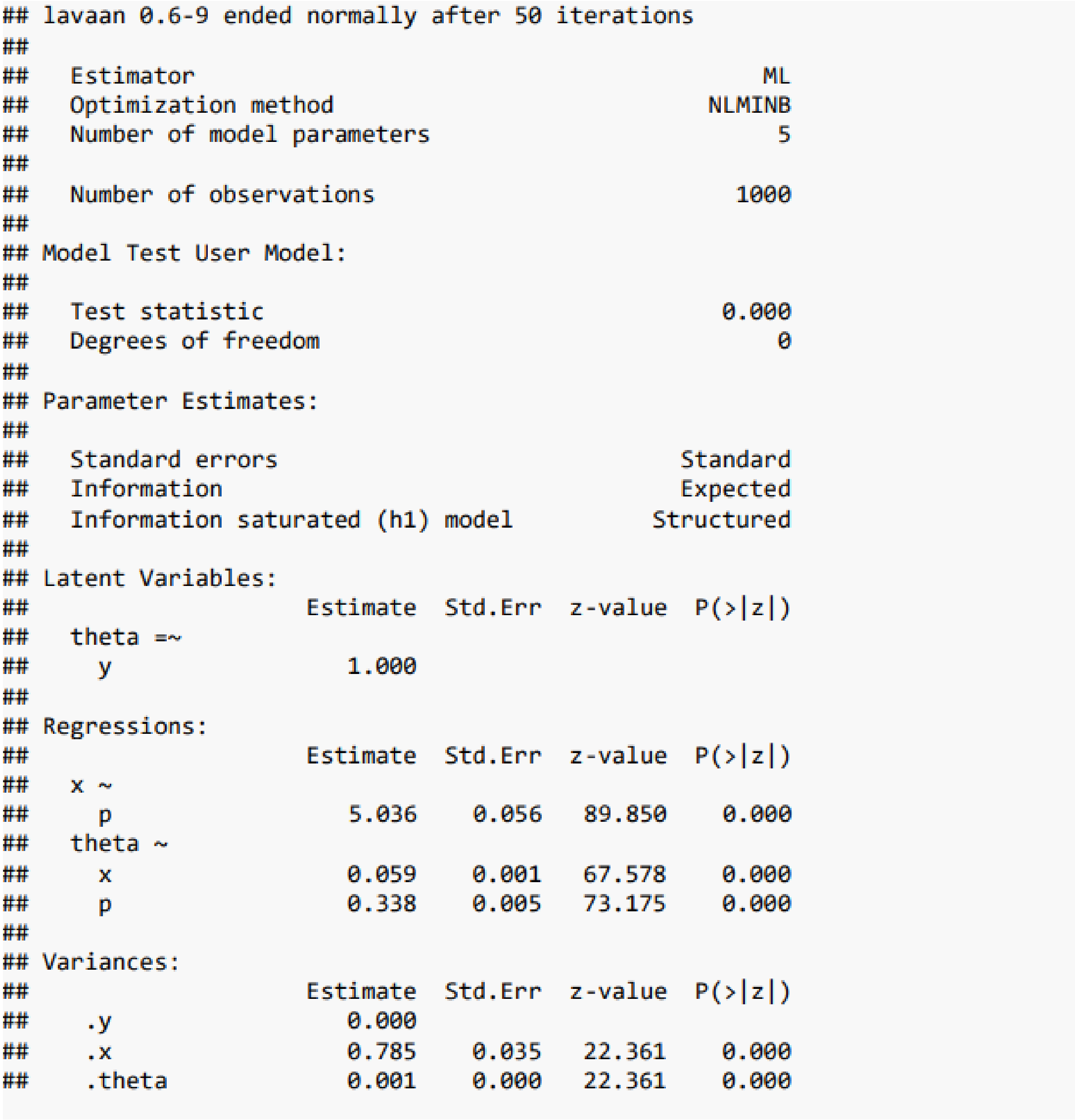
This simulation will be done in R see anexe but can also be done using tetrad.
Next step is to use the previous method to test the causal relation, this method as privously indicated can be automated, so using R we gonna test the model, see anexe and for result, see Figure 3.6.

We can also use tetrad to test our simulated data.
When using tetrad's algorithm what is the most powerful part is to add knowledge, and then be able to search from the same data sets multiple outcome based on this knowledge.

- Testing with tetrad without knowledge:

  Based on figure 3.7 we can see that the tetrad algorithm was able to find the causal indirected relationships between the simulated variables.

- Now we will test with tetrad while adding multiple tiers based on this data and we will find a directed relation 3.8:
  Tier 1: P
  Tier 2: X
  Tier 3: θ
  Tier 4: Y



```
## lavaan 0.6-9 ended normally after 50 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         5
##
##   Number of observations                          1000
##
## Model Test User Model:
##
##   Test statistic                                0.000
##   Degrees of freedom                                0
##
## Parameter Estimates:
##
##   Standard errors                            Standard
##   Information                                Expected
##   Information saturated (h1) model         Structured
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   theta =~
##     y                 1.000
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   x ~
##     p                 5.036    0.056   89.850    0.000
##   theta ~
##     x                 0.059    0.001   67.578    0.000
##     p                 0.338    0.005   73.175    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .y                 0.000
##    .x                 0.785    0.035   22.361    0.000
##    .theta             0.001    0.000   22.361    0.000
```

Figure 3.6: R test Output

3.5 Using Path specific effect to quantify causality

3.5.1 Causal inference

Definition 6 ***Structural causal model***

$\mathcal{M} =< U, V, F, P(U) >$ is a structural causal model with:

- ***U**: exogenous variables determined by factors outside the model.*
- ***P(U)**: probability distribution defined over U.*
- ***V**: endogenous variables determined by variables in $U \cup \mathbf{V}$.*

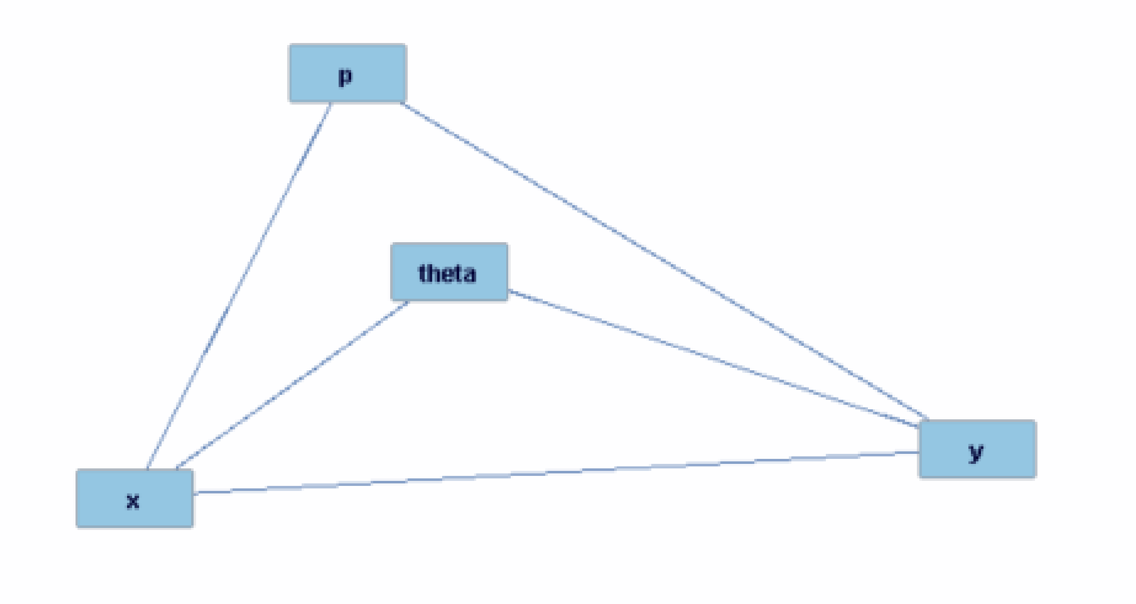


Figure 3.7: Tetrad test Output

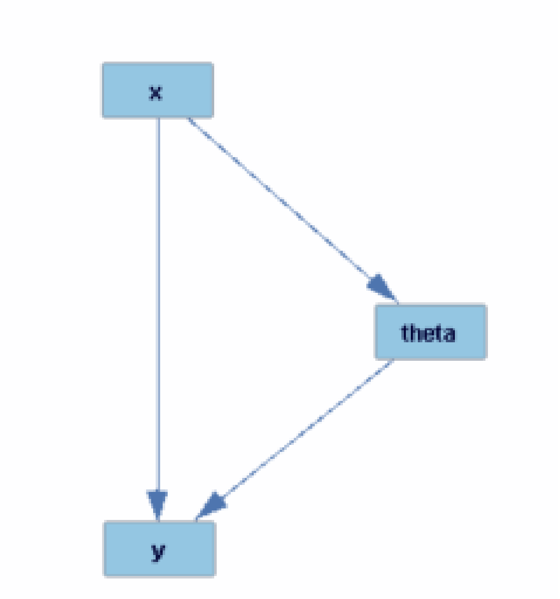


Figure 3.8: Tetrad test Output with knowlege

- ***F**: set of functions **from** $U \cup V$ **to***

$$\forall v \in V \; \exists f_v \in F \; that \; v = f_v(pa_v, u_v) \tag{3.1}$$

with

  - pa_v *one realisation of endogenous variables* $Pa_v \in V|v$ *that* *determine directly* v.

- u_v *realisation of exogenous variables that determine directly* v.

This causal models are associated with a directed graph.

Definition 7 *Markovian model: a causal model where all exogenous variants are independent of each other.*

Definition 8 *Semi-Markovian model:a causal model where all exogenous variants are non-independent..*

Property 1 *A directed acyclic causal graph is associated with a Markovian model.*
An acyclic causal graph with dotted bi-directed edges is associated with a Markovian model.

- *Directed Acyclic Graph (DAG)*

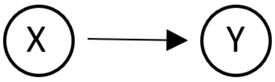


Figure 3.9: Causal graphs of a Markovian model

- *Acyclic Graph (AG)*

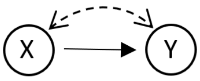


Figure 3.10: Causal graphs of a semi-Markovian model

With a Markovian model $P(X)$ joint distribution is decomposed into conditional probability.

$$P(x) = \prod_{x_i \in X} P(x_i | pa_{Xi}) \tag{3.2}$$

whith $P(x_i|pa_{Xi})$: conditional probability associated with X_i

3.5.1.1 Intervention to a causal model

Consider a causal model $\mathcal{M}$ and a graph $\mathcal{G}$ related to $\mathcal{M}$, and X an endogenous variable.
We define $d_0(\mathbf{X} = \mathbf{x})$ the intervention that will forces the value of $\mathbf{X}$ to become $\mathbf{x}$.

Then after this intervention:

- The original equation 3.1 $X = f(Pa_x, U_X)$ will be substituted with $X = x, \forall X \in \mathbf{X}$
- The causal model $\mathcal{M}$ becomes a sub-model $\mathcal{M}_{\mathbf{x}}$
- $\mathcal{G}_{\mathbf{x}}$, causal graph of $\mathcal{M}_{\mathbf{x}}$,is a variant of $\mathcal{G}$ with all the edges coming toward $\mathbf{X}$ are deleted and $\mathbf{X}$ is set as $\mathbf{x}$
- $\mathbf{Y}_{\mathbf{x}}$ in the model $\mathcal{M}_{\mathbf{x}}$ is the post-interventional variant of $\mathbf{Y} \in \mathbf{V}|\mathbf{X}$ affected by the intervention
- $P(\mathbf{Y} = \mathbf{y}|d_0(\mathbf{X} = \mathbf{x})) = P(\mathbf{y}|d_0(\mathbf{x})) = P(\mathbf{y}_{\mathbf{x}})$ is the distribution of $\mathbf{Y}_{\mathbf{x}}$ called post-intervention distribution of $\mathbf{Y}$ under $d_0(\mathbf{x})$

3.5.1.2 Causal inference

Definition 9 ***Causal inference****: is the method of estimating a causal quantities.*

- ***Post-interventional**, from data and causal graph:*

  *For a **Markovian model**:*

$$P(\mathbf{y}|d_0(\mathbf{x})) = \prod_{Y \in \boldsymbol{Y}} P(y|\boldsymbol{pa}_Y)\delta_{\boldsymbol{X}=\mathbf{x}} \tag{3.3}$$

  with $\delta_{\boldsymbol{X}=\mathbf{x}}$: affecting $\boldsymbol{X}$ with the corresponding

  For a single variable Y with intervention X:

$$P(y|d_0(x)) = \sum_{\mathbf{V}'} \prod_{V \in \boldsymbol{V}|\{X\}} P(v|\boldsymbol{pa}_V)\delta_{X=x} \tag{3.4}$$

  with $\boldsymbol{V}' = \boldsymbol{V}|\{X, Y\}$

Definition 10 ***Identifiable Causal quantity:***

A causal quantity is identifiable if its estimator is unique knowing observational data compatible with any causal model.

Definition 11 ***Total causal effect***

A value that measure the effect of a shift of X from x_1 to x_2 in $Y = y$

$$TE(x_2, x_1) = P(y|d_o(x_2)) - P(y|d_o(x_1)) \tag{3.5}$$

3.5.2 Causal discrimination

Consider a Markovian model.

Definition 12 ***Direct discrimination***

Causal effect transmitted directly from the sensitive attribute to the decision.

Definition 13 ***Indirect discrimination:***

Causal effect transmitted through the non-direct paths from the sensitive attribute to the decision.

To be able to quantify this two type of discrimination the technique of path-specific effect was adopted in literature [Wu(2020)]

Note that unidentifiability of the path-specific effect can occur and is when path-specific effect is not computable from the observational data this only occurs in the case of indirect paths.

3.5.2.1 Computing techniques for path specific effect

With:

- The effect of X on Y with the intervention transmitted through π
- The effect of X on Y without the intervention transmitted through $\bar{\pi}$
- $P(y|d_o(x_2|\pi, x_1|\bar{\pi})$ distribution of Y after the intervention on X from x_1 to x_2

Definition 14 ***Path-specific effect****: or $\pi-$specific effect measures the effect of changing X from x_1 to x_2 on Y:*

$$PSE_\pi(x_2, x_1) = P(y|d_o(x_2|\pi, x_1|\bar{\pi})) - P(y|d_o(x_1)) \tag{3.6}$$

Property 2 *$PSE_\pi(x_2, x_1)$ is identifiable $\iff$ P (y / $d_o(x_2|\pi, x1|\bar{\pi}))$ is computable*

Definition 15 ***Recanting Witness Criterion****:*

Consider:

- *π path from X to Y*
- *W a node in $\mathcal{G}$*

With:

1. *$\exists$ a segment of a path in π that is a path from X to W*
2. *$\exists$ a segment of a path in π that is a path from W to Y*
3. *$\exists$ a path from W to Y that is not a segment of a path in π*

Then

*For the π-specific effect the **recanting witness criterion** is satisfied with W witness.*

And the causal graph is called "kite".

Theorem 2 *$P(y|d_o(x_2|\pi, x1|\bar{\pi}))$ is computable $\iff$ the recanting witness criterion is not satisfied*

Computing $P(y|d_o(x_2|\pi, x1|\bar{\pi}))$ steps:

1. Express $P(y|do(x_1))$ as the truncated factorization formula based of 3.4.
2. Divide children of X exept Y on to :
   - S_π={S child of X such as path from X to S is a segment from π}
   - $\bar{S}_\pi$ ={S child of X, that is not included in any path or the path from X to S is a segment of a path not in π}

   with $Ch_X|Y = S_\pi \cup \bar{S}_\pi$ and $S_\pi \cap \bar{S}_\pi = \emptyset$

3. Replace x_1 with x_2 for the terms corresponding to nodes in S_π, and keep values x_1 unchanged for the terms corresponding to nodes in $\bar{S}_\pi$

3.5.2.2 Modeling Direct/Indirect Discrimination as Path-Specific Effects

$\mathcal{D}$: data-set with attributes $\mathbf{V}$ that include:

- The sensitive attributes
- The decision,
- The non-sensitive attributes

Assumption: The data contains **R** redlining attributes i.e. non sensitive that can't be justified objectively when used in the decision making process.

Notation:

- $C \in \{c^-, c^+\}$: sensitive attribute
- $E \in \{e^-(\textit{negative decision}), e^+(\textit{positive decision})\}$: the decision

Assumption:

(a) C has no parent in a causal graph $\mathcal{G}$

(b) E has no child in a causal graph $\mathcal{G}$

(c) The causal graph $\mathcal{G}$ can be built to represent correctly the data-set $\mathcal{D}$

Considering these assumptions we can model causal effects transmitted along different paths.

I. **Direct Discrimination from C to E**

Consider:

π_d: the path set that contains only $C \to E$

Then:

$PSE_{\pi_d}(c^+, c^-)$: π_d-specific effect the causal effect caused by changing C from c^- to c^+, which means the expected change in decisions for individuals of group c^- if this individual were from group c^+ according to the decision maker, and **can be considered as a measure of direct discrimination.**

II. **Indirect discrimination**

Consider:

π_i: the path set that contains all the causal paths from C to E that path through R.

Then

$PSE_{\pi_i}(c^+, c^-)$: π_i-specific effect the causal effect transmitted through indirect paths, which means the expected change in decisions for individuals of group c^- if there redlining attributes were altered as if they are part of the group c^+ according to the decision maker, and **can be considered as a measure of indirect discrimination.**

Criterion for discrimination

- Direct discrimination exists if $PSE_{\pi_d}(c^+, c^-) > \tau$ or $PSE_{\pi_d}(c^-, c^+) > \tau$
- Indirect discrimination exists if $PSE_{\pi_i}(c^+, c^-) > \tau$ or $PSE_{\pi_i}(c^-, c^+) > \tau$

whith $\tau > 0$ threshold defined by the user depending on laws.

Theorem 3 ***Calculating the PSE***

$$PSE_{\pi_d}(c^+, c^-) = \sum_{Q}(P(e^+|c^+, \mathbf{q})P(\mathbf{q}|c^-)) - P(e^+|c^-) \tag{3.7}$$

with $Q = Pa_E|\{C\}$

If $\boldsymbol{S}_{\pi_i} \cap \bar{\boldsymbol{S}}_{\pi_i} = \emptyset$ ***Then***

$$PSE_{\pi_i}(c^+, c^-) = \sum_{V'}(P(e^+|c^-, \mathbf{q}) \prod_{G \in S_{\pi_i}} P(g|c^+, pa_G \setminus \{C\}) \\ \times \prod_{H \in \bar{S}_{\pi_i} \setminus \{E\}} P(h|c^-, pa_H \setminus \{C\}) \prod_{O \in (V)} P(o|pa_O)) - P(e^+|c^-) \tag{3.8}$$

where $\boldsymbol{V'} = \boldsymbol{V} \setminus \{C, E\}$

Proposition 1 *If* π *contains all causal paths from* C *to* E *except direct edge* $C \rightarrow E$*: Then:*

$$PSE_{\pi}(c^+, c^-) = TE(c^+, c^-) = P(e^+|c^+) - P(e^+|c^-) \tag{3.9}$$

This proposition is not valid for all path π_i and π_d

$$PSE_{\pi_d}(c^+, c^-) + PSE_{\pi_i}(c^+, c^-) = PSE_{\pi_d \cup \pi_i} \tag{3.10}$$

3.5.2.3 Algorithm used to discover discrimination

Based on the criterion for discrimination the path specific effect Discrimination Discovery (PSE-DD) shown in algorithm 1 is built.

Step 1: Build a causal graph from the historical data-set.

Step 2: Compute $PSE_d(.)$ based on 3.7.

Step 3: Compute $PSE_i\ (.)$ based on 3.8.

A critical part is to separate the indirect paths between the $\mathbf{S}_{\pi i}$ and the $\bar{\mathbf{S}}_{\pi_i}$. An uncomplicated way to achieve this goal is to:

- Find every path in π_i
- For every children S of a sensitive attribute C verify whether the path $C \rightarrow S$ is contained in any path π_i or not

Property 3 *The separation method used is based on the paths from S to E that passes by $\boldsymbol{R}$:*

- $S \in \boldsymbol{S}_{\pi_i} \iff \exists$ *a path $S \rightarrow E$ that passes through $\boldsymbol{R}$*
- $S \in \bar{\boldsymbol{S}}_{\pi_i} \iff$ *there is not an existing path $S \rightarrow E$ that passes through $\boldsymbol{R}$*

Property 4 *This method can be complex due to the exponential number of paths between two nodes in a DAG (direct acyclic graph).*

Algorithm 1 *PSE-DD*

```
    Input: $\mathcal{D}$, $C$, $E$, $\boldsymbol{R}$, and $\tau$
    Output: $judge_d$ (direct discrimination), $judge_i$ (indirect discrimination),
    $\mathcal{G}$: buildCausalNetwork( $\mathcal{D}$)
    $judge_d = judge_i = false$
    Use 3.7 to find $PSE_{\pi_d}(\cdot)$
    If $PSE_{\pi_d}(c^+, c^-) > \tau || PSE_{\pi_d}(c^-, c^+) > \tau$
    Then:$judge_d = true$
    Call subroutine $[\boldsymbol{S}_{\pi_i}, \bar{\boldsymbol{S}}_{\pi_i}] = DivideChildren(\mathcal{G}, C, E, \boldsymbol{R})$
    If $\boldsymbol{S}_{\pi_i} \cap \bar{\boldsymbol{S}}_{\pi_i} = \emptyset$
    Then $judge_i = unknown$ Return $[judge_d, judge_i]$
    Compute $PSE_{\pi_i}(\cdot)$ according to 3.8
    If $PSE_{\pi_i}(c^+, c^-) > \tau || PSE_{\pi_i}(c^-, c^+) > \tau$
    Then: $judge_i = true$
    Return $[judge_d, judge_i]$;
```

A subroutine presented in algorithm 2 held finding $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$, through checking the existence of a node $R \in \mathbf{R}$, such as $R \to S \to E$.

Property 5 *All the involved nodes can be obtained if we pass the network starting with C and with $O(|\epsilon|)$ time of travel $\Longrightarrow$ The complexity of the subroutine is $O(|V^2| + |\epsilon|)$*

Algorithm 2 *: Subroutine DivideChildren*

```
    Input: $\mathcal{G}$, $C$, $E$, $\mathbf{R}$
    Output: $\mathbf{S}_{\pi_i}$

If $R \in De_S \cup \{S\} E \in De_R$ Then $\mathbf{S}_{\pi_i} = \mathbf{S}_{\pi_i} \cup \{S\}$;
Else $\bar{\mathbf{S}}\pi_i = \bar{\mathbf{S}}_{\pi_i} \cup \{S\}$;

Return$[\mathbf{S}_{\pi_i}, \bar{\mathbf{S}}_{\pi_i}]$.
```

3.5.2.4 Algorithm used to remove discrimination

I. **Naive Discrimination Removal Approach**: Delete sensitive attribute.

   By doing that we are removing valuable information, and solving only direct discrimination.

II. **Path-Specific Effect based Discrimination Removal (PSE-DR)** :

This method modify the causal and generate new data:

Modify $P(e|\mathbf{pa}_E)$ into $P'(e|\mathbf{pa}_E)$ such as direct and indirect discrimination are bellow the threshold.

This can be achieved by maximizing the utility for the modified data-set.

With:

$P(v)$, respectively $P'(v)$: joint distribution of the original graph respectively modified graph computed according to 3.2 with $P(e|pa_E)$ respectively $P'(e|pa_E)$,

Maximizing the utility for a modified data-set $\Longrightarrow$minimize $\sum_v (P'(v) - P(v))^2$

Under:

- PSE for direct and indirect discrimination are lower or equal to the threshold τ computed using 3.7 and 3.8
- $\forall pa_E$, $P'(e^+|pa_E) + P'(e^-|pa_E) = 1$,

This optimization problem is achieved with the solution of a quadratic programming problem. And afterward the modified data set is generated based on our new causal graph.

Algorithm 3 *PSE-DR*

```
    **Input**: D, C, E, R,and $\tau$
    **Output:**D' modified data
    $[judge_d, judge_i] = PSE - DD(\mathcal{D}, C, E, \boldsymbol{R}, \tau)$
    **If** $[judge_d, judge_i] == [false, false]$ **thenReturn** $\mathcal{D}$
    $\mathcal{G}$ = buildCausalNetwork($\mathcal{D}$)
    **If** $judge_i == unknown$ **Then** Call subroutine GraphPreprocess;
    Obtain the modified $P(e|\boldsymbol{pa}_E)$ by solving the quadratic programming problem;
    Calculate $P'(v)$ according to 3.2 using the modified $P(e|\boldsymbol{pa}_E)$;
    Generate D' based on $P'(v)$
    Return $D'$
```

A pseudo-code that is added to the algorithm 3 after building to the causal graph G is called a GraphPreprocess, and is the procedure of modifying the causal graph to remove the “kite” pattern by:

$\forall S \in \mathbf{S}_{\pi i} \cap \bar{\mathbf{S}}_{\pi_i}$ cut all causal graph through $\mathbf{R}$ then S will be $\notin \mathbf{S}_{\pi i} \Rightarrow \mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} = \emptyset$

Algorithm 4 *subroutine GraphPreprocess*

```
    Input: G, C, E, $\boldsymbol{R}$
     for every $S \in \boldsymbol{S}_{\pi_i} \cap \bar{\boldsymbol{S}}_{\pi_i}$:
      for every $Q \in \boldsymbol{Pa}_E$:
        for every $R \in \boldsymbol{R}$:
          If $R \in \boldsymbol{De}_S$ & $Q \in \boldsymbol{De}_S$ then
           Remove $Q \rightarrow E$ from $\mathcal{G}$
           Break
```

3.5.3 Causal discrimination in an Unidentifiable Situation

In this subsection the main idea is to define an upper and lower bounds to be able to discover an unidentifiable indirect discrimination.

- **If** the upper bound $< \tau$, **then** there is no indirect discrimination,
- **If** the lower bound $> \tau$, **then** there is indirect discrimination,

This method is known as the refined removal algorithm, and preserve the data utility as much as PSE-DR.

Notation:

- $P(\mathrm{y_x}) \triangleq P(\mathbf{Y}_\mathrm{x} = \mathrm{y}) \triangleq P(\mathbf{y}|do(\mathrm{x}))$

  With: $\mathbf{Y}_\mathrm{x}$: $\mathbf{Y}$ under intervention $do(\mathrm{x})$ or in other terms the $\mathbf{Y}_\mathrm{x}$ is interpreted as the counterfactual statement "the value of $\mathbf{Y}$ if $\mathbf{X}$ was $\mathbf{x}$" and it's worth noting that if all $\mathbf{U}$ exogenous variable are known $\mathbf{Y}_\mathrm{x}$ are fixed variables.

- $\mathbf{Y}_\mathrm{x}(\mathbf{u})$: $\mathbf{Y}_\mathrm{x}$ in the context of $\mathbf{U}$

Property 6 *Regarding the counterfactual statement*

1. $Y_{\boldsymbol{Pa}_Y} \perp\!\!\!\perp$ *of the counterfactual statements of all Y 's non-descendants*
2. $P(y_{pa_{\boldsymbol{Y}}}) = P(y|\boldsymbol{pa}_Y)$
3. $P(\mathrm{y}_{pa_\mathrm{Y},\mathrm{x}}) = P(\boldsymbol{y}_{pa_{\boldsymbol{Y}}})$ *with* X *endogenous disjoint from* $\{\mathrm{Y}, \mathrm{Pa_Y}\}$
4. $\boldsymbol{Z_x}(\boldsymbol{u}) = \boldsymbol{z} \Rightarrow \boldsymbol{Y_x}(\boldsymbol{u}) = \boldsymbol{Y_{x,z}}(\boldsymbol{u})$ *with* $\boldsymbol{X, Y, Z}$ *endogenous.*

Formulas:

$$P(\mathbf{y_x}) = \sum_{\mathbf{u}:\mathbf{Y_x}(\mathbf{u})=\mathbf{y}} P(\mathbf{u}) \tag{3.11}$$

Joint distribution of multiple counterfactual statements

$$P(\mathbf{y_x}, \mathbf{y}'_{\mathbf{x}'}) = P(\mathbf{Y_x} = \mathbf{y}, \mathbf{Y}_{\mathbf{x}'} = \mathbf{y}') = \sum_{\{\mathbf{u}:\mathbf{Y_x}(\mathbf{u})=\mathbf{y}, \mathbf{Y}'_\mathbf{x}(\mathbf{u})=\mathbf{y}'\}} P(\mathbf{u})$$

(3.12)

If $\mathbf{x} \neq \mathbf{x}'$ then $\mathbf{Y_x}$ and $\mathbf{Y'_x}$ cannot be computed together and as a result $P(\mathbf{y_x}, \mathbf{y'_{x'}})$ is not identifiable, which leads to the unidentifiability of the path-specific effect satisfying the recanting witness criterion.

In that case: $P(\mathbf{y_x}, \mathbf{y'}_{x'})$ is bounded by

$$P(\mathbf{y_x}) = \sum_{\mathbf{y}'} P(\mathbf{y_x}, \mathbf{y'_{x'}})$$

Probability of $E = e^+$ under the mediation of C from c^- and c^+ transmitted through the path π_i: $P(e^+|do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})$

Computation techniques:

Denote Y_{c^+} respectively E_{c^+}: value of Y and E after intervention c^- to c^+ through the path π_i, in the sense that if there is no path from C to Y then the intervention on C will not affect Y and the value of Y will stay the same as if $C = c^-$.

To get the value of Y_{c^+} for every ancestor W of Y, get the value of the one affected by the intervention (i.e. path $W \to Y$ part of π_i) and the ancestor that the intervention doesn't affect (i.e. $W \to Y$ segment of π_i).

Note that when W is part of two edges one part of the path π_i and one that is not a part of π_i, get both the value of W affected by the intervention and not affected by the intervention.

Discerning between the two edges types of W:

W_{c^+}: value former to intervention.

W_{c^-}: value of W after intervention.

W in that case is a **witness variable/node**.

This analysis for W witness variable between the two sets of realisation is not the case for Y the non witness variable where we only consider Y_{c^+}

Property 7 *Consider:*

$\boldsymbol{X},\ \boldsymbol{W},\ \boldsymbol{Y}$ *endogenous variable,*

$\boldsymbol{W}$ *a witness variable,*

x, x' *realisation of* X

w, w' *realisation of* W

For every π*-specific effect of* $\boldsymbol{X}$

$$W_x(u) = w, W'_x(u) = w' \Rightarrow Y_x(\boldsymbol{u}) = Y_{x,w^*}(\boldsymbol{u})$$

with

$$w^* = \begin{cases} w, & \textit{if the path } W \to Y \textit{is part of } \pi \\ w', & \textit{otherwise} \end{cases}$$

Property 8 *Consider:*

- Y*: an endogenous variable*
- π_i*-specific effect* PSE_{π_i} *for* Y

Then, the realisation of the parents of Y *is:*

- pa_Y^+ *which means if witness node* W *or* $C \subset Pa_Y$*, its value will be*
$$\begin{cases} w^+ \textit{or } c^+, & \textit{if the path } W \to Y \textit{is part of } \pi_i \\ w^- \textit{or } c^-, & \textit{otherwise} \end{cases}$$
- pa_Y^- *which means if witness node* W *or* $C \subset Pa_Y$*, its value will be* w^- *or* c^-

  And:

  If Y *is not a witness variable* ***Then*** $\begin{cases} P(y_{c^+}, ...), & \textit{if } Y \textit{ is part of } W \to Y \textit{is part} \\ w^- \textit{or } c^-, & \textit{otherwise} \end{cases}$

  Else $P(y_{c^+}, ...) = P(y_{pa_Y^+}, ...)$ *and* $P(y_{c^-}, ...) = P(y_{pa_Y^-}, ...)$

...: all the other variables.

Theorem 4 *Under the recanting witness criterion:*

$$P(e^+|do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) = \sum_{a,b,w^+,w^-} P(e^+|c^-, \boldsymbol{q}) \prod_{A \in \boldsymbol{A}} P(a|pa_A^+) \prod_{B \in \boldsymbol{B}} P(b|pa_B^+) \prod_{W \in \boldsymbol{W}} P(w^+_{pa_W^+}$$

Theorem 5

- *Upper Bound of $P(e^+|do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$:*

$$\sum_{a_2,w^-} \max_{a_1,w^+} \{P(e^+|c^-, \boldsymbol{q})\} \prod_{A \in \boldsymbol{A}_2} P(a|\boldsymbol{pa}_A^+) \prod_{B \in \boldsymbol{B}} P(b|\boldsymbol{pa}_B^-) \prod_{W \in \boldsymbol{W}} P(w^-|\boldsymbol{pa}_W^-) \tag{3.13}$$

- *Lower Bound*

$$\sum_{a_2,w^-} \min_{a_1,w^+} \{P(e^+|c^-, \boldsymbol{q})\} \prod_{A \in \boldsymbol{A}_2} P(a|\boldsymbol{pa}_A^+) \prod_{B \in \boldsymbol{B}} P(b|\boldsymbol{pa}_B^-) \prod_{W \in \boldsymbol{W}} P(w^-|\boldsymbol{pa}_W^-) \tag{3.14}$$

3.5.3.1 Algorithm used to discover discrimination

The algorithm used in that situation, is based on the algorithm 1 PSE-DD with some slight changes.

Algorithm 5 *PSE-DD**

```
    Input: $\mathcal{D}$, C,E,R,and $\tau$
    Output: $judge_d$ (direct discrimination), $judge_i$ (indirect discrimination),
    $\mathcal{G}$: buildCausalNetwork( $\mathcal{D}$)
    $judge_d = judge_i = false$
    Use 3.7 to find $SE_{\pi_d}(.)$
    If $SE_{\pi_d}(c^+, c^-) > \tau || SE_{\pi_d}(c^-, c^+) > \tau$
    Then:$judge_d = true$
```

```
Call subroutine $[\boldsymbol{S}_{\pi_i}, \bar{\boldsymbol{S}}_{\pi_i}] = DivideChildren(\mathcal{G}, C, E, \boldsymbol{R})$
If $\boldsymbol{S}_{\pi_i} \cap \bar{\boldsymbol{S}}_{\pi_i} = \emptyset$
Then
    Use 3.13 and 3.14 to find:
    $lb(SE_{\pi_i}(c^+, c^-)$, $ub(SE_{\pi_i}(c^+, c^-)$, $lb(SE_{\pi_i}(c^-, c^+)$ and $ub(SE_{\pi_i}(c^-, c^+)$
    If $ub(SE_{\pi_i}(c^+, c^-) \leq \tau \;\; ub(SE_{\pi_i}(c^-, c^+) \leq \tau$
    Then $judge_i = false$
    Else
        If $lb(SE_{\pi_i}(c^+, c^-) > \tau || lb(SE_{\pi_i}(c^-, c^+) > \tau$
        Then $judge_i = true$
        Else $judge_i = unknown$
Return $[judge_d, judge_i]$
Use 3.8 to find $SE_{\pi_i}(.)$
If $SE_{\pi_i}(c^+, c^-) > \tau || SE_{\pi_i}(c^-, c^+) > \tau$
Then:$judge_i = true$
Return [$judge_d, judge_i$];
```

3.5.3.2 Algorithm used to remove discrimination

This algorithm is also based on the algorithm used in the identified situation, with SE being replaced by its upper bound.

Algorithm 6 *PSE-DR**

```
Input: D, C, E, R, and $\tau$
Output:D*$modifieddata$
$[judge_d, judge_i] = PSE - DD(\mathcal{D}, C, E, \boldsymbol{R}, \tau)$
    If $[judge_d, judge_i] == [false, false]$
    Then Return $\mathcal{D}$
```

```
$\mathcal{G}$ = buildCausalNetwork($\mathcal{D}$)
    If $judge_i$ == unknown
    Then Solve the the adjust quadratic programming problem to find the modified $P(e|\boldsymbol{pa}_E)$
    Else Solve the the quadratic programming problem to find the modified $P(e|\boldsymbol{pa}_E)$
Calculate $P^*(v)$ using the modified $P(e|\boldsymbol{pa}_E)$, Generate $D^*$
Return $D^*$
```

Note that one of the feasible solutions of the adjusted programming problem is the the modified $P(e|\mathbf{pa}_E)$ obtained from the quadratic programming after performing GraphPreprocess. Which leads to consider that algorithm 6 will perform, at least as good as algorithm 1.

3.5.4 Experimentation

Using path specific effect to test discrimination on the data set has been used on the Adult data set in the "Achieving Causal Fairness in Machine Learning" by Yongkai Wu [Wu(2020)] to summarize these data set it is a data set extracted from the census data containing 48842 variables and include 11 attributes including age, education, sex, occupation, income, marital status, native country, race, relationship, hour, occupation and work-class.
According to the article, if we take the sex as sensitive attribute, and income as the decision and marital status as our redlining attribute and by computing the path specific effect we can have that $SE_{\pi_d} = 0.025$ and $SE_{\pi_i} = 0.175$ with π_d the direct path between sex and income and π_i the path between sex and income that contains the marital status, thus if take

the threshold $\tau = 0.05$ we can see no direct discrimination and significant direct discrimination against women.

This result is an important result because if we drive our insurance premium from the income and other variables our result will be affected by the unfairness detected in this data base.

CONCLUSION

This report treated the fairness in insurance data and credit scoring. Before discussing the different notion of fairness, we gave detailed information about the reasons that motivate us to choose this subject from different legislation to previous proofs of inequality and the causes of unfairness.
To then move on to the different fairness notions and tested data based on them.
To finally move to causality where we showcased different way we can calculate causal dependency based already existing algorithm. Starting with d-separation moving to Structural Equation Model and ending with path specific effect.
The question at the end will be how can we develop the causal detection method explained in order to serve fairness detection and how this model can be compered between each others.

ANNEXE A
CODING TO TEST

Loading German Data

```
 german=read.csv("german.csv")
 DGerman=german[which
 (german$personal_status_sex!="female : non-single or male : single"),]
 n=nrow(DGerman)
 Genre=ifelse(DGerman$personal_status_sex=="female : single","Female","Male")
```

- Fairness through unawareness

  Creating the regressions:

```
Reg=glm(formula=factor(credit_risk)~. ,  family = binomial ( link ="logit"),
data = DGerman)
Predic=predict(Reg,newdata=DGerman,type="response")
```

```
Reg1=glm(formula=factor(credit_risk)~. ,  family = binomial ( link ="logit"),
data =subset(DGerman, select =-c(personal_status_sex) ))
Predic1=predict(Reg1,newdata=subset(DGerman,select=-c(personal_status_sex)),
type="response")
```

Boostrap.

```
ns=200
Pre = matrix(NA,ns,1)
for(i in 1:ns){
    idx_s = sample(1:n,size=n, replace=TRUE)
    Reg=glm(formula=factor(credit_risk)~. ,  family = binomial ( link ="logit"), data = DGerman[idx_s,])
    Predic=predict(Reg,newdata=DGerman[idx_s,],type="response")
    Reg_s1 = glm(formula=factor(credit_risk)~. ,  family = binomial ( link ="logit"), data = subset(DGerman[idx_s,], select = -c(personal_status_sex) ))
    Predic1=predict(Reg1,newdata=subset(DGerman[idx_s,], select = -c(personal_status_sex) ),type="response")
    Pre[i,] = sum((Predic1-Predic)^2)
    }
plot(density(Pre))
abline(v=mean(Pre),lty=2,col="red")
```

- fairness

First, we calculate the number of privileged and nonprivelegier group of individuals

```

Sum_genre=as.matrix(summary(DGerman$personal_status_sex))
(n_male=Sum_genre[3]+Sum_genre[4])
(n_fmale=Sum_genre[1]+Sum_genre[2])
```

Then we calculate the number of non-privileged individuals with a favorable result and the number of privileged individuals with a favorable result we consider the result coming from Reg:

```
Predic_0=ifelse(Predic>0.5,"bad","good")
DGerman_1=data.frame(DGerman,Genre,Predic,Predic_0)
FavorableGroup=DGerman_1[which(DGerman_1$Predic_0=="good"),]
Sum_genreFG=as.matrix(summary(FavorableGroup$personal_status_sex))
(n_male_FG=Sum_genreFG[3]+Sum_genreFG[4])
(n_fmale_FG=Sum_genreFG[1]+Sum_genreFG[2])
```

And compare the probability for an individual to be assigned the favourable outcome for each group.

```
(P_0_Yhat1=n_fmale_FG/n_fmale)
```

```r
(P_1_Yhat1=n_male_FG/n_male)
P_0_Yhat1==P_1_Yhat1
```

Boostrap method.

```r
ns=200
P = matrix(NA,ns,2)
for(i in 1:ns){
    idx_s = sample(1:n,size=n, replace=TRUE)
    Reg_s=glm(formula=credit_risk~., family=binomial(link ="logit"),data= DGerman[idx_s,])
    S_s = predict(Reg_s,type="response")
    P[i,] = c(mean(S_s[DGerman[idx_s,"personal_status_sex"]=="female :
single"]),mean(S_s[DGerman[idx_s,"personal_status_sex"]!="female :
single"]))
}
plot(density(P[,2]-P[,1]))
abline(v=0,lty=2,col="red")
```

Epsilon group fairness

```r
epsilon=0.05
abs(P_0_Yhat1-P_1_Yhat1)<epsilon
```

Normalised Difference

```r
d_max=min(nrow(FavorableGroup)/n_male,(n-nrow(FavorableGroup))/n_fmale)
(delta1=(P_0_Yhat1-P_1_Yhat1)/d_max)
```

- Predictive parity We have to compare the PPV for both groups:

```r
TP_ind=(DGerman_1[which(DGerman_1$credit_risk=="good" &
DGerman_1$Predic_0=="good"),])
FP_ind=(DGerman_1[which(DGerman_1$credit_risk=="bad" &
DGerman_1$Predic_0=="good"),])
FN_ind=(DGerman_1[which(DGerman_1$credit_risk=="good" &
DGerman_1$Predic_0=="bad"),])
TN_ind=(DGerman_1[which(DGerman_1$credit_risk=="bad" &
DGerman_1$Predic_0=="bad"),])
```

```r
Sum_genrePvA=cbind(as.matrix(summary(TP_ind$personal_status_sex)),
as.matrix(summary(FP_ind$personal_status_sex)),
Sum_FNgenre=as.matrix(summary(FN_ind$personal_status_sex)),
Sum_TNgenre=as.matrix(summary(TN_ind$personal_status_sex)))
colnames(Sum_genrePvA)=c("TP","FP","FN","TN")
Sum_genrePvA
PPV_male=(Sum_genrePvA["male : divorced/separated","TP"]+
Sum_genrePvA["male : married/widowed","TP"])/
(Sum_genrePvA["male : divorced/separated","TP"]+
Sum_genrePvA["male : married/widowed","TP"]+
Sum_genrePvA["male : divorced/separated","FP"]+
Sum_genrePvA["male : married/widowed","FP"])
PPV_Fmale=(Sum_genrePvA["female : single","TP"])/
(Sum_genrePvA["female : single","TP"]+Sum_genrePvA["female : single","FP"])
PPV_male==PPV_Fmale
```

Epsilon predictive parity

```r
abs(PPV_male-PPV_Fmale)<epsilon
```

To be able to do the comparison of PPV for both group we will be using the Boostrap method.

```r
ns=200
PPV_S=matrix(NA,ns,2)
colnames(PPV_S)=c("Male","Female")
Sum_genrePvAS=array(dim = c(4,4,ns))
colnames(Sum_genrePvAS)=c("TP","FP","FN","TN")
row.names(Sum_genrePvAS)=row.names(Sum_genrePvA)
P = matrix(NA,ns,2)
for(i in 1:ns){
    idx_s = sample(1:n,size=n, replace=TRUE)
    Reg_s=glm(formula=credit_risk~. , family = binomial ( link="logit"), data = DGerman[idx_s,])
    S_s = predict(Reg_s,type="response")
    Predic_s=ifelse(S_s>0.5,"bad","good")

    TP_inds=(DGerman[idx_s,][which(DGerman[idx_s,]$credit_risk=="good" & Predic_s[idx_s]=="good"),])
    FP_inds=(DGerman[idx_s,][which(DGerman[idx_s,]$credit_risk=="bad" & Predic_s[idx_s]=="good"),])
    FN_inds=(DGerman[idx_s,][which(DGerman[idx_s,]$credit_risk=="good" & Predic_s[idx_s]=="bad"),])
    TN_inds=(DGerman[idx_s,][which(DGerman[idx_s,]$credit_risk=="bad" & Predic_s[idx_s]=="bad"),])
    Sum_genrePvAS[,,i]=cbind(as.matrix(summary(TP_inds$personal_status_sex)),as.matrix(summary(FP_inds$personal_status_sex)),Sum_FNgenre=as.matrix(summary(FN_inds$personal_status_sex)),Sum_TNgenre=as.matrix(summary(TN_inds$personal_status_sex)))
}
PPV_S[,"Male"]=(Sum_genrePvAS["male : divorced/separated","TP",]+Sum_genrePvAS["male : married/widowed","TP",])/(Sum_genrePvAS["male : divorced/separated","TP",]+Sum_genrePvAS["male : married/widowed","TP",]+Sum_genrePvAS["male : divorced/separated","FP",]+Sum_genrePvAS["male : married/widowed","FP",])
PPV_S[,"Female"]=(Sum_genrePvAS["female : single","TP",])/(Sum_genrePvAS["female : single","TP",]+Sum_genrePvAS["female : single","FP",])

plot(density(PPV_S[,"Male"]-PPV_S[,"Female"]))
abline(v=0,lty=2,col="red"
```

1. Predictive Equality: We have to compare the FPR for both groups:

```r
FPR_male=(Sum_genrePvA["male : divorced/separated","FP"]+
Sum_genrePvA["male : married/widowed","FP"])/
(Sum_genrePvA["male : divorced/separated","FP"]+
Sum_genrePvA["male : married/widowed","FP"]+
Sum_genrePvA["male : divorced/separated","TN"]+
Sum_genrePvA["male : married/widowed","TN"])
FPR_Fmale=(Sum_genrePvA["female : single","FP"])/
(Sum_genrePvA["female : single","TP"]+
Sum_genrePvA["female : single","TN"])
```

```
FPR_male
FPR_Fmale
```

Epsilon Predictive Equality

```
abs(FPR_male-FPR_Fmale)<epsilon
```

To be able to do the comparison of FPR for both group we will be using the Boostrap method.

```
FPR_S=matrix(NA,ns,2)
colnames(FPR_S)=c("Male","Female")
FPR_S[,"Male"]=(Sum_genrePvAS["male : divorced/separated","FP",]+Sum_genrePvAS["male : married/widowed","FP",])/(Sum_genrePvAS["male : divorced/separated","FP",]+Sum_genrePvAS["male : married/widowed","FP",]+Sum_genrePvAS["male : divorced/separated","TN",]+Sum_genrePvAS["male : married/widowed","TN",])
FPR_S[,"Female"]=(Sum_genrePvAS["female : single","FP",])/(Sum_genrePvAS["female : single","FP",]+Sum_genrePvAS["female : single","TN",])

plot(density(FPR_S[,"Male"]-FPR_S[,"Female"]))
abline(v=0,lty=2,col="red")
```

2. Equality of Opportunity We have to compare the FNR for both groups:

```
FNR_male=(Sum_genrePvA["male : divorced/separated","FN"]+
Sum_genrePvA["male : married/widowed","FN"])/
(Sum_genrePvA["male : divorced/separated","FN"]+
Sum_genrePvA["male : married/widowed","FN"]+
Sum_genrePvA["male : divorced/separated","TP"]+
Sum_genrePvA["male : married/widowed","TP"])
FNR_Fmale=(Sum_genrePvA["female : single","FN"])/
(Sum_genrePvA["female : single","FN"]+
Sum_genrePvA["female : single","TP"])
FNR_male
FNR_Fmale
```

Epsilon Equality of Opportunity

```r
abs(FNR_male-FNR_Fmale)<epsilon
```

To be able to do the comparison of FNR for both group we will be using the Boostrap method.

```r
FNR_S=matrix(NA,ns,2)
colnames(FNR_S)=c("Male","Female")
FNR_S[,"Male"]=(Sum_genrePvAS["male : divorced/separated","FN",]+Sum_genrePvAS["male : married/widowed","FN",])/(Sum_genrePvAS["male : divorced/separated","FN",]+Sum_genrePvAS["male : married/widowed","FN",]+Sum_genrePvAS["male : divorced/separated","TP",]+Sum_genrePvAS["male : married/widowed","TP",])
FNR_S[,"Female"]=(Sum_genrePvAS["female : single","FN",])/(Sum_genrePvAS["female : single","FN",]+Sum_genrePvAS["female : single","TP",])

plot(density(FNR_S[,"Male"]-FNR_S[,"Female"]))
abline(v=0,lty=2,col="red")
```

3. Equality of odds: We have to compare the FPR and FNR for both groups:

```r
abs(FPR_male-FPR_Fmale)<epsilon && abs(FNR_male-FNR_Fmale)<epsilon
```

4. Conditional Use Accuracy Equality:

   We have to compare the PPV and NPV for both groups:

```r
Sum_genrePvA
(NPV_male=(Sum_genrePvA["male : divorced/separated","TN"]+
Sum_genrePvA["male : married/widowed","TN"])/
(Sum_genrePvA["male : divorced/separated","TN"]+
Sum_genrePvA["male : married/widowed","TN"]+
Sum_genrePvA["male : divorced/separated","FN"]+
Sum_genrePvA["male : married/widowed","FN"]))
(NPV_Fmale=(Sum_genrePvA["female : single","TN"])/
(Sum_genrePvA["female : single","TN"]+
Sum_genrePvA["female : single","FN"]))
\begin{lstlisting}[language=R]
```

```
abs(PPV_male-PPV_Fmale)<epsilon && abs(NPV_male-NPV_Fmale)<epsilon
```

Boostraping:

```
NPV_S=matrix(NA,ns,2)
colnames(NPV_S)=c("Male","Female")
NPV_S[,"Male"]=((Sum_genrePvAS["male : divorced/separated","TN",]+Sum_genrePvAS["male : married/widowed","TN",])/(Sum_genrePvAS["male : divorced/separated","TN",]+Sum_genrePvAS["male : married/widowed","TN",]+Sum_genrePvAS["male : divorced/separated","FN",]+Sum_genrePvAS["male : married/widowed","FN",]))
NPV_S[,"Female"]=((Sum_genrePvAS["female : single","TN",])/(Sum_genrePvAS["female : single","TN",]+Sum_genrePvAS["female : single","FN",]))
plot(density(NPV_S[,"Male"]-NPV_S[,"Female"]))
abline(v=0,lty=2,col="red")
```

5. Overall Accuracy Equality: We have to compare the $P_1(Y = \hat{Y})$ and $P_0(Y = \hat{Y})$

```
P_mal=(Sum_genrePvA["male : divorced/separated","TN"]+
Sum_genrePvA["male : divorced/separated","TP"]+
Sum_genrePvA["male : married/widowed","TN"]+
Sum_genrePvA["male : married/widowed","TP"])/n_male
P_Fmal=(Sum_genrePvA["female : single","TN"]+
Sum_genrePvA["female : single","TP"])/n_fmale
abs(P_Fmal-P_mal)<epsilon
```

Boostraping:

```
P_Y_Yh=matrix(NA,ns,2)
colnames(P_Y_Yh)=c("Male","Female")
P_Y_Yh[,"Male"]=((Sum_genrePvAS["male : divorced/separated","TN",]+Sum_genrePvAS["male : divorced/separated","TP",]+Sum_genrePvAS["male : married/widowed","TN",]+Sum_genrePvAS["male : married/widowed","TP",])/(Sum_genrePvAS["male : divorced/separated","TP",]+Sum_genrePvAS["male : divorced/separated","FP",]+Sum_genrePvAS["male : divorced/separated","FN",]+Sum_
genrePvAS["male : divorced/separated","TN",]+Sum_genrePvAS["male : married/widowed","TP",]+Sum_genrePvAS["male : married/widowed","FP",]+Sum_genrePvAS["male : married/widowed","FN",]+Sum_genrePvAS["male : married/widowed","TN",]))
P_Y_Yh[,"Female"]=((Sum_genrePvAS["female : single","TN",]+Sum_genrePvAS["female : single","TP",])/(Sum_genrePvAS["female : single","TP",]+Sum_genrePvAS["female : single","FP",]+Sum_genrePvAS["female : single","FN",]+Sum_genrePvAS["female : single","TN",]))
plot(density(P_Y_Yh[,"Male"]-P_Y_Yh[,"Female"]))
abline(v=0,lty=2,col="red")
```

6. Treatment Equality:

   We have to test equal ratio of false positives and false negatives.

```
FN1_FP1=(Sum_genrePvA["male : divorced/separated","FN"]+
Sum_genrePvA["male : married/widowed","FN"])/
(Sum_genrePvA["male : divorced/separated","FP"]+
Sum_genrePvA["male : married/widowed","FP"])
FN0_FP0=(Sum_genrePvA["female : single","FN"])/
(Sum_genrePvA["female : single","FP"])
abs(FN1_FP1-FN0_FP0)<epsilon
```

   Booatraping

```
FN_FP_s=matrix(NA,ns,2)
colnames(FN_FP_s)=c("Male","Female")
FN_FP_s[,"Male"]=(Sum_genrePvAS["male : divorced/separated","FN",]+Sum_genrePvAS["male : married/widowed","FN",])/(Sum_genrePvAS["male : divorced/separated","FP",]+Sum_genrePvAS["male : married/widowed","FP",])
FN_FP_s[,"Female"]=(Sum_genrePvAS["female : single","FN",])/(Sum_genrePvAS["female : single","FP",])
plot(density(FN_FP_s[,"Male"]-FN_FP_s[,"Female"]))
abline(v=0,lty=2,col="red")
```

- Calibration We have to compare $P_1(Y=1|S=s)=P_0(Y=1|S=s),\ \forall s\in[0,1$

```r
DF_Ger=subset(DGerman, select = -c(personal_status_sex,credit_risk) )
Y = ifelse(DGerman$credit_risk=="good",1,0)
DGerman2=data.frame(Y,DF_Ger)
install.packages(locfit)
Reg2=glm(formula=Y~. ,  family = binomial ( link ="logit"), data = DGerman2)
S = predict(Reg2,type="response")
Loc_F = locfit.raw(x=S[DGerman$personal_status_sex == "female : single"],
                   y=DGerman2$Y[DGerman$personal_status_sex == "female : single"],
                   family="binomial",
                   kern="rect",deg=0)
vs=(0:100)/100
probaF = predict(Loc_F,newdata=vs)
Loc_M = locfit.raw(x=S[DGerman$personal_status_sex != "female : single"],
                   y=DGerman2$Y[DGerman$personal_status_sex != "female : single"],
                   family="binomial",
                   kern="rect",deg=0)
probaM = predict(Loc_M,newdata=vs)
plot(vs,probaF,col="red",ylim=0:1,type="l")
lines(vs,probaM,col="blue")
```

- Testing d-seperation statemnents

```r
german=read.csv("german.csv")
DF_German=german[which
(german$personal_status_sex!="female : non-single or male : single"),]
Genre=ifelse(DF_German$personal_status_sex=="female : single","Female","Male")
DF_German=data.frame(DF_German,Genre)
```

```r
CreditGender=table(DF_German$credit_risk,DF_German$Genre)
chisq.test(CreditGender)$expected
chisq.test(CreditGender)
```

```

CreditGenderProp=table(DF_German$credit_risk,
DF_German$Genre,DF_German$property)
CreditGenderProp[,,1]
chisq.test(CreditGenderProp[,,1])$expected
chisq.test(CreditGenderProp[,,2])$expected
chisq.test(CreditGenderProp[,,3])$expected
chisq.test(CreditGenderProp[,,4])$expected
```

```
chisq.test(CreditGenderProp[,,1])
chisq.test(CreditGenderProp[,,2])
chisq.test(CreditGenderProp[,,3])
chisq.test(CreditGenderProp[,,4])
```

```
CreditGendersaving=table(DF_German$credit_risk,DF_German$Genre,
   DF_German$savings)
```

```
chisq.test(CreditGendersaving[,,1])$expected
chisq.test(CreditGendersaving[,,2])$expected
chisq.test(CreditGendersaving[,,3])$expected
chisq.test(CreditGendersaving[,,4])$expected
chisq.test(CreditGendersaving[,,5])$expected
```

```
chisq.test(CreditGendersaving[,,1])
chisq.test(CreditGendersaving[,,2])
chisq.test(CreditGendersaving[,,3])
chisq.test(CreditGendersaving[,,4])
chisq.test(CreditGendersaving[,,5])
```

```
CreditGenderpurpose=table(DF_German$credit_risk,DF_German$Genre,
DF_German$purpose)
CreditGenderpurpose=table(DF_German$credit_risk,DF_German$Genre,
DF_German$purpose)
purposeG=ifelse(DF_German$purpose=="others","others",
ifelse(DF_German$purpose=="furniture/equipment","househld",
ifelse(DF_German$purpose=="radio/television","househld",
ifelse(DF_German$purpose=="domestic appliances",
"househld","persona"))))
dim(CreditGenderpurpose)
DF_German=data.frame(DF_German,purposeG)
summary(DF_German$purposeG)
```

```
TBLE=table(DF_German$credit_risk,DF_German$Genre,DF_German$purposeG)
```

```
chisq.test(TBLE[,,1])$expected
chisq.test(TBLE[,,2])$expected
chisq.test(TBLE[,,3])$expected
```

```
chisq.test(TBLE[,,1])
chisq.test(TBLE[,,2])
chisq.test(TBLE[,,3])
```

```
TBLE2=table(DF_German$credit_risk,DF_German$Genre,DF_German$housing)
chisq.test(TBLE2[,,1])
chisq.test(TBLE2[,,2])

chisq.test(TBLE2[,,3])
```

```
nbcr=ifelse(DF_German$number_credits=="1","1",">1")
DF_German=data.frame(DF_German,nbcr)

TBLE3=table(DF_German$credit_risk,DF_German$Genre,DF_German$nbcr)
chisq.test(TBLE3[,,1])
chisq.test(TBLE3[,,2])
```

```
TBLE5=table(DF_German$credit_risk,DF_German$Genre,DF_German$people_liable)
chisq.test(TBLE5[,,1])
chisq.test(TBLE5[,,2])
```

- Simulating credit score data

```
n <- 1000
set.seed(10)
#exogenous variables
p <- rbinom(size=1, n=1000, prob=0.5)
#endogenous variables
x <- 5*p+rnorm(n,0,sqrt(1-0.5^2))
theta <- 0.35*x+0.5*p
y <- exp(theta)/(1+exp(theta))+exp(p)/(1+exp(p))
```

```
#We create a data.frame without the latent variable (theta)
dat <- data.frame(x,p,y)
#We made the model save lavaan
```

- Testing the simulated data

```
library(lavaan)
mod<-"
x~p
theta~x+p
theta=~y
"
fit<-sem(model=mod,data=dat)
summary(fit)
```

BIBLIOGRAPHY

[Act(1976)] Equal Pay Act. Sex discrimination act 1975. *Race Relations Act*, 1976.

[Bellhouse(2016)] David Bellhouse. How our days became numbered: Risk and the rise of the statistical individual. by dan bouk (chicago, university of chicago press, 2015) 304 pp., 2016.

[Conseil de l'Union Européenne (2004)()] Conseil de l'Union Européenne (2004). Directive 2004/113/ce du conseil du 13 décembre 2004 mettant en œuvre le principe de l'égalité de traitement entre les femmes et les hommes dans l'accès ' des biens et services et la fourniture de biens et services. jo l 373, p. 37-43.

[Datta et al.(2014)Datta, Tschantz, and Datta] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

[Du Bois(1896)] WEB Du Bois. Review of race traits and tendencies of the american negro. *Annals of the American Academy*, pages 127–33, 1896.

[Dwork et al.(2012)Dwork, Hardt, Pitassi, Reingold, and Zemel] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings* *of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Elisburg(1978)] Donald Elisburg. Equal pay in the united states: The development and implementation of the equal pay act of 1963. *Labor Law Journal*, 29(4):195, 1978.

[European Union Legislation(2000)] European Union Legislation. Council directive 2000/78/ec of 27 november 2000 establishing a general framework for equal treatment in employment and occupation, 2000.

[European Union Legislation(2006)] European Union Legislation. Council directive 2006/54/ec of 5 july 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation. *Official Journal of the European Communities*, L 204/23, 2006.

[Feast and Hand(2015)] Pat Feast and James Hand. Enigmas of the equality act 2010—"three uneasy pieces". *Cogent Social Sciences*, 1(1): 1123085, 2015.

[Hedges(1977)] Bob A Hedges. Gender discrimination in pension plans: Comment. *The Journal of Risk and Insurance*, 44(1):141–144, 1977.

[Heen(2009)] Mary L Heen. Ending jim crow life insurance rates. *Nw. JL & Soc. Pol'y*, 4:360, 2009.

[Hoffman(1896)] Frederick Ludwig Hoffman. *Race traits and tendencies of the American Negro*, volume 11. American Economic Association, 1896.

[Ingold and Soper(2016)] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it? bloomberg, april 21, 2016, 2016.

[Legislation(1985a)] Canada Legislation. Canada labour code, 1985a. c. L-2.

[Legislation(1985b)] Canada Legislation. Canadian human rights act, 1985b. c. H-6.

[Legislation(1995)] Canada Legislation. Employment equity act, 1995. c. 44.

[Legislation(2000)] European Union Legislation. Council directive 2000/43/ec of 29 june 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, 2000.

[Legislation(1968)] U.S. Federal Legislation. Civil rights act, 1968. pub.l. 90–284, 82 stat. 73.

[Martin(1977)] Gerald D Martin. Gender discrimination in pension plans: Author's reply. *The Journal of Risk and Insurance*, 44(1):145–149, 1977.

[Merry and Edwards(2002)] AJ Merry and DM Edwards. Disability part 1: the disability discrimination act (1995)–implications for dentists. *British dental journal*, 193(4):199–201, 2002.

[Murphy(1976)] WT Murphy. The british race relations act of 1976. *Comp. Lab. L.*, 1:137, 1976.

[Myers(1977)] Robert J Myers. Gender discrimination in pension plans:

Further comment. *The Journal of Risk and Insurance*, 44(1):144–145, 1977.

[O'neil(2016)] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[Pearl(1988)] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[Rebert and Van Hoyweghen(2015)] Lisa Rebert and Ine Van Hoyweghen. The right to underwrite gender: The goods & services directive and the politics of insurance pricing. *Tijdschrift Voor Genderstudies*, 18 (4):413–431, 2015.

[Rovatsos et al.(2019)Rovatsos, Mittelstadt, and Koene] Michael Rovatsos, Brent Mittelstadt, and Ansgar Koene. Landscape summary: Bias in algorithmic decision-making: what is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it? 2019.

[Schmeiser et al.(2014)Schmeiser, Störmer, and Wagner] Hato Schmeiser, Tina Störmer, and Joël Wagner. Unisex insurance pricing: consumers' perception and market implications. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 39(2):322–350, 2014.

[Shipley(2016)] Bill Shipley. *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press, 2016.

[Smith(1977)] James F Smith. The equal credit opportunity act of 1974: A cost/benefit analysis. *The Journal of Finance*, 32(2):609–622, 1977.

[Thiery and Van Schoubroeck(2006)] Yves Thiery and Caroline Van Schoubroeck. Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 31(2):190–211, 2006.

[Verma and Rubin(2018)] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[Wiggins(2013)] Benjamin Alan Wiggins. *Managing Risk, Managing Race: Racialized Actuarial Science in the United States, 1881–1948.* University of Minnesota, 2013.

[Wolff(2006)] Megan J Wolff. The myth of the actuary: life insurance and frederick l. hoffman's race traits and tendencies of the american negro. *Public Health Reports*, 121(1):84, 2006.

[Wu(2020)] Yongkai Wu. *Achieving Causal Fairness in Machine Learning.* University of Arkansas, 2020.