UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN COMPUTER ENGINEERING

Investigation of the Usage of Rademacher Averages for Causal Rule Discovery

Relatore: Prof. Fabio Vandin

Laureanda: Paola Donolato

Correlatori: Ing. Dario Simionato
Ing. Antonio Collesei

ANNO ACCADEMICO 2022 – 2023

Data di laurea 21/04/2023

Fai ciò che ti rende felice

Abstract

The aim of this thesis is to investigate the usage of Rademacher Averages in causal rule discovery, to overcome two main issues of the state of the art approach. In causal rule discovery, a rule, defined as a clause on the variables of the dataset, is assessed as causal or not through the computation of a given statistic. Since the statistic is computed from observational data, it is necessary to consider its confidence interval. The state of the art approach has two main limitations: (i) it computes the confidence interval modeling the error with a normal distribution; (ii) it tests several rules but without accounting for multiple hypothesis testing correction. We propose an approach based on Rademacher Averages to compute a confidence interval that: (i) depends on the features of the chosen statistic; (ii) directly accounts for MHT correction. Furthermore the interval provides rigorous probabilistic guarantees. We tested three different statistics using simulated data, and found that one provides good performance when the dataset is sufficiently big, but it has fairly good performance also for smaller sample sizes.

Sommario

Lo scopo di questa tesi è studiare la possibilità di utilizzo delle Rademacher Averages nella ricerca di regole causali, al fine superare due limitazioni del corrente stato dell'arte. Nella ricerca di regole causali, una regola, definita come una clausola sulle variabili del dataset, è valutata come causale o meno attraverso il calcolo di una statistica. Dal momento che la statistica è calcolata da dati osservazionali, è necessario considerare il suo intervallo di confidenza. L'approccio stato dell'arte ha due limitazioni principali: (i) l'intervallo di confidenza è calcolato modellando l'errore con una distribuzione normale; (ii) vengono testate diverse regole ma senza tenere conto della correzione per test multipli. Noi proponiamo un approccio basato sulle Rademacher Averages per calcolare un intervallo di confidenza che: (i) dipende dalle caratteristiche della statistica scelta; (ii) tiene conto della correzione per test multipli. Inoltre l'intervallo ha garanzie probabilistiche. Abbiamo testato tre diverse statistiche su dati simulati, e trovato che una dà buoni risultati quanto il dataset è sufficientemente grande, e abbastanza buoni anche per dataset più piccoli.

Contents

Chapter 1

Introduction

The aim of this thesis is to investigate the usage of Rademacher Averages for causal rule discovery. Two main concepts needs to be introduced, what are Rademacher Averages and in what causal rule discovery consists, and of course how they are combined to reach the final purpose. Rademacher Averages are a statistical tool and they require a formal definition, thus in this introduction I give only the intuition of their meaning and usage and I focus more on motivating why causal rule discovery is important.

To introduce this latter topic, let start from an example, reported in Figure 1.1. The plot shows, on a time window of ten years, the divorce rate in Maine, in red, and the per capita consumption of margarine, in black. The two trends are very similar, therefore one could think, for example, that divorcing causes a reduction on the consumption of margarine, even if, intuitively, it does not make much sense. This is because this is not a *causal* relation, but rather a *correlation*.

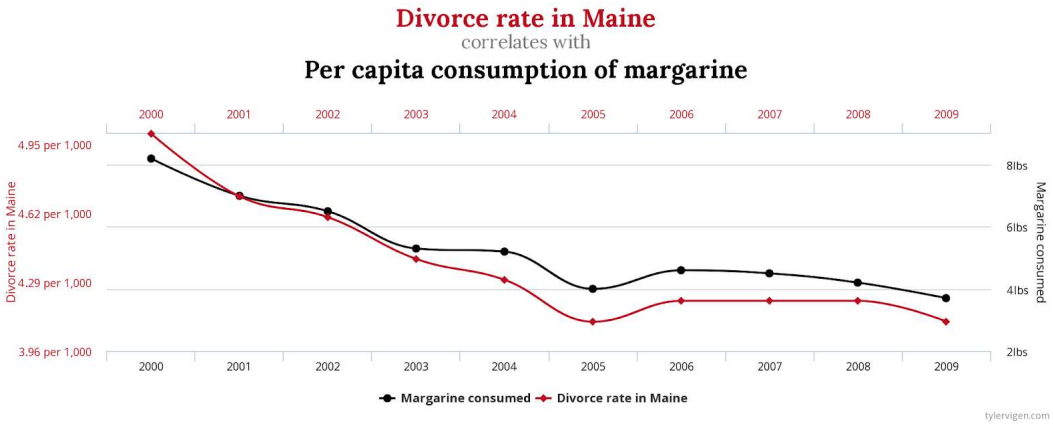

Figure 1.1: Correlation between divorce rate in Maine and margarine consumption (taken from [11]).

Let us now formalize better these two concepts; I will refer to the work of Person, who was one of the main developers of causation and correlation theory. Pearson, in "The Grammar of Science" [8] defines causation as follows: "If a sequence of events D, E, F are always preceded by the event C, then C is a *cause* of D, E, F". *Correlation* instead, according to him, is given by "all the relationship between two events,

both dependence and independence ones". Causation is therefore a particular case of correlation, that holds when there are no *spurious* relations, which, intuitively, are associations that make seem an event as a cause of another, even if it is not the case. At this point should be more clear that the example of Figure 1.1 is affected by a spurious relation and it is not a real causal one.

I want to present now another example, more focused on the aim of this thesis and which should better characterise the concept of spurious relation; it is taken from "Introduction To Causal Inference", by Brady Neal [6]. Let us focus first on Figure 1.2 (a). Suppose that someone told us that from a study emerged that sleeping with the shoes on is associated with headache, but without specifying anything else. One could therefore think that there is a causal relation between the two, but clearly this does not make much sense. Suppose now that we have access to another piece of information, see Figure 1.2 (b), that before was hidden, which is the fact that the night before the person drank. It is now evident that that the two events "go to bed with shoes on" and "wake up with headache" have a common cause, namely drinking alcohol. This latter is called *confounder*, because it is an event that influences both events involved in the association and leads to a wrong conclusion, what we called a spurious relation. Once we know the presence and the role of the confounder, we can act to remove its influence and found true causal relations.

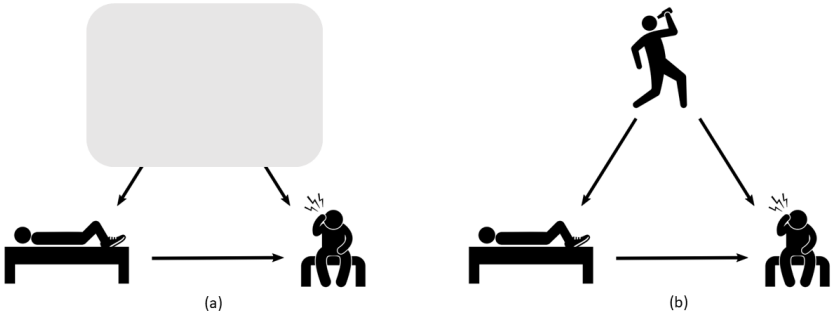

Figure 1.2: Example of a spurious relation (going to bed with shoes causes headache) due to the presence of a confounder (drink alcohol the night before), (taken from [6]).

One way to remove spurious dependencies is through a randomized controlled trial, a type of experiment in which the population under study is split into two, one half will receive a *treatment*, while the other half no; the aim is to infer if the treatment causes a particular *outcome*. The treatments are called also *actionable variables*. Recalling the previous example, the treatment consists in going to bed with the shoes on and the outcome to observe is waking up with headache or not. The outcome of the experiment will be that there is no evidence of a relation between sleeping with shoes and headache.

Randomized control trials are however, in general, difficult to perform in practise and we would like instead to develop an approach relying on the data available.

In most scenarios we have access only to observational data, which, for the way they are collected, contains any kind of association, included spurious ones. However, under

certain conditions, that will be exhaustively discussed in section 2.3, it is possible to extract true causal relations from observational data.

The causal framework on which this thesis is based is the one proposed by [3]. The main idea is the following. For every causal association of the type "event A causes event B", is computed by taking properly account for the presence of the confounding events, a "score", thus assessing whether such rule is causal or not, and, in case of positive answer, the magnitude of the relation is considered. Since this score, called *effect* by [3], is estimated from the data, it is necessary to provide also a confidence interval for it; the authors computed it assuming the error follows a normal distribution.

The contribution of this thesis is the application of Rademacher Averages in causal rule discovery, to provide a different framework for dealing with this topic. Rademacher Averages requires the definition of a statistic, that is what I called "score" before, to assess the presence of a causal relation or not. The major difference with respect to the work of [3] is that Rademacher Averages allows the computation of a confidence interval that depends on the statistic itself, without resort to an approximation of the error with a known distribution.

Chapter 2

Preliminaries

In this chapter I provide an overview of the key concepts used in this thesis. Based on what was presented in [3], I start introducing in Section 2.1 the type of dataset of interest and what kind of information we want to extract from it. Then I define formally the fundamental mathematical concepts for causal rule discovery: bayesian network and causal graph, in Section 2.2, and the causal framework in Section 2.3.

In Section 2.4 I formalize the concept of statistical hypothesis testing [4] and the multiple hypothesis testing correction, necessary when more than one test is performed at the same time. Then, in Section 2.5 I define the Rademacher Averages [9], which are the statistical tool at the base of the contributions of this thesis. Finally, Section 2.6 is dedicated to the formalization of the problem analysed.

2.1 Basic Definitions and General Meaning

In this section I provide the basic definitions that will allow me, in the following paragraphs, to formally characterize the problem analyzed in this thesis. In particular, I describe the characteristics of the datasets of our interest and which kind of relations we want to extract from them.

The dataset D is divided into three sets of variables $D = (\mathbf{Z}, \mathbf{X}, \mathrm{Y})$. $\mathbf{Z}$ is the confounders' set, composed by the variables that may lead to spurious associations, $\mathbf{X}$ the set of the actionable variables and Y is the target variable; all the variables are discrete.

The aim is to extract causal relations from the data, eliminating the effect of $\mathbf{Z}$. There are different frameworks for dealing with causality, the one analyzed in this thesis focuses on the search of conditions σ, defined on the set $\mathbf{X}$, that cause a specific value of the target $y \in \mathrm{Y}$. Such conditions are defined as a conjunction over a subset $X = \{\mathrm{X}_1, \ldots, \mathrm{X}_i, \ldots, \mathrm{X}_\ell\} \subset \mathbf{X}$; for each variable is expressed a clause, for example $\gamma_i = \mathrm{X}_i \leq 3$, thus the resulting formula can be expressed as:

$$\sigma = \gamma_1 \wedge \ldots \wedge \gamma_l. \tag{2.1}$$

In order to find causal rules $\pi : \sigma \to y$ the dependence of $\mathbf{X}$ from $\mathbf{Z}$ must be removed. The mathematical concept that formalises such independence is the Pearl's do-notation $P(\mathrm{Y} = y|do(\mathbf{X} := \mathbf{x}))$, which represents the probability of observing a particular value y for the target once all the actionable variables are fixed to assume value $\mathbf{x}$. Such probability, in general, is different from the standard conditional probability $P(\mathrm{Y} = y|\mathbf{X} = \mathbf{x})$ because the former forces $\mathbf{X}$ to assume a particular value $\mathbf{x}$ a priori, without any external influence, while in the second case variables from the set $\mathbf{Z}$ may have influenced directly or indirectly, thus though $\mathbf{X}$, the outcome.

In practice, as anticipated in Chapter 1, we have access to observational data, collected without control on the values assumed by $\mathbf{X}$, that therefore follows the standard conditional probability distribution. However, under certain conditions, it is possible to derived from them the desired Pearl's do-notation, but before explaining how it can be done I need to define the concepts of Bayesian Network and causal graph.

2.2 Bayesian Network and Causal Graph

A Bayesian Network is a DAG (Direct Acyclic Graph) G = ($\mathbf{V}$,$\mathbf{E}$), where $\mathbf{V}$= $\{\mathrm{V}_1, \ldots \mathrm{V}_n\}$ is the set of vertices and $\mathbf{E}$ is the set of edges, that do not contain direct cycles. Each node represents a variable (e.g. X, Y, Z,$\ldots$, X_1,$\ldots$, X_n) and the arcs between the nodes represent the conditional dependencies among variables, defined accordingly to a conditional probability distribution p.

In particular the structure of the network satisfies the Markov condition [5]:

Definition 2.2.1. $\forall \mathrm{V} \in \mathbf{V}$, V is independent from all its non-descendant, for a given value of its parents (which are the variables that directly influence V).

Thanks to the Markov assumption it is possible to infer some statements about the data starting from the graphical structure.

However in causal inference we would like to do the opposite, that is starting from data trying to infer something about their dependency structure. Thus an implication in the opposite direction is needed, which is expressed in the faithfulness assumption [10]:

Definition 2.2.2. A graph G is *faithful* to a joint conditional distribution p if and only if every independence in p is represented in G.

Thanks to this we can search for dependencies in the data and those will tell something about the structure of the graph. The missing step is how to search for relations in the data that are causal, in fact we are only interested in them; in other words, how we can approximate the Pearl's do-notation to extract the desired information.

Before entering in the details I need to define the three key structures of a graph, chain, fork and collider, and discuss the dependence relations among the variables;

X represent the actionable variable, Z the confounder and Y the target. Figure 2.1 reports the three structures.

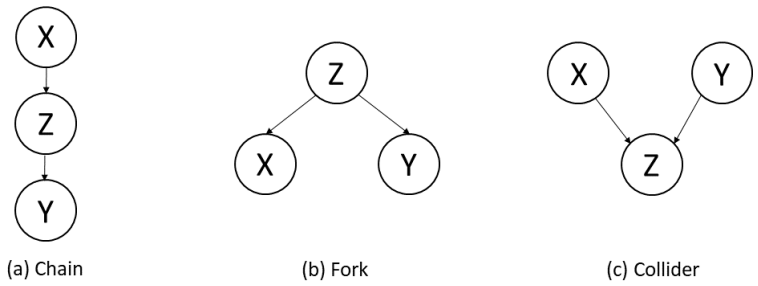

Figure 2.1: The three fundamental structures in a graph, from left to right: chain, fork, collider.

The first two entail the same dependency/independence relations among variables X and Y. Let us start from the chain. X and Y are dependent, in fact X influences Z, that in turn influences Y; however if it could be possible to fix the value of the confounder Z, then the dependence is removed. In fact, in this case, Z will not depend on X, thus the only factor that determines Y is Z.

For the fork, instead, the dependence between X and Y is due to the fact that Z is a common cause for both and this leads to the identification of a spurious correlation. However fixing the value of Z, that corresponds somehow to observe that value, removes the dependence. Such observation translates, from a mathematical point of view, into conditioning on the value of Z.

The formal proof of the independence between X and Y for the fork is reported below.

Proof. We want to show the independence between X and Y conditioning on Z in the fork, that is:

$$P(\mathrm{X},\mathrm{Y}|\mathrm{Z}) = P(\mathrm{X}|\mathrm{Z})P(\mathrm{Y}|\mathrm{Z}) \tag{2.2}$$

Recalling the definition of conditional probability, we have that:

$$P(\mathrm{X},\mathrm{Y}|\mathrm{Z}) = \frac{P(\mathrm{X},\mathrm{Y},\mathrm{Z})}{P(\mathrm{Z})} \tag{2.3}$$

We then apply the bayesian network factorization to the numerator since, for the Markov assumption, every child variable depends only from its non-descendant, obtaining:

$$P(\mathrm{X},\mathrm{Y},\mathrm{Z}) = [P(\mathrm{X}|\mathrm{Z})P(\mathrm{Y}|\mathrm{Z})]P(\mathrm{Z}) \tag{2.4}$$

Substituting in Eq.(2.3) we get:

$$P(\mathrm{X},\mathrm{Y}|\mathrm{Z}) = \frac{P(\mathrm{X}|\mathrm{Z})P(\mathrm{Y}|\mathrm{Z})P(\mathrm{Z})}{P(\mathrm{Z})} \tag{2.5}$$

Therefore simplifying P(Z) we derive:

$$P(\mathrm{X},\mathrm{Y}|\mathrm{Z}) = P(\mathrm{X}|\mathrm{Z})P(\mathrm{Y}|\mathrm{Z}) \tag{2.6}$$

□

A similar reasoning can be applied also for the chain, while for collider the relations among variables are opposite. In this case X and Y influence Z, so conditioning on Z, X and Y are dependent because they are forced to assume a value according to their "effect" Z, while without conditioning they are independent.

2.3 The Causal Framework

As stated in Section 2.1, to find causal relations $\pi : \sigma \rightarrow y$ the dependence of $\mathbf{X}$ from $\mathbf{Z}$ must be removed; in other words it is necessary to find a way to compute $P(\mathrm{Y} = y|do(\mathbf{X} := \mathbf{x}))$ from the data, which instead follows the conditional probability distribution $P(\mathrm{Y} = y|\mathbf{X} := \mathbf{x})$.

A condition under which these two probabilities are equal, meaning that there is independence from the set $\mathbf{Z}$, is expressed by the Back-Door Adjustment Theorem ([7], Theorem 3.3.2), but before giving its formulation it is necessary to introduce the definition of (blocked) spurious path and the Back-Door Criterion. Both definitions below are from [3]

Definition 2.3.1. Given a pair of nodes X and Y, a *spurious path* is any undirected path connecting them with an incoming edge towards X.

Definition 2.3.2. A spurious path is *blocked* by $\mathbf{Z}$ if one of these two is satisfied: i) the path contains a collider not in the set $\mathbf{Z}$ ii) the path has a non-collider in $\mathbf{Z}$.

At this point the Back-Door Criterion can be introduced (from [3]):

Definition 2.3.3. Given a set of nodes $\mathbf{X}$ and a node Y, the set of nodes $\mathbf{Z}$ satisfies the *Back-Door Criterion* if the following holds: i) $\forall$ Z in $\mathbf{Z}$ and $\forall$ X in $\mathbf{X}$, Z is not a descendant of X; ii) all the spurious path between any X in $\mathbf{X}$ and Y are blocked by the set $\mathbf{Z}$.

It is now possible to state the Back-Door Adjustment theorem (from [3]):

Theorem 2.3.1. If the set $\mathbf{Z}$ satisfies the Back-Door Criterion with respect to $\mathbf{X}$ and Y, in the subset of the input dataset for which $\mathbf{Z}$ assumes value $\mathbf{z}$ (called $\mathbf{z}$-stratum) it holds:

$$p(y|do(\mathbf{X} := \mathbf{x})) = p(y|\mathbf{x},\mathbf{z}). \tag{2.7}$$

To summarise, the crucial implication of the Back-Door Adjustment theorem is that if the Back-Door Criterion holds and the analysis is restricted to the **z**-stratum because we are fixing Z, then the conditional probability $p(y|\mathbf{x},\mathbf{z})$ is exactly the "desired" one.

Before going on, I want to make an important observation. Let us consider, for example, the fork. As said before, conditioning on Z, so restricting to a **z**-stratum, X and Y are independent. But looking at the collider if we restrict to the stratum, X and Y are dependent, in fact the fixed value of Z, forces X and Y to assume values according to it and not freely. This is an intuition to say that this structure does not satisfy the Back-Door Criterion, thus it is not possible to exploit Eq.(2.7). As a consequence of this, not for all graphs we can use conditional probabilities to assess the presence or not of a causal rule.

Now that they key concepts have been formalized, it is possible to provide the definition of effect of a rule $\pi : \sigma \rightarrow y$, the metric proposed by [3] to assess the causality of the rule.

In general there can be several assignment to the vector $\mathbf{x}$ that satisfy σ; for example if a clause is defined as $X_i \leq 4$ it is true for $X_i = 4$, $X_i = 3$, and so on. This is formalized by the stochastic policy Q_σ, that is a family of conditional probability distributions. A typical area of application is reinforcement learning, in which it is defined as the conditional probability $\pi_S(A|S)$ of all the possible actions given all possible states. In our case we want the policy to take into account the fraction of times we observe the value $\mathbf{X} = \mathbf{x}$ in the **z**-stratum we defined:

$$Q_\sigma(do(\mathbf{x})) = p(\mathbf{X} = \mathbf{x}|\sigma = T, \mathbf{Z} = \mathbf{z}) \tag{2.8}$$

It is now possible to define the probability of observing a particular value y of the target, given a rule σ; given an input dataset $D = (\mathbf{Z},\mathbf{X},\mathrm{Y})$ with the desired properties, we have:

$$\begin{aligned} p(y|do(Q_\sigma) &= \sum_{\mathbf{x}:\sigma(\mathbf{x})=T} \sum_{\mathbf{z}\in\mathbf{Z}} p(y|\mathbf{x},\mathbf{z})p(\mathbf{z})Q_\sigma(do(\mathbf{x})) \\ &= \sum_{\mathbf{z}\in\mathbf{Z}} p(\mathbf{z}) \sum_{\mathbf{x}:\sigma(\mathbf{x})=T} p(y|\mathbf{x},\mathbf{z})Q_\sigma(do(\mathbf{x})) \end{aligned} \tag{2.9}$$

The summation on $\mathbf{z} \in \mathbf{Z}$ is quite obvious since it is necessary to consider all the contributions from the set $\mathbf{Z}$ to compute the overall probability, each weighted by the probability $p(\mathbf{z})$ of the stratum, while the other one takes into account what explained above about the policy.

At this point it is possible to give the definition of the effect of a rule σ:

$$e(\sigma) = p(y|do(Q_\sigma)) - p(y|do(Q_{\bar{\sigma}})) \tag{2.10}$$

where $\bar{\sigma}$ is the negation of the rule. The first term is the probability of observing

the desired target variable y, given a certain rule σ that is a candidate cause for y; the second one instead is the probability of observing the same target but considering false the cause. The intuition behind this definition is the following. If the effect is positive it means that is more probable to observe that target under the conditions expressed by σ, rather than its negation, so σ is a cause of y. The magnitude of the effect will give an idea of how strong the relation is. If the effect is 0 there is the same probability of observing y in the sample restricted to σ and in all the other cases, so the relation is not informative. Finally a negative effect tells that the rule under analysis makes no sense, since the negation of the rule, that is quite general, has an higher probability than the rule itself.

According to the policy defined in Eq.(2.8) we can rewrite the effect as:

$$e(\sigma) = \mathbb{E}[p(y|\sigma, \mathbf{Z})] - \mathbb{E}[p(y|\bar{\sigma}, \mathbf{Z})]. \tag{2.11}$$

Since we are dealing with finite datasets, we can only give an estimation of the probability p, computed directly on the sample, that is we estimate the empirical probability $\hat{p}$ as follows:

$$\hat{p}(y|\sigma, \mathbf{z}) = \frac{\#\text{ of times in which } (\sigma = \mathrm{T} \wedge y) \text{ is observed in the } \mathbf{z}\text{-stratum}}{\text{total } \#\text{ of instances in the } \mathbf{z}\text{-stratum}}. \tag{2.12}$$

We can define the estimator with respect to the empirical distribution $\hat{p}$ as:

$$\begin{aligned} \hat{e}(\sigma) =& \mathbb{E}\big[\hat{p}(y|\sigma, \mathbf{z}) - \hat{p}(y|\bar{\sigma}, \mathbf{z})\big] \\ &= \sum_{\mathbf{z}\in\mathbf{Z}} (\hat{p}(y|\sigma, \mathbf{z}) - \hat{p}(y|\bar{\sigma}, \mathbf{z}))\hat{p}(\mathbf{z}) \\ &= \sum_{z\in\mathbf{Z}} (\hat{p}_{\sigma,\mathbf{z}} - \hat{p}_{\bar{\sigma},\mathbf{z}})\hat{p}(\mathbf{z}) \end{aligned} \tag{2.13}$$

where I expressed $\hat{p}(y|\sigma, \mathbf{z})$ and $\hat{p}(y|\bar{\sigma}, \mathbf{z})$ more compactly as $\hat{p}_{\sigma,\mathbf{z}}$ and $\hat{p}_{\bar{\sigma},\mathbf{z}}$.

In order to compute the empirical probabilities $\hat{p}_{(\sigma,\mathbf{z})}$ and $\hat{p}_{(\bar{\sigma},\mathbf{z})}$, for every rule $\pi : \sigma \to y$ is derived the correspondent contingency table, that stores the following counts (the subscript is to stress the fact that we are restricting to a particular $\mathbf{z}$-stratum):

- $\mathrm{A}_{\mathbf{z}}$ = # of instances of the dataset that satisfies σ and for which $\mathrm{Y} = y$;
- $\mathrm{B}_{\mathbf{z}}$ = # of instances of the dataset that satisfies σ and for which $\mathrm{Y} \neq y$;
- $\mathrm{C}_{\mathbf{z}}$ = # of instances of the dataset that do not satisfies σ and for which $\mathrm{Y} = y$;
- $\mathrm{D}_{\mathbf{z}}$ = # of instances of the dataset that do not satisfies σ and for which $\mathrm{Y} \neq y$.

Furthermore I define the following quantities:

- $m_{\sigma,\mathbf{z}}$ = #of instances satisfying σ ;
- $m_{\bar{\sigma},\mathbf{z}}$ = #of instances not satisfying σ ;
- $m_{\mathbf{z}}$ = total # of instances of $\mathbf{z}$-stratum.

The contingency table thus is:

	$Y = y$	$Y \neq y$	
$\sigma = T$	$A_{\mathbf{z}}$	$B_{\mathbf{z}}$	$m_{\sigma,\mathbf{z}}$
$\bar{\sigma} = T$	$C_{\mathbf{z}}$	$D_{\mathbf{z}}$	$m_{\bar{\sigma},\mathbf{z}}$
			$m_{\mathbf{z}}$

With these definitions it is possible to derive the expression for $\hat{p}_{(\sigma,\mathbf{z})}$ and $\hat{p}_{(\bar{\sigma},\mathbf{z})}$, that are respectively:

$$\hat{p}_{(\sigma,\mathbf{z})} = \frac{A_{\mathbf{z}}}{A_{\mathbf{z}} + B_{\mathbf{z}}} = \frac{A_{\mathbf{z}}}{m_{\bar{\sigma},\mathbf{z}}} \tag{2.14}$$

$$\hat{p}_{(\bar{\sigma},\mathbf{z})} = \frac{C_{\mathbf{z}}}{C_{\mathbf{z}} + D_{\mathbf{z}}} = \frac{C_{\mathbf{z}}}{m_{\bar{\sigma},\mathbf{z}}} \tag{2.15}$$

Finally, defining:

$$\hat{p}(\mathbf{z}) = \frac{\#\text{ of times Z takes value } \mathbf{z}}{m} \tag{2.16}$$

we can express $\hat{e}(\sigma)$ as:

$$\begin{aligned} \hat{e}(\sigma) &= \sum_{z \in \mathbf{Z}} (\hat{p}_{\sigma,\mathbf{z}} - \hat{p}_{\bar{\sigma},\mathbf{z}}) \cdot \hat{p}(\mathbf{z}) \\ &= \sum_{z \in \mathbf{Z}} \left(\frac{A_{\mathbf{z}}}{m_{\bar{\sigma},\mathbf{z}}} - \frac{C_{\mathbf{z}}}{m_{\bar{\sigma},\mathbf{z}}} \right) \cdot \hat{p}(\mathbf{z}) \end{aligned} \tag{2.17}$$

The problem of this formulation is that the denominator can be equal to zero; to avoid zero division it is introduced the Laplace correction, a simple smoothing technique that consists in adding one to all the counts of the contingency table.

Applying the correction $\hat{e}(\sigma)$ becomes:

$$\hat{e}(\sigma) = \sum_{z \in \mathbf{Z}} \left(\frac{A_{\mathbf{z}} + 1}{m_{\sigma,\mathbf{z}} + 2} - \frac{C_{\mathbf{z}} + 1}{m_{\bar{\sigma},\mathbf{z}} + 2} \right) \cdot \hat{p}(\mathbf{z}) \tag{2.18}$$

An issue with this estimator is that it shows an high variance, resulting thus in overfitting, a well-known phenomenon for which the model fits too much with the used dataset but is not able to generalize to other ones.

Before explaining how this was fixed by [3], I would like to provide an high level description of overfitting and then focus on what this implies in our framework.

Suppose we want to perform a regression task to learn the model that best fits our data; assume that the data are the orange dots of figure 2.2. The most natural choice of a models to fit them is a polynomial of a certain degree. The figure shows two different situations; on the left the model chosen is a polynomial of high degree, which fits really well the data, the one on the right instead is simpler, thus also the fit is more rough.

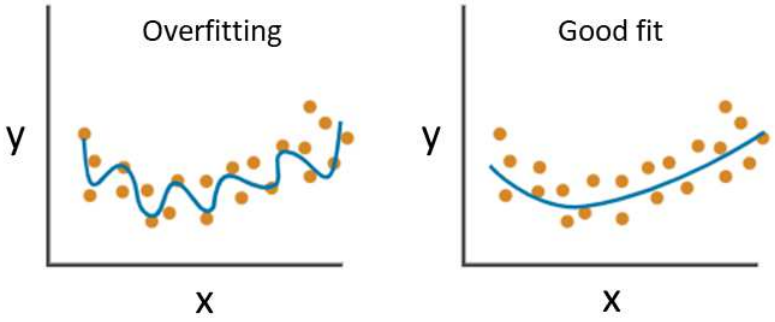

Figure 2.2: Regression task. Left: the model chosen overfitts, because it is too complex. Right: the model is simpler but it is able to generalize better (the image is taken from [1]).

The first situation correspond to overfitting; in this scenario the learned model is data-dependent, in fact it fits almost perfectly all the dots, however if it is applied on a different dataset, generated with a different distribution, it will be unsatisfactory. The simpler model instead, even if it is less accurate, generalises better on a different domain of application.

When data are spread, as in this case, if we do not include any prior knowledge what we obtain is a model that tries to capture all of them, resulting in a high complexity.

In the domain of our interest, the fact that the estimate of the effect of a rule shows an high variance, implies that it takes its value, for example 0.2, only considering that particular dataset, but in another one may be that the average value is 0.3. So the importance of the rule is dataset-dependent. This is evident in particular for dataset of small size. We do not want to learn rules whose relative effect is specific to a sub-population, thus that overfits, but instead we would like to infer quite general ones.

To avoid this, a bias is introduced, in terms of a correction factor applied directly to the estimator. The prior knowledge that was chosen is the confidence interval, that expresses how confident we are about the estimate; it defines a range of values around the one assumed by the estimator, which represents the oscillations of the estimator itself. The narrower the interval, the more accurate the estimation. The estimator (Eq.(2.13)) is composed by two blocks and an interval is computed for both. Inside the $\mathbf{z}$-stratum the number of instances that satisfy σ and $\bar{\sigma}$, denoted by $m_{\sigma,\mathbf{z}}$ and $m_{\bar{\sigma},\mathbf{z}}$ respectively, are two binomial random variables with success probability given by $p(y|\sigma,\mathbf{z})$ and $p(y|\bar{\sigma},\mathbf{z})$. If the size of the dataset is high, it is possible to approximate

them with two normal random variables, thus the confidence intervals for the two terms are, by definition:

$$\mathrm{CI}_{\sigma,\mathbf{z}} = \frac{\beta}{2\sqrt{m_{\sigma,\mathbf{z}}}} \tag{2.19}$$

$$\mathrm{CI}_{\bar{\sigma},\mathbf{z}} = \frac{\beta}{2\sqrt{m_{\bar{\sigma},\mathbf{z}}}} \tag{2.20}$$

where β is a value, called Z-score, that depends on the distribution and on the error rate α, a parameter that controls the maximum tolerated error. A more complete characterization of these quantities can be found in Section 2.4.

Given that, [3] defined the reliable estimator, in a stratum, as:

$$\tau(\mathbf{z}) = \left(\hat{p}_{\sigma,\mathbf{z}} - \hat{p}_{\bar{\sigma},\mathbf{z}}\right) - \left(\frac{\beta}{2\sqrt{m_{\sigma,\mathbf{z}}}} + \frac{\beta}{2\sqrt{m_{\bar{\sigma},\mathbf{z}}}}\right). \tag{2.21}$$

From thus formulation, it is easy to see that $\tau(\mathbf{z})$ is a consistent estimator of the difference of the two estimated probabilities; in fact as the number of instances grows, the correction term becomes negligible:

$$\lim_{\min(m_{\sigma,\mathbf{z}}, m_{\bar{\sigma},\mathbf{z}})\to\infty} \frac{\beta}{2\sqrt{m_{\sigma,\mathbf{z}}}} + \frac{\beta}{2\sqrt{m_{\bar{\sigma},\mathbf{z}}}} = 0. \tag{2.22}$$

Substituting the values for $\hat{p}_{\sigma,\mathbf{z}}$ and $\hat{p}_{\bar{\sigma},\mathbf{z}}$ derived in Eq.(2.14) and Eq.(2.15) and applying the Laplace correction, we obtain:

$$\tau(\mathbf{z}) = \frac{\mathrm{A}+1}{m_{\sigma,\mathbf{z}}+2} - \frac{\mathrm{C}+1}{m_{\bar{\sigma},\mathbf{z}}+2} - \frac{\beta}{2\sqrt{m_{\sigma,\mathbf{z}}+2}} + \frac{\beta}{2\sqrt{m_{\bar{\sigma},\mathbf{z}}+2}}. \tag{2.23}$$

Finally we obtain the reliable estimator of a rule simply averaging the corrected estimators within each $\mathbf{z}$-stratum:

$$\hat{r}(\sigma) = \sum_{\mathbf{z}\in\mathbf{Z}} \tau(\mathbf{z})\hat{p}(\mathbf{z}). \tag{2.24}$$

2.4 Statistical Hypothesis Testing

So far I described the approach presented by [3] to compute the effect of a rule, which is a measure of its causality. It was explained that a positive effect means that there is a causal relation, a null one that there is no relation and a negative one means the rule is not informative. All this discussion was performed at a high level, but the mathematical reason for which from the value of the effect we can draw the above conclusions about dependence or independence, is that a statistical test is performed. This section is dedicated to the definition of statistical hypothesis testing and to the corrections that needs to be applied when multiple tests are performed at the same time. The main source that I used to write this session is [4].

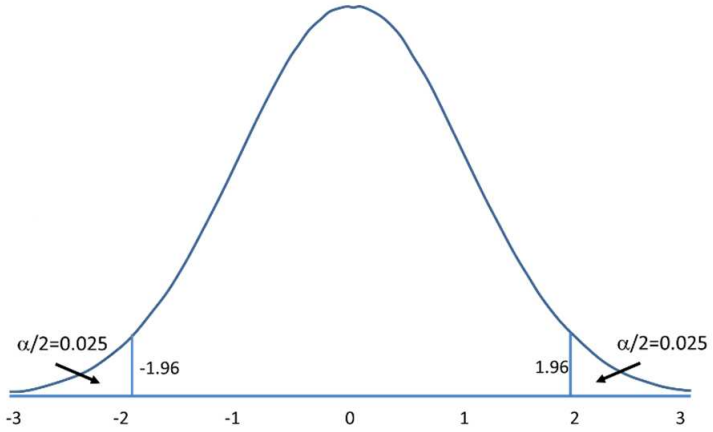

Figure 2.3: Example of a two-sided test for a normal distribution, with confidence level $\alpha = 0.05$ (the image is taken from [2]).

Statistical hypothesis testing is a statistical procedure whose aim is to decide whether the data supports an hypothesis. In particular two hypothesis are compared, the null hypothesis H_0 and the alternative hypothesis H_1. The possible outcomes are accept or reject the null hypothesis, basing the decision on the comparison of a statistic computed on the data with a threshold, or better a rejection region. Intuitively, when the statistic falls in this region the null hypothesis is rejected. Two types of error can be made: rejecting H_0 when it is true, also known as type I error, and accepting H_0 when it is false, called also type II error.

The aim is to bound the type I error, which is regulated by the error rate α, a parameter that represents the probability of committing such error. A typical choice is $\alpha = 0.05$, meaning that in expectation in 5% of the cases we are making a false positive discovery.

According to the value of α is computed the boundary of the rejection region (the limit value for which the null hypothesis is not accepted), called also Z-score. It depends also on the distribution considered and typically are available conversion tables that reports the value it takes for various α.

There are two types of test, depending on the way the null hypothesis is built: two-sided or one-sided. In the first is tested H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$; in the second is tested H_0: $\theta \leq \theta_0$ versus H_1: $\theta > \theta_0$ or H_0: $\theta \geq \theta_0$ versus H_1: $\theta < \theta_0$.

Figure 2.3 shows an example of a two-sided test for a normal distribution. The overall α is 0.05, thus to have such overall error rate, the two Z-scores, corresponding to the two tails, are computed setting $\alpha_{\text{tail}} = \frac{\alpha}{2} = 0.025$. their values are, respectively, -1.96 and 1.96 and are the boundaries of the two rejecting regions. Thus, for example, if the value of the sample statistic is 2, it falls in the critical region and therefore H_0 is rejected.

In our framework we are interested in this type of tests. For a one-tailed test the differences are that the rejecting region will be only the left or the right tail; thus only one Z-score is computed and using the overall α.

At this point I want to give a more formal and general definition of the test procedure. First it is necessary to define H_0, H_1 and the statistics. Then the distribution followed by the data should be approximated with a "standard" one in order to perform the test, like a normal distribution or a chi-squared. For asymmetric distributions the test must be one-sided, while in the other cases the choice depends on which property is tested; once this is fixed is computed the rejection region according to the chosen α. Finally from the comparison of the observed statistics, that is the one computed on the sample, and the rejection region, it is decided whether the null hypothesis is accepted or rejected.

This is the general procedure when only one test is performed. However when multiple tests are carried on, in order to guarantee an overall small error some corrections needs to be applied. This issue is discussed in the next section.

2.4.1 Multiple Hypothesis Testing (MHT)

In the following I give some details about MHT (Multiple Hypothesis Testing) correction, a technique to control the number of false discoveries when multiple tests are performed.

In particular, the quantity kept under control is the Family-Wise Error Rate (FWER), that is the probability of rejecting the null hypothesis when it is true or, in other words, of making a type I error:

$$\text{FWER} = P(\text{making one type I error}) = 1 - P(\text{no type I error}). \tag{2.25}$$

The probability of making a type I error is exactly α, so the one of making no error is 1 - α. Since we are considering multiple tests, say N, and we assume they are independent the overall probability of making no error is $(1-\alpha)^N$, thus:

$$\text{FWER} = 1 - (1-\alpha)^{\text{N}}. \tag{2.26}$$

When only one test is performed, so N = 1, the FWER is exactly α, in fact:

$$\text{FWER} = 1 - (1-\alpha)^{\text{N}} = 1 - (1-0.05)^1 = 1 - (0.95) = 0.05. \tag{2.27}$$

Considering for example N = 5, without applying any correction for multiple hypothesis testing, so using the value α in all the tests independently, the FWER becomes:

$$\text{FWER} = 1 - (1-\alpha)^{\text{N}} = 1 - (1-0.05)^5 = 1 - (0.95)^5 = 0.226. \tag{2.28}$$

A simple correction to keep this quantity under control is the Bonferroni one; it simply divides α by the number N of tests performed:

$$\alpha_{corr} = \frac{\alpha}{\text{N}}. \tag{2.29}$$

Coming back to the previous example, the computation is:

$$\alpha_{corr} = \frac{\alpha}{\mathrm{N}} = \frac{0.05}{5} = 0.01 \tag{2.30}$$

$$\mathrm{FWER} = 1 - (1 - \alpha_{corr})^{\mathrm{N}} = 1 - (1 - 0.01)^5 = 1 - (0.99)^5 = 0.049. \tag{2.31}$$

that is even more restrictive than the desired value.

The problem with Bonferroni correction is that is assumes independence between all tests and it is conservative since the error rate used in each test is quite small and this leads to an higher number of false negatives. An alternative approach, developed in this thesis, exploits Rademacher Averages to compute better bounds, while considering all the hypothesis together.

2.5 Rademacher Averages

In this section I introduce the concept of Rademacher Averages, basing the discussion on what was presented in [9]. Their usage in the perspective of this thesis will be developed in the next section.

Rademacher Averages allow to compute a bound to the error of an estimate, computed with respect to a family of functions. Given a family of functions F defined on a domain X, f: $X \to$ [a,b] $\subset \mathbb{R}$, $\forall\ f \in F$; given a set $S = \{s_1,..., s_m\}$ of m i.i.d. values taken from an unknown distribution μ, the sample mean $\hat{\mathbb{E}}_S[f]$ and its expectation $\mathbb{E}_\mu[f]$, are:

$$\hat{\mathbb{E}}_S[f] = \frac{1}{m} \sum_{s_i \in S} f(s_i) \tag{2.32}$$

$$\mathbb{E}_\mu[f] = \mathbb{E}_\mu[f(s_i)] = \mathbb{E}_\mu\left[\frac{1}{m} \sum_{s_i \in S} f(s_i)\right]. \tag{2.33}$$

We are interested in the error committed estimating the true mean $\mathbb{E}_\mu[f]$ with $\hat{\mathbb{E}}_S[f]$, computed on the sample S, that means giving a bound to the Supremum Deviation *D(F,S)*, the maximum deviation of the empirical mean from the true one with respect to all the functions of the family F:

$$D(F,S) = \sup_{f \in F} \left|\hat{\mathbb{E}}_S[f] - \mathbb{E}_\mu[f]\right|. \tag{2.34}$$

A possible approach to obtain this bound is given by Rademacher Averages. A Rademacher random variable takes values in {-1, 1} with the same probability. Given a vector λ of m i.i.d. Rademacher random variables, $\lambda = \{\lambda_1, ..., \lambda_m\}$ the Empirical Rademacher Average (ERA), with respect to the family F and the sample S is:

$$\hat{R}(F,S) = \mathbb{E}_\lambda\left[\sup_{f\in F}\frac{1}{m}\sum_{s_i\in S}\lambda_i f(s_i)\right]. \tag{2.35}$$

The intuition behind this definition is the following. Multiplying by λ_i is like splitting the sample in two groups, the one one for which $\lambda_i = 1$ and $\lambda_i = -1$, respectively; since λ_i takes both these values with probability $\frac{1}{2}$, as m increases, more the two groups will have the same size. Furthermore, as m increases the more the two groups will take homogeneous values, so the averages computed on them separately will be similar and thus the expectation will be small. To conclude the reasoning, the ERA gives a measure about how well, considering the current sample size m, the sample captures the true mean. If, for a fixed m, the ERA is small, then the sample is sufficient to provide a good estimate for all the functions of the family F. The problem of such definition is that the expectation is computed over 2^m possible values and this is often infeasible. An approach to estimate it is through n Monte-Carlo trials and averaging the obtained values. Given a matrix $\Lambda = \{-1,1\}^{n\times m}$ where each column λ_j is a m-vector of Rademacher random variables, the n-sample Monte-Carlo Empirical Rademacher Average (n-MCERA) $\hat{R}^n_m(F,S,\Lambda)$ is:

$$\hat{R}^n_m(F,S,\Lambda) = \frac{1}{n}\sum_{j=1}^{n}\sup_{f\in F}\frac{1}{m}\sum_{s_i\in S}\lambda_{j,i}f(s_i). \tag{2.36}$$

Thanks to n-MCERA it is possible to provide a formulation of a first bound of the Supremum Deviation. Defining $z = \max\{|a|,|b|\}$, $c = |b-a|$ and denoting as $\tilde{R}$:

$$\tilde{R} = \hat{R}^n_m(F,S,\Lambda) + 2z\sqrt{\frac{\ln\frac{4}{\delta}}{2nm}} \tag{2.37}$$

it has been shown ([9], Theorem 3.1) the following.

Theorem 2.5.1. With probability ≥ 1 - δ over the choices of S and Λ, for a fixed $\delta \in$ (0,1):

$$D(F,S) \leq 2\tilde{R} + \frac{\sqrt{c(4m\tilde{R} + c\ln\frac{4}{\delta})\ln\frac{4}{\delta}}}{m} + \frac{c\ln\frac{4}{\delta}}{m} + c\sqrt{\frac{\ln\frac{4}{\delta}}{2m}} \tag{2.38}$$

This is the bound on the error of interest.

2.5.1 Improved Version of the Bound

In [9] (Theorem. 3.5) is presented also a tighter bound to the Supremum Deviation, that requires, to be applied, more knowledge about the family F and in particular about its variance. Its formulation is reported in Eq(4.15).

Theorem 2.5.2. With δ, z, c as before, define the following terms:

$$\rho = \hat{R}^n_m(F,S,\Lambda) + 2z\cdot\sqrt{\frac{\log\frac{4}{\delta}}{2nm}} \tag{2.39}$$

$$\mathrm{r} = \rho + \frac{1}{m} \cdot \left(\sqrt{c \cdot \left(4m\rho + c \log \frac{4}{\delta} \right) \cdot \log \frac{4}{\delta}} + c \log \frac{4}{\delta} \right) \tag{2.40}$$

Let v be an upper bound of the variance of the family of functions F, and define:

$$\epsilon = 2\mathrm{r} + \sqrt{\frac{2 \log \frac{4}{\delta} \cdot (v + 4c\mathrm{r})}{m}} + \frac{c \log \frac{4}{\delta}}{3m} \tag{2.41}$$

then, with probability $\geq 1 - \delta$, over the choices of S and Λ, the bound is given by:

$$D(F,S) \leq \epsilon. \tag{2.42}$$

2.6 Formalization of the Problem

In this Section I formalize the problem analyzed in this thesis. The objective is to study a new framework to extract causal relations from a dataset, overcoming two issues of the approach presented in [3].

Recalling the previous work [3], given a dataset $D = \{\mathbf{Z}, \mathbf{X}, \mathrm{Y}\}$ with the structure described in Section 2.1 and such that the underlying causal graph satisfied the Back-Door Adjustment Theorem (Theorem. 2.3.3), the goal is to extract causal rules from the data. A rule π is defined by a clause σ, expressed as a conjunction over the set of actionable variables $\mathbf{X}$, that is a candidate *cause* for a value y of the target variable Y: $\pi : \sigma \rightarrow y$. To assess whether a rule represents a true causal relation or not it is computed a score, that the authors called effect (Eq.(2.13)), and since it is estimated from the data, it is necessary to compute a confidence interval. It is then performed a statistical test: if the effect, corrected for the interval, is greater than 0, then σ and y are causally dependent, otherwise not.

The two limitations of the state of the art approach are that (i) in the computation of the confidence interval the error is modeled with a normal distribution and (ii) several rules are tested, but without correcting for multiple hypothesis testing.

The new approach proposed in this thesis makes use of Rademacher Averages (presented in Section 2.5) to overcome these two issues. For a given family of functions $f \in F$, Rademacher Averages allow the computation of a confidence interval to bound the empirical estimation of each function f. The bound is computed reflecting the property of the family itself, thus not approximating the error with a known distribution, and considering all the functions in F simultaneously, directly accounting for MHT correction. In addition, the bound is provided with rigorous probabilistic guarantees.

My contributions are the following. I defined and tested three families of functions F, each of which is a different statistic to assess the causality of a rule. In particular, each function f of the family is the statistic computed with respect to the rule π under analysis. I then bounded such quantities with the confidence interval computed with

Rademacher Averages; the bound defines the acceptance region of the null hypothesis for the statistical test, which is formally defined below.

Definition 2.6.1. For a rule $\pi : \sigma \to y$, the null hypothesis H_0 states the independence between σ and y.

I performed the experimental evaluation of my statistics on synthetic datasets. A first analysis, presented in Chapter 4, shows that only the last one is really promising; the detailed results can be found instead in Chapter 5.

Now that the key concepts and the focus have been explained, it is possible to move on to the technical discussion.

Chapter 3

Methods

In this chapter I present the statistics that I defined to assess the causality of a rule. As required by Rademacher Averages, for each statistic I provide the overall definition and the sample version one. Furthermore, all these definitions are provided both in the conditional and unconditional case. Finally, for each statistic I specified the relative test procedure.

3.1 Statistic that mimics the "overall effect"

The first approach is a statistic which behaviour mimics the one of the effect, presented in [3]. I therefore called this first class of functions for which the Rademacher Averages are computed, F_e.

Recalling the definition (Eq.(2.35)), Rademacher Averages depend on the dataset $S = \{s_1, \cdots, s_m\}$ and a family of functions $f \in F$, each of which is computed on all the elements $s_i \in S$. F represents a class of functions for which we want to compute the error committed in estimating its true expectation through the empirical sample mean, computed on S. For our purposes, F is a class such that its empirical mean captures a particular data property: the presence or not of a causal relation.

As anticipated, the first choice is a class F_e of functions whose mean mimics the behaviour the effect. To be more clear, since each $f_e \in F_e$ is relative to a specific rule $\pi : \sigma \to y$, I will denote it as $f_{e,\pi}$ for all possible rules π, therefore: $F_e = \{f_{e,\pi} \mid \forall \pi\}$.

For sake of clarity, I report the definition of the effect on a contingency table (Eq.(3.1)) and the table itself (Table 3.1). The following discussion will be developed in the unconditional case; in the final part of this section I will give the definition of the statistic $f_{e,\pi,\text{cond}}$ in the conditional case, considering the contributions of each **z**-stratum. As a reminder, the effect $e(\sigma)$ of a rule is:

$$e(\sigma) = \frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}} - \frac{\mathrm{C}}{\mathrm{C}+\mathrm{D}}. \tag{3.1}$$

It is necessary to find a way to express which is the contribution of a sample s_i (that is a row of the dataset) to the effect. The reasoning that I made was the following. I

	$Y = y$	$Y \neq y$	
$\sigma = T$	A	B	m_σ
$\bar{\sigma} = T$	C	D	$m_{\bar{\sigma}}$
	m_y	$m_{\bar{y}}$	m

Table 3.1: General form of the contingency table for a rule $\pi : \sigma \rightarrow y$

considered the denominators (A+B) and (C+D) known constants (they can be easily derived from the data once the rule is fixed) and I focused on numerators, noticing that only counts A and C appear in the formula. Therefore, in correspondence of a particular rule π, each element s_i can contributes with only one of the followings: $\frac{1}{\text{A+B}}$ if the current observation satisfies σ and the target Y assumes the desired value y, $-\frac{1}{\text{C+D}}$ if it does not satisfy σ but the target is the desired one and 0 in the other two cases.

The definition of the sample statistic $f_{e,\pi}(s_i)$ thus is:

$$f_{e,\pi}(s_i) = \begin{cases} \frac{1}{\text{A+B}} & if\ \sigma = T \text{ and } Y = y; \\ -\frac{1}{\text{C+D}} & if\ \bar{\sigma} = T \text{ and } Y = y; \\ 0 & if\ \sigma = T \text{ and } Y \neq y; \\ 0 & if\ \bar{\sigma} = T \text{ and } Y \neq y; \end{cases} \tag{3.2}$$

that can be expressed in compact form (Eq.(3.5)) introducing the two following indicator functions:

$$\mathbf{1}_\sigma = \begin{cases} 1 & if\ \sigma = T \\ 0 & if\ \bar{\sigma} = T \end{cases} \tag{3.3}$$

$$\mathbf{1}_Y = \begin{cases} 1 & if\ Y = y \\ 0 & if\ Y \neq y \end{cases} \tag{3.4}$$

$$f_{e,\pi}(s_i) = \mathbf{1}_\sigma \cdot \mathbf{1}_Y \cdot \frac{1}{\text{A} + \text{B}} - (1 - \mathbf{1}_\sigma) \cdot \mathbf{1}_Y \cdot \frac{1}{\text{C} + \text{D}}. \tag{3.5}$$

The summation of $f_{e,\pi}(s_i), \forall s_i \in S$ is exactly the effect, thus the empirical sample mean, which is the overall statistic, $\hat{\mathbb{E}}_S[f_{e,\pi}(s_i)] \doteq f_{e,\pi}$, is the effect divided by m (Eq.(3.6)):

$$f_{e,\pi} = \frac{1}{m} \sum_{s_i \in S} f_{e,\pi}(s_i) = \frac{1}{m}\left[\frac{\text{A}}{(\text{A+B})} - \frac{\text{C}}{(\text{C+D})}\right] \tag{3.6}$$

therefore the codomain of $f_{e,\pi}$ is $[-\frac{1}{m}, \frac{1}{m}]$ (recall that the effect is in $[-1, 1]$).

At this point I want to derive the codomain of the sample statistic $f_{e,\pi}(s_i)$, because its value directly influences the bound to the Supremum Deviation. The maximum positive value assumed by $f_{e,\pi}(s_i)$ is 1, in fact the positive contribution is given by $\frac{1}{\text{A+B}}$ and it reaches it maximum for the minimum possible value for which the denominator is defined, that is 1. Following a similar reasoning, the minimum value assumed by

$f_{e,\pi}(s_i)$ is -1, that derives from $-\frac{1}{\mathrm{C+D}}$. So $f_{e,\pi}(s_i) : s_i \to [-1, 1]$.

Such discrepancy in the order of magnitude of the codomain of the sample statistic and its empirical sample mean gave some problems, in particular the bound is too large; the details are discussed in Section 4.3.1.

3.1.1 Extension of $f_{e,\pi}$ in the Conditional Case

In the conditional case the way of proceeding is very similar. For each value $\mathbf{z}$ that the confounder under analysis Z can assume, the contingency table is (Table 3.2):

	$\mathrm{Y} = y$	$\mathrm{Y} \neq y$	
$\sigma = \mathrm{T}$	$\mathrm{A}_\mathbf{z}$	$\mathrm{B}_\mathbf{z}$	$m_{\sigma,\mathbf{z}}$
$\bar{\sigma} = \mathrm{T}$	$\mathrm{C}_\mathbf{z}$	$\mathrm{D}_\mathbf{z}$	$m_{\bar{\sigma},\mathbf{z}}$
	$m_{y,\mathbf{z}}$	$m_{\bar{y},\mathbf{z}}$	$\mathrm{m}_\mathbf{z}$

Table 3.2: General form of the contingency table for a rule $\pi : \sigma \to y$ in a $\mathbf{z}$-stratum.

Defining $S_\mathbf{z}$ as the subset of the dataset S for which the $\mathrm{Z} = \mathbf{z}$, and $s_j \in S_\mathbf{z}$, the function $f_{e,\pi,\mathbf{z}}(s_j)$ defined on an element of the restricted dataset $S_\mathbf{z}$ is:

$$f_{e,\pi,\mathbf{z}}(s_j) = \left[\mathbf{1}_\sigma \cdot \mathbf{1}_\mathrm{Y} \cdot \frac{1}{\mathrm{A}_\mathbf{z} + \mathrm{B}_\mathbf{z}} - (1 - \mathbf{1}_\sigma) \cdot \mathbf{1}_\mathrm{Y} \cdot \frac{1}{\mathrm{C}_\mathbf{z} + \mathrm{D}_\mathbf{z}}\right] \cdot \hat{p}(\mathbf{z}). \tag{3.7}$$

Summing the contributions of all the points, it is obtained the overall effect in that stratum, already weighted by the relative probability, and dividing by $m_\mathbf{z}$ the empirical sample mean $\hat{\mathbb{E}}_{S_\mathbf{z}}[f_{e,\pi,\mathbf{z}}(s_j)] \doteq f_{e,\pi,\mathbf{z}}$ of that $\mathbf{z}$-stratum:

$$\begin{aligned} f_{e,\pi,\mathbf{z}} &= \frac{1}{m_\mathbf{z}} \sum_{s_j \in S_\mathbf{z}} f_{e,\pi,\mathbf{z}}(s_j) = \\ &= \frac{1}{m_\mathbf{z}} \left(\frac{\mathrm{A}_\mathbf{z}}{\mathrm{A}_\mathbf{z} + \mathrm{B}_\mathbf{z}} - \frac{\mathrm{C}_\mathbf{z}}{\mathrm{C}_\mathbf{z} + \mathrm{D}_\mathbf{z}}\right) \cdot \hat{p}(\mathbf{z}) = \\ &= \frac{1}{m_\mathbf{z}} \left(\frac{\mathrm{A}_\mathbf{z}}{\mathrm{A}_\mathbf{z} + \mathrm{B}_\mathbf{z}} - \frac{\mathrm{C}_\mathbf{z}}{\mathrm{C}_\mathbf{z} + \mathrm{D}_\mathbf{z}}\right) \cdot \frac{m_\mathbf{z}}{m} \\ &= \frac{1}{m} \left(\frac{\mathrm{A}_\mathbf{z}}{\mathrm{A}_\mathbf{z} + \mathrm{B}_\mathbf{z}} - \frac{\mathrm{C}_\mathbf{z}}{\mathrm{C}_\mathbf{z} + \mathrm{D}_\mathbf{z}}\right). \end{aligned} \tag{3.8}$$

Finally, from the summation of all the contributions $f_{e,\pi,\mathbf{z}}$ it is obtained the total empirical sample mean $\hat{\mathbb{E}}_{S,\mathrm{cond}}[f_{e,\pi,\mathbf{z}}] \doteq f_{e,\pi,\mathrm{cond}}$:

$$f_{e,\pi,\mathrm{cond}} = \sum_{\mathbf{z} \in \mathrm{Z}} f_{e,\pi,\mathbf{z}}. \tag{3.9}$$

Recalling the worst case analysis in the unconditional case, the codomain of the sample statistic $f_{e,\pi,\mathbf{z}}(s_j)$ is $[-1, 1] \cdot \hat{p}(\mathbf{z})$. The worst possible case in when $\hat{p}(\mathbf{z}) \to 1$, thus we obtain again the range $[-1, 1]$. The empirical sample mean, instead, as before is the effect divided by m, so it lies in $[-\frac{1}{m}, \frac{1}{m}]$.

Before defining the test procedure, I prove the following theorem.

Theorem 3.1.1. The value of the statistic in the conditional and unconditional case is different: $f_{e,\pi} \neq f_{e,\pi,\text{cond}}$.

Proof of Theorem 3.1.1. The proof is trivial. In the conditional case the computation of the statistic in a stratum, $f_{e,\pi,\mathbf{z}}$, is based on a subset $S_{\mathbf{z}}$ of the original dataset S.

Considering the formulation of Eq.(3.8), due to the asymmetry introduced by the denominators $(\mathrm{A}_{\mathbf{z}}+\mathrm{B}_{\mathbf{z}})$ and $(\mathrm{C}_{\mathbf{z}}+\mathrm{D}_{\mathbf{z}})$, the only way to obtain $f_{e,\pi}$, summing the strata contributions, is when these two holds:

- $(\mathrm{A}_{\mathbf{z}} + \mathrm{B}_{\mathbf{z}}) = (\mathrm{A} + \mathrm{B}),\ \forall \mathbf{z}$;
- $(\mathrm{C}_{\mathbf{z}} + \mathrm{D}_{\mathbf{z}}) = (\mathrm{C} + \mathrm{D}),\ \forall \mathbf{z}$.

However, this implies that there is only one stratum, thus we are in the unconditional case. □

3.1.2 Test Procedure for F_e

Once the value of the statistic is computed, conditioning or not, the independence test is performed. For a rule $\pi : \sigma \to y$, the null hypothesis H_0 states the independence between σ and y (see Definition 2.6.1). It is therefore necessary to compute also the expected value of $f_{e,\pi}$ and $f_{e,\pi,\text{cond}}$ under the null hypothesis.

Theorem 3.1.2 states which is the expected value under the null hypothesis.

Theorem 3.1.2. The expected value of statistics $f_{e,\pi}$ and $f_{e,\pi,\text{cond}}$ under the null hypothesis is 0.

Below is reported the proof of this statement, but before I propose a rewriting of the contingency tables of Tables 3.1 and 3.2, needed for the proof, and that will be used also later in this thesis. I start form Table 3.1. Observe that the values A, B, C, D can be modeled with a Binomial random variable Bin(n,p), which, I recall, counts the number of successes, defined by the probability p, out of n trials. For example, A is the number of times, out of the total number of samples m, in which σ = T and Y = y, thus is is modeled by: $X_{\mathrm{A}} \sim \mathrm{Bin}(m,\, p(\sigma, y))$. Similarly, B, C and D are modeled, respectively, by: $X_{\mathrm{B}} \sim \mathrm{Bin}(m,\, p(\sigma, \bar{y}))$, $X_{\mathrm{C}} \sim \mathrm{Bin}(m,\, p(\bar{\sigma}, y))$ and $X_{\mathrm{D}} \sim \mathrm{Bin}(m,\, p(\bar{\sigma}, \bar{y}))$.

At this point it is possible to rewrite the contingency table. The number of times we "expect" to fall in case A, is given by the expected value of X_{A}, that is: $\hat{\mathbb{E}}[X_{\mathrm{A}}] = m \cdot p(\sigma, y)$. Following the same reasoning for B, C and D, we obtain the contingency of Table 3.3.

In the conditional case the way of proceeding is very similar. The counts to be modeled are relative to a $\mathbf{z}$-stratum, thus the total number of samples is $m_{\mathbf{z}}$ and the probabilities of success take into account for the presence of the confounder, conditioning on it. For example, $\mathrm{A}_{\mathbf{z}}$ is modeled by: $X_{\mathrm{A}_{\mathbf{z}}} \sim \mathrm{Bin}(m_{\mathbf{z}}, p(\sigma, y|\mathbf{z}))$.

	$Y = y$	$Y \neq y$	
$\sigma = T$	$m \cdot p(\sigma, y)$	$m \cdot p(\sigma, \bar{y})$	$m \cdot p(\sigma)$
$\bar{\sigma} = T$	$m \cdot p(\bar{\sigma}, y)$	$m \cdot p(\bar{\sigma}, \bar{y})$	$m \cdot p(\bar{\sigma})$
	$m \cdot p(y)$	$m \cdot p(\bar{y})$	m

Table 3.3: Rewriting of the contingency table in the unconditional case using probabilities.

$Z = \mathbf{z}$	$Y = y$	$Y \neq y$	
$\sigma = T$	$m_{\mathbf{z}} \cdot p(\sigma, y\|\mathbf{z})$	$m_{\mathbf{z}} \cdot p(\sigma, \bar{y}\|\mathbf{z})$	$m_{\mathbf{z}} \cdot p(\sigma\|\mathbf{z})$
$\bar{\sigma} = T$	$m_{\mathbf{z}} \cdot p(\bar{\sigma}, y\|\mathbf{z})$	$m_{\mathbf{z}} \cdot p(\bar{\sigma}, \bar{y}\|\mathbf{z})$	$m_{\mathbf{z}} \cdot p(\bar{\sigma}\|\mathbf{z})$
	$m_{\mathbf{z}} \cdot p(y\|\mathbf{z})$	$m_{\mathbf{z}} \cdot p(\bar{y}\|\mathbf{z})$	$m_{\mathbf{z}}$

Table 3.4: Rewriting of the contingency table in the conditional case using probabilities.

Repeating the previous reasoning we obtain the rewriting showed in Table 3.4

At this point it is possible to prove Theorem 3.1.2.

Proof of Theorem 3.1.2. Unconditional case Let us start form the unconditional case; in Eq.(3.10) is defined the random variable that represents the value assumed by the sample statistic $f_{e,\pi}(s_i)$, according to what was derived in Table 3.3.

$$X_{e,\pi,s_i} = \begin{cases} \frac{1}{m \cdot p(\sigma)} & \text{with } p(\sigma, y) \\ -\frac{1}{m \cdot p(\bar{\sigma})} & \text{with } p(\bar{\sigma}, y) \\ 0 & \text{with } p(\sigma, \bar{y}) \\ 0 & \text{with } p(\bar{\sigma}, \bar{y}). \end{cases} \tag{3.10}$$

Its expected value is:

$$\hat{\mathbb{E}}[X_{e,\pi,s_i}] = \frac{1}{m \cdot p(\sigma)} \cdot p(\sigma, y) - \frac{1}{m \cdot p(\bar{\sigma})} \cdot p(\bar{\sigma}, y). \tag{3.11}$$

Under the null hypothesis (Definition 2.6.1), the joint probability can be split, obtaining the following expression for the expectation:

$$\begin{aligned} \hat{\mathbb{E}}_{H_0}[X_{e,\pi,s_i}] &= \frac{1}{m \cdot p(\sigma)} \cdot p(\sigma) \cdot p(y) - \frac{1}{m \cdot p(\bar{\sigma})} \cdot p(\bar{\sigma}) \cdot p(y) = \\ &= \frac{1}{m} \cdot p(y) - \frac{1}{m} \cdot p(y) = 0. \end{aligned} \tag{3.12}$$

From this it can be easily derived that also the expected value of the overall statistic $f_{e,\pi}$ is 0. In fact, it is computed by summing all the single contributions $f_{e,\pi}(s_i), \forall s_i \in S$ (Eq.(3.6)) and by the property of linearity of expectation: $\hat{\mathbb{E}}_{H_0}[f_{e,\pi}] = \sum_{s_i \in S} \hat{\mathbb{E}}_{H_0}[X_{e,\pi,s_i}] = 0$.

Conditional case. Let us now switch to the conditional case. The random variable $X_{e,\pi,s_i,\mathbf{z}}$, representing the values assumed by the sample statistic in a $\mathbf{z}$-stratum, is

obtained as in the previous case, with values taken from Table 3.4.

$$X_{e,\pi,s_i,\mathbf{z}} = \begin{cases} \frac{1}{m_{\mathbf{z}} \cdot p(\sigma|\mathbf{z})} & \text{with } p(\sigma, y|\mathbf{z}); \\ -\frac{1}{m_{\mathbf{z}} \cdot p(\bar{\sigma}|\mathbf{z})} & \text{with } p(\bar{\sigma}, y|\mathbf{z}); \\ 0 & \text{with } p(\sigma, \bar{y}|\mathbf{z}); \\ 0 & \text{with } p(\bar{\sigma}, \bar{y}|\mathbf{z}). \end{cases} \tag{3.13}$$

The expected value is given by:

$$\hat{\mathbb{E}}[X_{e,\pi,s_i,\mathbf{z}}] = \frac{1}{m_{\mathbf{z}} \cdot p(\sigma|\mathbf{z})} \cdot p(\sigma, y|\mathbf{z}) - \frac{1}{m_{\mathbf{z}} \cdot p(\bar{\sigma}|\mathbf{z})} \cdot p(\bar{\sigma}, y|\mathbf{z}) \tag{3.14}$$

which under the null hypothesis becomes:

$$\begin{aligned} \hat{\mathbb{E}}_{H_0}[X_{e,\pi,s_i,\mathbf{z}}] &= \frac{1}{m_{\mathbf{z}} \cdot p(\sigma|\mathbf{z})} \cdot p(\sigma|\mathbf{z}) \cdot p(y|\mathbf{z}) - \frac{1}{m_{\mathbf{z}} \cdot p(\bar{\sigma}|\mathbf{z})} \cdot p(\bar{\sigma}|\mathbf{z}) \cdot p(y|\mathbf{z}) = \\ &= \frac{1}{m_{\mathbf{z}}} \cdot p(y|\mathbf{z}) - \frac{1}{m_{\mathbf{z}}} \cdot p(y|\mathbf{z}) = 0. \end{aligned} \tag{3.15}$$

This is exactly the same situation as before. The overall expected value is obtained combining all the strata contributions, but since in each stratum it is 0 we can conclude that also in the conditional case the expected value of the statistic under the null hypothesis is 0.

□

Now that Theorem 3.1.2 has been proved, it is possible to define the test procedure for the class F_e:

- compute $f_{e,\pi}$ (or $f_{e,\pi,\text{cond}}$) from the dataset S;
- compare this value with the one expected under H_0: if the confidence interval built around the statistic traps the 0, then the null hypothesis is accepted (because we fall "close" the value expected under independence); otherwise the null hypothesis is rejected, thus there is dependence between σ and y.

Figure 3.1 reports a summary of the test procedure. In orange are reported the statistic and the confidence interval, while in green the expected value under H_0, which is 0. It the example reported, in the unconditional case there is dependence, while in the conditional one no.

3.2 Statistic that mimics the "broken down effect"

In this paragraph I present another statistic, or better a family of statistics, F_4, which aims at the estimation of the effect, but in a different way with respect to the previous

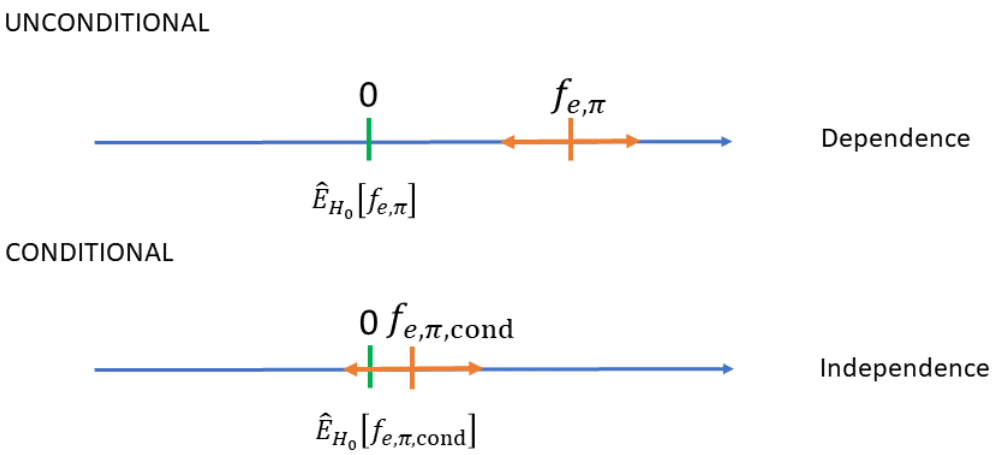

Figure 3.1: Graphical summary of the test procedure for F_e

one. In fact, instead of considering the overall effect, it is broken down into four pieces, hence the name, estimated separately.

In order to better understand the idea behind F_4, I report below the general definition of the effect, based on a contingency table, and a manipulation of the formula needed for our purposes:

$$\begin{aligned} e(\sigma) &= \frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}} - \frac{\mathrm{C}}{\mathrm{C}+\mathrm{D}} = \\ &= \frac{\frac{\mathrm{A}}{m}}{\frac{\mathrm{A}+\mathrm{B}}{m}} - \frac{\frac{\mathrm{C}}{m}}{\frac{\mathrm{C}+\mathrm{D}}{m}} = \\ &= \frac{p(\sigma, y)}{p(\sigma)} - \frac{p(\bar{\sigma}, y)}{p(\bar{\sigma})}. \end{aligned} \tag{3.16}$$

The idea is to estimate each probability $p(\sigma, y)$, $p(\bar{\sigma}, y)$, $p(\sigma)$, $p(\bar{\sigma})$, respectively with the statistics $f_{4.1,\pi}$, $f_{4.2,\pi}$, $f_{4.3,\pi}$, $f_{4.4,\pi}$. The advantage of doing so is that the issue regarding the discrepancy in the order of magnitude between the statistic and the relative bound, which arose with $f_{e,\pi}$ (the other statistic that mimics the effect), is solved. In fact, each of the four statistic has a magnitude bigger than the relative bound, but other problems arose when the four statistics are combined to obtain the overall one $f_{4,\pi}$. This will be exhaustively discussed in Section 4.3.2.

Let us start by the formalization of the sample statistics in the unconditional case (Eqs.(3.17)-(3.20)). For example, since $f_{4.1,\pi}$ estimates the probability $p(\sigma, y)$, its sample version $f_{4.1,\pi}(s_i)$ will be 1 only when σ = T and the target Y is the desired one y. Given this explanation, the derivation of the other three is trivial.

$$f_{4.1,\pi}(s_i) = \begin{cases} 1 & \text{when } \sigma = T \text{ and } \mathrm{Y} = y \\ 0 & \text{when } \sigma = T \text{ and } \mathrm{Y} \neq y \\ 0 & \text{when } \bar{\sigma} = T \text{ and } \mathrm{Y} = y \\ 0 & \text{when } \bar{\sigma} = T \text{ and } \mathrm{Y} \neq y \end{cases} \tag{3.17}$$

$$f_{4.2,\pi}(s_i) = \begin{cases} 0 & \text{when } \sigma = T \text{ and } \mathrm{Y} = y \\ 0 & \text{when } \sigma = T \text{ and } \mathrm{Y} \neq y \\ 1 & \text{when } \bar{\sigma} = T \text{ and } \mathrm{Y} = y \\ 0 & \text{when } \bar{\sigma} = T \text{ and } \mathrm{Y} \neq y \end{cases} \tag{3.18}$$

$$f_{4.3,\pi}(s_i) = \begin{cases} 1 & \text{when } \sigma = T \\ 0 & \text{when } \bar{\sigma} = T \end{cases} \tag{3.19}$$

$$f_{4.4,\pi}(s_i) = \begin{cases} 0 & \text{when } \sigma = T \\ 1 & \text{when } \bar{\sigma} = T \end{cases} \tag{3.20}$$

Making use of the indicator functions of Eqs.(3.3) and (3.4), they can be rewritten in a compact way (Eqs.(3.21)-(3.24)).

$$f_{4.1,\pi}(s_i) = \mathbf{1}_\sigma \cdot \mathbf{1}_\mathrm{Y} \tag{3.21}$$

$$f_{4.2,\pi}(s_i) = (1 - \mathbf{1}_\sigma) \cdot \mathbf{1}_\mathrm{Y} \tag{3.22}$$

$$f_{4.3,\pi}(s_i) = \mathbf{1}_\sigma \tag{3.23}$$

$$f_{4.4,\pi}(s_i) = (1 - \mathbf{1}_\sigma) \tag{3.24}$$

At this point it is possible to define the overall statistics; let us start from $f_{4.1,\pi}$. The sum of the values $f_{4.1,\pi}(s_i), \forall s_i \in S$ gives A and dividing by the dataset size m we obtain the probability $p(\sigma, y)$, thus:

$$f_{4.1,\pi} = \hat{\mathbb{E}}_S[f_{4.1,\pi}(s_i)] = \frac{1}{m} \sum_{s_i \in S} f_{4.1,\pi}(s_i) = \frac{\mathrm{A}}{m} = p(\sigma, y) \tag{3.25}$$

Similarly, summing $f_{4.2,\pi}(s_i)$, $f_{4.3,\pi}(s_i)$, $f_{4.4,\pi}(s_i), \forall s_i \in S$ we get, respectively, C, (A+B) and (C+D) and dividing by m the corresponding probabilities:

$$f_{4.2,\pi} = \hat{\mathbb{E}}_S[f_{4.2,\pi}(s_i)] = \frac{1}{m} \sum_{s_i \in S} f_{4.2,\pi}(s_i) = \frac{\mathrm{C}}{m} = p(\bar{\sigma}, y) \tag{3.26}$$

$$f_{4.3,\pi} = \hat{\mathbb{E}}_S[f_{4.3,\pi}(s_i)] = \frac{1}{m}\sum_{s_i \in S} f_{4.3,\pi}(s_i) = \frac{\mathrm{A}+\mathrm{B}}{m} = p(\sigma) \tag{3.27}$$

$$f_{4.4,\pi} = \hat{\mathbb{E}}_S[f_{4.4,\pi}(s_i)] = \frac{1}{m}\sum_{s_i \in S} f_{4.4,\pi}(s_i) = \frac{\mathrm{C}+\mathrm{D}}{m} = p(\bar{\sigma}) \tag{3.28}$$

The codomain of each sample statistic, which influences the bound, is $\{0,1\}$ like the one of the overall statistics (observe that they are probabilities). Therefore family F_4 solves the problem due to the discrepancy in the order of magnitude between the bound and the statistic, which arose with F_e.

The estimation of the effect, thus $f_{4,\pi}$, is given by the combination of the four statistics:

$$f_{4,\pi} = \hat{e}(\sigma) = \frac{f_{4.1,\pi}}{f_{4.3,\pi}} - \frac{f_{4.2,\pi}}{f_{4.4,\pi}}. \tag{3.29}$$

In this case, we are estimating four different bounds, one for each statistic, thus the procedure to compute the confidence interval is more complicated than just removing and adding the bound to the overall estimation. The lower bound is given by the following expression:

$$lb_4 = \begin{cases} \frac{f_{4.1,\pi}-b_1}{f_{4.3,\pi}+b_3} - \frac{f_{4.2,\pi}+b_2}{f_{4.4,\pi}-b_4} & \text{when } 0 \le \frac{f_{4.1,\pi}-b_1}{f_{4.3,\pi}+b_3} \le 1, 0 \le \frac{f_{4.2,\pi}+b_2}{f_{4.4,\pi}-b_4} \le 1 \\ \frac{f_{4.1,\pi}-b_1}{f_{4.3,\pi}+b_3} - 1 & \text{when } 0 \le \frac{f_{4.1,\pi}-b_1}{f_{4.3,\pi}+b_3} \le 1, f_{4.4,\pi} - b_4 \le 0 \\ 0 - \frac{f_{4.2,\pi}+b_2}{f_{4.4,\pi}-b_4} & \text{when } f_{4.1,\pi} - b_1 \le 0, 0 \le \frac{f_{4.2,\pi}+b_2}{f_{4.4,\pi}-b_4} \le 1 \\ 0 - 1 & \text{when } f_{4.1,\pi} - b_1 \le 0, f_{4.4,\pi} - b_4 \le 0. \end{cases} \tag{3.30}$$

The first line correspond to the case in which both ratios, which are probabilities, are well defined. Since it is the lower bound, we need to minimize the first ratio composing $f_{4,\pi}$, which has positive sign; to do that we reduce the numerator, subtracting b_1 and we increment the numerator adding b_3. The second ratio, which instead has negative sign, needs to be maximised; for this purpose the numerator is increased adding b_2, while the denominator is decreased of the quantity b_4.

However it can be that $f_{4.4,\pi} - b_4$ is negative, changing the sign of the second ratio and this could lead to a lower bound greater that the statistic itself (it was observed also experimentally). To avoid that one could think to bring the negative value to the lowest possible for a probability, but this implicates a zero division, which gives infinity. For this reason, I decided to set the second term equal to 1, which is the maximum significant value, thus the corresponding infinity. It could happen that also $f_{4.1,\pi} - b_1$ is negative. In this case the approach is easier because it appears in the numerator, thus it is sufficient to set it equal to zero and all the first term becomes null. Finally, the fourth row is relative to the case in which both the critical terms are negative.

To obtain the upper bound it is necessary to combine the bounds and the statistics in the opposite way. In the following I report its definition:

$$ub_4 = \begin{cases} \frac{f_{4.1,\pi}+b_1}{f_{4.3,\pi}-b_3} - \frac{f_{4.2,\pi}-b_2}{f_{4.4,\pi}+b_4} & \text{when } 0 \le \frac{f_{4.1,\pi}+b_1}{f_{4.3,\pi}-b_3} \le 1, 0 \le \frac{f_{4.2,\pi}-b_2}{f_{4.4,\pi}+b_4} \le 1 \\ 1 - \frac{f_{4.2,\pi}-b_2}{f_{4.4,\pi}+b_4} & \text{when } f_{4.3,\pi} - b_3 \le 0, 0 \le \frac{f_{4.2,\pi}-b_2}{f_{4.4,\pi}+b_4} \le 1 \\ \frac{f_{4.1,\pi}+b_1}{f_{4.3,\pi}-b_3} - 0 & \text{when } 0 \le \frac{f_{4.1,\pi}+b_1}{f_{4.3,\pi}-b_3} \le 1, f_{4.2,\pi} - b_2 \le 0 \\ 1 - 0 & \text{when } f_{4.2,\pi} - b_2 \le 0, f_{4.3,\pi} - b_3 \le 0. \end{cases} \tag{3.31}$$

3.2.1 Extension of $f_{4,\pi}$ in the conditional case

In the conditional case it is necessary to combine the results coming from all the strata, however within a **z**-stratum the estimation procedure is exactly the same of the unconditional case. The quantity to be estimated is:

$$\begin{aligned} \hat{e}(\sigma) &= \sum_{\mathbf{z}\in Z} \left(\frac{A_{\mathbf{z}}}{A_{\mathbf{z}} + B_{\mathbf{z}}} - \frac{C_{\mathbf{z}}}{C_{\mathbf{z}} + D_{\mathbf{z}}} \right) \cdot \hat{p}(\mathbf{z}) = \\ &= \sum_{\mathbf{z}\in Z} \left(\frac{\frac{A_{\mathbf{z}}}{m_{\mathbf{z}}}}{\frac{A_{\mathbf{z}}+B_{\mathbf{z}}}{m_{\mathbf{z}}}} - \frac{\frac{C_{\mathbf{z}}}{m_{\mathbf{z}}}}{\frac{C_{\mathbf{z}}+D_{\mathbf{z}}}{m_{\mathbf{z}}}} \right) \cdot \hat{p}(\mathbf{z}) = \\ &= \sum_{\mathbf{z}\in Z} \left(\frac{p(\sigma, y|\mathbf{z})}{p(\sigma|\mathbf{z})} - \frac{p(\bar{\sigma}, y|\mathbf{z})}{p(\bar{\sigma}|\mathbf{z})} \right) \cdot \hat{p}(\mathbf{z}) \end{aligned} \tag{3.32}$$

though the four functions: $f_{4.1,\pi,\mathbf{z}}$, $f_{4.2,\pi,\mathbf{z}}$, $f_{4.3,\pi,\mathbf{z}}$ and $f_{4.4,\pi,\mathbf{z}}$.

The sample statistics are defined in Eqs.(3.33)-(3.36) and as usual, s_j are the element of S for which the value of the confounder is the analysed one **z**.

$$f_{4.1,\pi,\mathbf{z}}(s_j) = \begin{cases} 1 & \text{when } \sigma = T \text{ and } Y = y, \text{ given } \mathbf{z} \\ 0 & \text{when } \sigma = T \text{ and } Y \neq y, \text{ given } \mathbf{z} \\ 0 & \text{when } \bar{\sigma} = T \text{ and } Y = y, \text{ given } \mathbf{z} \\ 0 & \text{when } \bar{\sigma} = T \text{ and } Y \neq y, \text{ given } \mathbf{z} \end{cases} \tag{3.33}$$

$$f_{4.2,\pi,\mathbf{z}}(s_j) = \begin{cases} 0 & \text{when } \sigma = T \text{ and } Y = y, \text{ given } \mathbf{z} \\ 0 & \text{when } \sigma = T \text{ and } Y \neq y, \text{ given } \mathbf{z} \\ 1 & \text{when } \bar{\sigma} = T \text{ and } Y = y, \text{ given } \mathbf{z} \\ 0 & \text{when } \bar{\sigma} = T \text{ and } Y \neq y, \text{ given } \mathbf{z} \end{cases} \tag{3.34}$$

$$f_{4.3,\pi,\mathbf{z}}(s_j) = \begin{cases} 1 & \text{when } \sigma = T, \text{ given } \mathbf{z} \\ 0 & \text{when } \bar{\sigma} = T, \text{ given } \mathbf{z} \end{cases} \tag{3.35}$$

$$f_{4.4,\pi,\mathbf{z}}(s_j) = \begin{cases} 1 & \text{when } \sigma = T, \text{ given } \mathbf{z} \\ 0 & \text{when } \bar{\sigma} = T, \text{ given } \mathbf{z} \end{cases} \tag{3.36}$$

Differently from the previous cases, the stratum probability $\hat{p}(\mathbf{z})$ is not already in-

cluded in the sample statistics. In fact, if this would be the case when $f_{4,\pi}$ is computed, $\hat{p}(\mathbf{z})$ will simplify in the ratios.

Now I define the four overall statistics. Summing the values $f_{4.1,\pi,\mathbf{z}}(s_j)$, $f_{4.2,\pi,\mathbf{z}}(s_j)$, $f_{4.3,\pi,\mathbf{z}}(s_j)$ and $f_{4.4,\pi,\mathbf{z}}(s_j)$ $\forall s_j \in S_{\mathbf{z}}$, we obtain the counts relative to the $\mathbf{z}$-stratum under analysis, respectively $\mathrm{A}_{\mathbf{z}}$, $\mathrm{C}_{\mathbf{z}}$, $(\mathrm{A}_{\mathbf{z}} + \mathrm{B}_{\mathbf{z}})$ and $(\mathrm{C}_{\mathbf{z}} + \mathrm{D}_{\mathbf{z}})$ and dividing by $m_{\mathbf{z}}$ we obtain the desired probabilities:

$$f_{4.1,\pi,\mathbf{z}} = \mathbb{E}_{S_{\mathbf{z}}}[f_{4.1,\pi,\mathbf{z}}(s_j)] = \frac{1}{m_{\mathbf{z}}} \sum_{s_j \in S_{\mathbf{z}}} f_{4.1,\pi,\mathbf{z}}(s_j) = \frac{\mathrm{A}_{\mathbf{z}}}{m_{\mathbf{z}}} = p(\sigma, y|\mathbf{z}) \tag{3.37}$$

$$f_{4.2,\pi,\mathbf{z}} = \mathbb{E}_{S_{\mathbf{z}}}[f_{4.2,\pi,\mathbf{z}}(s_j)] = \frac{1}{m_{\mathbf{z}}} \sum_{s_i \in S_{\mathbf{z}}} f_{4.2,\pi,\mathbf{z}}(s_j) = \frac{\mathrm{C}_{\mathbf{z}}}{m_{\mathbf{z}}} = p(\bar{\sigma}, y|\mathbf{z}) \tag{3.38}$$

$$f_{4.3,\pi,\mathbf{z}} = \mathbb{E}_{S_{\mathbf{z}}}[f_{4.3,\pi,\mathbf{z}}(s_j)] = \frac{1}{m_{\mathbf{z}}} \sum_{s_j \in S_{\mathbf{z}}} f_{4.3,\pi,\mathbf{z}}(s_j) = \frac{\mathrm{A}_{\mathbf{z}} + \mathrm{B}_{\mathbf{z}}}{m_{\mathbf{z}}} = p(\sigma|\mathbf{z}) \tag{3.39}$$

$$f_{4.4,\pi,\mathbf{z}} = \mathbb{E}_{S_{\mathbf{z}}}[f_{4.4,\pi,\mathbf{z}}(s_j)] = \frac{1}{m_{\mathbf{z}}} \sum_{s_j \in S_{\mathbf{z}}} f_{4.4,\pi,\mathbf{z}}(s_j) = \frac{\mathrm{C}_{\mathbf{z}} + \mathrm{D}_{\mathbf{z}}}{m_{\mathbf{z}}} = p(\bar{\sigma}|\mathbf{z}) \tag{3.40}$$

The estimation of the effect in a $\mathbf{z}$-stratum $\hat{e}(\sigma|\mathbf{z})$, for us $f_{4,\pi,\mathbf{z}}$, is given by the combination of the four statistics above, as in the unconditional case:

$$f_{4,\pi,\mathbf{z}} = \hat{e}(\sigma|\mathbf{z}) = \frac{f_{4.1,\pi,\mathbf{z}}}{f_{4.3,\pi,\mathbf{z}}} - \frac{f_{4.2,\pi,\mathbf{z}}}{f_{4.4,\pi,\mathbf{z}}} \tag{3.41}$$

while the lower and upper bounds have the same formulation of before, but restricted to the stratum of interest:

$$lb_{4,\mathbf{z}} = \begin{cases} \frac{f_{4.1,\pi,\mathbf{z}} - b_1}{f_{4.3,\pi,\mathbf{z}} + b_3} - \frac{f_{4.2,\pi,\mathbf{z}} + b_2}{f_{4.4,\pi,\mathbf{z}} - b_4} & \text{when } 0 \leq \frac{f_{4.1,\pi,\mathbf{z}} - b_1}{f_{4.3,\pi,\mathbf{z}} + b_3} \leq 1, 0 \leq \frac{f_{4.2,\pi,\mathbf{z}} + b_2}{f_{4.4,\pi,\mathbf{z}} - b_4} \leq 1 \\ \frac{f_{4.1,\pi,\mathbf{z}} - b_1}{f_{4.3,\pi,\mathbf{z}} + b_3} - 1 & \text{when } 0 \leq \frac{f_{4.1,\pi,\mathbf{z}} - b_1}{f_{4.3,\pi,\mathbf{z}} + b_3} \leq 1, f_{4.4,\pi,\mathbf{z}} - b_4 \leq 0 \\ 0 - \frac{f_{4.2,\pi,\mathbf{z}} + b_2}{f_{4.4,\pi,\mathbf{z}} - b_4} & \text{when } f_{4.1,\pi,\mathbf{z}} - b_1 \leq 0, 0 \leq \frac{f_{4.2,\pi,\mathbf{z}} + b_2}{f_{4.4,\pi,\mathbf{z}} - b_4} \leq 1 \\ 0 - 1 & \text{when } f_{4.1,\pi,\mathbf{z}} - b_1 \leq 0, f_{4.4,\pi,\mathbf{z}} - b_4 \leq 0 \end{cases} \tag{3.42}$$

$$ub_{4,\mathbf{z}} = \begin{cases} \frac{f_{4.1,\pi,\mathbf{z}} + b_1}{f_{4.3,\pi,\mathbf{z}} - b_3} - \frac{f_{4.2,\pi,\mathbf{z}} - b_2}{f_{4.4,\pi,\mathbf{z}} + b_4} & \text{when } 0 \leq \frac{f_{4.1,\pi,\mathbf{z}} + b_1}{f_{4.3,\pi,\mathbf{z}} - b_3} \leq 1, 0 \leq \frac{f_{4.2,\pi,\mathbf{z}} - b_2}{f_{4.4,\pi,\mathbf{z}} + b_4} \leq 1 \\ 1 - \frac{f_{4.2,\pi,\mathbf{z}} - b_2}{f_{4.4,\pi,\mathbf{z}} + b_4} & \text{when } f_{4.3,\pi,\mathbf{z}} - b_3 \leq 0, 0 \leq \frac{f_{4.2,\pi,\mathbf{z}} - b_2}{f_{4.4,\pi,\mathbf{z}} + b_4} \leq 1 \\ \frac{f_{4.1,\pi,\mathbf{z}} + b_1}{f_{4.3,\pi,\mathbf{z}} - b_3} - 0 & \text{when } 0 \leq \frac{f_{4.1,\pi,\mathbf{z}} + b_1}{f_{4.3,\pi,\mathbf{z}} - b_3} \leq 1, f_{4.2,\pi,\mathbf{z}} - b_2 \leq 0 \\ 1 - 0 & \text{when } f_{4.2,\pi,\mathbf{z}} - b_2 \leq 0, f_{4.3,\pi,\mathbf{z}} - b_3 \leq 0. \end{cases} \tag{3.43}$$

Finally, the overall effect estimation is obtained summing the single strata contri-

butions, defined in Eq.(3.41), each multiplied by the relative probability:

$$f_{4,\pi,\text{cond}} = \hat{e}_{\text{cond}}(\sigma) = \sum_{\mathbf{z}\in Z}\left(\frac{f_{4.1,\pi,\mathbf{z}}}{f_{4.3,\pi,\mathbf{z}}} - \frac{f_{4.2,\pi,\mathbf{z}}}{f_{4.4,\pi,\mathbf{z}}}\right)\cdot \hat{p}(\mathbf{z}) \tag{3.44}$$

while the lower and upper bounds, combining the values obtained in the single strata, each weighted by the relative probability:

$$lb_{4,\text{cond}}(\sigma) = \sum_{\mathbf{z}\in Z} lb_{4,\mathbf{z}}\cdot \hat{p}(\mathbf{z}) \tag{3.45}$$

$$ub_{4,\text{cond}}(\sigma) = \sum_{\mathbf{z}\in Z} ub_{4,\mathbf{z}}\cdot \hat{p}(\mathbf{z}). \tag{3.46}$$

The analysis of the bounds is reported in paragraph 4.3.2.

3.2.2 Test procedure for F_4

Once the statistic and the relative lower and upper bounds are computed, the test procedure is the same that for f_e.

In principle I should formally prove that the expected value of $f_{4,\pi}$ and $f_{4,\pi,\text{cond}}$ under the null hypothesis is 0, but it can be avoided making the following observation. Let us consider the expression for the statistic in the unconditional and conditional case, derived respectively in Eqs.(3.29) and (3.44), they are exactly equal to the empirical effect. From previous literature ([3]) we know that the expected value of the effect, conditioning and not, is always 0, thus I can avoid the proof.

To conclude, the test procedure for the family F_4, for a given rule π, is:

- compute the value of the statistic, $f_{4,\pi}$ if it is considered the unconditional version of the rule, $f_{4,\pi,\text{cond}}$ otherwise;
- compute the lower and the upper bounds;
- if the confidence interval built around the statistic traps the 0, then the null hypothesis is accepted (thus independence holds), otherwise it is rejected (there is dependence).

3.3 Statistic "Ratio"

The first two families of functions, F_e and F_4, do not lead to satisfactory results; the details can be found in section 4.3. I therefore defined a new statistic that I called "Ratio" because its definition is simply a ratio; I denote the relative family of functions F_R.

The idea is to construct a function such that if it is computed on a sample s_i it lies in $[-1, 1]$, but such that its average still lies in $[-1, 1]$, unlike with $f_{e,\pi}$. More formally, the requirements are: $f_{R,\pi}(s_i) : s_i \to [-1, 1]$ and $f_{R,\pi} = \frac{1}{m} \sum_{s_i \in S} f_{R,\pi}(s_i) \in [-1, 1]$.

I start from the unconditional case; the new statistic $f_{R,\pi}$, with respect to the rule $\pi : \sigma \to y$ under analysis is:

$$f_{R,\pi} = \frac{\mathrm{A} + \mathrm{D} - \mathrm{B} - \mathrm{C}}{m}. \tag{3.47}$$

Each sample s_i contributes to the overall statistic adding or removing one unit, therefore the sample statistic is:

$$f_{R,\pi}(s_i) = \begin{cases} 1 & \text{if } \sigma = \mathrm{T} \text{ and } \mathrm{Y} = y \\ 1 & \text{if } \bar{\sigma} = \mathrm{T} \text{ and } \mathrm{Y} \neq y \\ -1 & \text{if } \sigma = \mathrm{T} \text{ and } \mathrm{Y} \neq y \\ -1 & \text{if } \bar{\sigma} = \mathrm{T} \text{ and } \mathrm{Y} = y \end{cases} \tag{3.48}$$

or in a more compact form ($\oplus$ is the logical XOR function):

$$f_{R,\pi}(s_i) = sign\left\{\frac{1}{2} - (\sigma \oplus y)\right\}. \tag{3.49}$$

Averaging the values $f_{R,\pi}(s_i),\ \forall s_i \in S$, we obtain exactly Eq.(3.47). It can be easily seen that both $f_{R,\pi}(s_i)$ and $f_{R,\pi}$ lie in $[-1, 1]$, solving the issue due to the discrepancy in the order of magnitude that affects $f_{e,\pi}$.

Before extending the statistic in the conditional case, I want to give an intuition of how it works. Recalling the definition of effect, $e(\sigma) = \frac{\mathrm{A}}{(\mathrm{A+B})} - \frac{\mathrm{C}}{(\mathrm{C+D})}$, we have that it takes a positive value when A or D, or both are high. This behaviour is reflected also by $f_{R,\pi}$, in fact A and D have a positive sign, so the highest, the bigger the statistic is. Similarly the effect is negative when B and C are high and again this holds for $f_{R,\pi}$. Finally, the only missing case is when the effect is null. This does not have a counterpart for the new statistic, not in the sense that it cannot be 0, but because for $f_{R,\pi}$ being 0 does not mean independence. This will be discussed in detail in the rest of the chapter.

3.3.1 Extension of $f_{R,\pi}$ in the Conditional Case

Let $\mathrm{A}_{\mathbf{z}}$, $\mathrm{B}_{\mathbf{z}}$, $\mathrm{C}_{\mathbf{z}}$ and $\mathrm{D}_{\mathbf{z}}$ be the entries of the contingency table relative to a certain $\mathbf{z}$-stratum and $\hat{p}(\mathbf{z})$ the stratum probability. Then, in the conditional case, the statistic relative to that stratum $f_{R,\pi,\mathbf{z}}$ is:

$$f_{R,\pi,\mathbf{z}} = \frac{\mathrm{A}_{\mathbf{z}} + \mathrm{D}_{\mathbf{z}} - \mathrm{B}_{\mathbf{z}} - \mathrm{C}_{\mathbf{z}}}{m_{\mathbf{z}}} \cdot \hat{p}(\mathbf{z}) \tag{3.50}$$

For $s_j \in S_{\mathbf{z}}$ ($S_{\mathbf{z}}$ as before is the dataset restricted to the entries for which $\mathrm{Z} = \mathbf{z}$), the statistic computed on an element of the sample is:

$$f_{R,\pi,\mathbf{z}}(s_j) = \begin{cases} 1 \cdot p(\mathbf{z}) & \text{if } \sigma = \mathrm{T} \text{ and } \mathrm{Y} = y, \text{ given } \mathbf{z} \\ 1 \cdot p(\mathbf{z}) & \text{if } \bar{\sigma} = \mathrm{T} \text{ and } \mathrm{Y} \neq y, \text{ given } \mathbf{z} \\ -1 \cdot p(\mathbf{z}) & \text{if } \sigma = \mathrm{T} \text{ and } \mathrm{Y} \neq y, \text{ given } \mathbf{z} \\ -1 \cdot p(\mathbf{z}) & \text{if } \bar{\sigma} = \mathrm{T} \text{ and } \mathrm{Y} = y, \text{ given } \mathbf{z} \end{cases} \tag{3.51}$$

or, equivalently, in a more compact form:

$$f_{R,\pi,\mathbf{z}}(s_j) = sign\left\{\frac{1}{2} - (\sigma \oplus y)\right\} \cdot \hat{p}(\mathbf{z}). \tag{3.52}$$

The sample statistic averaged across all s_j gives exactly Eq.(3.50). As final step, to obtain the expression of the statistic in the conditional case, it is sufficient to sum the values $f_{R,\pi,\mathbf{z}}$ for all the $\mathbf{z}$-strata:

$$f_{R,\pi,cond} = \sum_{\mathbf{z} \in \mathrm{Z}} f_{R,\pi,\mathbf{z}}. \tag{3.53}$$

In this case, the sample statistic is in $\{-1, 1\} \cdot \hat{p}(\mathbf{z})$; the worst case is when $\hat{p}(\mathbf{z}) \to 1$, thus the codomain is $\{-1, 1\}$ as in the unconditional case. The empirical sample mean in a stratum $f_{R,\pi,\mathbf{z}}$ is in $[-1, 1] \cdot \hat{p}(\mathbf{z})$, therefore the overall one, which is obtained combining all the strata contributions, is in $[-1, 1]$.

The detailed analysis of the bounds is reported in paragraph 4.3.3.

3.3.2 Analysis of Statistic "Ratio"

The first tests that I performed on F_R underlined that, given a rule and its conditional form, the statistic is the same in the two cases; this behaviour is different from the one of F_e and F_4. In fact for the former it was proven in Theorem 3.1.1 that the statistic is different in the conditional and unconditional case. For the latter it was not formally proved, but since it is exactly the effect, the fact follows.

Theorem 3.3.1 states the behaviour of F_R.

Theorem 3.3.1. Given a rule and its conditional form, the value of statistic in the conditional and unconditional case is the same: $f_{R,\pi} = f_{R,\pi,\mathrm{cond}}$.

Before proving it formally I present an example; it is relative to the first chain dataset of size $m = 1000$. In particular, the rule considered is $\mathrm{X} = 0 \to \mathrm{Y} = 0$ and its conditional form $\mathrm{X} = 0 \to \mathrm{Y} = 0 \mid \mathrm{Z}$ (the confounder values are binary). I start from the unconditional case, the contingency is reported in Table 3.5.

Applying Eq.(3.47) we obtain: $f_{R,\pi} = \frac{\mathrm{A+D-B-C}}{m} = \frac{457+194-329-20}{1000} = \frac{302}{1000} = 0.302$.

In Table 3.6, instead, are reported the contingency tables relative to the conditional case.

	$\mathrm{Y} = y$	$\mathrm{Y} \neq y$	
σ = T	457	329	786
$\bar{\sigma}$ = T	20	194	214
	477	523	1000

Table 3.5: Contingency table for the first chain dataset of size $m = 1000$, for the rule X = 0 → Y = 0 in the unconditional case.

$\mathrm{Z} = z_0$	$\mathrm{Y} = y$	$\mathrm{Y} \neq y$	
σ = T	441	82	523
$\bar{\sigma}$ = T	12	0	12
	453	82	535

$\mathrm{Z} = z_1$	$\mathrm{Y} = y$	$\mathrm{Y} \neq y$	
σ = Y	16	247	263
$\bar{\sigma}$ = T	8	194	202
	24	441	465

Table 3.6: Contingency table in the conditional case.

The application of Eq.(3.50) to the two strata gives: $f_{R,\pi,z_0} = \frac{\mathrm{A}_{z_0}+\mathrm{D}_{z_0}-\mathrm{B}_{z_0}-\mathrm{C}_{z_0}}{m_{z_0}} \cdot \hat{p}(z_0) = \frac{441+0-82-12}{535} \cdot \frac{535}{1000} = \frac{347}{1000} = 0.347$ and $f_{R,\pi,z_1} = \frac{\mathrm{A}_{z_1}+\mathrm{D}_{z_1}-\mathrm{B}_{z_1}-\mathrm{C}_{z_1}}{m_{z_1}} \cdot \hat{p}(z_1) = \frac{16+194-247-8}{465} \cdot \frac{465}{1000} = -\frac{45}{1000} = -0.045$. Combining the two to compute the overall contribution we obtain the same value of the unconditional case: $f_{R,\pi,\mathrm{cond}} = 0.347 - 0.045 = 0.302$.

I will now prove this fact formally, focusing on a rule π.

Proof of Theorem 3.3.1. The proof is really simple, since it naturally follows from Eq.(3.53); expanding the summand we get:

$$\begin{aligned} f_{R,\pi,\mathrm{cond}} &= \sum_{\mathbf{z}\in\mathrm{Z}} f_{2,\pi,\mathbf{z}} = \\ &= \sum_{\mathbf{z}\in\mathrm{Z}} \frac{\mathrm{A}_{\mathbf{z}} + \mathrm{D}_{\mathbf{z}} - \mathrm{B}_{\mathbf{z}} - \mathrm{C}_{\mathbf{z}}}{m_{\mathbf{z}}} \cdot \frac{m_{\mathbf{z}}}{m} = \\ &= \sum_{\mathbf{z}\in\mathrm{Z}} \frac{\mathrm{A}_{\mathbf{z}} + \mathrm{D}_{\mathbf{z}} - \mathrm{B}_{\mathbf{z}} - \mathrm{C}_{\mathbf{z}}}{m}. \end{aligned} \tag{3.54}$$

Observe that the summation on all the $\mathbf{z}$-strata of $A_{\mathbf{z}}$ is exactly A (in fact no sample is missing when all the contingency tables are considered), thus $\sum_{\mathbf{z}\in\mathrm{Z}} A_{\mathbf{z}} = \mathrm{A}$; similarly for B, C and D. As a consequence of this:

$$\begin{aligned} f_{R,\pi,\mathrm{cond}} &= \sum_{\mathbf{z}\in\mathrm{Z}} \frac{\mathrm{A}_{\mathbf{z}} + \mathrm{D}_{\mathbf{z}} - \mathrm{B}_{\mathbf{z}} - \mathrm{C}_{\mathbf{z}}}{m} \\ &= \frac{\mathrm{A} + \mathrm{D} - \mathrm{B} - \mathrm{C}}{m} = f_{2,\pi} \end{aligned} \tag{3.55}$$

which is equal to the statistic in the unconditional case. □

At this point, it remains to derive the expected value of the statistic under the null hypothesis H_0, that is the key value to perform the independence test.

3.3.3 Test Procedure for F_R

The first experiments that I performed suggested that the expected value of the statistic under the null hypothesis, in the unconditional and conditional case are different. This

fact is formalized in Theorem 3.3.2.

Theorem 3.3.2. For family F_R, the expected value of statistic under the null hypothesis H_0 in the unconditional and conditional case is different: $\mathbb{E}_{H_0}[f_{R,\pi}] \neq \mathbb{E}_{H_0}[f_{R,\pi,\text{cond}}]$.

Before proving it I want to stress the differences between family F_R and the previous two. For these latter, given a rule, the statistic takes different values conditioning or not, while the expected value under the null hypothesis is the same (and it is equal to 0). For the new class F_R, instead the statistic takes the same value, for a given rule, but the expected value under H_0, which is the reference quantity for the independence test, is different conditioning or not. This suggests that the test procedure for F_R will be different.

Proof of Theorem 3.3.2. Unconditional case. The derivation in the unconditional case is trivial. Let X_{R,π,s_i} be a random variable that represents the value assumed by the sample statistic $f_{R,\pi}(s_i)$ with the relative probabilities (Eq.(3.56)):

$$X_{R,\pi,s_i} = \begin{cases} 1 & \text{with } p(\sigma, y) \\ 1 & \text{with } p(\bar{\sigma}, \bar{y}) \\ -1 & \text{with } p(\sigma, \bar{y}). \\ -1 & \text{with } p(\bar{\sigma}, y). \end{cases} \tag{3.56}$$

Its expected value is given by:

$$\hat{\mathbb{E}}[X_{R,\pi,s_i}] = p(\sigma, y) + p(\bar{\sigma}, \bar{y}) - p(\bar{\sigma}, y) - p(\sigma, \bar{y}). \tag{3.57}$$

Under H_0, σ and y are independent, thus the joint probability can be decomposed into the product $p(\sigma) \cdot p(y)$, leading to:

$$\hat{\mathbb{E}}_{H_0}[X_{R,\pi,s_i}] = p(\sigma)p(y) + p(\bar{\sigma})p(\bar{y}) - p(\bar{\sigma})p(y) - p(\sigma)p(\bar{y}). \tag{3.58}$$

At this point, the only missing step is to compute the expected value of the overall statistic:

$$\begin{aligned} \hat{\mathbb{E}}_{S,H_0}[f_{R,\pi}] &= \frac{1}{m} \sum_{s_i \in S} \hat{\mathbb{E}}_{S,H_0}[f_{R,\pi}(s_i)] = \\ &= \frac{1}{m} \cdot m \cdot \hat{\mathbb{E}}_{S,H_0}[f_{R,\pi}(s_i)] = \\ &= p(\sigma)p(y) + p(\bar{\sigma})p(\bar{y}) - p(\sigma)p(\bar{y}) - p(\bar{\sigma})p(y). \end{aligned} \tag{3.59}$$

I will now switch to the conditional case.

Conditional case Let $X_{R,\pi,s_i,\mathbf{z}}$ be the random variables representing the values assumed by the sample statistic $f_{R,\pi,\mathbf{z}}(s_i)$ in a $\mathbf{z}$-stratum:

$$X_{R,\pi,s_i,\mathbf{z}} = \begin{cases} 1 \cdot \hat{p}(\mathbf{z}) & \text{with } p(\sigma, y|\mathbf{z}) \\ 1 \cdot \hat{p}(\mathbf{z}) & \text{with } p(\bar{\sigma}, \bar{y}|\mathbf{z}) \\ -1 \cdot \hat{p}(\mathbf{z}) & \text{with } p(\sigma, \bar{y}|\mathbf{z}) \\ -1 \cdot \hat{p}(\mathbf{z}) & \text{with } p(\bar{\sigma}, y|\mathbf{z}). \end{cases} \tag{3.60}$$

Its expected value is given by the following expression:

$$\begin{aligned} \hat{\mathbb{E}}[X_{R,\pi,s_i,\mathbf{z}}] = \hat{p}(\mathbf{z}) \cdot p(\sigma, y|\mathbf{z}) + \hat{p}(\mathbf{z}) \cdot p(\bar{\sigma}, \bar{y}|\mathbf{z}) - 1 \cdot \hat{p}(\mathbf{z}) \cdot p(\sigma, \bar{y}|\mathbf{z}) - 1 \cdot \hat{p}(\mathbf{z}) \cdot p(\bar{\sigma}, y|\mathbf{z}) = \\ = \big[p(\sigma, y|\mathbf{z}) + p(\bar{\sigma}, \bar{y}|\mathbf{z}) - p(\sigma, \bar{y}|\mathbf{z}) - p(\bar{\sigma}, y|\mathbf{z})\big] \cdot \hat{p}(\mathbf{z}). \end{aligned} \tag{3.61}$$

which under the null hypothesis becomes:

Applying the independence assumption to Eq.(3.61), which states $\sigma \perp\!\!\!\perp y \mid \mathbf{z}$, we get the following expression for the expectation under H_0:

$$\hat{\mathbb{E}}_{S,H_0}[f_{R,\pi,\mathbf{z}}(s_j)] = \big[p(\sigma|\mathbf{z})p(y|\mathbf{z}) + p(\bar{\sigma}|\mathbf{z})p(\bar{y}|\mathbf{z}) - p(\sigma|\mathbf{z})p(\bar{y}|\mathbf{z}) - p(\bar{\sigma}|\mathbf{z})p(y|\mathbf{z})\big] \cdot \hat{p}(\mathbf{z}) \tag{3.62}$$

(as usual I indicate with s_j the elements in $S_\mathbf{z}$, the subset of the dataset restricted to the rows in which Z = $\mathbf{z}$).

By the linearity of the expectation, the expected value in a $\mathbf{z}$-stratum is:

$$\begin{aligned} \hat{\mathbb{E}}_{S,H_0}[f_{R,\pi,\mathbf{z}}] = \frac{1}{m_\mathbf{z}} \sum_{s_j \in S_\mathbf{z}} \hat{\mathbb{E}}_{S,H_0}[f_{R,\pi,\mathbf{z}}(s_j)] = \\ = \frac{1}{m_\mathbf{z}} \cdot m_\mathbf{z} \cdot \hat{\mathbb{E}}_{S,H_0}[f_{R,\pi,\mathbf{z}}(s_j)] = \\ = \big[p(\sigma|\mathbf{z})p(y|\mathbf{z}) + p(\bar{\sigma}|\mathbf{z})p(\bar{y}|\mathbf{z}) - p(\sigma|\mathbf{z})p(\bar{y}|\mathbf{z}) - p(\bar{\sigma}|\mathbf{z})p(y|\mathbf{z})\big] \cdot \hat{p}(\mathbf{z}). \end{aligned} \tag{3.63}$$

The only missing step is to sum all the contributions coming from the different strata; this is:

$$\begin{aligned} \hat{\mathbb{E}}_{S,H_0}[f_{R,\pi,\text{cond}}] = \sum_{\mathbf{z} \in Z} \hat{\mathbb{E}}_{S,H_0}[f_{R,\pi,\mathbf{z}}] = \\ = \sum_{\mathbf{z} \in Z} \big[p(\sigma|\mathbf{z})p(y|\mathbf{z}) + p(\bar{\sigma}|\mathbf{z})p(\bar{y}|\mathbf{z}) - p(\sigma|\mathbf{z})p(\bar{y}|\mathbf{z}) - p(\bar{\sigma}|\mathbf{z})p(y|\mathbf{z})\big] \cdot \hat{p}(\mathbf{z}) = \\ = \sum_{\mathbf{z} \in Z} \big[p(\sigma|\mathbf{z})p(y|\mathbf{z}) \cdot \hat{p}(\mathbf{z})\big] + \big[p(\bar{\sigma}|\mathbf{z})p(\bar{y}|\mathbf{z}) \cdot \hat{p}(\mathbf{z})\big] - \big[p(\sigma|\mathbf{z})p(\bar{y}|\mathbf{z}) \cdot \hat{p}(\mathbf{z})\big] - \big[p(\bar{\sigma}|\mathbf{z})p(y|\mathbf{z}) \end{aligned} \tag{3.64}$$

To show that Eq.(3.64) is different from Eq.(3.59) the reasoning that has been made starts from a wrong derivation. Let us consider the general form of the expected value in the conditional case and suppose that we first marginalize (passage (i)) and then we

apply the independence assumption (passage (ii)):

$$\begin{aligned}\hat{\mathbb{E}}_S[f_{R,\pi,\text{cond}}] &= \sum_{\mathbf{z}\in Z}\big[p(\sigma,y|\mathbf{z}) + p(\bar{\sigma},\bar{y}|\mathbf{z}) - p(\sigma,\bar{y}|\mathbf{z}) - p(\bar{\sigma},y|\mathbf{z})\big]\cdot\hat{p}(\mathbf{z}) = \\ &= \sum_{\mathbf{z}\in Z} p(\sigma,y,\mathbf{z}) + \sum_{\mathbf{z}\in Z} p(\bar{\sigma},\bar{y},\mathbf{z}) - \sum_{\mathbf{z}\in Z} p(\sigma,\bar{y},\mathbf{z}) - \sum_{\mathbf{z}\in Z} p(\bar{\sigma},y,\mathbf{z}) = \\ &= (i) = p(\sigma,y) + p(\bar{\sigma},\bar{y}) - p(\sigma,\bar{y}) - p(\bar{\sigma},y) \\ &= (ii) = p(\sigma)p(y) + p(\bar{\sigma})p(\bar{y}) - p(\sigma)p(\bar{y}) - p(\bar{\sigma})p(y).\end{aligned} \tag{3.65}$$

We obtain the same expression of the unconditional case, so one could think that Theorem 3.3.2 is wrong. However the problem is that we applied the wrong independence assumption; in fact in the conditional case it states $\sigma \perp\!\!\!\perp y \mid \mathrm{Z}$ but marginalizing we have removed the confounder.

Thus the correct derivation is:

$$\begin{aligned}\hat{\mathbb{E}}_{S,H_0}[f_{R,\pi,\text{cond}}] &= \sum_{\mathbf{z}\in Z}\big[p(\sigma,y|\mathbf{z}) + p(\bar{\sigma},\bar{y}|\mathbf{z}) - p(\sigma,\bar{y}|\mathbf{z}) - p(\bar{\sigma},y|\mathbf{z})\big]\cdot\hat{p}(\mathbf{z}) = \\ &= \sum_{\mathbf{z}\in Z}\big[p(\sigma|\mathbf{z})p(y|\mathbf{z}) + p(\bar{\sigma}|\mathbf{z})p(\bar{y}|\mathbf{z}) - p(\sigma|\mathbf{z})p(\bar{y}|\mathbf{z}) - p(\bar{\sigma}|\mathbf{z})p(y|\mathbf{z})\big]\cdot\hat{p}(\mathbf{z})\end{aligned} \tag{3.66}$$

and there is no way of marginalizing to obtain $\hat{\mathbb{E}}_{S,H_0}[f_{R,\pi}]$ thus we can conclude the two are different.

□

Given these results, in the next paragraph I define the test procedure for the family F_R.

To summarize, so far it was shown that:

- the value of the statistic in the conditional and unconditional case is the same;
- under H_0, the expected value of the statistic conditioning or not is different.

As anticipated, the behaviour is different with respect to $f_{e,\pi}$ and $f_{4,\pi}$. In fact, in this case the threshold to be used to assess the independence, which is the expected value under H_0 is different conditioning or not, while the statistic is fixed.

The test procedure, for a rule π, is formalized in the following.

- Compute $f_{R,\pi}$ (or $f_{R,\pi,\text{cond}}$ even if it is the same);
- compute the the expected value under the null hypothesis, $\hat{\mathbb{E}}_{S,H_0}[f_{R,\pi}]$ for a non conditional rule or $\hat{\mathbb{E}}_{S,H_0}[f_{R,\pi,\text{cond}}]$ for a conditional one;
- if the confidence interval built around the statistic traps the expected value under H_0, then the null hypothesis is accepted (because the statistic takes the same value it would take in case of independence); otherwise, the null hypothesis is rejected (because the statistic is far from its value in case of independence).

Figure 3.2 is reports a graphical representation of the test procedure. In orange are reported the value of the statistic and the relative confidence interval, while in green the expected value under if independence holds. It is reported an example in which not conditioning there is dependence, while conditioning it is removed.

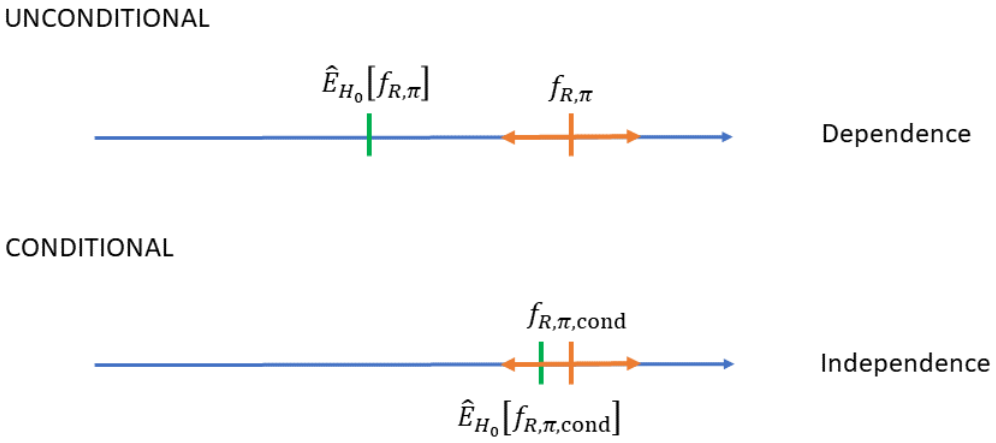

Figure 3.2: Graphical illustration of the test procedure using family F_R.

This concludes the formal definition and discussion about this statistic. In section 4.3.3 is reported the analysis of the order of magnitude of the bound to the Supremum Deviation when the family F_R is used.

Chapter 4

Implementation, Experimental Setup and Bound Evaluation

In this chapter are discussed the technical details of the thesis. I start with the description of the procedure for the computation of the n-MCERA (Eq.(2.36)), since it is not trivial. I then switch to the definition of the synthetic datasets that I used to test the statistics. Finally, I perform the analysis of the magnitude of the two bounds to the Supremum Deviation for each statistic, when it is applied to my datasets.

4.1 n-MCERA Implementation

I start explaining how the computation of n-MCERA was performed, since it is a not trivial step of the code; for simplicity, I report its definition:

$$\hat{R}_m^n(F, S, \Lambda) = \frac{1}{n} \sum_{j=1}^{n} \sup_{f \in F} \frac{1}{m} \sum_{s_i \in S} \lambda_{j,i} f(s_i). \tag{4.1}$$

Let us start from the summation to the right: it requires the computation of the sample statistic $f(s_i), \forall s_i \in S$. Moving to the left, there is a superior computed on all the functions f of the family F: from this it is possible to infer that we need the values $f(s_i)$ not only $\forall s_i \in S$, but also $\forall f \in F$. All this information will be stored in a matrix A, of size $m \times k$ with k cardinality of F, where each row is relative to an element s_i of the dataset, while each column corresponds to a function $f \in F$. To give a better intuition, each function f is labeled by a number, from 1 to k, to stress the fact they are different functions. Here I report the matrix:

$$A = \begin{bmatrix} f_1(s_1) & \dots & f_k(s_1) \\ \vdots & \ddots & \vdots \\ f_1(s_m) & \dots & f_k(s_m) \end{bmatrix} \tag{4.2}$$

Each value $f_\ell(s_i)$, with $\ell \in \{1, \dots, k\}$, is multiplied by a Rademacher random variable $\lambda_{j,i}$, with $j \in \{1, \dots, n\}$ and $i \in \{1, \dots, m\}$. Thus a second matrix Λ of size $n \times m$,

containing all such random variables is needed:

$$\Lambda = \begin{bmatrix} \lambda_{1,1} & \dots & \lambda_{1,m} \\ \vdots & \ddots & \vdots \\ \lambda_{n,1} & \dots & \lambda_{n,m} \end{bmatrix} \tag{4.3}$$

We perform the matrix product between Λ and A, obtaining $B = \Lambda \cdot A$, with size $(n \times k)$, where each row is relative to one of the n-trials of the Monte Carlo approach and each column is relative to a function f_ℓ. An entry, say (r,t) of B, stores the following information: $\sum_{s_i \in S} \lambda_{r,i} f_t(s_i)$. I denote a row of matrix B as $B[j,:]$; the final computation of the n-MCERA thus is:

$$\hat{R}^n_m = \frac{1}{m \cdot n} \sum_{j=1}^{n} \max B[j,:] \tag{4.4}$$

in which I substituted the sup with a max, since we are dealing with a finite set of functions.

4.2 Synthetic Datasets Definition

As anticipated at the beginning of the chapter, the proposed statistics were tested on synthetic datasets; in this section I provide the details about how I constructed them.

For each fundamental structure of a graph (chain, fork, collider) and for each size $m \in \{1000, 100000\}$ I created 100 datasets. Each variable, corresponding to a node, takes values according to a Bernoulli distribution and such value is combined with a certain noise, also modeled with a Bernoulli random variable. The relations among the variables are defined through the logical or function ($\vee$), while the noise is added through the xor function ($\oplus$).

Figure 4.1 shows the relations among variables in the three structures, together with the choice of the parameters of the Bernoulli random variables.

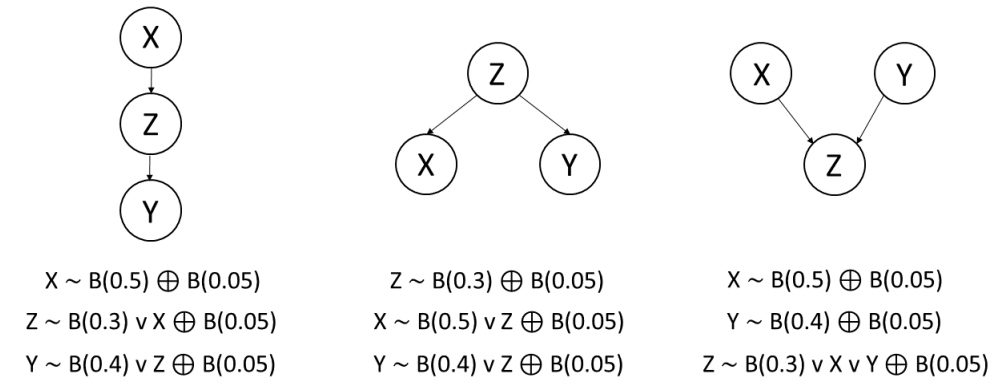

Figure 4.1: Relation among the variables X, Z and Y in the three fundamental structures (from left to right, chain, fork and collider), that I chose to create the synthetic datasets.

Before moving on, I want to motivate why the noise was added through the xor function, instead than the or. Consider for example the chain and suppose that the noise is added everywhere though the or function: when X is 1 (or true) there is no way for Z or Y to be 0. Thanks to the xor instead, the value of Z or Y can flip from 1 to 0, simulating a real-world behaviour.

4.3 Bound Estimation

In this section I develop the estimation of the order of magnitude of the bounds to the Supremum Deviation (Eqs.(2.38) and (2.42)) for the three families F_e, F_4 and F_R, varying some of the parameters defining them.

4.3.1 Estimation of the Bounds for F_e

Let us start with F_e. As anticipated in paragraph 3.1, there is discrepancy in the order of magnitude of the overall statistic $f_{e,\pi}$ and its sample version $f_{e,\pi}(s_i)$ and this gives some problems when bounding it. Their codomain are, respectively, $[-\frac{1}{m}, \frac{1}{m}]$ and $[-1, 1]$, and the extreme values of the latter are involved in the definition of the bound for $f_{e,\pi}$, resulting in a too large interval. Given this evidence, the major aim of this analysis is to see if varying some parameters, like the number of Monte-Carlo trials n, the issue can be fixed.

I decomposed the first bound (Eq.(2.38)) in the following four factors:

$$\tilde{R} \doteq \hat{R}_m^n(F, S, \Lambda) + 2z\sqrt{\frac{\log\frac{\delta}{4}}{2nm}} \tag{4.5}$$

$$\mathrm{f}_1 \doteq \frac{\sqrt{c(4m\tilde{R} + c\log(\frac{4}{\delta}))\log(\frac{4}{\delta})}}{m} \tag{4.6}$$

$$\mathrm{f}_2 \doteq \frac{c\log\frac{\delta}{4}}{m} \tag{4.7}$$

$$\mathrm{f}_3 \doteq c\sqrt{\frac{\log\frac{\delta}{4}}{2m}}. \tag{4.8}$$

Recalling that $f : X \to [a, b] \subset \mathbb{R}$, $z = \max\{|a|, |b|\}$, $c = |b - a|$ and according to what was derived in paragraph 3.1, we have that $f_{e,\pi}(s_i) : X \to [-1, 1]$ and therefore $z = 1$ and $c = 2$. I fixed the significance level δ equal to 0.05.

In the following I do the analysis of the order of magnitude of each factor, for $m =$

$n = 1000$; with this setting the n-MCERA is order of 10^{-4}.

$$\begin{aligned}\tilde{R} &\sim 10^{-4} + 10^{0} \cdot \sqrt{\frac{1}{10^{3} \cdot 10^{3}}} \\ &\sim 10^{-4} + \sqrt{10^{-6}} \\ &\sim 10^{-3}\end{aligned} \tag{4.9}$$

$$\begin{aligned}\mathrm{f}_1 &\sim \frac{\sqrt{10^{0} \cdot (10^{3} \cdot 10^{-3} + 10^{0})}}{10^{3}} \\ &\sim \frac{\sqrt{10^{0} \cdot (10^{0})}}{10^{3}} \\ &\sim 10^{0} \cdot 10^{-3} \\ &\sim 10^{-3}\end{aligned} \tag{4.10}$$

$$\begin{aligned}\mathrm{f}_2 &\sim 10^{0} \cdot \frac{1}{10^{3}} \\ &\sim 10^{-3}\end{aligned} \tag{4.11}$$

$$\begin{aligned}\mathrm{f}_3 &\sim 10^{0} \cdot \sqrt{\frac{1}{10^{3}}} \\ &\sim \sqrt{10^{-3}} \\ &\sim 10^{-1.5}\end{aligned} \tag{4.12}$$

The first three terms have the same order of magnitude of the maximum values the statistic $f_{e,\pi}$ can take, in fact it lies in $[-0.001, 0.001]$ for the considered sample size. However in general it is order of 10^{-4} so this is not sufficient to provide an informative bound. In any case the biggest problem is given by f_3 that will completely erase the information of the statistic.

Since the factors composing the bound depends on m one could think that increasing m the issue could in part be solved. Nevertheless also the codomain of $f_{e,\pi}$ is reduced in magnitude, thus the problem still remains.

It is possible to increase n, however this will affect only the factor $\tilde{R}$, which is not the major source of problems.

Table 4.1 presents a summary of the previous reasoning. I report the magnitude of the bound for the two sample sizes m and for three different values of n, together with the range of values (column "Range") assumed by the statistic. (Note: for $m = 100000$ the order of the n-MCERA is 10^{-7} while the other parameters are the same; I do not report the derivation of the order of magnitude of the bound for this size, but only its value).

Looking at the result it is clear that this approach is unsatisfactory. In fact, subtracting and adding to the statistic a bound which is at least three order of magnitude higher, will completely erase any information.

m	n	Bound	Range
1000	1000	0.1221	$[-0.001, 0.001]$
1000	10000	0.1151	$[-0.001, 0.001]$
1000	100000	0.1127	$[-0.001, 0.001]$
100000	1000	0.0104	$[-0.00001, 0.00001]$
100000	10000	0.0098	$[-0.00001, 0.00001]$
100000	100000	0.0096	$[-0.00001, 0.00001]$

Table 4.1: Value of the bound to computed for various values of m and n; $z = 1$.

However, before switching to the improved bound, another observation was made. The codomain of $f_{e,\pi}(s_i)$ is $[-1, 1]$, but this was obtained in the worst case scenario in which there is only one sample that satisfies σ (alternatively, (A+B) is 1) and only one satisfying $\bar{\sigma}$ (or (C+D) is 1). This is conservative, but in practise having that σ is satisfied (or not) by only one sample is likely to be an error. Therefore I fixed a minimum support, that is a minimum number of samples that have to satisfy σ (or $\bar{\sigma}$) to consider the rule $\pi : \sigma \rightarrow y$ valid. I arbitrarily set this threshold to $\frac{m}{100}$. More formally, a rule π is acceptable if: $\min = \{(\text{A+B}), (\text{C+D})\} \geq \frac{m}{100}$, where, as usual, A, B, C, D are the entries of the relative contingency table.

Given that, the largest possible codomain of $f_{e,\pi}(s_i)$ is $\left[-\frac{1}{\frac{m}{100}}, \frac{1}{\frac{m}{100}} \right]$ and therefore $z = 0.1$ and $c = 0.2$ for $m = 1000$, while $z = 0.001$ and $c = 0.002$ for $m = 100000$.

Table 4.2 shows the results obtained in this setting. With respect to the previous case, they are smaller of a factor 10; the range of values assumed by the statistic is the same, the only difference is that rules with low support are not considered. In any case they the bound is still not informative (only setting the support equal to m it becomes informative, but this makes no sense).

m	n	Bound	Range
1000	1000	0.01249	$[-0.001, 0.001]$
1000	10000	0.01183	$[-0.001, 0.001]$
1000	100000	0.01160	$[-0.001, 0.001]$
100000	1000	0.00104	$[-0.00001, 0.00001]$
100000	10000	0.00098	$[-0.00001, 0.00001]$
100000	100000	0.00096	$[-0.00001, 0.00001]$

Table 4.2: Value of the bound to computed for various values of m and n; $z = 0.1$.

At this point I switched to the improved version of the bound of Eq.(2.42), keeping the assumption of the minimum support fixed to $\frac{m}{100}$ for reasons of significance of the results.

The bound ϵ depends on other two quantities, ρ and r, and on v, the upper bound of the variance of the family of functions F. In the following I analyse the order of

magnitude of each factor and for sake of simplicity I report their definition.

$$\rho = \hat{R}_m^n(F, S, \Lambda) + 2z \cdot \sqrt{\frac{\log \frac{4}{\delta}}{2nm}} \tag{4.13}$$

$$\mathrm{r} = \rho + \frac{1}{m} \cdot \left(\sqrt{c \cdot \left(4m\rho + c \log \frac{4}{\delta}\right) \cdot \log \frac{4}{\delta}} + c \log \frac{4}{\delta} \right) \tag{4.14}$$

$$\epsilon = 2\mathrm{r} + \sqrt{\frac{2 \log \frac{4}{\delta} \cdot (v + 4c\mathrm{r})}{m}} + \frac{c \log \frac{4}{\delta}}{3m} \tag{4.15}$$

However, before proceeding I need to derive the expression of v.

The variance of a random variable X is: $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Let $X_{e,\pi}$ be the random variable that represents the values assumed by the statistic $f_{e,\pi}(s_i)$, for a rule π:

$$X_{e,\pi} = \begin{cases} \frac{1}{\mathrm{A+B}} & \text{with } p(\sigma, y) \\ -\frac{1}{\mathrm{C+D}} & \text{with } p(\bar{\sigma}, y); \\ 0 & \text{with } p(\sigma, \bar{y}) \\ 0 & \text{with } p(\bar{\sigma}, \bar{y}) \end{cases} \tag{4.16}$$

Its expected value is given by:

$$\begin{aligned} \mathbb{E}[X_{e,\pi}] &= \sum_{x \in X_{e,\pi}} x \cdot p(x) = \\ &= \frac{1}{\mathrm{A+B}} \cdot p(\sigma, y) - \frac{1}{\mathrm{C+D}} \cdot p(\bar{\sigma}, y) = \\ &= \frac{1}{p(\sigma) \cdot m} \cdot p(\sigma, y) - \frac{1}{p(\bar{\sigma}) \cdot m} \cdot p(\bar{\sigma}, y) \end{aligned} \tag{4.17}$$

where it was exploited the fact that (A+B) is the number of elements satisfying σ, that can be rewritten as the probability of such event, $p(\sigma)$, multiplied by the total number of elements m; similarly for (C+D).

We develop the calculation of the variance under the null hypothesis; thanks to independence assumption it is possible to split the joint probabilities as the product of the probabilities of the single events, obtaining:

$$\begin{aligned} \mathbb{E}[X_{e,\pi}] &= \frac{1}{p(\sigma) \cdot m} \cdot p(\sigma)p(y) - \frac{1}{p(\bar{\sigma}) \cdot m} \cdot p(\bar{\sigma})p(y) = \\ &= \frac{p(y)}{m} - \frac{p(\bar{y})}{m} = 0 \end{aligned} \tag{4.18}$$

The reason for which we can do this assumption is that in the considered framework, the null hypothesis H_0 states the independence between σ and y. Thus, since the procedure consists in testing H_0 versus H_1, we compute the bound that defines the accepting region for the null hypothesis under such hypothesis.

Finally, the variance is:

$$
\begin{aligned}
Var(X_{e,\pi}) &= \mathbb{E}[(X_{e,\pi})^2] - (\mathbb{E}[X_{e,\pi}])^2 = \\
&= \Big(\frac{1}{\mathrm{A+B}}\Big)^2 \cdot p(\sigma) \cdot p(y) + \Big(-\frac{1}{\mathrm{C+D}}\Big)^2 \cdot p(\bar{\sigma}) \cdot p(y) - 0 = \\
&= \Big(\frac{1}{(p(\sigma))^2 \cdot m^2}\Big) \cdot p(\sigma) \cdot p(y) + \Big(\frac{1}{(p(\bar{\sigma}))^2 \cdot m^2}\Big) \cdot p(\bar{\sigma}) \cdot p(y) = \\
&= \frac{1}{m^2} \cdot p(y) \cdot \Big(\frac{1}{p(\sigma)} + \frac{1}{p(\bar{\sigma})}\Big) = \\
&= \frac{1}{m^2} \cdot p(y) \cdot \Big(\frac{p(\sigma) + p(\bar{\sigma})}{p(\sigma) \cdot p(\bar{\sigma})}\Big) = \\
&= \frac{1}{m^2} \cdot p(y) \cdot \Big(\frac{1}{p(\sigma) \cdot p(\bar{\sigma})}\Big) = \\
&= \frac{1}{m^2} \cdot \frac{\mathrm{A+C}}{m} \cdot \frac{1}{\frac{\mathrm{A+B}}{m} \cdot \frac{\mathrm{C+D}}{m}} = \\
&= \frac{\mathrm{A+C}}{m \cdot (\mathrm{A+B}) \cdot (\mathrm{C+D})}
\end{aligned}
\tag{4.19}
$$

Since the term v is the upper bound to the variance, to compute its value first I performed the analysis of its denominator, to find the minimum, and then I combined this information with the one relative to the numerator in order to maximize the overall value.

I define $x \doteq (\mathrm{A+B})$, therefore $(\mathrm{C+D}) = m - x$, thus the denominator can be rewritten as: $t(x) = m \cdot x \cdot (m - x) = m^2 \cdot x - m \cdot x^2$. Studying the sign of the inequality $-m \cdot x^2 + m^2 \cdot x \geq 0$, it can be easily seen that in the range of values of interest, that is $[\frac{m}{100}, \frac{99}{100} \cdot m]$ it is always positive (as it should). Then I studied the first derivative to search for maximal and minimal points: $\frac{dt}{dx} = -2m \cdot x + m^2 \geq 0$. It is strictly positive, therefore increasing, when $x < \frac{m}{2}$, and negative for $x > \frac{m}{2}$, therefore the point $x = \frac{m}{2}$ is a maximum. The minima are thus at the extreme of the interval of interest: $x_{min,1} = \frac{m}{100}$ and $x_{min,2} = \frac{99}{100} \cdot m$.

Since the goal it to maximise v, we are interested in minimizing the denominator; given the symmetry of the problem, I will develop the rest of the reasoning considering $x = x_{min,1} = \frac{m}{100}$.

At this point the rows of the contingency table are fixed (in fact x is (A+B)), what misses is to fix also the columns. This is trivial since we are under the null hypothesis, then the two rows must be in the same proportion to have a null effect. The choice that will maximise the numerator (A+C) is obtained in the extreme case in which all the samples fall in the first column, therefore we obtain: A = $\frac{m}{100}$, B = 0, C = $\frac{99}{100} \cdot m$, D = 0.

I report the contingency tables that maximises v (Tables 4.3 and 4.4) and the relative computations (Eqs.(4.20) and (4.21)), for dataset sizes $m = 1000$ and $m = 100000$.

	y	$\bar{y}$	$m_{\sigma/\bar{\sigma}}$
σ	10	0	10
$\bar{\sigma}$	990	0	990
$m_{y/\bar{y}}$	1000	0	1000

Table 4.3: Contingency table that maximise the variance; dataset size $m = 1000$.

	y	$\bar{y}$	$m_{\sigma/\bar{\sigma}}$
σ	1000	0	1000
$\bar{\sigma}$	99000	0	99000
$m_{y/\bar{y}}$	100000	0	100000

Table 4.4: Contingency table that maximise the variance; dataset size $m = 100000$.

$$v_{1k} = \frac{10^3}{10^3 \cdot 10 \cdot 990} = 1.01 \cdot 10^{-4} \tag{4.20}$$

$$v_{100k} = \frac{10^5}{10^5 \cdot 10^3 \cdot 99 \cdot 10^3} = 1.01 \cdot 10^{-8} \tag{4.21}$$

It is now possible to analyse the order of magnitude of the improved bound. I start the analysis for size $m = 1000$; in this case the n-MCERA is order of 10^{-4}, while the other parameters are: $n = 10^3$, $z = \frac{100}{m} = 0.1$, $c = \frac{200}{m} = 0.2$ and $v = 1.01 \cdot 10^{-4}$.

$$\begin{aligned}
[\rho] &\sim 10^{-4} + 10^{-1} \cdot \sqrt{\frac{1}{10^3 \cdot 10^3}} \\
&\sim 10^{-4} + 10^{-1} \cdot \sqrt{10^{-6}} \\
&\sim 10^{-4} + 10^{-1} \cdot 10^{-3} \\
&\sim 10^{-4}
\end{aligned} \tag{4.22}$$

$$\begin{aligned}
[\mathrm{r}] &\sim 10^{-4} + 10^{-3} \cdot \left(\sqrt{10^{-1} \cdot (10^3 \cdot 10^{-4} + 10^{-1})} + 10^{-1} \right) \\
&\sim 10^{-4} + 10^{-3} \cdot \left(\sqrt{10^{-1} \cdot (10^{-1} + 10^{-1})} + 10^{-1} \right) \\
&\sim 10^{-4} + 10^{-3} \cdot \left(\sqrt{10^{-1} \cdot (10^{-1})} + 10^{-1} \right) \\
&\sim 10^{-4} + 10^{-3} \cdot \left(\sqrt{10^{-2}} + 10^{-1} \right) \\
&\sim 10^{-4} + 10^{-3} \cdot \left(10^{-1} + 10^{-1} \right) \\
&\sim 10^{-4} + 10^{-3} \cdot 10^{-1} \\
&\sim 10^{-4} + 10^{-4} \\
&\sim 10^{-4}
\end{aligned} \tag{4.23}$$

$$
\begin{aligned}
[\epsilon] &\sim 10^{-4} + \sqrt{\frac{10^{-4} + 10^{-1} \cdot 10^{-4}}{10^3}} + \frac{10^{-1}}{10^3} \\
&\sim 10^{-4} + \sqrt{\frac{10^{-4} + 10^{-5}}{10^3}} + 10^{-4} \\
&\sim 10^{-4} + \sqrt{10^{-4} \cdot 10^{-3}} + 10^{-4} \\
&\sim 10^{-4} + \sqrt{10^{-7}} + 10^{-4} \\
&\sim 10^{-4} + 10^{-3.5} + 10^{-4} \\
&\sim 10^{-3.5}
\end{aligned}
\tag{4.24}
$$

The improved bound is definitely closer to the statistic in magnitude, but it is still not suitable for this sample size (recall that the statistic lies in $[-0.001, 0.001]$, but in practice, since it is given by the effect (which is order of 10^{-1}) divided by m, it is order of 10^{-4}).

Even trying to adjust the parameters there is no relevant improvement. I started by the consideration that the dominant term is always the n-MCERA, except in ϵ in which it is determined by v. At best it is possible to equal the order of the n-MCERA, in fact further refinements will be useless. For example, fixing the minimum support for a rule to 100, we have $z = 0.01$, $v = 1.11 \cdot 10^{-5}$ and ϵ becomes:

$$
\begin{aligned}
[\epsilon] &\sim 10^{-4} + \sqrt{\frac{10^{-5} + 10^{-2} \cdot 10^{-4}}{10^3}} + \frac{10^{-2}}{10^3} \\
&\sim 10^{-4} + \sqrt{\frac{10^{-5} + 10^{-6}}{10^3}} + 10^{-5} \\
&\sim 10^{-4} + \sqrt{10^{-5} \cdot 10^{-3}} + 10^{-5} \\
&\sim 10^{-4} + \sqrt{10^{-8}} + 10^{-5} \\
&\sim 10^{-4} + 10^{-4} + 10^{-5} \\
&\sim 10^{-4}
\end{aligned}
\tag{4.25}
$$

which equals the order of magnitude of $f_{e,\pi}$; however in practice the bound is still higher and the requirements on the support are too strict.

An exact computation of the factors and the bound, for the two support settings, is reported in Table 4.5, together with the value of the statistic. This latter value is computed for the first chain dataset and is relative to the rule X = 0 → Y = 0. What emerges is that even under the most conservative hypothesis on the minimum support, the bound is higher than the statistic, or at best, following the order of magnitude analysis of the same order. Thus it is not informative.

The behaviour for size $m = 100000$ is different. The statistic in this case is order of 10^{-6} and the n-MCERA is order of 10^{-7} (this is a conservative assumption, it is around

Support	ρ	r	ϵ	Statistic
10	$4.360 \cdot 10^{-4}$	$1.632 \cdot 10^{-3}$	$7.067 \cdot 10^{-3}$	$4.879 \cdot 10^{-4}$
100	$1.696 \cdot 10^{-4}$	$3.430 \cdot 10^{-4}$	$1.296 \cdot 10^{-3}$	$4.879 \cdot 10^{-4}$

Table 4.5: Values of the quantities ρ, r, ϵ that define the improved bound of the Supremum Deviation, computed for the first chain dataset of size $m = 1000$, for $n = 1000$ repetitions of the Monte Carlo approach and for the two minimum supports of 10 and 100.

$8 \cdot 10^{-8}$), while the other parameters are: $n = 10^3$, $z = \frac{100}{m} = 10^{-3}$, $c = \frac{200}{m} = 2 \cdot 10^{-3}$ and $v = 1.01 \cdot 10^{-8}$. From the analysis reported below it is possible to see that in this case the bound is smaller than the statistic.

$$\begin{aligned}
[\rho] &\sim 10^{-7} + 10^{-3} \cdot \sqrt{\frac{1}{10^5 \cdot 10^3}} \\
&\sim 10^{-7} + 10^{-3} \cdot \sqrt{10^{-8}} \\
&\sim 10^{-7} + 10^{-3} \cdot 10^{-4} \\
&\sim 10^{-7}
\end{aligned} \tag{4.26}$$

$$\begin{aligned}
[\mathrm{r}] &\sim 10^{-7} + 10^{-5} \cdot \left(\sqrt{10^{-3} \cdot (10^5 \cdot 10^{-7} + 10^{-3})} + 10^{-3} \right) \\
&\sim 10^{-7} + 10^{-5} \cdot \left(\sqrt{10^{-3} \cdot (10^{-2} + 10^{-3})} + 10^{-3} \right) \\
&\sim 10^{-7} + 10^{-5} \cdot \left(\sqrt{10^{-3} \cdot (10^{-2})} + 10^{-3} \right) \\
&\sim 10^{-7} + 10^{-5} \cdot \left(\sqrt{10^{-5}} + 10^{-3} \right) \\
&\sim 10^{-7} + 10^{-5} \cdot \left(10^{-2.5} + 10^{-3} \right) \\
&\sim 10^{-7} + 10^{-5} \cdot 10^{-2.5} \\
&\sim 10^{-7} + 10^{-7.5} \\
&\sim 10^{-7}
\end{aligned} \tag{4.27}$$

$$\begin{aligned}
[\epsilon] &\sim 10^{-7} + \sqrt{\frac{10^{-8} + 10^{-3} \cdot 10^{-7}}{10^5}} + \frac{10^{-3}}{10^5} \\
&\sim 10^{-7} + \sqrt{\frac{10^{-8} + 10^{-10}}{10^5}} + 10^{-8} \\
&\sim 10^{-7} + \sqrt{10^{-8} \cdot 10^{-5}} + 10^{-8} \\
&\sim 10^{-7} + \sqrt{10^{-13}} + 10^{-8} \\
&\sim 10^{-7} + 10^{-6.5} + 10^{-8} \\
&\sim 10^{-6.5}
\end{aligned} \tag{4.28}$$

As before, I report in Table 4.6 the exact computation of such quantities, computed for the first chain datasets and for the rule X = 0 → Y = 0, together with the value of the statistic.

n	ρ	r	ϵ	Statistic
1000	$3.760 \cdot 10^{-7}$	$6.066 \cdot 10^{-7}$	$2.383 \cdot 10^{-6}$	$3.374 \cdot 10^{-6}$

Table 4.6: Values of the quantities ρ, r, ϵ that define the improved bound of the Supremum Deviation, computed for the first chain datasets of size $m = 100000$, for $n = 1000$ repetitions of the Monte Carlo approach and with support 1000.

The bound provided by ϵ is of the same order of the statistic and it is even lower than it, but still close. In this case the order is not determined by v, but rather by the factor 10^{-7} coming from ρ. With $n = 1000$, the order of ρ is given by the sum of two terms of order 10^{-7}; since the second one depends on n, it is possible to slightly improve the bound increasing its value. In Table 4.7 are reported the results with $n = 10000$ and it is possible to see that with this setting ϵ turns to be half of the statistic.

n	ρ	r	ϵ	Statistic
10000	$1.736 \cdot 10^{-7}$	$3.483 \cdot 10^{-7}$	$1.784 \cdot 10^{-6}$	$3.374 \cdot 10^{-6}$

Table 4.7: Values of the quantities ρ, r, ϵ that define the improved bound of the Supremum Deviation, computed for the first chain dataset of size $m = 100000$, for $n = 10000$ repetitions of the Monte Carlo approach and with support 1000.

However the bound cannot be improved continuously, because at a certain point the second addend of ρ will become order of magnitudes lower than the first, which is the n-MCERA, that will drive the overall bound. Furthermore n affects significantly the computational time. In Table 4.8 is reported a summary of trend of improvement for various values of n and it is possible to see the saturation.

n	ρ	r	ϵ
1000	$3.769 \cdot 10^{-7}$	$6.078 \cdot 10^{-7}$	$2.386 \cdot 10^{-6}$
10000	$1.745 \cdot 10^{-7}$	$3.496 \cdot 10^{-7}$	$1.787 \cdot 10^{-6}$
100000	$1.105 \cdot 10^{-7}$	$2.621 \cdot 10^{-7}$	$1.583 \cdot 10^{-6}$
1000000	$9.029 \cdot 10^{-8}$	$2.333 \cdot 10^{-7}$	$1.516 \cdot 10^{-6}$
10000000	$8.389 \cdot 10^{-8}$	$2.240 \cdot 10^{-7}$	$1.494 \cdot 10^{-6}$
100000000	$8.1864 \cdot 10^{-8}$	$2.211 \cdot 10^{-7}$	$1.487 \cdot 10^{-6}$

Table 4.8: Values of the factors that compose the improved bound, computed for $m = 100000$, $z = 0.001$ and for various n.

At this point since this approach seems promising, I extended the experiments, always on the first chain dataset of size $m = 100000$, considering all the possible rules, both in the conditional and unconditional case. In Table 4.9 are reported the rules, the value of the statistic $f_{e,\pi}$, the magnitude bound ϵ computed for $n = 10000$ and support $\frac{m}{100} = 1000$, the confidence interval (obtained subtracting and adding ϵ to $f_{e,\pi}$)

and if the null hypothesis is accepted or rejected (column"H_0"), according to the test procedure defined in paragraph 3.1.2.

Rule	Statistic	ϵ	Interval	H_0
X = 0 → Y = 0	$3.37 \cdot 10^{-6}$	$1.78 \cdot 10^{-6}$	$[1.59 \cdot 10^{-6}, 5.15 \cdot 10^{-6}]$	rejected
X = 0 → Y = 1	$-3.37 \cdot 10^{-6}$	$1.78 \cdot 10^{-6}$	$[-5.15 \cdot 10^{-6}, -1.59 \cdot 10^{-6}]$	rejected
X = 1 → Y = 0	$-3.37 \cdot 10^{-6}$	$1.78 \cdot 10^{-6}$	$[-5.15 \cdot 10^{-6}, -1.59 \cdot 10^{-6}]$	rejected
X = 1 → Y = 1	$3.37 \cdot 10^{-6}$	$1.78 \cdot 10^{-6}$	$[1.59 \cdot 10^{-6}, 5.15 \cdot 10^{-6}]$	rejected
X = 0 → Y = 0 | Z	$-3.36 \cdot 10^{-8}$	$1.78 \cdot 10^{-6}$	$[-1.81 \cdot 10^{-6}, 1.75 \cdot 10^{-6}]$	accepted
X = 0 → Y = 1 | Z	$3.36 \cdot 10^{-8}$	$1.78 \cdot 10^{-6}$	$[-1.75 \cdot 10^{-6}, 1.81 \cdot 10^{-6}]$	accepted
X = 1 → Y = 0 | Z	$3.36 \cdot 10^{-8}$	$1.78 \cdot 10^{-6}$	$[-1.75 \cdot 10^{-6}, 1.81 \cdot 10^{-6}]$	accepted
X = 1 → Y = 1 | Z	$-3.36 \cdot 10^{-8}$	$1.78 \cdot 10^{-6}$	$[-1.81 \cdot 10^{-6}, 1.75 \cdot 10^{-6}]$	accepted

Table 4.9: Summary table for one of the chain datasets of size $m = 100000$. The table reports, for each rule, the value of the statistic, the bound ϵ computed for $n = 10000$, the confidence interval, if the null hypothesis is accepted or rejected.

From the prior knowledge we have about the chain structure, we know that the test outcome is always correct. In fact without conditioning there is dependence, thus the null hypothesis is always rejected, while conditioning the opposite.

However we are in a limit situation because the maximum effect detectable is around 0.18. In fact a rule is detected if the statistic is grater than ϵ: $f_{e,\pi} - \epsilon > 0$. Thus $f_{e,\pi} > \epsilon = 1.78 \cdot 10^{-6}$ which corresponds to the minimum effect $e_{min}(\sigma) = f_{e,\pi,min} \cdot m = 1.78 \cdot 10^{-6} \cdot 10^5 = 0.178$.

To summarise, family F_e is suitable only for big sample sizes and it is able to detect only effect of a certain magnitude. The details of the results are presented in Section 5.1.

4.3.2 Estimation of the Bounds for F_4

In this paragraph I analyse the bounds for the approach presented in paragraph 3.2, which makes use of four statistics. It was shown that each of the sample statistics has codomain $\{0, 1\}$ thus the parameters z and c involved in the bound are both equal to 1.

Let us start by the first bound (Eq.(2.38)). The n-MCERA for size $m = 1000$ is order of 10^{-2} and I will consider as usual $n = 1000$; since we are estimating four probabilities, in order to have an overall error rate of at most 0.05, for the bound estimation of each statistic I set $\delta = 0.0125$. The order of the first factor is:

$$\begin{aligned} \tilde{R} &\sim 10^{-2} + 10^{0} \cdot \sqrt{\frac{1}{10^3 \cdot 10^3}} \\ &\sim 10^{-2} + \sqrt{10^{-6}} \\ &\sim 10^{-2} + 10^{-3} \\ &\sim 10^{-2}. \end{aligned} \tag{4.29}$$

so the dominant term is the n-MCERA; the other three terms, instead, give:

$$\begin{aligned} \mathrm{f}_1 &\sim \frac{\sqrt{10^0 \cdot (10^3 \cdot 10^{-2} + 10^0)}}{10^3} \\ &\sim \frac{\sqrt{10^0 \cdot (10^1 + 10^0)}}{10^3} \\ &\sim \frac{\sqrt{10^0 \cdot (10^1)}}{10^3} \\ &\sim 10^{0.5} \cdot 10^{-3} \\ &\sim 10^{-2.5} \end{aligned} \tag{4.30}$$

$$\begin{aligned} \mathrm{f}_2 &\sim 10^0 \cdot \frac{1}{10^3} \\ &\sim 10^{-3} \end{aligned} \tag{4.31}$$

$$\begin{aligned} \mathrm{f}_3 &\sim 10^0 \cdot \sqrt{\frac{1}{10^3}} \\ &\sim \sqrt{10^{-3}} \\ &\sim 10^{-1.5}. \end{aligned} \tag{4.32}$$

Each statistic $f_{4.i,\pi}$ is a probability, thus it lies in $[0, 1]$ and in general the value taken is order of 10^{-1}; therefore the order of bound seems promising.

In Table 4.10 are reported the bounds relative to the statistic $f_{4.1,\pi}$ and the range of values the statistic itself assumes, for the usual values of m and for various n (for the biggest sample size the n-MCERA is order of 10^{-3}). Since the values of the n-MCERA is very similar for all the four statistics, once the size m is fixed, and the other parameters are the same, I will not report the exact computation of the bound of the other three statistics.

m	n	bound	range
1000	1000	0.1176	$[0, 1]$
1000	10000	0.1127	$[0, 1]$
1000	100000	0.1111	$[0, 1]$
100000	1000	0.0108	$[0, 1]$
100000	10000	0.0103	$[0, 1]$
100000	100000	0.0102	$[0, 1]$

Table 4.10: Value of the bound to $f_{4.1,\pi}$, computed for various values of m and n; $z = 1$.

The magnitude of the bound could is good to bound a single probability, even for the smaller sample size. However in this latter case, when the four probabilities and

the relative bounds are combined together, the result is unsatisfactory. To show this is the case, in Table 4.11 are reported, for each rule, the overall statistic $f_{4,\pi}$ and the corresponding confidence interval, for the first chain dataset of size $m = 1000$.

Rule	Statistic	Interval
X = 0 → Y = 0	0.488	[-0.986, 0.856]
X = 0 → Y = 1	-0.488	[-2.834, 0.435]
X = 1 → Y = 0	-0.488	[-0.855, 1.003]
X = 1 → Y = 1	0.488	[-0.433, 2.886]
X = 0 → Y = 0 | Z	-0.074	[-0.385, 0.741]
X = 0 → Y = 1 | Z	0.074	[-0.843, 0.583]
X = 1 → Y = 0 | Z	0.074	[-0.740, 0.385]
X = 1 → Y = 1 | Z	-0.074	[-0.581, 1.012]

Table 4.11: Value of the statistic $f_{4,\pi}$ and relative lower bound, computed for all the rules for the first chain dataset of size $m = 1000$.

The interval, for most of the rules, has a magnitude that covers all the range of possible values for the statistic, or it is even bigger.

For dataset size $m = 100000$, instead, the results are good. Table 4.12 reports the same information of the previous one, again for the first chain dataset (the parameters defining the bound are the same as for the smaller sample size, expect for the n-MCERA which is order of 10^{-3}). It is possible to see that in this case the confidence interval covers an adequate and quite narrow range of values.

Rule	Statistic	Interval
X = 0 → Y = 0	0.337	[0.283, 0.392]
X = 0 → Y = 1	-0.337	[-0.413, -0.262]
X = 1 → Y = 0	-0.337	[-0.392, -0.283]
X = 1 → Y = 1	0.337	[0.262, 0.413]
X = 0 → Y = 0 | Z	-0.003	[-0.155, 0.122]
X = 0 → Y = 1 | Z	0.003	[-0.166, 0.151]
X = 1 → Y = 0 | Z	0.003	[-0.123, 0.155]
X = 1 → Y = 1 | Z	-0.003	[-0.152, 0.166]

Table 4.12: Value of the statistic $f_{4,\pi}$ and relative confidence interval, computed for all the rules for the first chain dataset of size $m = 100000$.

Let us now consider the improved bound (Eq.(2.42)). It is necessary to compute the upper bound to the variance $v_i, i \in 1, ...4$ for all the four sample statistics. Starting from $f_{4.1,\pi}(s_i)$, let $X_{4.1,\pi}$ be the random variable that represents the values taken by statistic (for the rule under analysis), with the relative probabilities:

$$X_{4.1,\pi} = \begin{cases} 1 & \text{with } p(\sigma, y) \\ 0 & \text{with } p(\sigma, \bar{y}) \\ 0 & \text{with } p(\bar{\sigma}, y) \\ 0 & \text{with } p(\bar{\sigma}, \bar{y}) \end{cases} \tag{4.33}$$

its expected value is:

$$\mathbb{E}[X_{4.1,\pi}] = 1 \cdot p(\sigma, y) = p(\sigma, y). \tag{4.34}$$

As usual, we work under the null hypothesis, thus we can decompose the expected value.

$$\begin{aligned} \mathbb{E}_{H_0}[X_{4.1,\pi}] = p(\sigma, y) &= \frac{\mathrm{A}}{m} && (1) \\ &= p(\sigma)p(y) = \frac{(\mathrm{A+B})}{m} \cdot \frac{(\mathrm{A+C})}{m} && (2) \end{aligned} \tag{4.35}$$

The other term that contributes to the variance is:

$$\begin{aligned} \mathbb{E}_{H_0}[(X_{4.1,\pi})^2] = 1^2 \cdot p(\sigma, y) &= \frac{\mathrm{A}}{m} && (1) \\ &= 1^2 \cdot p(\sigma)p(y) = \frac{(\mathrm{A+B})}{m} \cdot \frac{(\mathrm{A+C})}{m} && (2) \end{aligned} \tag{4.36}$$

thus its final expression is:

$$\begin{aligned} Var(X_{4.1,\pi}) = \mathbb{E}_{H_0}[(X_{4.1,\pi})^2] - (\mathbb{E}_{H_0}[X_{4.1,\pi}])^2 = \\ = \frac{\mathrm{A}}{m} - \left(\frac{\mathrm{A}}{m}\right)^2 && (1) \\ = \frac{(\mathrm{A+B})}{m} \cdot \frac{(\mathrm{A+C})}{m} - \left(\frac{(\mathrm{A+B})}{m} \cdot \frac{(\mathrm{A+C})}{m}\right)^2 && (2) \end{aligned} \tag{4.37}$$

To compute the upper bound to the variance, I focused on the maximization of the term denoted by (1) of Eq.(4.37) because it was easier to deal with it.

Defining $x \doteq \frac{\mathrm{A}}{m}$, the expression to analyse becomes $t(x) = x - x^2$; it is positive for $x \in [0, 1]$, which is our domain of interest. To search for maximal and minimal point it is necessary to study the sign of the first derivative: $\frac{dt(x)}{dx} = 1 - 2x$. It is strictly positive, therefore increasing, for $x < \frac{1}{2}$ and decreasing for $x > \frac{1}{2}$; the point $x = \frac{1}{2}$ is thus a maximum. This value for x corresponds to put half samples in A: $\mathrm{A} = \frac{m}{2}$. At this point it is possible to compute v_1, however first a clarification must be made. The formula denoted by (1) do not explicitly consider the independence assumption; the clause derived concerns only entry A of the contingency, but to be precise, the entries B, C and D should not be assigned randomly, but in a way such that the independence assumption holds.

However since this in practice does not affect the computation, it is possible to substitute $\mathrm{A} = \frac{m}{2}$ in Eq.(4.37), obtaining upper bound v_1:

$$v_1 = \frac{\frac{m}{2}}{m} - \left(\frac{\frac{m}{2}}{m}\right)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \tag{4.38}$$

The computation for $f_{4.2,\pi}(s_i)$ is very similar; let $X_{4.2,\pi}$ be the random variable that

represents the values taken by the statistic for the considered rule.

$$X_{4.2,\pi} = \begin{cases} 0 & \text{with } p(\sigma, y) \\ 0 & \text{with } p(\sigma, \bar{y}) \\ 1 & \text{with } p(\bar{\sigma}, y) \\ 0 & \text{with } p(\bar{\sigma}, \bar{y}) \end{cases} \tag{4.39}$$

The expected value is:

$$\mathbb{E}[X_{4.2,\pi}] = 1 \cdot p(\bar{\sigma}, y) = p(\bar{\sigma}, y). \tag{4.40}$$

which under the null hypothesis becomes:

$$\begin{aligned} \mathbb{E}_{H_0}[X_{4.2,\pi}] = p(\bar{\sigma}, y) &= \frac{\mathrm{C}}{m} && (1) \\ = p(\bar{\sigma})p(y) &= \frac{(\mathrm{C+D})}{m} \cdot \frac{(\mathrm{A+C})}{m} && (2) \end{aligned} \tag{4.41}$$

The other contribution to the variance is given by:

$$\begin{aligned} \mathbb{E}_{H_0}[(X_{4.2,\pi})^2] = 1^2 \cdot p(\bar{\sigma}, y) &= \frac{\mathrm{C}}{m} && (1) \\ = 1^2 \cdot p(\bar{\sigma})p(y) &= \frac{(\mathrm{C+D})}{m} \cdot \frac{(\mathrm{A+C})}{m} && (2) \end{aligned} \tag{4.42}$$

that in this case is:

$$\begin{aligned} Var(X_{4.2,\pi}) &= \mathbb{E}_{H_0}[(X_{4.2,\pi})^2] - (\mathbb{E}_{H_0}[X_{4.2,\pi}])^2 = \\ &= \frac{\mathrm{C}}{m} - \left(\frac{\mathrm{C}}{m}\right)^2 && (1) \\ &= \frac{(\mathrm{C+D})}{m} \cdot \frac{(\mathrm{A+C})}{m} - \left(\frac{(\mathrm{C+D})}{m} \cdot \frac{(\mathrm{A+C})}{m}\right)^2. && (2) \end{aligned} \tag{4.43}$$

Focusing on Eq.(4.43)(1) and repeating the previous reasoning, it can be easily seen that the maximum is reached when $\mathrm{C} = \frac{m}{2}$, leading to the following upper bound:

$$v_2 = \frac{\frac{m}{2}}{m} - \left(\frac{\frac{m}{2}}{m}\right)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \tag{4.44}$$

The reasoning for $f_{4.3,\pi}(s_i)$ and $f_{4.4,\pi}(s_i)$ is even simpler. Starting from the first, the corresponding random variable, for a given rule, is:

$$X_{4.3,\pi} = \begin{cases} 1 & \text{with } p(\sigma) \\ 0 & \text{with } p(\bar{\sigma}) \end{cases} \tag{4.45}$$

The expected value, in its general form and under H_0 is the same:

$$\mathbb{E}[X_{4.3,\pi}] = \mathbb{E}_{H_0}[X_{4.3,\pi}] = 1 \cdot p(\sigma) = p(\sigma) = \frac{\text{A+B}}{m}. \tag{4.46}$$

As usual, it is necessary to derive the expression of $\mathbb{E}[(X_{4.3,\pi})^2]$ to compute the variance; their values are derived, respectively, in Eqs.(4.47) and (4.48).

$$\mathbb{E}_{H_0}[(X_{4.3,\pi})^2] = 1^2 \cdot p(\sigma) = \frac{\text{A+B}}{m} \tag{4.47}$$

$$\begin{aligned} Var(X_{4.3,\pi}) &= \mathbb{E}_{H_0}[(X_{4.3,\pi})^2] - (\mathbb{E}_{H_0}[X_{4.3,\pi}])^2 = \\ &= \frac{\text{A+B}}{m} - \left(\frac{\text{A+B}}{m}\right)^2 \end{aligned} \tag{4.48}$$

Denoting $x = \frac{(\text{A+B})}{m}$, the maximum is, as usual, for $x = \frac{1}{2}$, so when half of the samples falls in A and B. Independently of how the values are arranged, the upper bound v_3 is 0.25.

The computation for $f_{4.4,\pi}(s_i)$ is exactly the same, but with (C+D) instead of (A+B).

At this point it is possible to compute the order of magnitude of the bound and the bound itself; the parameters for the four statistics are the same: $v_i = 0.25$, the n-MCERA $\hat{R}$ is order of 10^{-2} for size 1000, $c = z = 1$ and, as usual, $n = 1000$. I start from the magnitude analysis; the first term, ρ, is dominated by the n-MCERA:

$$\begin{aligned} [\rho] &\sim 10^{-2} + 10^0 \cdot \sqrt{\frac{1}{10^3 \cdot 10^3}} \\ &\sim 10^{-2} + \sqrt{10^{-6}} \\ &\sim 10^{-2} + 10^{-3} \\ &\sim 10^{-2} \end{aligned} \tag{4.49}$$

but is not true for the second one r, as can be seen in the derivation below.

$$\begin{aligned} [\text{r}] &\sim 10^{-2} + 10^{-3} \cdot \left(\sqrt{10^0 \cdot (10^3 \cdot 10^{-2} + 10^0)} + 10^0\right) \\ &\sim 10^{-2} + 10^{-3} \cdot \left(\sqrt{(10^1 + 10^0)} + 10^0\right) \\ &\sim 10^{-2} + 10^{-3} \cdot \left(10^{0.5} + 10^0\right) \\ &\sim 10^{-2} + 10^{-3} \cdot 10^{0.5} \\ &\sim 10^{-2} + 10^{-1.5} \\ &\sim 10^{-1.5} \end{aligned} \tag{4.50}$$

$$
\begin{aligned}
[\epsilon] &\sim 10^{-1.5} + \sqrt{\frac{10^{-1} + 10^{0} \cdot 10^{-1.5}}{10^{3}}} + \frac{10^{0}}{10^{3}} \\
&\sim 10^{-1.5} + \sqrt{\frac{10^{-1} + 10^{-1.5}}{10^{3}}} + 10^{-3} \\
&\sim 10^{-1.5} + \sqrt{10^{-1} \cdot 10^{-3}} + 10^{-3} \\
&\sim 10^{-1.5} + \sqrt{10^{-4}} + 10^{-3} \\
&\sim 10^{-1.5} + 10^{-2} + 10^{-3} \\
&\sim 10^{-1.5}
\end{aligned}
\tag{4.51}
$$

As usual, in Table 4.13 are reported the terms composing the bound, the bound itself and the range of values assumed by $f_{4.1,\pi}$, for the usual m and n. Even if the results are relative to $f_{4.1,\pi}$, they are representative also of the other statistics since all the parameters are the same. The bound is comparable with the previous one for both sizes m (thus we can conclude that for the smaller one is not satisfactory at all); for this reason I will no develop further the analysis in this context.

m	n	ρ	r	ϵ	range
1000	1000	0.0229	0.0354	0.1309	$[0, 1]$
1000	10000	0.0209	0.0329	0.1252	$[0, 1]$
1000	100000	0.0203	0.0322	0.1234	$[0, 1]$
100000	1000	0.0023	0.0027	0.0109	$[0, 1]$
100000	10000	0.0021	0.0025	0.0105	$[0, 1]$
100000	100000	0.0020	0.0024	0.0103	$[0, 1]$

Table 4.13: Values of the improved bound ϵ, together with the factors ρ and r defining it, and range of values assumed by the statistic $f_{4.1,\pi}$, for various m and n.

The detailed results for the two bounds can be found in section 5.2. In particular, since for the smaller sample size both bounds are completely unsatisfactory, the discussion will regard only the bigger sample size.

4.3.3 Estimation of the Bounds for F_R

In this section I perform the estimation of the bounds for family F_R, defined in paragraph 3.3.

Let us start by the first bound (Eq.(2.38)). It is necessary to derive the values of z and c relative to the family $F_R = \{f_{R,\pi} \mid \forall \pi\}$. Recalling the definition of the sample statistic $f_{R,\pi}(s_i)$ (see Eq.(3.48)), we have that $f_{R,\pi}(s_i) \in \{-1, 1\}$ thus $z = 1$ and $c = 2$. In the following I analyse the order of magnitude of the factors, both m and n are equal to 1000; the n-MCERA, instead, is order of 10^{-2}, while δ, as usual, is 0.05.

$$\begin{aligned}\tilde{R} &\sim 10^{-2} + 10^{0} \cdot \sqrt{\frac{1}{10^{3} \cdot 10^{3}}} \\ &\sim 10^{-2} + \sqrt{10^{-6}} \\ &\sim 10^{-2} + 10^{-3} \\ &\sim 10^{-2}\end{aligned} \tag{4.52}$$

$$\begin{aligned}\mathrm{f}_1 &\sim \frac{\sqrt{10^{0} \cdot (10^{3} \cdot 10^{-2} + 10^{0})}}{10^{3}} \\ &\sim \frac{\sqrt{10^{0} \cdot (10^{1})}}{10^{3}} \\ &\sim 10^{1} \cdot 10^{-3} \\ &\sim 10^{-2}\end{aligned} \tag{4.53}$$

$$\begin{aligned}\mathrm{f}_2 &\sim 10^{0} \cdot \frac{1}{10^{3}} \\ &\sim 10^{-3}\end{aligned} \tag{4.54}$$

$$\begin{aligned}\mathrm{f}_3 &\sim 10^{0} \cdot \sqrt{\frac{1}{10^{3}}} \\ &\sim \sqrt{10^{-3}} \\ &\sim 10^{-1.5}\end{aligned} \tag{4.55}$$

As before, the driving term is f_3, however in this case the statistic lies in $[-1, 1]$ and in practice its value is order of 10^{-1}, therefore a significant bound can be provided. In Table 4.14 are reported the bound, together with the range of values covered by the statistic, for the usual two sample sizes m and different n (for $m = 100000$ the n-MCERA is order of 10^{-3}; I do not report the derivation of the magnitude of the bound for this sample size).

m	n	Bound	Range
1000	1000	0.1883	$[-1, 1]$
1000	10000	0.1831	$[-1, 1]$
1000	100000	0.1814	$[-1, 1]$
100000	1000	0.0160	$[-1, 1]$
100000	10000	0.0156	$[-1, 1]$
100000	100000	0.0155	$[-1, 1]$

Table 4.14: Value of the bound for $f_{R,\pi}$, computed for various values of m and n; $z = 1$.

Differently from the previous two statistics, the bound in this case works much better, even if for the smaller sample size, the behaviour is not excellent. In fact, in this case, the maximum value of the statistic that is detectable, in the sense that it is not overwhelmed by the magnitude of the bound, is around 0.18. This means that if the statistic and its version computed under the null hypothesis are not at a distance of at least 0.18, then the null hypothesis is always accepted. However this is still a good results, since for the precedent ones for size $m = 1000$ the bound does not work at all. For the biggest sample size, instead, the results are really good, providing an interval even narrower than the one obtained using family F_4. Thus F_R is the best approach, not only because its works also for the smaller sample size, but also because it improves the results also for the bigger one.

In this case it is not possible to vary z, since the sample statistic takes binary values, -1 and 1, not in a continuum range.

I switch now to the analysis of the improved bound, which requires the computation of the upper bound of the variance v of the family F_R. For this purpose, let $X_{R,\pi}$ be the random variable that represents the values assumed by the sample statistic $f_{R,\pi}(s_i)$:

$$X_{R,\pi} = \begin{cases} 1 & \text{with } p(\sigma, y) \\ 1 & \text{with } p(\bar{\sigma}, \bar{y}) \\ -1 & \text{with } p(\sigma, \bar{y}) \\ -1 & \text{with } p(\bar{\sigma}, y). \end{cases} \tag{4.56}$$

As before, the calculation is performed under the null hypothesis; the expected value is:

$$\begin{aligned} \mathbb{E}_{H_0}[X_{R,\pi}] &= p(\sigma)p(y) + p(\bar{\sigma})p(\bar{y}) - p(\sigma)p(\bar{y}) - p(\bar{\sigma})p(y) \\ &= p(\sigma)\big[p(y) - p(\bar{y})\big] - p(\bar{\sigma})\big[p(y) - p(\bar{y})\big] = \\ &= \big[p(\sigma) - p(\bar{\sigma})\big]\cdot\big[p(y) - p(\bar{y})\big] = \\ &= \big[p(\sigma) - (1 - p(\sigma))\big]\cdot\big[p(y) - (1 - p(y))\big] = \\ &= \big[2p(\sigma) - 1\big]\cdot\big[2p(y) - 1\big] = \\ &= \left[2 \cdot \frac{\mathrm{A} + \mathrm{B}}{m} - 1\right]\cdot\left[2 \cdot \frac{\mathrm{A} + \mathrm{C}}{m} - 1\right]. \end{aligned} \tag{4.57}$$

Recalling the definition of variance of a random variable X, $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, to maximise the variance it is necessary to minimise the expected value of X.

The minimum possible value for $\mathbb{E}_{H_0}[X_R]$ is 0 and it is achieved when at least one between the two factors composing it is 0. The first is null when (A+B) = $\frac{m}{2}$, so for example we can decide to put half of the samples in "A"; at this point we are not free to distribute the other half samples between "C" and "D". In fact, since we are working under H_0, not all the distributions will give a statistic that equals its value

computed under the null hypothesis (which is the condition for the independence). A possible one is to put the other half in "C"; another example of a table satisfying the condition and with null expected value is the one with all entries equal to $\frac{m}{4}$.

The other term that contributes to the variance is $\mathbb{E}_{H_0}[(X_{R,\pi})^2]$:

$$\begin{aligned}\mathbb{E}_{H_0}[(X_{R,\pi})^2] &= p(\sigma)p(y) + p(\bar{\sigma})p(\bar{y}) + p(\sigma)p(\bar{y}) + p(\bar{\sigma})p(y) = \\ &= p(\sigma)\big[p(y) + p(\bar{y})\big] + p(\bar{\sigma})\big[p(y) + p(\bar{y})\big] \\ &= \big[p(\sigma) + p(\bar{\sigma})\big]\cdot\big[p(y) + p(\bar{y})\big] \\ &= 1.\end{aligned} \tag{4.58}$$

Thus the maximum variance is $v = 1$; it is now possible to perform the analysis of the bound. For $m = n = 1000$ we get (δ as before is equal to 0.05 and the n-MCERA is order of 10^{-2}):

$$\begin{aligned}[\rho] &\sim 10^{-2} + 10^{0}\cdot\sqrt{\frac{1}{10^3\cdot 10^3}} \\ &\sim 10^{-2} + \sqrt{10^{-6}} \\ &\sim 10^{-2} + 10^{-3} \\ &\sim 10^{-2}\end{aligned} \tag{4.59}$$

$$\begin{aligned}[\mathrm{r}] &\sim 10^{-2} + 10^{-3}\cdot\Big(\sqrt{10^0\cdot(10^3\cdot 10^{-2} + 10^0)} + 10^0\Big) \\ &\sim 10^{-2} + 10^{-3}\cdot\Big(\sqrt{(10^1 + 10^0)} + 10^0\Big) \\ &\sim 10^{-2} + 10^{-3}\cdot\Big(10^{0.5} + 10^0\Big) \\ &\sim 10^{-2} + 10^{-3}\cdot 10^{0.5} \\ &\sim 10^{-2} + 10^{-1.5} \\ &\sim 10^{-1.5}\end{aligned} \tag{4.60}$$

$$\begin{aligned}[\epsilon] &\sim 10^{-1.5} + \sqrt{\frac{10^0 + 10^0\cdot 10^{-1.5}}{10^3}} + \frac{10^0}{10^3} \\ &\sim 10^{-1.5} + \sqrt{\frac{10^0 + 10^{-1.5}}{10^3}} + 10^{-3} \\ &\sim 10^{-1.5} + \sqrt{10^0\cdot 10^{-3}} + 10^{-3} \\ &\sim 10^{-1.5} + \sqrt{10^{-3}} + 10^{-3} \\ &\sim 10^{-1.5} + 10^{-1.5} + 10^{-3} \\ &\sim 10^{-1.5}\end{aligned} \tag{4.61}$$

The asymptotic analysis seems promising to bound the statistic: In Table 4.15 are

reported the results for the usual sample sizes and various n and it is possible to see that they are comparable with the ones obtained with the other formulation of the bound.

m	n	ρ	r	ϵ	range
1000	1000	0.0229	0.0422	0.1956	$[-1, 1]$
1000	10000	0.0209	0.0396	0.1894	$[-1, 1]$
1000	100000	0.0203	0.0387	0.1875	$[-1, 1]$
100000	1000	0.0023	0.0028	0.0151	$[-1, 1]$
100000	10000	0.0021	0.0026	0.0146	$[-1, 1]$
100000	100000	0.0020	0.0025	0.0145	$[-1, 1]$

Table 4.15: Values of the improved bound ϵ, together with the factors ρ and r defining it, and range of values assumed by the statistic $f_{R,\pi}$, for various m and n.

The detailed description of the results on synthetic datasets can be found in paragraph 5.3.

Chapter 5

Experimental Results

In this Chapter I discuss the results for family of statistics, F_e, F_4, F_R, tested on the synthetic datasets I defined in Section 4.2. As already discussed, for each fundamental structure (chain, fork and collider) and for sizes in $\{1000, 100000\}$ I created 100 datasets; for each (structure, size) configuration the results are the average across all the 100 datasets. For each family were pre-tested the two bounds of Eqs.(2.38) and (2.42) under different parameters settings; the detailed analysis is reported in Section 4.3. For simplicity, in this Chapter I report and discuss exhaustively only the results relative to the best parameter configuration.

To simplify the analysis, I make a short recap of the test procedure. For all the statistics given a rule $\pi : \sigma \to y$, the null hypothesis states the independence between σ and y (Definition 2.6.1). However the way in which this is assessed is different for the three, so I report in Table 5.1 a short summary.

Family	Condition
F_e	The confidence interval traps the 0
F_4	The confidence interval traps the 0
F_R	The confidence interval traps the value of the statistic computed under the null hypothesis

Table 5.1: Recap of the test procedure; for each family is reported the condition under which the null hypothesis is accepted.

I report also, in Table 5.2, the true answer to the independence test for chain, fork and collider.

Structure	Conditional case	Unconditional case
Chain	Independence	Dependence
Fork	Independence	Dependence
Collider	Dependence	Independence

Table 5.2: True answer to the independence test for the three fundamental structures of a graph.

5.1 Results for Family F_e

A key parameter for statistic $f_{e,\pi}(s_i) : X \to [a, b]$ is the maximum codomain absolute value $z = \max\{|a|, |b|\}$. Recalling the discussion of paragraph 4.3.1, the most conservative choice was $z = 1$, that was completely unsatisfactory. Trying to improve, this condition was relaxed, setting $z = \frac{m}{100}$, but without success; the results that I discuss in the following are relative to this latter choice of z and to $n = 1000$ Monte Carlo trials.

I start from the results relative to the smaller sample size $m = 1000$; Tables 5.3-5.5 are relative, respectively, to the chain, the fork and the collider. For each rule are reported the following average quantities, across the 100 datasets:

- Statistic, the value assumed by the overall statistic $f_{e,\pi}$ (or $f_{e,\pi,\text{cond}}$ if the rule is conditional);
- Interval, the average length of the confidence interval (which extremes are obtained subtracting and adding the bound to the Supremum Deviation to the statistic value);
- N rejects, the number of times the null hypothesis is rejected, out of 100.

F_e: Chain, $m = 1000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	$3.4138 \cdot 10^{-4}$	$2.2489 \cdot 10^{-2}$	0
X = 0 → Y = 1	$-3.4138 \cdot 10^{-4}$	$2.2489 \cdot 10^{-2}$	0
X = 1 → Y = 0	$-3.4138 \cdot 10^{-4}$	$2.2489 \cdot 10^{-2}$	0
X = 1 → Y = 1	$3.4138 \cdot 10^{-4}$	$2.2489 \cdot 10^{-2}$	0
X = 0 → Y = 0 | Z	$5.2076 \cdot 10^{-6}$	$2.2489 \cdot 10^{-2}$	0
X = 0 → Y = 1 | Z	$-5.2076 \cdot 10^{-6}$	$2.2489 \cdot 10^{-2}$	0
X = 1 → Y = 0 | Z	$-5.2076 \cdot 10^{-6}$	$2.2489 \cdot 10^{-2}$	0
X = 1 → Y = 1 | Z	$5.2076 \cdot 10^{-6}$	$2.2489 \cdot 10^{-2}$	0

Table 5.3: Results for family F_e on the chain datasets of size 1000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

The results confirm what was already discussed in section 4.3.1, that is the fact that the bound is not adequate since it is order of magnitude higher than the statistic. Given that and since the statistic is really small, the 0 is always trapped by the confidence interval, thus for all the rules the outcome of the independence test is to accept the null hypothesis. However, we know that for the chain and the fork in the unconditional case there is dependence, while for the collider this happens conditioning. We can therefore conclude that for this sample size the information brought by the statistic is erased by the bound.

Also with size $m = 100000$ the results are not satisfactory, for the same reason as before; they are reported in Tables 5.6 - 5.8.

F_e: Fork, $m = 1000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	$2.3114 \cdot 10^{-4}$	$2.4906 \cdot 10^{-2}$	0
X = 0 → Y = 1	$-2.3114 \cdot 10^{-4}$	$2.4906 \cdot 10^{-2}$	0
X = 1 → Y = 0	$-2.3114 \cdot 10^{-4}$	$2.4906 \cdot 10^{-2}$	0
X = 1 → Y = 1	$2.3114 \cdot 10^{-4}$	$2.4906 \cdot 10^{-2}$	0
X = 0 → Y = 0 | Z	$1.7682 \cdot 10^{-6}$	$2.4906 \cdot 10^{-2}$	0
X = 0 → Y = 1 | Z	$-1.7682 \cdot 10^{-6}$	$2.4906 \cdot 10^{-2}$	0
X = 1 → Y = 0 | Z	$-1.7682 \cdot 10^{-6}$	$2.4906 \cdot 10^{-2}$	0
X = 0 → Y = 0 | Z	$1.7682 \cdot 10^{-6}$	$2.4906 \cdot 10^{-2}$	0

Table 5.4: Results for family F_e on the fork datasets of size 1000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_e: Collider, $m = 1000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	$-4.9839 \cdot 10^{-6}$	$2.4784 \cdot 10^{-2}$	0
X = 0 → Y = 1	$4.9839 \cdot 10^{-6}$	$2.4784 \cdot 10^{-2}$	0
X = 1 → Y = 0	$4.9839 \cdot 10^{-6}$	$2.4784 \cdot 10^{-2}$	0
X = 1 → Y = 1	$-4.9839 \cdot 10^{-6}$	$2.4784 \cdot 10^{-2}$	0
X = 0 → Y = 0 | Z	$-1.1929 \cdot 10^{-4}$	$2.4784 \cdot 10^{-2}$	0
X = 0 → Y = 1 | Z	$1.1929 \cdot 10^{-4}$	$2.4784 \cdot 10^{-2}$	0
X = 1 → Y = 0 | Z	$1.1929 \cdot 10^{-4}$	$2.4784 \cdot 10^{-2}$	0
X = 0 → Y = 0 | Z	$-1.1929 \cdot 10^{-4}$	$2.4784 \cdot 10^{-2}$	0

Table 5.5: Results for family F_e on the collider datasets of size 1000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_e: Chain, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	$3.4026 \cdot 10^{-6}$	$2.1156 \cdot 10^{-5}$	0
X = 0 → Y = 1	$-3.4026 \cdot 10^{-6}$	$2.1156 \cdot 10^{-5}$	0
X = 1 → Y = 0	$-3.4026 \cdot 10^{-6}$	$2.1156 \cdot 10^{-5}$	0
X = 1 → Y = 1	$3.4026 \cdot 10^{-6}$	$2.1156 \cdot 10^{-5}$	0
X = 0 → Y = 0 | Z	$1.0238 \cdot 10^{-9}$	$2.1156 \cdot 10^{-5}$	0
X = 0 → Y = 1 | Z	$-1.0238 \cdot 10^{-9}$	$2.1156 \cdot 10^{-5}$	0
X = 1 → Y = 0 | Z	$-1.0238 \cdot 10^{-9}$	$2.1156 \cdot 10^{-5}$	0
X = 1 → Y = 1 | Z	$1.0238 \cdot 10^{-9}$	$2.1156 \cdot 10^{-5}$	0

Table 5.6: Results for family F_e on the chain datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

From the analysis of paragraph 4.3.1 emerged that the improved version of the bound was more promising, but only for size $m = 100000$. For example, in Table 5.9, which is relative to the chain dataset, we can see that without conditioning the null hypothesis is correctly always rejected.

Table 5.10 shows instead the results for the fork datasets. In this case the null

F_e: Fork, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	$2.2988 \cdot 10^{-6}$	$2.1168 \cdot 10^{-5}$	0
X = 0 → Y = 1	$-2.2988 \cdot 10^{-6}$	$2.1168 \cdot 10^{-5}$	0
X = 1 → Y = 0	$-2.2988 \cdot 10^{-6}$	$2.1168 \cdot 10^{-5}$	0
X = 1 → Y = 1	$2.2988 \cdot 10^{-6}$	$2.1168 \cdot 10^{-5}$	0
X = 0 → Y = 0 | Z	$-5.2832 \cdot 10^{-9}$	$2.1168 \cdot 10^{-5}$	0
X = 0 → Y = 1 | Z	$5.2832 \cdot 10^{-9}$	$2.1168 \cdot 10^{-5}$	0
X = 1 → Y = 0 | Z	$5.2831 \cdot 10^{-9}$	$2.1168 \cdot 10^{-5}$	0
X = 1 → Y = 1 | Z	$-5.2832 \cdot 10^{-9}$	$2.1168 \cdot 10^{-5}$	0

Table 5.7: Results for family F_e on the fork datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

$f_{e,\pi}$: Collider, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	$2.4327 \cdot 10^{-9}$	$2.1081 \cdot 10^{-5}$	0
X = 0 → Y = 1	$-2.4327 \cdot 10^{-9}$	$2.1081 \cdot 10^{-5}$	0
X = 1 → Y = 0	$-2.4327 \cdot 10^{-9}$	$2.1081 \cdot 10^{-5}$	0
X = 1 → Y = 1	$2.4327 \cdot 10^{-9}$	$2.1081 \cdot 10^{-5}$	0
X = 0 → Y = 0 | Z	$-1.1596 \cdot 10^{-6}$	$2.1081 \cdot 10^{-5}$	0
X = 0 → Y = 1 | Z	$1.1596 \cdot 10^{-6}$	$2.1081 \cdot 10^{-5}$	0
X = 1 → Y = 0 | Z	$1.1596 \cdot 10^{-6}$	$2.1081 \cdot 10^{-5}$	0
X = 1 → Y = 1 | Z	$-1.1596 \cdot 10^{-6}$	$2.1081 \cdot 10^{-5}$	0

Table 5.8: Results for family F_e on the collider datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_e: Chain, improved bound, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	$3.4026 \cdot 10^{-6}$	$4.7838 \cdot 10^{-6}$	100
X = 0 → Y = 1	$-3.4026 \cdot 10^{-6}$	$4.7838 \cdot 10^{-6}$	100
X = 1 → Y = 0	$-3.4026 \cdot 10^{-6}$	$4.7838 \cdot 10^{-6}$	100
X = 1 → Y = 0	$3.4026 \cdot 10^{-6}$	$4.7838 \cdot 10^{-6}$	100
X = 0 → Y = 0 | Z	$1.0239 \cdot 10^{-9}$	$4.7838 \cdot 10^{-6}$	0
X = 0 → Y = 1 | Z	$-1.0239 \cdot 10^{-9}$	$4.7838 \cdot 10^{-6}$	0
X = 1 → Y = 0 | Z	$-1.0239 \cdot 10^{-9}$	$4.7838 \cdot 10^{-6}$	0
X = 1 → Y = 1 | Z	$1.0239 \cdot 10^{-9}$	$4.7838 \cdot 10^{-6}$	0

Table 5.9: Results for family F_e on the chain datasets of size 100000, relative to the improved bound. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

hypothesis is always accepted, even without conditioning; in fact, halving the bound we obtain a value greater than the statistic, thus the 0 will be always captured. In section 4.3.1 I showed that the bound could be improved a little bit increasing the number of Monte Carlo trials n; in particular I am referring to the results of Table 4.8.

F_e: Fork, improved bound, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	$2.2988 \cdot 10^{-6}$	$4.7966 \cdot 10^{-6}$	0
X = 0 → Y = 1	$-2.2988 \cdot 10^{-6}$	$-4.695 \cdot 10^{-6}$	0
X = 1 → Y = 0	$-2.2988 \cdot 10^{-6}$	$4.7966 \cdot 10^{-6}$	0
X = 1 → Y = 1	$2.2988 \cdot 10^{-6}$	$4.7966 \cdot 10^{-6}$	0
X = 0 → Y = 0 | Z	$-5.2832 \cdot 10^{-9}$	$4.7966 \cdot 10^{-6}$	0
X = 0 → Y = 1 | Z	$5.2832 \cdot 10^{-9}$	$4.7966 \cdot 10^{-6}$	0
X = 1 → Y = 0 | Z	$5.2832 \cdot 10^{-9}$	$4.7966 \cdot 10^{-6}$	0
X = 1 → Y = 1 | Z	$-5.2832 \cdot 10^{-9}$	$4.7966 \cdot 10^{-6}$	0

Table 5.10: Results for family F_e on the fork datasets of size 100000, relative to the improved bound. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

However, with this fixing, the maximum detectable statistic is order of $1.8 \cdot 10^{-6}$, thus even if increasing n the results for the fork will be satisfactory, this will not be the case for the collider for which the value is around $1.2 \cdot 10^{-6}$ (given this evidence the results for this latter structure are not reported).

Thus we can conclude that for size $m = 100000$, the results are satisfactory only using the improved version of the bound and if the value of the statistic is sufficiently high with respect to the bound itself, that cannot be improved arbitrarily.

Instead, regarding the dataset size $m = 1000$, the improved bound was bigger than the statistic. Its computation is reported in Table 4.5; it is relative to $n = 1000$, but since the dominant term of the bound does not depend by this quantity and there are no other parameters on which we can intervene, I do not report the results because they are meaningless.

5.2 Results for Family F_4

In this section I will discuss the results relative to the family F_4.

In paragraph 4.3.2 it was shown that for sample size $m = 1000$ this family leads to completely unsatisfactory results. In particular, in Table 4.11, which is relative to the first chain dataset, it was shown that the interval covers almost all the range of possible values for the effect, that is what we are estimating. Even if it was derived only for one of the datasets, it is clear that for the others the outcome will be the same, for this reason for family F_4 I discuss only the results for the bigger sample size, which instead are promising.

First of all, I recall which are the values assumed by the parameters involved; $z = c = 1$, the size m is 100000, the number n of Monte Carlo trials is, as usual, 1000, while the n-MCERA is order of 10^{-3}.

I report, in Tables 5.11 - 5.13, the results relative to the first bound, applied,

respectively, to the chain, the fork and the collider. For each rule I report these average quantities, across the 100 datasets:

- Statistic, the average value of $f_{4,\pi}$ (or $f_{4,\pi,\text{cond}}$);
- Interval, the average confidence interval length;
- N rejects, the number of times, out of 100, the null hypothesis is rejected.

It is possible to notice that, for this family, the average confidence interval length is different for each rule. This is due to the different way, with respect to family F_e, in which the lower and upper bounds are computed.

F_4: Chain, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.3403	0.1084	100
X = 0 → Y = 1	-0.3403	0.1504	100
X = 1 → Y = 0	-0.3403	0.1084	100
X = 1 → Y = 1	0.3403	0.1504	100
X = 0 → Y = 0 | Z	0.0001	0.2773	0
X = 0 → Y = 1 | Z	-0.0001	0.3186	0
X = 1 → Y = 0 | Z	-0.0001	0.2773	0
X = 1 → Y = 1 | Z	0.0001	0.3185	0

Table 5.11: Results for family F_4 on the chain datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_4: Fork, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.2299	0.1396	100
X = 0 → Y = 1	-0.2299	0.1426	100
X = 1 → Y = 0	-0.2299	0.1397	100
X = 1 → Y = 1	0.2299	0.1427	100
X = 0 → Y = 0 | Z	-0.0005	0.2547	0
X = 0 → Y = 1 | Z	0.0005	0.3760	0
X = 1 → Y = 0 | Z	0.0005	0.2549	0
X = 1 → Y = 1 | Z	-0.0005	0.3762	0

Table 5.12: Results for family F_4 on the fork datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

For all the three structures and for all the rules, the outcome of the test is always correct, in fact there is dependence without conditioning for chain and fork and dependence, conditioning, for the collider.

The conclusion is the same when the improved version of the bound is considered, as it is possible to see in Tables 5.14 - 5.16, relative, as usual, to chain, fork and collider respectively.

F_4: Collider, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.0002	0.1369	0
X = 0 → Y = 1	-0.0002	0.1219	0
X = 1 → Y = 0	-0.0002	0.1369	0
X = 1 → Y = 1	0.0002	0.1219	0
X = 0 → Y = 0 | Z	-0.1159	0.1880	100
X = 0 → Y = 1 | Z	0.1159	0.1846	100
X = 1 → Y = 0 | Z	0.1159	0.1880	100
X = 1 → Y = 1 | Z	-0.1159	0.1846	100

Table 5.13: Results for family F_4 on the collider datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_4: Chain, improved bound, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.3403	0.1097	100
X = 0 → Y = 1	-0.3403	0.1522	100
X = 1 → Y = 0	-0.3403	0.1097	100
X = 1 → Y = 1	0.3403	0.1522	100
X = 0 → Y = 0 | Z	0.0001	0.2808	0
X = 0 → Y = 1 | Z	-0.0001	0.3226	0
X = 1 → Y = 0 | Z	-0.0001	0.2807	0
X = 1 → Y = 1 | Z	0.0001	0.3225	0

Table 5.14: Results for family F_4 with the improved version of the bound, on the chain datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_4: Fork, improved bound, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.2299	0.1413	100
X = 0 → Y = 1	-0.2299	0.1442	100
X = 1 → Y = 0	-0.2299	0.1412	100
X = 1 → Y = 1	0.2299	0.1443	100
X = 0 → Y = 0 | Z	-0.0005	0.2578	0
X = 0 → Y = 1 | Z	0.0005	0.3809	0
X = 1 → Y = 0 | Z	0.0005	0.2578	0
X = 1 → Y = 1 | Z	-0.0005	0.3804	0

Table 5.15: Results for family F_4 with the improved version of the bound, on the fork datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

Summarising this first result, family F_4 is suitable to detect causal relationships only for "big" dataset's sizes, of the order of 10^5, but it is useless for smaller ones.

F_4: Collider, improved bound, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.0002	0.1386	0
X = 0 → Y = 1	-0.0002	0.1234	0
X = 1 → Y = 0	-0.0002	0.1386	0
X = 1 → Y = 1	0.0002	0.1234	0
X = 0 → Y = 0 | Z	-0.1159	0.1936	100
X = 0 → Y = 1 | Z	0.1159	0.1869	100
X = 1 → Y = 0 | Z	0.1159	0.1904	100
X = 1 → Y = 1 | Z	-0.1159	0.1869	100

Table 5.16: Results for family F_4 with the improved version of the bound, on the collider datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

5.3 Results for family F_R

This section is dedicated to the discussion of the results obtained with family F_R, which is the most promising one.

As usual, the results relative to a pair (structure, size) are averaged across the 100 datasets built for that configuration. The quantities reported for each configuration and for each rule are:

- Statistic, the average value of the statistic;
- Interval, the length of the confidence interval (which is obtained removing and adding, respectively, the bound to the Supremum Deviation from the statistic);
- N rejects, the number of times the null hypothesis is rejected (out of 100).

Table 5.17 shows the results for the chain datasets of size 1000. The outcome of the tests is always correct: without conditioning there is dependence, and in fact the null hypothesis is rejected, while conditioning it is removed.

F_R: Chain, $m = 1000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.3404	0.3831	100
X = 0 → Y = 1	-0.3404	0.3831	100
X = 1 → Y = 0	-0.3404	0.3831	100
X = 1 → Y = 1	0.3404	0.3831	100
X = 0 → Y = 0 | Z	0.3404	0.3831	0
X = 0 → Y = 1 | Z	-0.3404	-0.3831	0
X = 1 → Y = 0 | Z	-0.3404	-0.3831	0
X = 1 → Y = 1 | Z	0.3404	0.3831	0

Table 5.17: Results for family F_R on the chain datasets of size 1000. For each rule are reported the following average quantities: value of the statistic, magnitude of the confidence interval and the number of times the null hypothesis is rejected.

F_R: Fork, $m = 1000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.2579	0.3827	74
X = 0 → Y = 1	-0.2579	0.3827	74
X = 1 → Y = 0	-0.2579	0.3827	74
X = 1 → Y = 1	0.2579	0.3827	74
X = 0 → Y = 0 | Z	0.2579	0.3827	0
X = 0 → Y = 1 | Z	-0.2579	0.3827	0
X = 1 → Y = 0 | Z	-0.2579	0.3827	0
X = 1 → Y = 1 | Z	0.2570	0.3827	0

Table 5.18: Results for family F_R on the fork datasets of size 1000. For each rule are reported the following average quantities: value of the statistic, magnitude of the confidence interval and the number of times the null hypothesis is rejected.

F_R: Collider, $m = 1000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	-0.0056	0.3826	0
X = 0 → Y = 1	0.0056	0.3826	0
X = 1 → Y = 0	0.0056	0.3826	0
X = 1 → Y = 1	-0.0056	0.3826	0
X = 0 → Y = 0 | Z	-0.0056	0.3826	12
X = 0 → Y = 1 | Z	0.0056	0.3826	12
X = 1 → Y = 0 | Z	0.0056	0.3826	12
X = 1 → Y = 1 | Z	-0.0056	0.3826	12

Table 5.19: Results for family F_R on the collider datasets of size 1000. For each rule are reported the following average quantities: value of the statistic, magnitude of the confidence interval and the number of times the null hypothesis is rejected.

In Table 5.18 are reported, instead, the results for the fork, for size 1000. In this case in the unconditional case not all the single instances make the correct decision, in fact, 74 out of 100 are correct, however this is still an high accuracy. An explanation for this is that the value of the statistic is closer to its expected value under H_0 in this case, meaning that some values are too small to be significant with respect to the magnitude of the bound.

To conclude for the smaller sample size, in Table 5.19 are reported the results relative to the collider. For this structure we expect dependence in the unconditional case, however it is possible to see from the last column that only 12 tests captures this. The reason for that is that the statistic is even closer to the expectation under the null hypothesis in this case.

Let us switch now to the bigger sample size, $m = 100000$. Tables 5.20 - 5.22 report, respectively, the results for chain, fork and collider datasets; for all rules, all the tests give the correct answer. Furthermore with this dataset size it is obtained a much tighter bound, with respect to the one relative to family F_4.

Given that this approach is the most promising, I tested in also on a series of

F_R: Chain, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.3403	0.0322	100
X = 0 → Y = 1	-0.3403	0.0322	100
X = 1 → Y = 0	-0.3403	0.0322	100
X = 1 → Y = 1	0.3403	0.0322	100
X = 0 → Y = 0 | Z	0.3403	0.0322	0
X = 0 → Y = 1 | Z	-0.3403	0.0322	0
X = 1 → Y = 0 | Z	-0.3403	0.0322	0
X = 1 → Y = 1 | Z	0.3403	0.0322	0

Table 5.20: Results for family F_R on the chain datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, magnitude of the confidence interval and the number of times the null hypothesis is rejected.

F_R: Fork, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.2586	0.0321	100
X = 0 → Y = 1	-0.2586	0.0321	100
X = 1 → Y = 0	-0.2586	0.0321	100
X = 1 → Y = 1	0.2586	0.0321	100
X = 0 → Y = 0 | Z	0.2586	0.0321	0
X = 0 → Y = 1 | Z	-0.2586	0.0321	0
X = 1 → Y = 0 | Z	-0.2586	0.0321	0
X = 1 → Y = 1 | Z	0.2586	0.0321	0

Table 5.21: Results for family F_R on the fork datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, magnitude of the confidence interval and the number of times the null hypothesis is rejected.

F_R: Collider, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.0002	0.0322	0
X = 0 → Y = 1	-0.0002	0.0322	0
X = 1 → Y = 0	-0.0002	0.0322	0
X = 1 → Y = 1	0.0002	0.0322	0
X = 0 → Y = 0 | Z	0.0002	0.0322	100
X = 0 → Y = 1 | Z	-0.0002	0.0322	100
X = 1 → Y = 0 | Z	-0.0002	0.0322	100
X = 1 → Y = 1 | Z	0.0002	0.0322	100

Table 5.22: Results for family F_R on the collider datasets of size 100000. For each rule are reported the following average quantities: value of the statistic, the magnitude of the confidence interval and the number of times the null hypothesis is rejected.

datasets of size $m = 10000$. The reason for this is that for the smaller size, it is good but not excellent and I want therefore to give the idea of the minimum dataset size that gives the correct results. Tables 5.23 - 5.25 reports the results for this new sample size. Looking in particular at the collider, which was the most critical one, for all the rules and for all the datasets the correct decision is made, in fact for this sample size

F_R: Chain, $m = 10000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.3397	0.1078	100
X = 0 → Y = 1	-0.3397	0.1078	100
X = 1 → Y = 0	-0.3397	0.1078	100
X = 1 → Y = 1	0.3397	0.1078	100
X = 0 → Y = 0 | Z	0.3397	0.1078	0
X = 0 → Y = 1 | Z	-0.3397	0.1078	0
X = 1 → Y = 0 | Z	-0.3397	0.1078	0
X = 1 → Y = 1 | Z	0.3397	0.1078	0

Table 5.23: Results for family F_R on the chain datasets of size 10000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_R: Fork, $m = 10000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.2593	0.1080	100
X = 0 → Y = 1	-0.2593	0.1080	100
X = 1 → Y = 0	-0.2593	0.1080	100
X = 1 → Y = 1	0.2593	0.1080	100
X = 0 → Y = 0 | Z	0.2593	0.1080	0
X = 0 → Y = 1 | Z	-0.2593	0.1080	0
X = 1 → Y = 0 | Z	-0.2593	0.1080	0
X = 1 → Y = 1 | Z	0.2593	0.1080	0

Table 5.24: Results for family F_R on the fork datasets of size 10000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_R: Collider, $m = 10000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	-0.0011	0.1081	0
X = 0 → Y = 1	0.0011	0.1081	0
X = 1 → Y = 0	0.0011	0.1081	0
X = 1 → Y = 1	-0.0011	0.1081	0
X = 0 → Y = 0 | Z	-0.0011	0.1081	100
X = 0 → Y = 1 | Z	0.0011	0.1081	100
X = 1 → Y = 0 | Z	0.0011	0.1081	100
X = 1 → Y = 1 | Z	-0.0011	0.1081	100

Table 5.25: Results for family F_R on the collider datasets of size 10000. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

the magnitude of the bound is quite narrow.

To summarise so far I showed that family F_R is the better, since with big and intermediate sizes works perfectly, while for the smaller one ($m = 1000$) is good.

At this point are missing only the results relative to the improved version of the bound. It was shown in Table 4.15, that such bound was comparable to the first one, for

this reason I report only the results for the chain datasets of sizes $m \in \{1000, 100000\}$, respectively in Tables 5.26 and 5.27.

F_R: Chain, improved bound, $m = 1000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.3404	0.4217	100
X = 0 → Y = 1	-0.3404	0.4217	100
X = 1 → Y = 0	-0.3404	0.4217	100
X = 1 → Y = 1	0.3404	0.4217	100
X = 0 → Y = 0 | Z	0.3404	0.4217	0
X = 0 → Y = 1 | Z	-0.3404	0.4217	0
X = 1 → Y = 0 | Z	-0.3404	0.4217	0
X = 1 → Y = 1 | Z	0.3404	0.4217	0

Table 5.26: Results for family F_R on the chain datasets of size 1000 and with the improved formulation of the bound. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

F_R: Chain, improved bound, $m = 100000$			
Rule	Statistic	Interval	N rejects
X = 0 → Y = 0	0.3403	0.0325	100
X = 0 → Y = 1	-0.3403	0.0325	100
X = 1 → Y = 0	-0.3403	0.0325	100
X = 1 → Y = 1	0.3403	0.0325	100
X = 0 → Y = 0 | Z	0.3403	0.0325	0
X = 0 → Y = 1 | Z	-0.3403	0.0325	0
X = 1 → Y = 0 | Z	-0.3403	0.0325	0
X = 1 → Y = 1 | Z	0.3403	0.0325	0

Table 5.27: Results for family F_R on the chain datasets of size 100000 for the improved formulation of the bound. For each rule are reported the following average quantities: value of the statistic, length of the confidence interval and the number of times the null hypothesis is rejected.

To conclude, family F_R works really well for datasets of size at least 10000 and achieves good results also for smaller ones, in particular around 1000. Furthermore this holds independently from the bound formulation, in fact it was shown that the results are comparable.

Chapter 6

Conclusions

The aim of this thesis was investigating the possibility of using Rademacher Averages in causal rule discovery, to overcome two main issues of the state of the art approach [3]. This latter, in fact, has two main limitations. First, it computes the confidence interval for the effect, the statistic defined to assess causality, modeling the error with a normal distribution. Second, are performed several independence tests, but without correcting for multiple hypothesis testing.

Our approach instead, thanks to the use of Rademacher Averages, compute a confidence interval that (i) depends on the property of the family of statistic defined; (ii) considers all the functions belonging to the family, directly accounting for the MHT correction; (iii) furthermore, the bound is provided with rigorous probabilistic guarantees.

I tested three different families of statistics on synthetic datasets. I implemented the three key structures for a graph, chain, fork and collider; I chose two different sample sizes $m \in \{1000, 100000\}$ and for each (structure, size) configuration I created 100 datasets.

One of them, F_e, is completely unsatisfactory since the bound is order of magnitude higher than the one of the statistic. F_4, instead, works perfectly for the big dataset size, but is meaningless for the small one, since the interval covers almost all the range of possible values for the statistic. Finally, F_R, works perfectly for the bigger sample size and in addition the interval is much tighter with respect to the one of F_4. Furthermore, for the smaller dataset sample size, if the statistic is "sufficiently far" from its expected value under the null hypothesis, it has an high accuracy. Since this family was more promising, I performed also some additional tests, on datasets of size $m = 10000$, to better understand which was the size threshold for which the results are always correct. With this latter choice of m, the outcome of the independence test is always correct.

To conclude, this work shows it is possible to use Rademacher Averages in causal rule discovery, with the advantages above mentioned, and in particular one of the families I tested, works really well in practice.

Bibliography

[1] *Overfitting*. https://it.mathworks.com/discovery/overfitting.html,

[2] *Two-sided-test*. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistest-means-proportions/bs704_hypothesistest-means-proportions3.html,

[3] Budhathoki, Kailash ; Boley, Mario ; Vreeken, Jilles: Discovering reliable causal rules. In: *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)* SIAM, 2021, S. 1–9

[4] Casella, George ; Berger, Roger L.: *Statistical inference*. Cengage Learning, 2021

[5] Eberhardt, Frederick: Introduction to the foundations of causal discovery. In: *International Journal of Data Science and Analytics* 3 (2017), S. 81–91

[6] Neal, Brady: *Introduction to Causal Inference*. Course Lecture Notes (draft), 2020

[7] Pearl, Judea: *Causality*. Cambridge university press, 2009

[8] Pearson, Karl: *The Grammar of Science*. Cambridge University Press, 2014 (Cambridge Library Collection - Physical Sciences). http://dx.doi.org/10.1017/CBO9781139878548. http://dx.doi.org/10.1017/CBO9781139878548

[9] Pellegrina, Leonardo ; Cousins, Cyrus ; Vandin, Fabio ; Riondato, Matteo: MCRapper: Monte-Carlo Rademacher averages for poset families and approximate pattern mining. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16 (2022), Nr. 6, S. 1–29

[10] Uhler, Caroline ; Raskutti, Garvesh ; Bühlmann, Peter ; Yu, Bin: Geometry of the faithfulness assumption in causal inference. In: *The Annals of Statistics* (2013), S. 436–463

[11] Vigen, Tyler: *Spurious Correlations*. Hachette UK, 2015