Introduction To Causality: A Modern Approach

Florian Brezina

Contents

Welcome

This is the HTML version of **"Introduction to Causality: A Moden Approach"**, a gentle but rigorous introduction into the art and science of causal inference. This book covers the basics of causal inference: you will learn

- how **causal inference** differs from **statistical inference** or **prediction**;
- how to express these differences in **unambiguous mathematical notation** and **causal graphs**;
- a variety of techniques to probe causal questions, from **randomized controlled experiments** to **structural equation models**;
- the current scientific edge on causal analysis, including reinforcement learning.

We approach this topic by closely examining most simple scenarios first and build upon those chapter by chapter. Throughout the book, we will use modern notation and language, primarily following Pearl (2000).

The book further contains an extensive appendix containing code snippets in the statistical programming language R as well as auxilliary material on statistics. We hope that this will allow the book to be a good standalone source for all those interested in causality, whether or not they have a solid foundation in statistics.

Chapter 1

Introduction

> Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there.'
> — Randall Munroe

Causal analysis is a fascinating field. It deals with the fundamental relation between cause and effect in complex environments. Being able to infer what the effect is going to be after doing A versus B is of utmost importance in a wide variety of applications, from policy analysis, drug prescription to marketing. Despite its ubiquity in all disciplines concerned with complex phenomena, the concept of **causality** has eluded a mathematically rigorous treatment for a long time, resulting in puzzling paradoxes and ambiguous statements. Only in recent decades a new formalism has emerged to solve these problems, with main contributors from the computer science and economics departments. The **causal revolution** (Pearl and Mackenzie 2018) swept away decades of experts and students arguing about the correct interpretation of phenomena such as **Simpson's paradox**, the nature and properties of the **error term in regression equations** and the interpretation of **structural parameters** in SEMs. The revolution established a new regime, which introduced new notation and a unified language, where causal and statistical concepts are finally separated. At last, **correlation** is never again to be confused with **causation**.

"Introduction To Causality" is a gentle introduction into this modern understanding of causality as it unfolded after the revolution. It will help you learn the fundamentals of this art and science of causal analysis. After reading this book, you'll have the tools to understand and communicate causal concepts and you will know how to tackle the common questions. The code in the appendix will help you to apply these methods using the R programming language.

1.1 What you will learn

First we will discuss causality in a trivial lab environment where we are able to control every important aspect of the environment. This will help us get familiar with the vocabulary and notation and will provide some insights in how to think about causality.

Once we have become familiar with the simple setting, we will loosen the assumption that we are able to fully control the environment. At this point, we will introduce causality as a probabilistic concept. Our inability to fully control and understand our environment forces us to settle for a less precise inference on the effects. We can't say what will happen, but we can still provide robust inference on what will happen **on average**. Most importantly, we will see why classical statistical concepts and notation are not sufficient for a rigorous and unambiguous treatment of causality and we will get a sense that there is a fundamental difference between what can be learned from passive observation ("correlation") and active intervention ("causation").

Once we mastered the probabilistic nature of causality we will discuss on how we can **measure** the effect of **actions** in a variety of settings. We will start with the easiest scenario, the randomized controlled trial. It is often considered the gold standard for clinical trials and applied across scientific disciplines. It will serve as a benchmark in our further discussion, where we will look at scenarios that violate the assumption behind the randomized controlled trial: we will discuss observational studies, synthetic cohorts and time series analysis.

After this tour de force, we shift gears and have a closer look at a couple of applications. We will discuss how to measure and interpret the placebo effect in clinical trials, how to optimize marketing using A/B tests and multi-armed bandits, and how to evaluate government interventions.

Finally, I will wrap up this book by providing some parting thoughts on epistemology and the importance of causality in the evolution of artificial intelligence and machine learning.

These chapters will hopefully provide you with a solid foundation and will allow you to find the right solution for your causal problem. But for most of your problems they will not be enough. Throughout this book we'll point you to resources where you can learn more. The appendix provides some additional material on statistics and programming.

1.2 What you won't learn

There are some important topics that this book doesn't cover. We hope that this book will leave you wanting more and that you will continue in your journey to master causality by going deeper into this topic or by exploring closly related fields and applications that we did not cover in sufficient length.

1.2.1 Statistics

The book focuses on causal inference rather than statistics. Some basic statistical concepts are discussed in the appendix, but they primarily serve as a refresher. We assume that the reader is (or has been) familiar with statistics as it is taught in most Statistics 101 classes. Details on estimation methods and properties of estimators (e.g. consistency) are not discussed. We will provide references that provide more details. We will, however, extensively discuss the differences between these two types of inferences and how they relate. Our discussion on causal inference will, except for the basic introduction, be probabilistic in nature and statistical notation will be used throughout the book. We have summarized information on notation and terminology in CHAPTER XX.

1.2.2 Machine Learning

We will address issues of machine learning where we see a connection to causal concepts. We do not go deep on any causal and non-causal ML algorithms. The discussion will focus on the discussion of supervised ML versus reinforcement learning.

1.2.3 Proofs

The book does not contain any proof or any heavy mathematical derivations. We will link to reference material. Despite that, we do intend to be rigorous in argumentation and notation and some discussions might seem overly verbose at first. We believe however that this is necessary, especially to avoid confusion between statistical and causal concepts.

1.2.4 Type Causality vs Actual Causality

When referring to causality, we will always mean what philosophers typically call *type causality* rather than *actual causality.* The former takes a forward-looking approach by inferring the *effects of causes.* This allows to predict outcome for interventions, e.g. it allows to answer questions like "what will be the outcome if we prescribe this new drug X to patients with heart disease". *Actual causality* instead mostly takes a backward look at a given instance and tries to infer *causes of effects.* This allows to answer questions such as "what caused person Y to die from heart disease". This type of inference is important if the goal is to assign responsibility, e.g. in a legal case. For a thorough introduction I recommend to take a look at (Halpern 2016).

1.3 How this book is organised

1.4 Prerequisites

To get the most out of this book, you should be familiar with basic concepts of statistical analysis, nomeclature and notation. If “expected value”, “conditional probability” or “hypothesis test” are only vaguely familiar to you, please review the appendix before digging into the main text.

The code snippets at the end of the book are purely optional. If you want to follow along on these, you need to have R on your computer. To download the software, go to CRAN, the **c**omprehensive **R** **a**rchive **n** etwork. CRAN is composed of a set of mirror servers distributed around the world and is used to distribute R and R packages. Don’t try and pick a mirror that’s close to you: instead use the cloud mirror, https://cloud.r-project.org, which automatically figures it out for you. RStudio is an integrated development environment, or IDE, for R programming. Download and install it from http://www.rstudio.com/download.

1.5 Acknowledgements

The book has been compiled from markdown documents using R package bookdown. This package has allowed me to adopt a very flexible workflow where the compilation and publication of an HTML version only takes seconds.

1.6 Links

A free HTML version of this book is available at https://flobrez.github.io/itc/. The markdown sources and supplementary material is available at https://github.com/flobrez/itc.

Chapter 2

Defining Causality

2.1 Causal Models

We assume the world can be modelled by *variables*. Variables can take various values. The variables themselves are denoted by upper-case latin letters, e.g. X, whereas we use lower-case letters for their values, e.g. x. In case X is *categorical*, different values will be denoted by a subscript x_j. Where X has two values only, we will encode them with 0 and 1.

2.1.1 Causal Graphs

Definition 2.1 (Causal Graph)**.** A graph is a mathematical structure. It consists of a set of nodes and and a set of edges, where edges connect ordered pairs of nodes. In *causal graphs*, nodes represent variables; edges represent the causal relation from cause to effect. Note that in a causal graph, an edge is an *ordered* pair of nodes, the edge therefore directed. In most graphs in this book, we will consider causal systems that can be represeted as directed acyclical graphs (DAGs)[1]. These DAGs have no feedback loops.

The causal graphs convey the qualitative pattern of causal relations. They do not quantify that relation, i.e. specify how two variables are related. A graph with relation $A \rightarrow B$ It The quantitative aspects are better represented in a set of structural equations.

[1] Readers familiar with DAGs data processing pipelines will recognize that these too describe causal mechanisms. Datasets are manipulated in an ordered sequence of steps to produce a final outcome where the result of each step is determined by the outcome of its parents steps (the input datasets) and the mechanism itself (the transformation of the datasets).

Definition 2.2 (Exogeneous and Endogeneous Variables)**.** An exogeneous variable in a graph G has no edges pointing into itself. An endogeneous variable in a graph G has at least one edge point into itself.

2.1.2 Structural Equations

Structural equations represent the causal relations between *variables.* The *absence* of a variable from the model assumes that it is not relevant for the causal description of the system. We will focus exposition on *categorical variables* which can assume a

$X \rightarrow Y$ means that X causes Y. Manipulating X determines the value of Y, but not the other way round. We call X the *cause* and Y the *outcome.* Others call Y the "*effect*", but we will use *effect* to denote changes in the outcome due to manipulations of the cause. This is in line with conventions in statistical literature (e.g. "average treatment effect") and its usage in everyday language (e.g. "tipping on that button had no effect on the brightness of the screen").

2.2 Causality In A Simple Environment

Let's first take a look at a maximally simple environment, shown in figure xxx. It represents a circuit diagram with a voltage source, a switch (X) and a lamp (Y). All elements of this environment can assume one of two states each, which we will conveniently encode as 0 and 1: * the switch can either be open (0) or closed (1) * the lamp can either be off (0) or on (1). Let's further assume that the voltage source has enough capacity to lighten the lamp if the switch is closed. Although this system is easy to understand and reason about, let's take an extra second to translate the circuit diagram into a *causal graph*, a representation that will become quite handy in more complex environments that will be discussed later in the book.

We can further represent a *causal graph* as a set of structural equations. Due to its simplicity, the circuit diagram can be represented with a single equation:

$$Y := f(X) = X \tag{2.1}$$

Note, that the equation uses operator ":="" rather than the usual "=". It reads "$f(X)$ is evaluated and *assigned* to Y" and therefore resembles variable assignment in many programming languages where, for example, `x = x + 1` is a valid expression. Crucially, it is asymmetric: if X is the air temperature and Y the reading of the temperature on a thermometer, the reading will change if we heat up the air; manipulating the reading, e.g. by exposing the thermometer to direct sunlight will not heat up the air around it.

Since the structural equation represents the *causal mechanism* relating the switch and the lamp, we can immediately read what happens if we intervene on the switch: when we close the switch, i.e. $do(X := 1)$, then the lamp will be on, $Y := 1$; if we open the switch, $do(X := 0)$, then the lamp dies, $Y := 0$.

Let's take it up a notch and create a more interesting environment by adding a second switch, connected in series, see figure XXX for the circuit diagram and figure xxx for the graph representation. The two switches allow the environment to be in four different states. Only if both switches are on, will the lamp be on, in the other three states it will be off:

switch X_1	switch X_2	lamp Y
0	0	0
0	1	0
1	0	0
1	1	1

We can easily spot that the structural equation representation is

$$Y := f(X_1, X_2) = X_1 \cdot X_2 \tag{2.2}$$

When we close switch 1, $X_1 := 1$, the state of the lamp is solely determined by the state of swith 2.

$$Y := f(1, X_2) = 1 \cdot X_2 = X_2 \tag{2.3}$$

At the same time, we also understand that, when $X_1 := 0$, the state of the lamp is independent of switch 2:

$$Y := f(0, X_2) = 0 \cdot X_2 = 0 \tag{2.4}$$

This might at first seem trivial, but here comes the twist: assume we know the system, but we are unable to observe the state of switch 1 (imagine it being hidden inside a plastic container or something). Imagine further that we also observe that switch 2 is open (and the lamp is therefore off). Will closing switch 2 turn on the light?

We cannot provide an answer to this question, at least not one with certainty. We know that the *effectiveness* of switch 2 depends on something that we don't know, the state of switch 1. *Only* through intervening with the system by closing switch 2 or by gathering information about the state of switch 1 we are able to answer this question.[2] Nevertheless, we might be able to provide a *probabilistic* answer to the question, an answer that quanitifies our uncertainty about switch 1. If we know that the likelihood that switch 1 is closed is 0.8 in all cases where

[2] Note, however, that this ambiguity disappears if we happen to observe switch 2 to be closed and therefore the light to be on. With our understanding of the system, it is clear that switch 1 also has to be closed. Opening switch 2 will consequently turn off the light, *with certainty*.

we encounter switch 2 to be open and and the light to be off, then closing switch 2 will turn on the light in 80% of the cases.

This example has shown that even in very simple causal systems, not being able to observe (and measure) a single variable, requires us to revert to inferences of a lesser kind, probabilistic rather than actual. Of course, most systems worth studying in fields outside of physics are far more complex as the one described here, and many variables of interest cannot be observed or measured. Causal analysis is therefore closely linked with statistics. From here on, we will consider these probabilistic use cases.

2.3 Causality In A Complex Environment

$$\Delta_i := Y_i^{S;do(Z_i:=1)} - Y_i^{S;do(Z_i:=0)} \tag{1}$$

As Y is binary, Δ can be one of $-1, 0, 1$ with $\Delta = 1$ being the desired outcome. As discussed in [causality], we are not able to measure this quantity directly, but need to resort to population-level quantities instead:

$$P(\Delta) = P^{S;do(Z:=1)}(Y) - P^{S;do(Z:=0)}(Y) \tag{2}$$

2.4 Causal Effects

The fundamental problem of causal inference

The definition of [causal effect] hints at a severe problem for its measurements. It involves two quantities which can never be observed at once. This poses a severe problem, often called **the fundamental problem of causal inference**. Nevertheless, it does not prevent us from inferring *average* causal effects. This might be counterintuitive at first. How could we measure the *average* of a quantity, if we can't measure the quantity itself? We will see that statistics comes to the rescue. The linearity of expectation states that

$$E(U - V) = E(U) - (V) \tag{2.5}$$

i.e. expected value of the difference of two random variables is the difference between the expected values of the individual random variables. Hence, even if $U - V$ cannot be observed, we can still calculate.[3]

[3] Imagine you are interested in the average *net income* of a certain population, i.e. $E(income - expenses)$. Even if you do not have access to individual-level data, say due to privacy concerns, you can calculate this value if you are given the population averages of income and expenses, i.e. $E(income) - E(expenses)$. Note that linearity is a property of the expected value, but not of other aggregate metris that might be of interest like the median value, where, in general, $Median(income - expenses) \neq Median(income) - Median(expenses)$.

2.4.1 Definition

bla

2.4.2 Causal Effect Statistics

Definition 2.3 (Average Treatment Effect)**.** The Average Treatment Effect, or ATE, is the expected value of xx in population x.

bla bla bla

Definition 2.4 (Average Treatment Effect on the Treated)**.** The Average Treatment Effect on the Treated, or ATT, is the expected value of xx in population x conditional on observing x.

It is often used in situations where the treatment effect is expected to be heterogeneous in a population. In a given environment, selection into treatment could yield treated individuals to have a different average treatment effect than the total population. For example, if university eduction has a higher effect on earnings for people with high intelligence and if people with high intelligence more often chose a university education than less intelligent ones, the average effect of university eduction of those who choose to go to university will be higher than in the overall population (and therefore than those choosing not to go to university).

Definition 2.5 (Intention To Treat Effect)**.** The Intention To Treat Effect, or ITT, is the expected value of xx in population x.

It is conceptually the same as the ATE, but often refers to a situation where the primary intervention cannot be manipulated directly, e.g. where a doctor can prescribe a drug but not enforce that the patient actually takes the drug.

Definition 2.6 (Local Average Treatment Effect)**.** The Average Treatment Effect, or ATE, is the expected value of xx in population x.

Chapter 3

Methods for Causal Inference

Causal relations can be inferred from **experiments** as well as **observational studies**. The randomized controlled experiment is a proven method to infer causal relations in complex environments. It involves full control over the assignment mechanism and the assignment is random. A common variation is the situation where the variable of interest cannot be directly intervened on, but a causal parent can. For example, a doctor can (randomly) prescribe a certain drug, but the patient still chooses to take the drug or not. The method of instrumental variables allows us to infer (local) causal effects nevertheless.

Afterwards, we switch to those methods that can be used even if we cannot intervene in the environment, but have to rely on passive observation only. Inferring causal relations in these situations requires a thorough understanding of the causal links from the variable of interest to the effect. We will study two different inference strategies which rely on different sets of assumptions.

Finally, we discuss methods for causal inference in samples of size 1. Given appropriate assumptions, we are able to infer causal relations by leveraging (dependent) observations over time.

3.1 Causal vs Statistical Inference

Causal inference is much harder than statistical inference.

3.2 Randomized Controlled Experiments

TODO

3.2.1 Assignment Mechanisms

Definition 3.1 (Complete Randomization)**.** (#def:rct_assignment) If intervention X is assigned through mechansim

$$X := U \tag{3.1}$$

where U *Bernoulli*(p) with $0 < p < 1$, the experiment is said to be completely randomized.

text

Definition 3.2 (Stratified Randomization)**.** (#def:strat_assignment) tbd

Stratified experiments first group individuals according to some observable attribute (e.g. by gender or by city). These groups are called strata. Within each stratum, treatment assignment follows a copmpletey randomized experiment. All methods for statistical inference can be used if the stratum is interpreted as the population for each sub-experiment. In many cases, however, we're not primarily interested in the effect in each stratum (although this can be informative) but in the population containing all strata. The statistics become more cumbersome, but stratification imposes no harm in the sense of additional assumptions as the stratification mechanism is fully known.

3.3 Instrumental Variables

In many practical applications, the assignment cannot be enforced, e.g. patients assigned to take a drug might choose to not follow through. In these cases, the effect of assignment and the effect of the actual treatment (the drug) will be different. A drug might be effective, if the application is difficult, many patients might choose not to follow through. A less effective drug that is easier to apply might have overall higher effectiveness of the assignment.

We can extend graph x from the previous section to show this mechanism explicitly. So far, we have focused on the effect of X on Y, ignoring the details of the mechanism, especially that Z, the patient's decision to follow the assignment, is a *mediator* of the effect of X on Y. This is not a problem if the effect of X on Y is our primary interest. However, we might be interested to split this mechanism into two sub-mechanisms in their own right. The mechanism $X- > Z$ explains how assignment of treatment is followed through by patients, whereas $Z- > Y$ is the biochemical of the drug. Often, researchers are interested in the latter, but aren't able to enforce assignment. As U is a confounder of $Z- > Y$, the correlation of Z and Y is not a valid method to estimate the causal effect. In comes the instrumental variable, in this case X. The intuition behind this method is as follows. We can reliably infer the effect of X on Y, as X is randomized and therefore there is no confounding. We can further also infer the effect of X on Z, again because X is randomized. In a sense, as X on Y is the combined effect of X on Z and Z on Y, there are ways we can get the latter from the former two (it is, in general, not just the difference of these two).

TODO

3.4 Propensity Score Matching

Many questions cannot be answered with deliberate experiments. Experimentation might be considered unethical or unfeasible; the intervention has already been done without randomization; TODO

3.5 Difference-in-Difference Estimator

TODO

3.6 Time Series Methods

TODO

Chapter 4

Applications

4.1 Marketing

Another common application where correlation is often confused with causation is in marketing. This might be because data scientist and business might not speak the same language. It might, however, also be a valid pragmatic assumption, which is later validated in an experiment. I will focus here on the (mis-)application of propensity modeling. Propensity modeling attempts to predict if a (potential) customers will perform a certain action, e.g. whether a new visitor on your site will register or whether a customer will buy a certain product. The model output is an estimated probability that the customer will do the action. Propensity models therefore fall into the class of binary regression.[1]

$$P(Y|I) = f(I) \tag{1}$$

where I is the information set available for prediction.[2]

[1] This application should not be considered a classification problem as we are not primarily interested in the prediction class, i.e. whether or not the customer will do the action, but we're mostly interested in the *likelihood* to do so. Most marketing application will have a success rate (often called conversion rate) far below .5 and a fairly limited information set. In that case for most or even all customer the best prediction will be to predict non-conversion for every customer. However, even if no customer might have a predicted likelihood greater than .5, it can be fruitful for marketing to know if that probability is, say, .01 or 0.4.

[2] Typically you will find a notation similar to $P(Y|X) = f(X)$ where X is called a feature set or a vector/matrix of features. I use the term information set to deliberately distinguish between the notion of all the information you have available for a customer. The information set is abstract and represents information in potentially very different formats, e.g. order data in your data base, the recorded call with the customer service team, or the customer's product reviews. Feature engineering is the step where this information is transformed into a format that can be used by ML algorithms: the order data could be transformed into multiple features, e.g. the total revenue in past 30 days, total revenue in past 180 days, the number of days the customer put an order in; the recorded call can transformed into days since the

There are plently of algorithms that could be used to estimate the function. Logistic regression is often chosen as it is easy to implement and the model itself might provide some insights. Here, we will focus not on the implementation part, but on the interpretation and (mis-) use of the model.

To see why the model might not be want you think it is, we will have a A naive usage of the model is to focus marketing on customers with high likelhood to do the action. This however, can be serverly misleading, as we will discuss next. To show this, we will start with an ideal assumption: our model is perfect. A perfect model means that we predict the customer action correctly for every customer and the model produces predicted probabilies which are well calibrated. This means that we only get two predictions, either a customer will do action with estimated probability 0 or whether they will do so with probability 1 - and the model is always right. However, although we have the best possible model, it illustrates well why the naive interpreation cannot be correct. If we focus our marketing on those customers with highest propensity (as there are only two values it is those with predicted probability 1), we focus our attention on a customer group that buys the product anyways. In fact, these are the customers that we should *least* focus on as the best we can do is to have no effect on those customers (but we still have the cost of the marketing intervention, which might be an opportunity cost) and in some case we might even affect the customer adversely (as they might be annoyed by the marketing). Hence, we're left with the second group of customers, those with predicted probability of 0. In this group, there might be some who will be convinced to buy the product after being exposed to the marketing, but we are not able to say which ones. Even after doing an experiment, we will not improve on our decision rule. Say, we run an A/B test on all customers who were predicted to not buy, and we estimate that 0.02 Depending on the situation, this model might not be very useful. The only thing that we learned from the model is that we should exclude those customers with highest probability from our marketing. This runs counter to our intuition. Furthermore, in many applications, the action might be a rare event, with likelihoods not much higher than 1%. Excluding these customers from marketing might not save a lot of money in the first place, and establishing a system where you're able to provide different marketing on customer-level might have some fixed costs (e.g. it might require to store and process customer-level data and deploy the model-outputs to production systems).

3

From the example it becomes clear that propensity modelling using a predictive model on passively observed data will at best be a proxy for the problem

latest customer service contanct, length of the call, length of waiting in line and whether or not the issue was resolved; the product reviews are transformed into word embeddings.

[3] Another issue that might arise in this context is the (monetary) value expected to be gained from the marketing intervention. As customers who are convinced by marketing to buy a product might fundamentally differ from customers who do so naturally, the might have e.g. higher return rates or more contacts with customer service. Hence, estimating the monetary value each gained customer might in itself be a non-trivial problem.

at hand. The goal of marketing optimization is to optimally trade-off the cost and effect of marketing, where the latter is a *causal* rather than an assosiative concept. Supervised machine learning models, regardless of their complexity, will fail at achieving the task since even a *ideal* falls short. To see this more clearly, let's restate the problem in causal notation first. Let Y denote the binary customer action we want to predict. Further, let S denote the default environment, which includes all relevant causal factors that determine customer decisions, their preferences and endowment, the offers available on the market and our own offer and current marketing strategy. For simplicity's sake, let's assume our marketing strategy is binary and denote it by M (e.g. whether or not we send the customer a marketing email). Assume the current marketing strategy is $M = 0$, i.e. we do currently not have an email program. The individual causal differential effect of sending an email to customer i is then

$$\Delta_i := Y_i^{S;do(Z_i:=1)} - Y_i^{S;do(Z_i:=0)} \tag{1}$$

As Y is binary, Δ can be one of $-1, 0, 1$ with $\Delta = 1$ being the desired outcome. As discussed in [causality], we are not able to measure this quantity directly, but need to resort to population-level quantities instead:

$$P(\Delta) = P^{S;do(Z:=1)}(Y) - P^{S;do(Z:=0)}(Y) \tag{2}$$

Both quantities on the right-hand side of equation (2) can be estimated. There are couple of ways to do so, many of which we discussed in [causality]. The most straigtforward way is to apply a randomized controlled trial (or "A/B test") where the population at hand is randomly split in two groups, one group being exposed to the marketing (i.e. $S; do(Z := 1)$) the other not being exposed (i.e. $S; do(Z := 0)$).[4] Conditioning the probability estimate on a set of features allows us to investigate whether the *differential causal effect* is co-related with observable information - which ultimately tells us who will be most affected by the marketing. In the most simple case, we condition by a single discrete attribute A, providing

$$P(\Delta|A) = P^{S;do(Z:=1)}(Y|A) - P^{S;do(Z:=0)}(Y|A) \tag{3}$$

This is superficially similar to equation (1), but note that the right-hand side in \$(3) describes two conditional probabilities drawn from two different environments. It will be helpful to rewrite this into a linear model form. Assume that

[4] If I_A and I_B denote the sets of individuals assigned to groups A and B, repectively, and n_A and n_B denote the size of these sets, an estimate for the average treatment effect is the group difference of the average success rate, i.e. $\hat{P}(\Delta) = \frac{1}{n_A}\sum_{i\in I_A} Y_i - \frac{1}{n_B}\sum_{i\in I_B} Y_i$.

A is binary. Then

$$\begin{aligned}
P(\Delta|A=1) &= P^{S;do(Z:=1)}(Y|A=1) - P^{S;do(Z:=0)}(Y|A=1) && (4.1)\\
&= a_1 - a_0 && (4.2)\\
&=: \Delta_1 && (4.3)\\
P(\Delta|A=0) &= P^{S;do(Z:=1)}(Y|A=0) - P^{S;do(Z:=0)}(Y|A=0) && (4.4)\\
&= a_3 - a_2 && (4.5)\\
&=: \Delta_0 && (4.6)
\end{aligned}$$

We can further represent $P(Y)$ as a function of Z

$$\begin{aligned}
P(Y) &= P^{S;do(Z:=0)}(Y) \cdot (1-Z) + P^{S;do(Z:=1)}(Y) \cdot Z && (4.7)\\
&= P^{S;do(Z:=0)}(Y) + (P^{S;do(Z:=1)}(Y) - P^{S;do(Z:=0)}(Y)) \cdot Z && (4.8)\\
&= P^{S;do(Z:=0)}(Y) + P(\Delta) \cdot Z && (4.9)\\
&= \beta_0 + \beta_1 Z && (4.10)
\end{aligned}$$

Further, replacing unconditional quantities with conditional ones, we can write

$$\begin{aligned}
P(Y|A) &= P^{S;do(Z:=0)}(Y|A) + P(\Delta|A) \cdot Z && (4.11)\\
&= \alpha_0 + \alpha_1 A + (\Delta_0 + \Delta_1 \cdot A) \cdot Z && (4.12)\\
&= \alpha_0 + \alpha_1 A + \Delta_0 Z + \Delta_1 AZ && (4.13)
\end{aligned}$$

which looks quite familiar as it is the conventional way to specify a logistic regression equation on A, Z and the interaction of both $A \cdot Z$.[^footnote-hte-nomenclature] If the treatment is ineffective $\Delta_0 = 0$ *and* $Delta_1 = 0$. The *differential causal effect* is said to be heterogeneous, if $\Delta_1 \neq 0$. [^footnote-hte-nomenclature]: The literature on *heterogeneous treatment effect* models often groups parameters of this equation regarding their role in application: α_1 is called "prognostic" as it shows if and how the success rate differs across attributes A *if no intervention/treatment* is provided; Δ_1 on the other hand is often called "predictive", meaning how predictive A is on the *effectiveness of the intervention/treatment*, i.e. whether and by how much treatment effects differ across values of A.

The model can be generalized to the case where not just a single (binary) attribute is considered, but a vector of attributes.

In a marketing context with binary treatment and outcome, table xxx can be restated as

Table 4.1: Caption here

	$Y^{\Gamma;do(X:=0)} = 0$	$Y^{\Gamma;do(X:=0)} = 1$
$Y^{\Gamma;do(X:=1)} = 0$	"Lost Cause"	"Do-Not-Disturb"
$Y^{\Gamma;do(X:=1)} = 1$	"Persuadable"	"Sure Things"

4.2 Drug Trial

4.3 Discrimination

In den letzten Jahren wurde in Politik, Medien und Wissenschaft häufiger über Diskriminierung diskutiert, einige Beispiele sind hierbei die Diskriminierung von Frauen im Berufsleben, der Wahl von Abegordneten zum Bundestag.[5] Ein Verbot von Diskriminierung hat in Deutschland im Jahre 2006 Gesetzescharakter angenommen mit der Verabschiedung des *Allgemeinen Gleichbehandlungsgesetzes (AGG)*[6], dass u.a. die Diskriminierung aus Gründen des Alters, der ethnischen Herkunft oder des Geschlechts unzulässig ist. Wie bereits des Gesetzesname andeutet, wird Diskriminierung als ein Vorgang angesehen, nicht als ein Zustand[7]. Dies wird insebesondere in §3 deutlich, wo eine (unmittelbare) Diskriminierung dann vorliegt "wenn eine Person wegen [Alter, Herkunft oder Geschlecht] eine weniger günstige Behandlung erfährt, als eine andere Person in einer vergleichbaren Situation erfährt". Die zunehmende Anwendung automatisierter Entscheidungsalgorithmen rückt die Frage nach Diskrimierung auch in den Blickpunkt der Forschung im Bereich des maschinellen Lernens. Es ist in diesem Kontext, dass dieses Thema eine mathematische Formalisierung erfahren hat. Dabei wurden zuletzt auch kausale Argumentationen und Notation eingeführt, siehe etwa Kilbertus etl al..

Wir werden uns im Folgenden an einem einfachen Beispiel diesem Problem nähern sowie einen Versuch der Die Identifikation von diskriminierenden Handlungen wird dadurch erschwert, dass die Handelnden Personen häufig nur einen Teil des gesamten Mechanismus kontrollieren und es begründete Sachzwänge gibt. Im Folgenden unterscheiden wir zwei Merkmalstypen

First, let's introduce some basic terminology first: * *protected feature*: this is a person's feature that is legally protected, e.g. age or gender. * *accepted feature*: these are features that are explicitely or implicitely accepted for discrimination, e.g. the job might (by law) require a certain certificate of qualification or the job requires a certain skillset like a specific programming language or the ability to lift weigths heavier than 30 kg.

While the protected features are usually unambiguous and explicitly stated in law, the accepted features can be reason for disagreement. It is usually the set of features that are deemed to be necessary to do the job, but of course "doing the job" can done in different qualities. A bar owner might believe it to be necessary for his waiters to be handsome and flirty with the predominantely

[5] Weitere prominente Beispiele, die aus den USA auch nach Deutschland Wirkung entfalteten war die Diskussion um die Vergabe von Filmpreisen an vorwiegend weiße und männliche Schauspieler, Produzenten und Autoren sowie die von Afroamerikanern bei der Oskar-verleihung.

[6] https://www.gesetze-im-internet.de/agg/index.html

[7] Die im allgemeinen Sprachgebrauch übliche Bezeichnung *Gleichstellungsgesetz* ist weniger klar und kann daher zu Fehlinterpretationen führen

female audience, but an applicant might think it is not. Getting the order right and providing the right servcie while being friendly he might consider to be sufficient.

- *unprotected feature*: these are all features that are not explicitly protected, e.g. eye color, ability to lift more than 30kg
- *zulässige Merkmale*: dies sind Merkmale, nach denen eine diskriminierende Handlung zulässig ist. So kann etwa für eine Tätigkeit in einem Warenlager voraussgesetzt werden, dass ein Bewerber in der Lage sein muss, Lasten über 20kg zu transportieren, da dies für die Erfüllung der Tätigkeit unabdingbar ist.
- *aufhebende Merkmale*: diese stellen eine Teilmenge der *zulässigen Merkmale* dar. Es sind diejenigen Merkmale, die auf dem kausalen Pfad zwischen den *geschützten* Merkmalen und dem Output liegen. Das Merkmal “kann Lasten über 20kg transportieren” ist vermutlich ein solches, da dieses kausal vom Geschlecht und/oder dem Alter eines Bewerbers beeinflusst ist.
- *unzulässige Merkmale*: sind Merkmale, nach denen eine diskriminierende Handlung nicht zulässig ist, die aber nicht ein *geschütztes Merkmal* sind. Dies könnte etwa die Augenfarbe des Bewerbers sein, die (etwa gemäß AGG nicht geschützt ist), jedoch für die Erfüllung der Tätigkeit irrelevant sein dürfte.
- *stellvertretende Merkmale*: diese stellen eine Teilmenge der *unzulässigen Merkmale* dar. Es sind diejenige Merkmale, die auf dem kausalen Pfad zwischen den *geschützten* Merkmalen und dem Output liegen. So kann die Augenfarbe eines Bewerbers ein *stellvertretendes Merkmal* sein, wenn die Augenfarbe etwa durch das Geschlecht oder die ethnische Herkunft beeinflusst wird. Sie werden *stellvertretend* genannt, da sie Information über das geschützte Merkmals beinhalten und somit zur Diskriminierung herangezogen werden könnten.

TODO: sollte *unzulässig* nicht anders bezeichnet werden, da ja eine Diskriminierung nach nicht-stellvertretern rechtlich nicht verboten ist.

Es gelten folgende Definitionen:

- kann ein Mechanismus der unmittelbaren Diskriminierung durch einen vom geschützten Merkmal unabhängigen Prozess (oder eine Konstante) ersetzt werden, ohne dass sich die Verteilung der Zielvariable ändert, so ist dieser Mechanismus nicht-diskriminierend

Chapter 5

Epistemology

5.1 Manipulation

Sometimes the cause cannot be directly manipulated: the doctor can prescribe a drug but not ensure the patient is taking the drug; a marketing manager might choose who to sent a mail to, but the delivery could fail; we dealt with problems like these in our application on instrumental variables. The problem is however, more subtle, and probably more important than we might have anticipated. The issue appears in situations where cannot directly intervene on a variable, but only on a mechanism that affects that variable.

Assume, you're an economist asked to evaluate the effect of university education on a persons lifetime income to inform a decision on whether or not to expand access to university education. As an expert in economics, statistics and causal inference, you immediately recognize the problems with this task. * knowing about treatment effect heterogeneity, you understand that the average treatment effect calculated for the current group of students is not an estimate for the treatment effect for those who will become available under the new policy. * you're asked to provide a causal effect, but the intervention is not specified. The government can't directly manipulate

Uncertainty about effectiveness of interventions can have various reasons: * *correct model*: we are uncertain about the assumptions on the causal structure used to make the causal inference * *extrapolation*: the causal inference relied on data drawn from other populations, either other individual or a different time. In dynamic environments such as the economy, causal knowledge can depreciate fast with new technology, a changing competition and ever-so slightly different regulation. *

In many circumstances the optimal decision will differ by perspective. Imagine a doctor having two different treatments at hand. Drug A is known to be most effective (100), but relies on the patient taking the drug reliably every day. If taken irregularly, the drug's effectiveness is severly lower (50). Drug B on the other hand works similarly if take regularly or irregularly (75). If the doctor has to make a decision whether to prefer A vs B, she will have to rely on her judgements of the patient's ability to follow through on a regular schedule. If it is known that most patients have a hard time doing this, the best treatment to prescribe from a doctor's perspective is drug B. The patient, however, might think differently. Knowing that they have the self-discipline they might choose to go with drug A. This is just one example on how a guideline with simple recommendations can be sub-optimal. If the patient was given the information, he could use it to make a better judgement (but of course the patient might be overconfident in his ability to stick to a rigid schedule). This is similar to giving someone advice to exercise and restrict to 2000kcal food intake to loose weight. This certainly works *if* one adhere's to it. The likelihood to fail will be great though. A less ambitious and rigid goal might be less effective, but might be substantially easier to follow-through. The *recommendation* for a diet will have to have a larger effect than the more rigid one.

5.2 Long-Term Effects

In examples so far we have considered situations where the intervention was one-shot, i.e. we only intervened once in the system. Although plausible in the applications considered, there are many applications where this is not the case. Especially when we are interested in long-term goals rather than short-term outcomes, it is implausible that we would not be able to intervene in the future. Of course, this can partly be alleviated estimating the likelihood of future actions and taken these into account, but that doesn't make sense if we are the actors ourselves. The British xx was tasked to evaluate the effect of Brexit on British GDP in the years after Brexit. This is difficult to establish as many important decisions are undetermined by Brexit itself (leaving out the problem that the terms and conditions of the Brexit are not fully specified themselves). Leaving the EU will allow Great Britain to explore policies that were not available within the EU: every country in the EU has to have a value-added tax system. Leaving the EU allows to choose a different policy. An economist trying to provide this estimate will have to assume a probability distribution on this topic, as it will certainly have non-trivial impact. A politician in charge, e.g. the prime minister, might disregard this recommendation as he might have quite different plans and the probability distribution might be quite different (it might not be fully centered at a single option as he might be unsure whether he might achieve this policy if he is willing to go for it).

5.3 Are all sciences equal?

The concept of causality as discussed in this book is a cornerstone of many scientific disciplines. Although in most cases not stated explicitly in those terms, the question **what will happen if i does x (versus y)** is implicit in most scientific work in diverse fields such as economics, psychology, medicine, and history[1]. The interventions contemplated and analyzed in these disciplines vary widely in terms of manipulability, scope and ethics: a randomized experiment on fiscal policy is neither operationally feasible nor ethically desirable - at least at a the level of entire nations; assessing how individuals provide different answers if questions are phrased differently is both manipulable in a controlled environment and too unintrusive to be ethically problematic. Throughout the book we discussed several methods to infer causal relationships ranging from randomized controlled trials to time series analysis. With every step away from the ideal properties of an RCT, the reliability and robustness of results crucially depended on the validity of the assumptions that we were willing to accept. In general, the more assumptions we make, the more likely the results will be unreliable. On a sidenote: of course, incorrect causal assumptions are not the only or even the main reason for scientific results to be unreliable. Many of those have to do with problems of *statisical inference* instead. Even in experiments following an RCT design, there are plenty of pitfalls to be aware of. Too small samples, p-hacking, and publishing bias are just a few reasons for the replication crisis in psychology, medicine and other social sciences.

Least reliable results are to be expected when the complexity of the subject matter is very high, but the methodoligal toolkit to investigate it is reduced to the least reliable ones

- macroeconomics. growth, business cycle, monetary and fiscal policy.
- epidemiology. including nutrition.
- social psychology.
- evolution. while the theory of evolution has strong empirical support, combinatorial explosion and complex dynamics do not allow for a reliable inference of effects of interventions in the biosphere.

Even if complexity is high, if some aspects of the problem can be thoroughly studied and combined using a consistent theory, the overall results can be reasonably reliable:

- climate science.

[1] For history, the question needs to be changed to **what would have happened if i did x (versus y)**. Note that *i* needs to be an actor who can *manipulate x* directly for this question to make sense. Statements such as *if it weren't for Mikhail Gorbachev, the Soviet Union wouldn't have been dissolved in '91* are therefore too ambiguous for a rigorous treatment: it will likely matter if Gorbachev wasn't elected by the Politburo in 1985 in the first place, being ousted from office by coup d'état or stepped down from office due to health reasons.

- microeconomics. pseudo-experiments on price controls (e.g. minimum wage or gas price cap) supported by strong theory of supply and demand. problem is often quantification of effects rather than the general direction. There is almost no disagreement across labor economists that there exists a general trade-off between higher minimum wages and lower unemployment. The debate today primarily focuses on the possibility of small negative or even positive effects on employment, but virtually everybody accepts the logic that unemployment will exist/increase above a certain threshold.
-
-

Appendix A

Notation and Terminology

Throughout the book we will follow notation and terminology similar to Peters, Janzik, and Schölkopf (2017). It deviates from Pearl (2009) primarily in the notation for intervention distributions. Where Pearl (2009) writes $P(Y|do(X = x))$, we will use $P^{\Gamma;do(X:=x)}(Y)$ instead: it avoids confusion with the notation for conditional distributions and it emphasizes that an intervention is an *assignment* (":=") rather than an *equality* ("="). Otherwise we follow conventional statistical notation.

symbol	represents
X, Y, U	(random) variable
x	value of X
$X = x$	X has value x
$X := x$	assignment of value x to variable X
i	sample index, observation
Γ	causal graph
$\Gamma; do(X := x)$	intervention on graph Γ by assigning value x to X
$P(Y)$	probability distribution of Y
$P(Y\|X = x)$	probability distribution of Y given $X = x$
$P(Y\|X)$	set of probability distributions $P(Y\|X = x)\forall x$
$P^{\Gamma;do(X:=x)}(Y)$	probability distributions of Y after intervention on graph Γ
$E(Y), \mu_Y$	expected value of Y
$Var(Y), \sigma_Y^2$	variance of Y
$\hat{\beta}$	estimate of parameter β
$X- > Y$	causal link from X to Y

A.1 Terminology Confusion

Causality has been studied for decades using only statistical notation and terminology. As the main text shows, this cannot be achieved. Trying to do so leads to imprecision, paradoxes, and outright confusion. Hence, it is important to always bear in mind that "causal and statistical concepts do not mix." (Pearl 2009, 332) The same author even goes so far as to proclaim:

> If I am remembered for no other contribution except for insisting on the causal-statistical distinction, I would consider my scientific work worthwhile.

He demonstrates this (on page 40) by acknolidging that statistical notation is not rich enough to discriminate between the causal relation between disease and symptoms

$$P(symptoms|disease)$$

and non-causal relationship

$$P(disease|symptoms)$$

If you already have some background on this topic, you might have heard many causal terms to be used in statistics text without the proper context. The following list will provide some guidance

Causal terminology: confounder, randomization, instrumental variable, experimental vs observational data, exogeneous vs endogeneous variables, structural equations, spurious correlation, explanation

statisical concepts: correlation, independence, regression (parameter), significance, variance, distribution, Granger causality, likelihood, propensity scores

> [B]ehind every causal conclusion there must lie some causal assumption that is not discernible from the distribution function. (Pearl 2009, 332)

And then there's "spurious correlation".

> In statistics, a spurious relationship or spurious correlation is a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor (referred to as a "common response variable", "confounding factor", or "lurking variable") [https://en.wikipedia.org/wiki/Spurious_relationship]

It is probably the worst of all of these. First, it obviously mixes statistical and causal concepts. Saying that a correlation is “spurious” because there is no causal relation is just weird. A spurious correlation should be an observation where the correlation in a sample is non-zero, although it is in the entire population. Second, “correlation” suggests that it’s just not the right statistical measure to convey the causal relationship - and that a conditional expectation or a regression coefficient might. They don’t.

Appendix B

Statistics

This section provides a brief introduction into concepts of probability theory and statistical inference that are essential for understanding the technical parts of the main text.

B.1 Distributions

B.1.1 Single Random Variable

The expected value of a discrete random variable X is the weighted avarage of the possible values x, with their probabilites as weights

$$E(X) = \sum_x p(x)x \tag{B.1}$$

The expected value of a sum of two random variables is the sum of their expected values

$$E(X+Y) = E(X) + E(Y) \tag{B.2}$$

and the expected value scales linearly with scaling factor a

$$E(aX) = aE(X) \tag{B.3}$$

Note, however, that this property does not hold for the product of two random variables

$$E(X \cdot Y) = E(X) \cdot E(Y) \tag{B.4}$$

only if X and Y are independent.

B.1.2 Multiple Random Variables

- Covariance
- Correlation
- Conditional Probability
- Regression Parameter

B.2 Statistical Inference

B.2.1 Estimators

B.2.2 Hypothesis Tests

Appendix C

Code

C.1 Equivalence of statistical properties and SQL

The probability distribution $P(X = x)$

```
SELECT
    X                                AS value_x
  , SUM(1.0) / SUM(SUM(1.0)) OVER () AS probability_of_x
FROM table
GROUP BY X
```

The expected value $E(X)$

```
SELECT
    AVG(X) AS expected_value_of_X
FROM table
```

The set of conditional expected values $E(X|Y)$

```
SELECT
      Y      AS value_y
    , AVG(X) AS cond_expected_value_X_given_y
FROM table
GROUP BY Y
```

The conditional expected values $E(X|Y = y)$

```
SELECT
      Y      AS value_y
    , AVG(X) AS cond_expected_value_X_given_y
FROM table
WHERE Y = y
GROUP BY Y
```

Halpern, Joseph. 2016. *Actual Causality.* MIT Press.

Pearl, Judea. 2009. *Causality.* 2nd ed. Cambridge University Press.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect.* Basic Books.

Peters, Jonas, Dominik Janzik, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press.