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

Causal Models

2.1 Causality, Asymmetry and Entropy

Causality is strongly linked to the concept of *time*. Cause precedes effect, never the other way round. Symptoms occur after infection This asymmetry is mirrored in the physical notion of entropy and (as an emergent property) time. While the past is determined we feel that we are able to act on the future, that we are able to choose one among many possible futures. This is due to the low entropy the universe had in the past. All fundamental physical laws are perfectly symmetrical and therefore reversable. Asymmetry is only introduced by a coarse-grained look at the world, and therefore an *emergent* property.

2.2 Basic Definitions

Definition 2.1 (Variable and State)**.** We assume the world can be modelled by **variables**. Variables can take various values, which we will call **states**. The variables themselves are denoted by upper-case latin letters, e.g. X, whereas we use lower-case letters for their states, 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.

2.2.1 Causal Graphs

Definition 2.2 (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 $A \to B$ It The quantitative aspects are better represented in a set of structural equations.

Definition 2.3 (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.

2.2.2   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

$X \to Y$ 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").

2.3   Deterministic Systems

2.3.1   An electric circuit with one switch

Let's first take a look at a most simple environment, shown in figure xxx. It represents a circuit diagram with a voltage source, a switch (X) and a lamp (Y). Both, X and Y, can assume one of two states. We will encode these as 0 and 1:

- switch is open (0) or closed (1)
- light bulb is off (0) or on (1).

This system is very easy to reason about. Assuming that the power source has enough capacity, the light bulb will be on if and only if the switch is closed. The system can be in one of only two states:

---

[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).
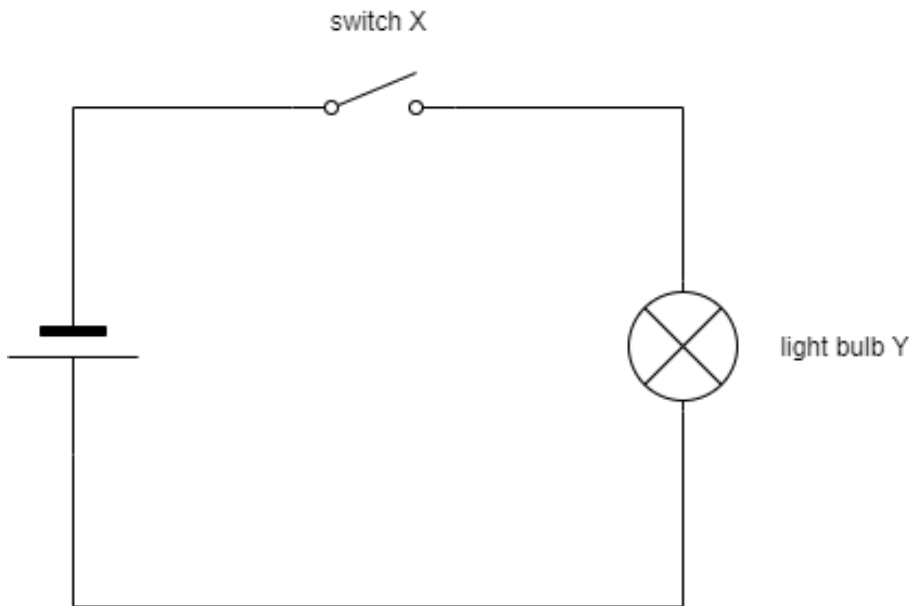
Figure 2.1: A simple circuit diagram with a single power source, a switch (X) and a light bulb (Y)

switch X	light bulb Y
0	0
1	1

This table, however, does not contain any information on the causal relationship between X and Y and will therefore not be sufficient to correctly reason about the system. Hence, let's take an extra second to translate this circuit diagram into a *causal graph*. This will become much more useful in more complex systems, but it's probably a good idea to get used to this representation from the start.
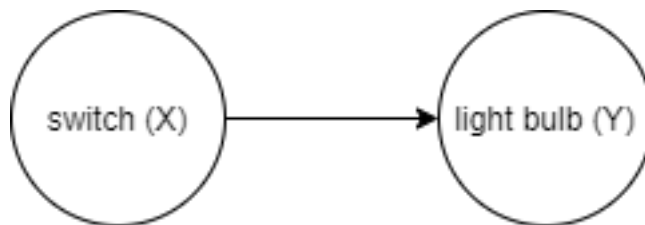

Figure 2.2: The simple circuit diagram converted to a causal graph, where X is the (only) cause of Y

A *causal graph* represents variables as *nodes* and causal relationships as *directed edges* between nodes. $X \to Y$ means that X causes Y. This in turn means that manipulating X determines the state of Y, but not the other way round. Here, the graph consists of two nodes, the switch X and the light bulb Y, connected by a directed edge from X to Y.

While the graphical representation of the causal structure makes it easy to *qualitatively* reason about the causal structure of the system, an algebraic representation will be needed for quantitative analysis. The algebraic representation of a causal relationship is called a *structural equation*. In any system we have one

infer $f()$ right away:

$$Y := f(X) = X \tag{2.2}$$

2.3.1.1   Intervention

Now that we managed to represent this system in various forms, we can start to reason about *interventions*. For now, we define an *intervention* as an operation in the system that fixes a variable to a certain value. Here, two interventions are of interest: we can open the switch, i.e. set $X := 0$, or close it, i.e. set $X := 1$. The *structural equation* is then

$$Y := f(X) = 0 \tag{2.3}$$

and

$$Y := f(X) = 1 \tag{2.4}$$

respectively.

This simple system has a nice property: for the light bulb to be on, the closed switch is a *necessary* and *sufficient* condition. This property is "nice" as it allows us to falsify the causal model from observation: a *single* observation where the light bulb is on but the switch is open (or the light bulb is off but the switch is closed) allows us to refute the causal model, e.g. the power source might not be strong enough or the circuit might have flaws.

2.3.2   An electric circuit with two switches

Let us now slightly increase the complexity of the system by adding a second switch. Both switches are connected in series as shown in the circuit diagram in figure XXX

Again, our understanding of the physical nature of this system allow us to derive the corresponding causal graph. Both switches cause the state of the light bulb whereas the light bulb does not cause any of the switches and switch X_1 does not cause switch X_2, nor vice versa. Hence our causal graph has three nodes and directed edges $X_1 \to Y$ and $X_2 \to Y$:

The structural equation for this system is also very simple:

$$Y := f(X_1, X_2) \tag{2.5}$$

but the functional form of $f()$ might not be obvious. The two switches allow the environment to be in four different states. Only if *both* switches are on, will the light bulb be on, in the other three states it will be off:
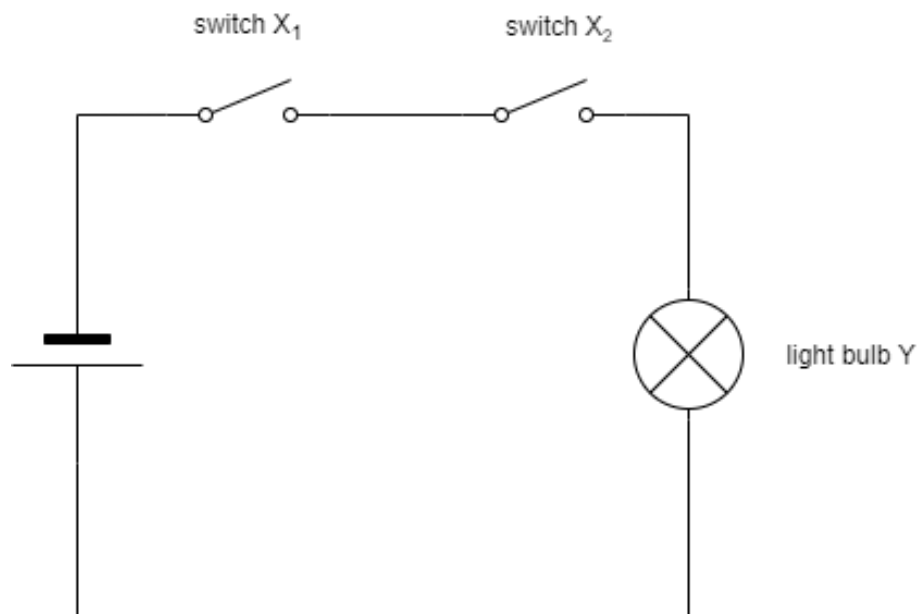
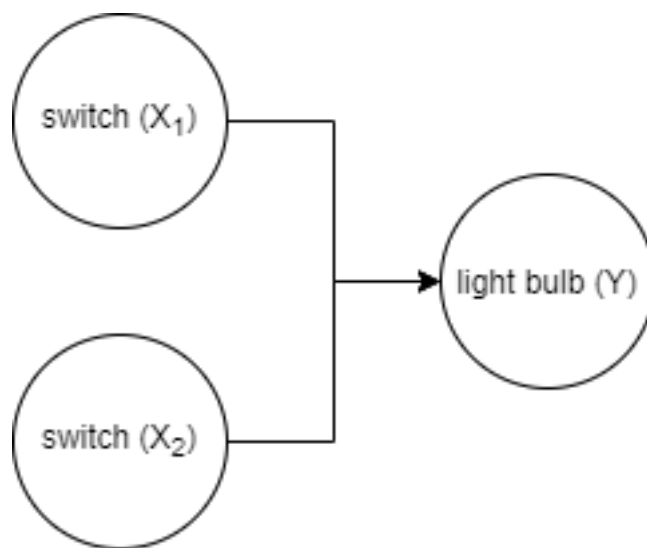Figure 2.3: A circuit diagram with a single power source, two switches (X1 and X2) and a light bulb (Y)

Figure 2.4: The simple circuit diagram converted to a causal graph, where X is the (only) cause of Y

switch X_1	switch X_2	light bulb Y
0	0	0
0	1	0
1	0	0
1	1	1

Hence, the structural equation has to be

$$Y := f(X_1, X_2) = X_1 \cdot X_2 \tag{2.6}$$

For the light bulb to be on, $X_1 = 1$ and $X_2 = 1$ are necessary conditions but neither is (on its own) sufficient.

2.3.2.1   Intervention

Let us now turn to interventions in this system. While in the system with just one switch, every intervention on X caused a *change* in the state of Y, this is not true in the case of interventions on a single switch now. In a state where $X_1 = 0$, we are unable to change the state of Y by intervening on X_2:

$$Y := f(0, X_2) = 0 \cdot X_2 = 0 \tag{2.7}$$

Conversely, if the first switch is closed, $X_1 = 1$, we're basically back in a system with just one switch, which solely determines by the state of the light bulb.

$$Y := f(1, X_2) = 1 \cdot X_2 = X_2 \tag{2.8}$$

2.4   Unobservability

So far, we were able to reason about the effectiveness of interventions due to our ability to fully specify the structural equations (i.e. we knew $f()$) *and* were able to *observe* the state of all causes. This allowed us to reason that closing switch 2 will *not* have an effect on the light bulb when switch 1 is open, but will change its state if switch 1 is closed.

Problems arise, when one of these conditions fails. Suppose we are still dealing with the system with two switches connected in series, but the state of switch 1 cannot be observed as the switch is hidden in a box, see figure XXX.

Imagine now that we observe that switch 2 is open (and the light bulb is off). Will closing switch 2 turn on the light?

Unfortunately, we are not any more able to answer this question. Our best answer is "It depends.". It depends on the state of switch 1, which is not observable. What were left is to resort to a different kind of reasoning, *probabilistic*
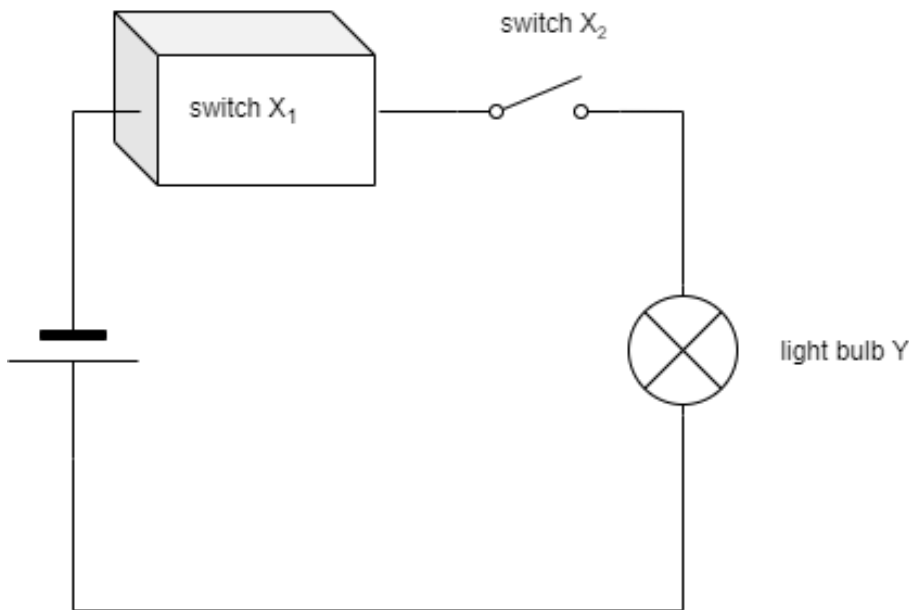
Figure 2.5: A circuit diagram with a single power source, two switches (X1 and X2) and a light bulb (Y), but switch 1 is hidden in a box.

reasoning. While we're not able to make any statements about any single system of that kind, we can still make statments on the likelihood of X_2 effectiveness based on probability distributions for X_1. If we knew that the likelihood that switch 1 is closed is 0.8 in all cases where 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.

2.4.1   Probabilistic Models of Causality

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

2.4.2   Interventions More Generally Defined

So far, we have used the term *intervention* when we fixed a variable of system to a certain value. More generally, an intervention is a change in one or more structural equations. An intervention that replaces the right-hand side of the structural equation with a *value* is called *hard intervention* (or *atomic*, *ideal*, *surgical*), whereas a replacement with another function with the same arguments is called a *soft intervention* (or *imperfect*, *parametric*) ((Peters, Janzik, and Schölkopf 2017, 35) and (Peters, Janzik, and Schölkopf 2017, 89))

We already encountered examples for hard interventions, like closing switch X_1:

$$X_1 := 1 \tag{2.9}$$

.

A soft intervention would be a rearrangment of the circuit, so X_1 and X_2 are placed are placed in parallel rather than in series. The state of the light bulb would still depend on both X_1 and X_2, but the assignment functions would be more like this:

$$Y := g(X_1, X_2) = I(X_1 + X_2 >= 0) \tag{2.10}$$

where $I(a)$ is the indicator function that returns 1 if a is true and 0 otherwise. In a parallel arrangement it is sufficient that *any* of the two switches is closed:

switch X_1	switch X_2	light bulb Y
0	0	0
0	1	1
1	0	1
1	1	1

2.5   Probabilistic Systems

In the previous chapter we have seen that our inability to observe the state of a variable introduces uncertainty. The model is then said to be *probabilistic* rather than *deterministic*. Uncertainty, however, can be introduced into the system for these reasons:

- unobservability;
- measurement error (i.e. we cannot prefectly observe the actual state of a variable but only an imperfect measure);
- inherent probabilistic behaviour of the physical mechanism ("quantum mechanics");
- modelling simplification, i.e. treating a determinsitc system as though it

were random, e.g. flipping a coin, generating "random numers" on a computer.

2.5.1 An electric circuit with two switches, one unobserved

Let us start with the problem above, where we model a system with two switches, but the state of switch one is unobserved. To make the distinction between the two switches more apparent, let's denote the unobserved switch by U and the observable switch simply by X:
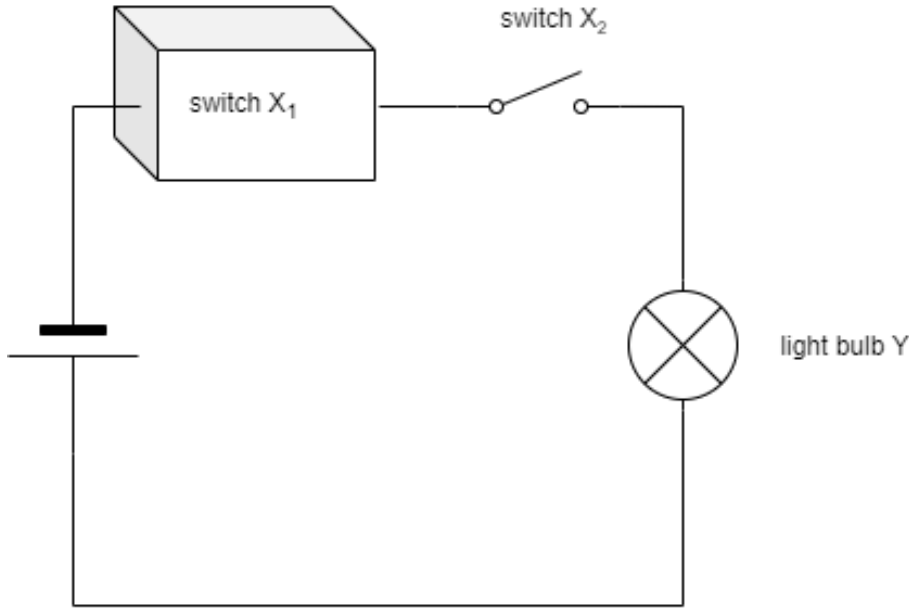
Figure 2.6: A circuit diagram with a single power source, a two switches (X and U) and a light bulb (Y). Switch U is unobserved.

Note that the *structure* of the causal graphs remains the same, as the box only hinders us to observe the *state* of the switch, but we still know that there is a switch. This allows us to further specify $f()$ of the structural equation, which after refactoring variable names is now

$$Y := f(U, X) = X \cdot U \tag{2.11}$$

Let us further introduce a symbol for the system itself, S, which not only denotes the structure of the system, but includes the probabilistic mechanism that determines U. We can now *derive* the *probability distributions* for Y from the structural equation for each possible interventions on X:

$$P^{S;do(X:=x)}(Y = y) = P^{S;do(X:=x)}(X{\cdot}U = y) = P^{S;do(X:=x)}(x{\cdot}U) = xP^{S;do(X:=x)}(U)$$
$$\tag{2.12}$$

where $do(X := x)$ denotes the intervention on X where its state is set to x. Note that, as the intervention fixes X to a specific value, it is non-random and can therefore be placed outside of $P()$.

Causal reasoning about the effect of the switch on the light bulb (the effect of setting the state of X on the state of Y) now depends not on the state of U, but its probability distribution. We can safely say that, if we open the switch, then the light bulb will never be on, regardless of the probability distribution of U:

$$P^{S;do(X:=0)}(Y = 1) = 0 \cdot P^{S;do(X:=0)}(U) = 0 \tag{2.13}$$

Again, if we were to observe a case where the switch is open but the light bulb is on, we would instantly know that our structural equation is invalid.

This is different however for the case where we close the switch:

$$P^{S;do(X:=1)}(Y = 1) = 1 \cdot P^{S;do(X:=1)}(U = 1) = P^S(U = 1) \tag{2.14}$$

Say, for example, system S refers to a population of circuits that are all created by the same machine. This machine is supposed to close the switch in the box, but fails to do so 1% of the cases. Hence, $P^S(U = 1) = 0.99$ and therefore $P^{S;do(X:=1)}(Y = 1) = 0.99$: only in 99% of the cases where we close the switch, we will see the light bulb go on, in 1% of the cases it will fail to do so.

2.5.2

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

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:

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

2.6   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

$$E(U - V) = E(U) - (V) \tag{2.15}$$

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.[2]

2.6.1 Definition

bla

2.6.2 Causal Effect Statistics

Definition 2.4 (Average Treatment Effect)**.** The Average Treatment Effect, or ATE, is the expected value of xx in population x.

bla bla bla

Definition 2.5 (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 2.6 (Intention To Treat Effect)**.** The Intention To Treat Effect, or ITT, is the expected value of xx in population x.

---

[2]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, $Median(income - expenses) \neq Median(income) - Median(expenses)$.

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 2.7 (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

Inferring causal mechansims from data is more difficult than learning about properties of the distribution of variables:

- Being able to infer causal mechanisms from data requires knowledge about the data generating process which cannot be inferred from the observed
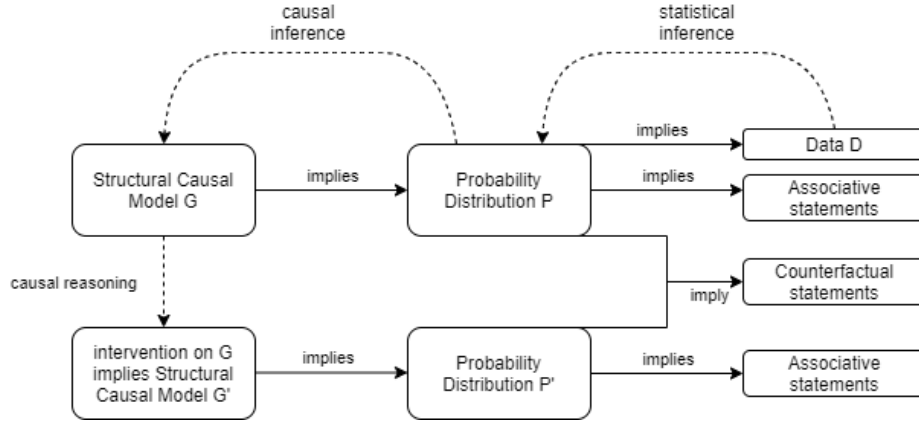
Figure 3.1: Causal vs Statistical Inference

data itself. Observed data can help us justifying assumptions on the causal structure we are willing to accept, but it cannot fully specify the causal model in any application of interest. Hence, we will always have to incorporate (domain) knowledge to enrich the observed data before we can make causal inferences.

- Causal inference does not get rid of the typical statistical problems. These are still there and need to be solved. The causal problems are just an additional layer on top.

The causal structure implies statistical dependencies. If observations violate the implied dependencies, the causal structure cannot be correct. In practice, observing two variables to be statistically dependent these variables either have a common cause or they directly cause each other. There are practical limitations though:

- this only holds true for i.i.d. variables. if the observed data is a time series, this property does not hold. hence, time series observations might be statistically dependent, although there is no causal link. a common phenomenon is that two variables are *trending* over time. As an effect, observations from such time series will likely have a very high correlation, despite them being causally independent.[1]

- our ability to correctly infer statistical dependence might not be straightforward:

---

[1] As most (macro-)economic data are time-series, the econometrics literature is full of methods to establish causality in time-series. Most of it goes beyond the description provided in section xxx but unfortunately there are currently on few sources that cover this material in the language and notation of causality adopted here. See also https://xkcd.com/925/.

    – if our analysis involves many variables, the number of tests necessary
to establish conditional dependence between any pair of two variables
increases swiftly. Maintaining a fixed type I error for independence
will become exceedingly difficult. We may then falsely identify two
variables as being dependent.[2]

    – we might not be able to observe (a random sample of) the outcomes
of a causal mechanism, but only a biased one. this is commonly
referred to as selection bias (or survivalship bias) in the literature.
See also https://xkcd.com/1827/

3.1.1 Example 1 Tutoring

Let us first explore a simple but insightful DAG and reason about interventions
and causal effects. Assume the model for the causal relationship between grade
Z, tutoring Y and passing the final exam Y is as shown in DAG XXX.

Figure 3.2: DAG1

More precisely, let's first look at a deterministic quantitative relation given by

---

[2]This problem is common in the search for heterogeneity in treatment effects, where e.g. the
effectiveness of drug is broken down by patient groups (e.g. male vs female, with prior indica-
tion vs without). If the number of groups available to the researcher increases, the likelihood
of a false discovery increases without proper control. See also https://xkcd.com/882/.

$$X := I(Z >= 4) \tag{3.1}$$

$$Y := I(Z < 5 \text{ or } (Z = 5 \text{ and } X = 1)) \tag{3.2}$$

i.e. all students with grades 4, 5 or 6 get tutoring (but no students with grades 3 or better); students with grades 4 or better are all passing the exam, whether or not getting tutored, as well as tutored students with grade 5; untutored students with grade 5 are failing the test as well as all students with grade 6 (tutored or not). This system is already insightful to show that the causal model entails *multiple* probabilistic models. First, note that we do not observe any students with grade 5 who fail the test and that every failed student had grade 6. The observed conditional expected values for $Y|Z$ are given in table xx. Second, we can tell what will happen if we intervene in one or more mechanisms. If we, for example, restrict tutoring to students with grade 6, i.e. $X := I(Z >= 6)$, all students with grade 5 will fail the exam as well as those with grade 6. But nothing will change for students with grade 4, they still will all pass the exam. Without knowing the causal mechanism, the different behavior of stundents with grade 4 and 5 could not have been predicted. In the observed data, both groups were not distinguishible, neither their relation to tutoring (all got tutored) nor by their exam performance (all passed the exam).

Let's now make this example a bit more interesting (and slightly more realistic) by abandoning the deterministic nature of the mechanism and using a probabilistic one instead. The graph is then to be xxx.

Let's assume the SCM to be given by

$$Z := f(U_z) \tag{3.3}$$

$$X := f(Z, U_x) \tag{3.4}$$

$$Y := f(X, Z, U_y) \tag{3.5}$$

This results int

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
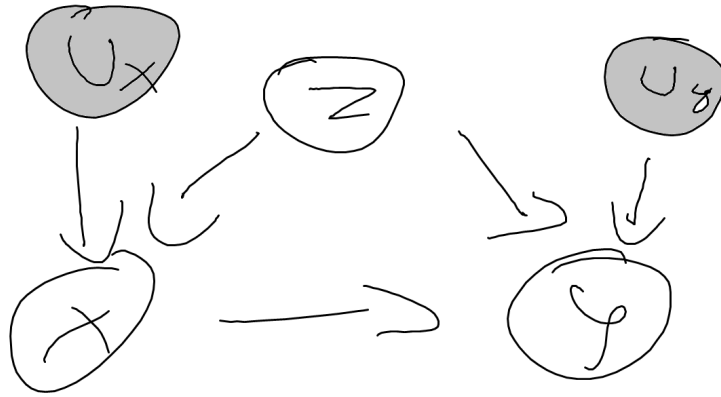
$$X := U \tag{3.6}$$

Figure 3.3: DAG2

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   Observational Data

We will take a close look at a classic problem that occurs when we try to infer causal statements from observational data. The basic model that we use is the
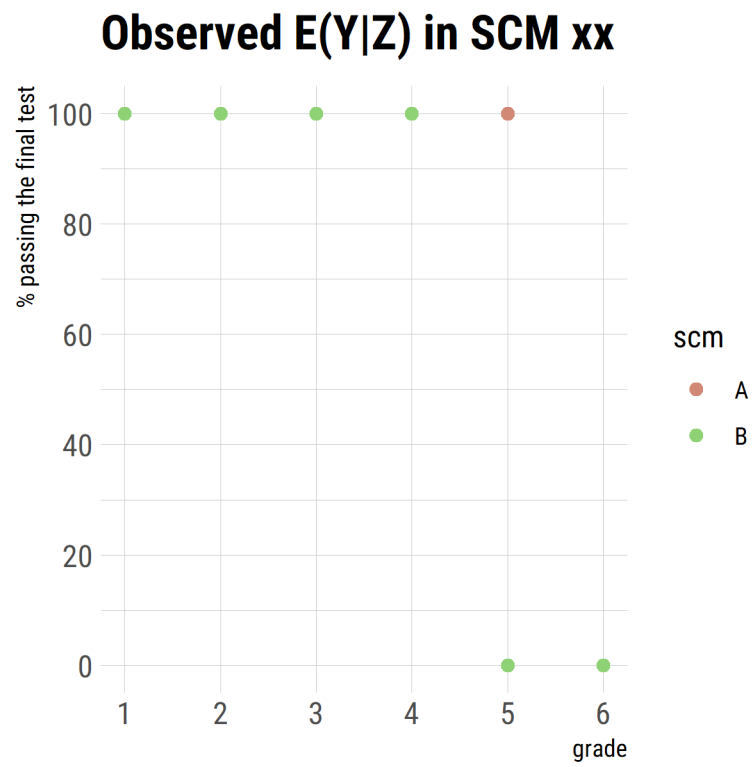
Figure 3.4: CE1

three-node model, where a variable Z is expected to be a confounder for the causal relation of X on Y, see graph xxx
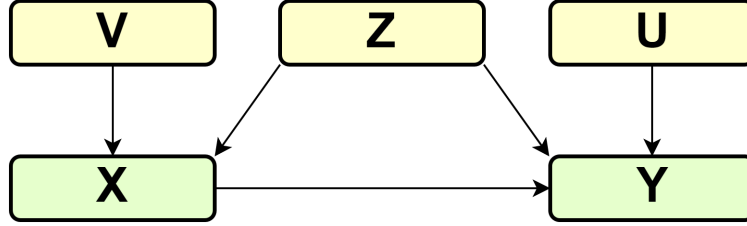
Figure 3.5: DAG

In the most simple case, Z, X and Y are all binary and the SCM Γ consists of

$$X := f_1(Z, V) \tag{3.7}$$
$$Y := f_2(X, Z, U) \tag{3.8}$$

and exogeneous variables Z, V, U are mutually independent.

A classic application of this problem occurs in epidemiology, where patients self-select into treatment X, where certain risk factors Z determine the health outcome Y as well as their decision to get treatment X. If observed data that was generated by such a process is used to make inferences about the effect of X on Y, comparing conditional expectations $E[Y|X = 1] - E[Y|X = 0]$ will lead to incorrect conclusions about the (average) causal effect.

To make this clear, let us walk through this example step-by-step to clarify this argument. First, note that the intervention we are interested in can be described by the graph below

Figure 3.6: DAG

and a change in the structural assignment of X

$$X := 1 \tag{3.9}$$

in the case of treatment of the entire population (Γ_1) and

$$X := 0 \tag{3.10}$$

in the case of no treatment (Γ_0).

What we're interested in is $E^{\Gamma_1}[Y] - E^{\Gamma_0}[Y]$, but the observed data was generated by Γ rather than Γ_1 or Γ_0. We not need to figure out how we can still get the right statistics.
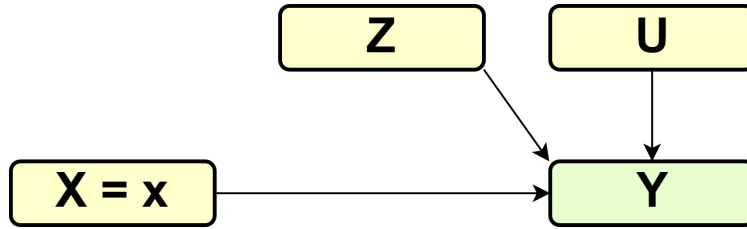
For this, follow

$$
\begin{aligned}
E^{\Gamma_1}[Y] &= P^{\Gamma_1}(Y = 1) \\
&= P^{\Gamma_1}(Y = 1, X = 1, Z) \\
&= P^{\Gamma_1}(Y = 1, X = 1, Z = 1)+ \\
&\quad P^{\Gamma_1}(Y = 1, X = 1, Z = 0) \\
&= P^{\Gamma_1}(Y = 1|X = 1, Z = 1) \cdot P^{\Gamma_1}(X = 1, Z = 1)+ \\
&\quad P^{\Gamma_1}(Y = 1|X = 1, Z = 0) \cdot P^{\Gamma_1}(X = 1, Z = 0) \\
&= P^{\Gamma_1}(Y = 1|X = 1, Z = 1) \cdot P^{\Gamma_1}(Z = 1)+ \\
&\quad P^{\Gamma_1}(Y = 1|X = 1, Z = 0) \cdot P^{\Gamma_1}(Z = 0)
\end{aligned}
$$

where the last equality holds because in Γ_1 everyone in the population gets the treatment and therefore

$$
\begin{aligned}
P^{\Gamma_1}(X = 1, Z = z) &= P^{\Gamma_1}(Z = z|X = 1) \cdot P^{\Gamma_1}(X = 1) \\
&= P^{\Gamma_1}(Z = z|X = 1) \cdot 1
\end{aligned}
$$

So far, we have rewritten $E^{\Gamma_1}[Y]$ using some tricks of probability theory. The key point now is to recognize that $P^{\Gamma_1}(Y = 1|X = 1, Z = 1)$ describes the mechanism of how the health output is determined by the treatment and the patient's risk factors. From earlier discussion, we know that mechanisms remain invariant under atomic interventions. Hence,

$$
\begin{aligned}
P^{\Gamma_1}(Y = 1|X = 1, Z = 1) &= P^{\Gamma}(Y = 1|X = 1, Z = 1) \\
P^{\Gamma_1}(Y = 1|X = 1, Z = 0) &= P^{\Gamma}(Y = 1|X = 1, Z = 0) \\
P^{\Gamma_1}(Z = 1) &= P^{\Gamma}(Z = 1) \\
P^{\Gamma_1}(Z = 0) &= P^{\Gamma}(Z = 0)
\end{aligned}
$$

and therefore

$$
\begin{aligned}
E^{\Gamma_1}[Y] &= P^{\Gamma}(Y = 1|X = 1, Z = 1) \cdot P^{\Gamma}(Z = 1)+ \\
&\quad P^{\Gamma}(Y = 1|X = 1, Z = 0) \cdot P^{\Gamma}(Z = 0)
\end{aligned}
$$

Following the same reasoning, we can also derive that

$$
\begin{aligned}
E^{\Gamma_0}[Y] &= P^{\Gamma}(Y = 1|X = 0, Z = 1) \cdot P^{\Gamma}(Z = 1)+ \\
&\quad P^{\Gamma}(Y = 1|X = 0, Z = 0) \cdot P^{\Gamma}(Z = 0)
\end{aligned}
$$

which finally produces

$$ATE = E^{\Gamma_1}[Y] - E^{\Gamma_0}[Y]$$
$$= P^\Gamma(Y = 1 | X = 1, Z = 1) \cdot P^\Gamma(Z = 1) + P^\Gamma(Y = 1 | X = 1, Z = 0) \cdot P^\Gamma(Z = 0) -$$
$$P^\Gamma(Y = 1 | X = 0, Z = 1) \cdot P^\Gamma(Z = 1) - P^\Gamma(Y = 1 | X = 0, Z = 0) \cdot P^\Gamma(Z = 0)$$
$$= (P^\Gamma(Y = 1 | X = 1, Z = 1) - P^\Gamma(Y = 1 | X = 0, Z = 1)) \cdot P^\Gamma(Z = 1) +$$
$$(P^\Gamma(Y = 1 | X = 1, Z = 0) - P^\Gamma(Y = 1 | X = 0, Z = 0)) \cdot P^\Gamma(Z = 0)$$

In words, the ATE can be calculated from the observational data by comparing the average outcome of treated and non-treated individuals *within each risk group Z*, and then weigthing these within-group comparisons by the proportion of individuals in each risk group.

Obsiously, this is different from the naive group comparison of treated and non-treated patients

$$E^\Gamma[Y|X = 1] - E^\Gamma[Y|X = 0] = P^\Gamma(Y = 1|X = 1) - P^\Gamma(Y = 1|X = 0) \quad (3.11)$$

which therefore does not provide the ATE.

In some cases, this can lead to the phenomenon known as Simpson's paradox, where the different metrics have differnt sign.

$Y = 1/Y = 0$	$Z = 0$	$Z = 1$
$X = 0$	260/400	55/100
$X = 1$	70/100	240/400

For the data shown int table xxx

$$E^{\Gamma_1}[Y] - E^{\Gamma_0}[Y] = (\frac{240}{400} - \frac{55}{100})\frac{500}{1000} + (\frac{70}{100} - \frac{260}{400})\frac{500}{1000} = (0.6 - 0.55)0.5 + (0.7 - 0.65)0.5 = 0.05$$
$$(3.12)$$

whereas

$$E^\Gamma[Y|X = 1] - E^\Gamma[Y|X = 0] = (\frac{310}{500} - \frac{315}{500}) = 0.62 - 0.63 = -0.01 \quad (3.13)$$

3.4 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

3.5   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

3.6   Difference-in-Difference Estimator

TODO

3.7   Time Series Methods

TODO

Chapter 4

Applications

4.1 Causality and Machine Learning

Machine learning has seen many successful applications in recent years. Methods most often employed today belong to the unsupervised and supervised learning paradigm. These methods assume that observed data are identically and independently distributed samples of a (super-) population. For example, supervised learning techniques attempt to learn Y from a set of variables X, i.e. trying to learn $P(Y|X) = f(X)$. From the discussion so far it should become clear that only few problems are fit for this paradigm. If X and Y are produced by the same causal/physical system, it should be able to learn the form of $f()$ by collecting lots of data points and fitting algorithms that function class. For example, the "hello world" example of machine learning consists of recognizing handwrittern letters. It is reasonable to assume that this is a rather stable relationship. There aren't so many different ways people write letters once you've seen houndreds of thousands of them. It is also likely that there is no evolution of that relationship, that observervations taken a year ago are not that different from those taken now. It should also be likely that you can transfer this relationship reasonably well between physical systems that do the same or a similar thing, e.g. handwritten letters from tablets vs phones. Transfer will not as well work if the physical system is quite different, e.g. writing with a pen on paper rathen than with a stylus on a digital device.

Many important systems, however, are constantly evolving. It is not reasonable to assume that movie recommendations will be stable over time. Recommending a new show that is currently all the buzz, might not be a good recommendation in a year. Customers who haven't chosen to see this show yet, might be well aware of the show, but have deliberately decided not to watch it. Using the same trained recommendation model over a prolonged time-period might lead to ever-decreasing model performance and will hence require constant re-training.

Models then do not always get better and better over time as information from past observations depreciates in value quite quickly.

Some systems are evolving much quicker than data becomes available. The (world) economy is such a system. Individual behavior and choices are continuously refined and can dramatically change on short notice, while most (macro-) economic data might only be available on a weekly or monthly basis. Innovation plays an important role. The action set is not fixed over time, but evolves constantly. It can neither be modelled nor predicted (it wouldn't be an invention then)

4.2   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]

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

---

[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 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.

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 likelihood 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 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

---

[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.

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 A is binary. Then

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

---

[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. $\hat{P}(\Delta) = \frac{1}{n_A} \sum_{i \in I_A} Y_i - \frac{1}{n_B} \sum_{i \in I_B} Y_i$.

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

$$P(Y) = P^{S;do(Z:=0)}(Y) \cdot (1-Z) + P^{S;do(Z:=1)}(Y) \cdot Z \qquad (4.7)$$

$$= P^{S;do(Z:=0)}(Y) + (P^{S;do(Z:=1)}(Y) - P^{S;do(Z:=0)}(Y)) \cdot Z \qquad (4.8)$$

$$= P^{S;do(Z:=0)}(Y) + P(\Delta) \cdot Z \qquad (4.9)$$

$$= \beta_0 + \beta_1 Z \qquad (4.10)$$

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

$$P(Y|A) = P^{S;do(Z:=0)}(Y|A) + P(\Delta|A) \cdot Z \qquad (4.11)$$

$$= \alpha_0 + \alpha_1 A + (\Delta_0 + \Delta_1 \cdot A) \cdot Z \qquad (4.12)$$

$$= \alpha_0 + \alpha_1 A + \Delta_0 Z + \Delta_1 AZ \qquad (4.13)$$

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"

4.3 Drug Trial

4.4 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 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

---

[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

- *zulässige Merkmale*: dies sind Merkmale, nach denen eine diskriminierende Handlung zulässig ist. So kann etwa für eine Tätigkeit in einem Warenlager vorausgesetzt 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

4.5 Epidemiology

Inspired by the SARS-CoV-2 epidemic of 2020, let's have a look at a causal model for aggregate epidemiological data. During the epidemic, the numbers of infected and deceased have been widely reported in news outlets throughout the world and (especially in the beginning) have been closely monitored by millions on a daily basis. Soon, the reported numbers have raised many questions

as they suggested wildly different conclusions about the virus' lethality across countries. While the WHO reported a case-fatality-rate (CFR) of 3.5%, some countries reported CFRs as low as 0.3%. Differences in pogress of the pandemic, demographics, and the availability of intensive-cars-units (ICU) would suggest *some* differences, but a deviation by an order of magnitude seemed unlikely

The differences in reported numbers and statistics are understood to be substantially driven by differences in the data generation process, especially the mechanism of who is subjected to a test for the virus. While South Korea adopted a strategy of widespread testing of the population, others only tested when patients showed (severe forms) of associated symptoms and/or recent contact with known infected people.

To illustrate this, we will model parts of the data generating process as causal mechanisms, allowing us to reason about the effects of different testing strategies on reported numbers, how we can estimate the number of patients who died *because* of the virus rather than died *with* the virus, and how government measures (lockdown) fit into this model.

In a second part, we swith focus a little bit and take a closer look at the dynamics of the pandemic. We show how dynamic systems can be represented in causal graphs and causal assigments. We will use this to have a closer look at different types of interventions in dynamical settings, primarily between one-time interventions and permanent interventions. Lastly, we shortly discuss the relation between unit-level models and models for aggregate variables.

4.5.1   The reporting mechanism

A crucial statistic in epidemiology is the CFR, the proportion of deaths from a certain infection compared to the total number of those diagnosed with the disease for a certain period of time. It is a measure risk measure as its definition guarantees that the value is always between 0 and 1. Although widely used in the medical literature and beyond, it is a problematic concept. Often used to compare severities of different diseases and viruses, its definition does not account for potentially large differences in the process generating the numerator and denominator of the statistic. During the SARS-CoV-2 outbreak, limited testing capabilities and large amount of asymptomatic cases dramatically suppressed the reported number of infected, distorting comparisons with similar better known viruses such as the influenza. A more robust statistic is the infection fatality rate (IFR), which uses the number of *infected* rather than *diagnosed* individuals in its denominator (but of course cannot be observed directly). We will discuss this in a minute.

The CFR (as well as the IFR) is furthermore problematic as it is often treated as an attribute of the virus itself (at least for a given population). This neglects the crucial fact that for many diseases, treatments might be available that save lives of many patients. If, as it was the case during the onset of the SARS-CoV-2

pandemic, a major concern is the availability of treatment resources (especially ICU units), distinguishing between a CFRs *with* and *without treatment* is important to avoid confusion.[8]

To dive deeper into this problem, let's take a look at a causal graph for the generation of these observable numbers and statistics. For simplicity, we will present the causal graph as a summary graph, where we neglect the inherent time structure of the mechanism. This does not cause any problems for causal reasoning, as the summary graph is still remains acyclic. However, to make inferences based on observed data, we will later on take a look at the full time graph as well.
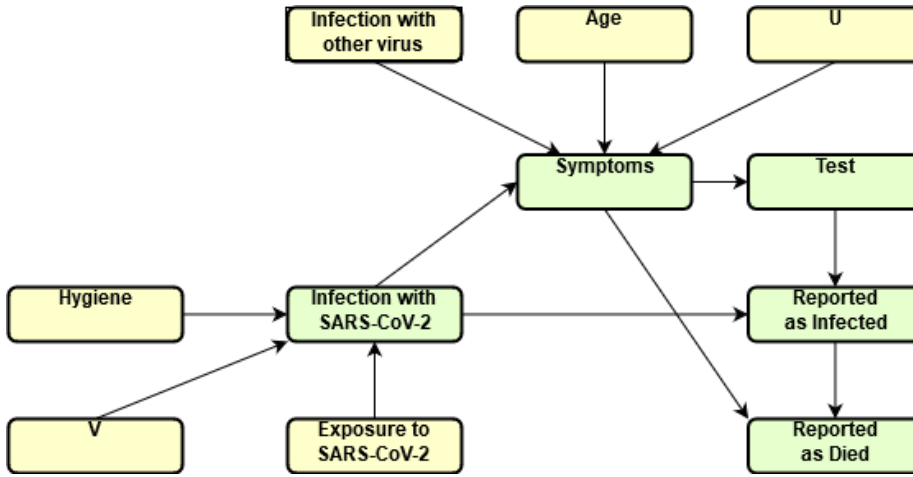

Figure 4.1: DAG of the process generating data on infected and deceased individuals due to SARS-CoV-2

While the causal graph might be a reasonable description for the reporting mechanism across countries and over time, the details of the causal relationships certainly were neither time-constant nor identical across coutries. Below we describe the situation as it likely existed in March of 2020 in Germany as an example:

- For an individual to be reported to be infected with SARS-CoV-2, there are two main causal influneces: whether the person is in fact infected with that virus and whether a PCR test was performed. If we assume that the PCR test used is perfect (i.e. has no type I error nor a type II error), we can model this relationship deterministically as

---

[8]Early numbers suggested that, after accounting for asymptomatic and untested cases, the SARS-CoV-2 virus might be similarly fatal as the influenza. This conclusion often neglected the fact that influenza-patients usually get the full treatment, but that faster spreading of SARS-CoV-2 might concentrate severe cases to a shorter time period, overwhelming the capaicities for treatment and therefore increasing the CFR.

$ReportedInfection := f(Test, Infection)$. Note that the assumption of a perfect test also allows us to disregard any causal influence from *OtherInfection* to *ReportedInfection*, which would be the case if the test accidently would show positive outcomes in cases where the person does not have SARS-CoV-2, but a similar infection (e.g. other corona viruses), a problem of early antibody tests.

- In Germany, an individual is tested for SARS-CoV-2 if they are symptomatic and/or have had recent contact with a person known to be SARS-CoV-2-positive. Again, for simplicity, we assume that this is a strict rule-based (and stable) mechanism.[9] The structural assigment is $Test := g(Symptoms, ContactInfected)$.

- A person's symptoms are determined by them being infected by the SARS-Cov-2 virus, other viruses causing respiratory diseases, age, pre-existing conditions, as well as other unspecified and unobserved factors (e.g. genetics). The assignment is

$$Symptoms := h(Infection, InfectionOther, Age, PreCondition, U)$$
$$(4.14)$$

and the assignment is probabilistic due to U being treated as a noise term. As we define symptoms to include the value "death", this mechanisms has to fully explain a person's likelihood to die with and without the virus.

- The infection mechanism is modelled as a probablistic function of personal hygiene and exposure to the virus:

$$Infection := h(Hygiene, Exposure, V) \qquad (4.15)$$

- finally, a person is reported to have died of SARS-CoV-2 if the person has been positively tested for the virus and has died (one of the possible values of the symptom):

$$ReportedDeath := h(ReportedInfection, Symptoms) \qquad (4.16)$$

We assume this to be a simple and deterministic mechanism.

4.5.1.1   Reasoning about interventions and counterfactuals

The model now allows us to reason about certain interventions that affect the reported numbers of infected and deceased.

---

[9]In reality, this assumption was violated. Over the course of the epidemic, testing capabilities have increased in most countries and tests have been expanded over time. In some countries and regions, asymptomatic persons have been tested as well, even if there was no known contact with another infected person. We discuss this in more detail when we take a look at the full time graph of the model.

4.5.1.1.1 No virus The most important one is certainly the counterfactual question of what would happen if there were no SARS-CoV-2 virus out there, which serves as a baseline model that allows us to compare the pandemic situation to a "normal" one. Technically, this amounts to replacing the assignment for variable *Infection* to

$$Infection := 0 \tag{4.17}$$

As this is just a hypothetical intervention, let us assume that it is atomic, i.e. we assume that the rest of the model is not changed by this intervention. Most importantly, the assignment $Symptoms :=$ $h(Infection, InfectionOther, Age, PreCondition, U)$ is still in place, it's just that the value of one of the inputs is 0 in the entire population. As $Symptoms$ was defined to be a categorical variable including the value "death", we can use the model to compare two distributions in the population, one with and one without the virus. The difference tells us the additional deaths due to SARS-CoV-2 virus rather than just the numer of deaths associated with SARS-CoV-2 (which is what $ReportedDeath$ is capturing). This information is important because we ultimately want to understand how deadly the disease acutally is, i.e. how many deaths it causes, not just how prevalent it is in people how are dying. With the median age of reported deaths with the virus being at around 80 and the probability of a person at 80 to die within 12 months being at around 5%, this is non-negligible.

4.5.1.1.2 Changing testing strategy expanding test capacities comparing numbers across countries with (very) different test strategies random testing

4.5.2 Evolving Symptoms

To model the evolution of symptoms over time, we need to switch from a time series summary graph to a full time graph.

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

Probabilistic Models and Reasoning

A probabilistic model is based on the theory of probability. It incorporates *random variables* that can assume multiple possible outcomes and maps these to *probability distributions*.

TODO:

B.1 Probabilistic Reasoning

TODO: joint distribution, conditional distribution, chain rule, independence

Appendix C

Statistical Inference

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.

C.1   Distributions

C.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

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

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

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

and the expected value scales linearly with scaling factor a

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

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

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

only if X and Y are independent.

C.1.2   Multiple Random Variables

- Covariance
- Correlation
- Conditional Probability
- Regression Parameter

C.2   Statistical Inference

C.2.1   Estimators

C.2.2   Hypothesis Tests

Appendix D

Code

D.1 Simulate Causal Models

The R programming language (see R Core Team (2020)) is a useful tool for simulating causal models and estimating casual parameters. It is particularly well suited to the task as its built-in statistical functions and its syntax allow us to easily translate causal models into code.

The simple causal model from section xxx with two switches and a light bulb can be written as

```r
X1 <- 1
X2 <- 0
Y  <- X1 * X2
```

It might be helpful to keep the skeleton of the causal model (as defined by the causal graph) and the functional form of the causal mechanism separately:

```r
# mechanisms
fu <- function(){}
fx <- function(u){}
fv <- function(){}
fy <- function(x, v){}
# skeleton
U <- fu()
X <- fx(U)
V <- fv()
Y <- fy(X, V)
```

A comparison of probability distributions from several interventions on a system can be simplified by first create a "function factory" (technically a *closure*) which returns a function.

```r
ff <- function(fu, fx, fv, fy){
  function(){
    U <- fu()
    V <- fv()
    X <- fx(U)
    Y <- fy(X, V)

    return(c(u=U, x=X, v=V, y=Y))
  }
}
```

The system before intervention is specified as follows:

```r
# system 1 is defined by these mechansims:
fu <- function(){rnorm(1)}
fx <- function(u){u}
fv <- function(){rnorm(1)}
fy <- function(x, v){1+0.5*x+v}

# s1 is a function 'running' system 1:
s1 <- m(fu=fu, fx=fx, fv=fv, fy=fy)
```

Note that `s1` is a function that takes no parameters

An intervention on the mechanism for X is a replacement of function $f_X(U) = 1$:

```r
# system 2 is a hard intervention on
# system 1 where the mechanism for x
# is changed to a fixed value (1):
fx <- function(u){1}

# s2 is a function 'running' system 2:
s2 <- m(fu=fu, fx=fx, fv=fv, fy=fy)
```

Each function call provides a draw from the distribution of its system:

```r
> s1()
        u         x         v         y
0.3272407 0.3272407 1.4474576 2.6110779
> s2()
         u          x          v          y
1.46572184 1.00000000 0.01263931 1.51263931
```

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.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.