Causal Inference in Statistics

With Exercises, Practice Projects, and R Code Notebooks

(Draft - Introduction)

Justin Belair

2024

Contents

Preface

Another Book on Causal Inference!?

This book stands on the shoulders of many great textbooks on causal inference.

A very gentle, short, and captivating non-technical book is [Rosenbaum, 2023]. A good introductory book which eases into technical material is [Pearl et al., 2016], yet it has some flaws in my opinion. Indeed, it does not go too far into the subject, as it is simply a primer and takes around a third of the space to go over basic probability theory and graph theory before really diving into causal inference. This short book thus leaves very little space to go deeper in causal inference.

My aim has been to take the next step from this book. That is, bring you, dear reader, up-to-date on current statistical methodology in causal inference! I made the book self-contained, so we will go over all the necessary theory, but we assume you, dear reader, is familiar with probability and statistics and leave all this basic theory in the Appendix 12.4, for your reference.

I attribute [Pearl, 2009] with sparking my curiosity to go deeper in the fascinating subject of causal inference - it's a true masterpiece and I recall fondly the intellectual awakening it sparked in me when I read it for the first time. It is a difficult book written with the goal of rigorously defining Pearl's approach to causal inference, which relies a lot on advanced mathematics–at least, more advanced than what a traditional applied (data) scientist is equipped to handle. [Imbens and Rubin, 2021] is another very elegant textbook which I loved reading, but it is also heavy on mathematics. [Morgan and Winship, 2014] is rich with references to literature and real-world examples taken from decades of causal debates in social sciences, but it can be a bit heavy. We will also discuss epidemiology as it has become increasingly technical, with much overlap with what was traditionally called biostatistics. Plus, epidemiology relies almost uniquely on observational data, where causal inference methodology is indispensable. As we will see, a major intellectual breakthrough in causal inference comes from our understanding of when and how we can analyze observational data as if it was experimental, so-to-speak. Great references for epidemiology and more specifically causal inference within the field, are [Rothman et al., 2008] and [Westreich, 2020]. There are even more specialized books like [VanderWeele, 2015] and [Shipley, 2000a], an ecology book, that will be referenced throughout. All this is without mentioning the 100s of papers I've read over the years that you will find in the bibliography. This sprawling literature on causal inference serves as a foundation to the book you're about to read.

As a trained mathematician and statistician who's taught dozens of university-level courses to business students, data scientists, engineers, and researchers, I have the necessary background to decipher these complex textbooks and bring out the essence for non-mathematicians, without having to go through long and rigorous mathematical proofs. I've also worked as a statistical consultant with dozens of scientists from around the world in many biomedical sciences, as well as researchers in psychology, psychiatry, agronomy, epidemiology, and even Research and Development (R&D) in biotechnology industries. I've been applying causal inference in applied scientific research for a long time now, through my services offered at a company I founded and run on a daily basis, JB Statistical Consulting[1]. I'm also a philosophy dilettante, who's been mostly interested in figuring out how this beautiful thing we call Science actually works and how to do it well, through reading philosophy of science, history of science, epistemology, and more. All this is to say that I (again, very humbly) believe I'm uniquely positioned to take the complex mathematical and philosophical ideas behind causal inference, decipher them and bring them to a level where an applied scientist can use them. Plus, I absolutely love sharing knowledge I'm passionate about!

Don't get me wrong, in this book, there will be some mathematics, and even some proofs, as you will see by skimming through its contents, especially in the first few Chapters, where we lay the conceptual foundation on which we will later build our causal inference toolbox. This is simply because the causal inference revolution we present in Chapter 1 is mainly one of finding a mathematical language to clearly and formally express ideas of causality. Rest

[1] www.justinbelair.ca

assured, mathematics will not be the main goal and kept to a minimum. All the necessary mathematical notations, definitions, and theorems will be stated and explained clearly. As I just described in the previous paragraph, this book is NOT aimed at mathematicians, although I sincerely hope they would still learn from it!

I don't believe in the so-called *dumbing down* of content; I'd rather elevate the reader. Learning anything worthwhile is usually difficult, requires patience and perseverance and is not always pure pleasure. Causal inference is not different–it is admittedly a difficult subject, but oh so interesting! My promise to you is to do my part to make things as clear and interesting as possible. And, believe me, there will definitely be lot's of fun amidst the difficulty, so don't give up when it gets tough!

Rather than being submerged by complex theorems and their proofs, you will be treated to elegant graphics representing the core ideas of the book, a suite of R notebooks with code that you can download for free to play around and practice with data, as well as a series of exercises to help you truly master the subject–something that is lacking in many causal inference textbooks. I also like to incorporate historical nuggets–they bring out the people and the stories behind these great ideas, giving them a bit more humanity. All the topics will be accompanied by extensive references to the literature, if ever you, dear reader, wish to go further!

Should I Read This Book? If So, Do I Have To Read Cover-to-Cover?

Before answering the title-question of this section, let me rapidly go over the layout of the book.

The first part of the book is formed by 4 chapters. The introductory chapter serves as a guiding light through the journey we are about to embark on together. Chapters 2 and 3 form the theoretical backbone of causal inference and Chapter 4 will discuss in detail the problem of uncovering causal effects when we cannot perform experimental manipulations. It is here that we will introduce the formal theories in sufficient detail for them to be understood clearly and rigorously, especially to later apply them.

The second part of the book forms the bulk of the tools that are used by applied researchers to study causal relationships in the data they care about, whether experimental scientists in pharmacology, social scientists using observational data for program evaluation, or psychologists trying to understand human behavior. We will tackle propensity score methods in Chapter 5, instrumental variables in Chapter 6, Structural Equation Models (SEM) in Chapter 7, analysis of mediation and interactions in Chapter 8, and quasi-experimental methods in Chapter 9, such as differences-in-differences. Chances are that if you're a scientist or researcher who routinely uses data, you've heard some of these terms before. You might even be familiar with some of these methods. In my humble opinion, this book can still be extremely valuable owing to the fact that, as far as I know, no book combines a thorough review of the theory with as much practical material through exercises and coding examples.

The final part of the book is dedicated to the frontier of causal inference in scientific research, tackling advanced methods such as longitudinal causal inference and time-varying confounding in Chapter 10. Methods related to machine learning and AI such as targeted learning, double machine learning, meta-learners, and causal discovery in Chapter 11. These recent and more advanced methods are extremely promising, as they build on the insights that have been gleaned from a solid few decades of research in causal inference.

Obviously, the second and third parts of the book are more application-focused. Their chapters can all be read individually and independently, even if they are loosely ordered by conceptual difficulty. If you are strictly interested in learning how to apply methods and use code to do so, you can tackle these chapters directly. On the other hand, the first part is mostly focused on developing a rigorous conceptual approach to causal inference and might seem difficult or even alien to you, depending on your background and your familiarity with mathematical ideas and abstract, conceptual thinking. In my opinion, I believe the first part of the book is very valuable, even to readers who already understand both the potential outcomes framework associated with the work of Donald Rubin and the graphical causal models (i.e. DAGs) attributed to Judea Pearl, or readers that don't have a particularly strong background in mathematics. If I did not think them valuable, I would not have put them in the book, as nobody has time to waste! Still, you might wish to skip over them, skim them rapidly, or go back to them to help you apply the tools of the two latter parts of the book. Multiple different approaches can work for you, depending on what kind of reader you are.

When writing the book, I was driven by the desire to make it easily navigable, both when reading cover-to-cover or when simply picking and choosing subjects of interest. For this reason, I also added an Appendix to cover some building blocks that you might not have studied before, such as certain concepts in probability theory or statistics.

Part I

Fundamentals of Causal Inference

Chapter1

The Causal Inference Revolution - Introduction and Examples

1.1 The Causal Inference Revolution : An Intellectual Landmark

We're experiencing nothing short of an intellectual revolution[1]. In a 2021 paper titled *What are the most important statistical ideas of the past 50 years?*[2], **counterfactual causal inference** is the first subject listed. Concepts of cause-and-effect, counterfactual storytelling, and learning about the world through its manipulation (i.e. actions or interventions) are, as far as we know, extremely intuitive and innate to human-beings. Consider the following examples.

- **Cause and Effect.** When it rains (sufficiently), the ground is wet.
- **Manipulating the world through actions and interventions.** When I drop something, it falls. Yet, if I drop something and hold it by a string, I can prevent it from falling all the way to the ground.
- **Counterfactual storytelling.** What if I hadn't eaten my leftover seafood? I probably wouldn't be sick right now.

[1] It has (almost) nothing to do with Large-Language Models (LLM), or chatGPT, a popular chatbot developed by American tech company OpenAI. But causality is a very active area of reasearch in Artificial Intelligence (AI), considered more broadly as the attempt to understand and emulate human intelligence. Indeed, there is no explicit causal modelling behind chatGPT, and some researchers believe that achieving true Artifical General Intelligence (AGI)–an ill-defined concept–will require some form of causality-based models.

[2] [Gelman and Vehtari, 2021]

These sorts of causality-related utterances seem to be used routinely and with ease by human beings. As a matter of fact, they seem to be indispensable to living as a human being in the world. Children are known to go through phases of extreme playfulness in which it is believed they learn about the world surrounding them through intervening on it and judging the effects of their causal actions.

Yet, for all this ease and simplicity of causality in everyday life, the same cannot be said scientifically and philosophically. It is only very recently that we have come to grips with formal (i.e mathematical) language to discuss this sort of reasoning clearly and rigorously. Computer scientist and mathematician Judea Pearl, a towering figure in the study of causality, puts it like this:

> In the last decade, owing partly to advances in graphical models, causality has undergone a major transformation: from a concept shrouded in mystery into a mathematical object with well-defined semantics and well-founded logic. Paradoxes and controversies have been resolved, slippery concepts have been explicated, and practical problems relying on causal information that long were regarded as either metaphysical or unmanageable can now be solve using elementary mathematics. Put simply, causality has been mathematized. [Pearl, 2009, p.xiii]

This is the fascinating causal inference revolution I wish to present in this book!

1.2 The Battle for The Soul of Causal Inference

To make things even more intellectually captivating, there is a full-scale battle being waged for the soul of causal inference. On one hand, there is the potential outcomes framework which was revitalized and developed through the work of statistician Donald Rubin beginning in the 1970s (sometimes called the Neyman-Rubin potential outcomes frameork, or the Rubin Causal Model (RCM)), which we will thoroughly discuss in Chapter 2. On the other hand, there is the theory of Structural Causal Models (SCM), a very visual graphical approach to causality, most strongly associated with Pearl. A major tool of this approach is the Directed Acyclic Graph (DAG). He developed this work through his long and productive career beginning also in the 1970s, for which he was even awarded the Turing prize

> For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning[3].

We will learn about this "calculus for probabilistic and causal reasoning", the so-called *do*-calculus, in Chapter 3.

For all these fascinating advances in causal inference, Pearl and Rubin disagree. Pearl has been more keen to engage with the ideas from the opposing camp, most notably discussing them at different places in his work[4]. Rubin, on the other hand, seems to want nothing to do with pearlian ideas related to causality. In his monograph on causal inference, [Imbens and Rubin, 2021], a major reference for the book you are about to read, there is exactly one paragraph, totalling around 100 words, devoted to Pearl's approach to causal inference.

> Pearl's work is interesting [...] In our work, [...] we have not found this approach to aid drawing of causal inferences, and we do not discuss it further in this text[5].

Rubin was also interviewed in 2024 to discuss his famous paper on the propensity score, an early landmark in causal inference (i.e. [Rosenbaum and Rubin, 1983]), a subject we tackle in Chapter 5. At a certain point in the conversation, Rubin is expressing his thoughts what he calls a "big idea", by which he more or less means mathematical simplicity and beauty. To him, the (Neyman-)Rubin Causal Model, the subject of Chapter 2, is in competition with the graphical approach of Chapter 3:

> For a while, I think there was this competition. I think the competition for RCM, for this Rubin Causal Model potential outcomes approach, was the graphical stuff. You know, it was the DAG stuff of Judea Pearl. And Pearl's a great character, but I always found just the DAG stuff to be clutter, because it added something that wasn't essential at all. [...] As soon as you get to something with experimental design, for example, split-plot models or things like that, complicated, not just simple randomization, the whole things falls apart. The whole simplicity of DAGs goes away. So why keep pursuing it[6]?

[3] [Russell, 2011]

[4] For example, in [Pearl, 2009], the entry for Rubin in the Name Index list 26 places where he is discussed by Pearl.

[5] [Imbens and Rubin, 2021, p.22]

[6] `https://youtu.be/eF12bVsvq2Q?si=n0LUEB3xoFLL5wWV&t=770`

While there is, in my opinion, an incredible beauty behind the potential outcomes approach, especially when it comes to discussing concepts like experimental design, treatment assignment mechanisms, and randomization, I argue from a purely subjective point of view that the graphical approach of Pearl contains itself unmistakable elegance, especially when trying to uncover **causal effects** from **observational studies** that contain complex relationships between many variables. We will discuss observational studies throughout, but especially in Chapter 4. Of course, these are both subjective opinions.

Despite this competition, so-to-speak, there is a growing number of enthusiastic researchers, in economics, sociology, political science, epidemiology, and more, which have embraced *both* causal inference traditions and have taken a pragmatic approach to choosing the best conceptual tools from each framework to answer their scientific questions with as much rigour and ease as possible. This is not merely a subjective appreciation of mathematical elegance, but rather pragmatic evidence that both approaches are useful, at least to certain camps of empirical researchers.

The overarching goal of this book will be to explore both these frameworks in depth (Chapters 2 and 3), but even more so to unite and compare their ideas when possible. It is my point of view that any user of causal inference methods is confronted with answering scientific research questions, rather than debating foundational issues. We can draw a parallel to the unsettled debates for the foundations of probability theory (Frequentism vs. Bayesianism), which do not bother any empirical researcher, even as they rely consistently on probability theory to do research and advance science[7]. This pragmatic approach to causal inference has led researchers to develop ideas, most notably through the notion of a *counterfactual*, that allow for a partial unification of these frameworks. We will devote a good deal of space to discussing this in Chapter 4.

1.3 Lord's Paradox–Decades of Statistical Debates Put To Bed?

Let's move on to important intellectual contributions that causal inference have brought to the table - solving of statistical paradoxes.

[7] If we believe Linkedin and Twitter, these debates are extremely important, as they garner much engagement and make for easy-to-publish content. I argue that in the real scientific literature, empirical researchers don't care, for better and for worse.

In 1967, Frederic M. Lord, considered by the National Council on Measurement in Education to be "the father of modern testing", published a famous paper titled *A Paradox in the Interpretation of Group Comparisons*[8] in which two hypothetical statisticians using two different, but defensible, analysis strategies come to contradictory conclusions. Lord also discussed the same type of problem in two other hypothetical examples in his 1968 paper *Statistical Adjustments When Comparing Preexisting Groups* [9]. This problem has come to be known as *Lord's Paradox.*

After explaining the nature of the paradox, we will simulate data to get a good grasp of the purported problem we have to face. We will then draw on modern insights to solve the paradox, leveraging causal inference approaches, namely those of Holland & Rubin in *On Lord's Paradox*, from a 1982 technical report [Holland and Rubin, 1982] and from Pearl's *Lord's Paradox Revisited* - (*Oh Lord! Kumbaya!*), published in 2016 [Pearl, 2016].

This example sits, perhaps due to coincidences, at the confluence of both the intellectual battle for the soul of causal inference being waged between the Neyman-Rubin potential outcomes framework[10] and Pearl's approach[11], and as a pivotal example at the heart of statistics history.

A brief note : throughout, we assume familiarity with Analysis-of-Covariance, (ANCOVA)[12].

1.3.1 Lord's Paradox In Brief

Lord's example is concerned with comparing preexisting groups (e.g. biological sex). He acknowledges that

> it is widely recognized that ideally the research worker should assign cases or individuals at random to the groups that are to be studied ... In behavioral research and in many other areas, such random assignment is usually difficult or impossible [13].

Precisely, the example is concerned with assessing

[8] [Lord, 1967]

[9] [Lord, 1968]

[10] See Chapter 2

[11] See Chapter 3

[12] https://en.wikipedia.org/wiki/Analysis_of_covariance

[13] [Lord, 1967, p.304]

> the effects on the students of the diet provided in the university dining halls and any sex differences in these effects[14].

The available data concerns the weight of each student gathered at the beginning of the school year in September (for which we will use X), and at the end, the following June (for which we will use Y). The sex of each student is recorded in our variable G.

The crux of the matter lies in analyzing either the so-called change scores, $Y - X$ in both sex groups, or rather regressing Y on X in both groups, through an ANCOVA-style approach.

1.3.2 Statistician 1 - Change Score Approach

In Lord's example, the average change for BOTH groups is 0, that is $\bar{Y} - \bar{X} = 0$. Therefore,

> the first statistician concludes that as far as these data are concerned, there is no evidence of any interesting effect of the school diet (or of anything else) on student weight. In particular, there is no evidence of any differential effect on the two sexes[15].

1.3.3 Statistician 2 - ANCOVA-style Approach

Lord states that the second statistician finds that

> the slope of the regression line of final weight [i.e. Y] on initial weight [i.e. X] is essentially the same for the two sexes. [Yet,] the difference between the intercepts is statistically highly significant. The second statistician concludes, as is customary in such cases, that the boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes[16].

Obviously, the statisticians cannot be both right. There cannot be no effect at all of diet on weight, and a significant effect of diet on weight. These statements are plainly contradictory!

[14] [Lord, 1967, p.304]
[15] [Lord, 1967, p.305]
[16] [Lord, 1967, p.305]

1.3.4 Simulation

Let's simulate data reproducing the main features of Lord's example. Here, $G \in \{A, B\}$, A for girls and B for boys, using Lord's terminology[17].

Next, Lord states

> although the weight of individual boys and girls has usually changed during the course of the year, [...] the group of girls considered as a whole has not changed in weight, nor has the group of boys[18].

Drawing the scatterplot of Y and X in Figure 1.1, we have in grey the girls and in black the boys. The large dots represent the mean values for each group, and the diamond represents the overall mean. The means all fall on the dashed $45°$ line, representing equal X and Y values, or no changes in the means from September to July, in both groups. Obviously, the data varies around this mean, as would be typical in a real-life setting[19].

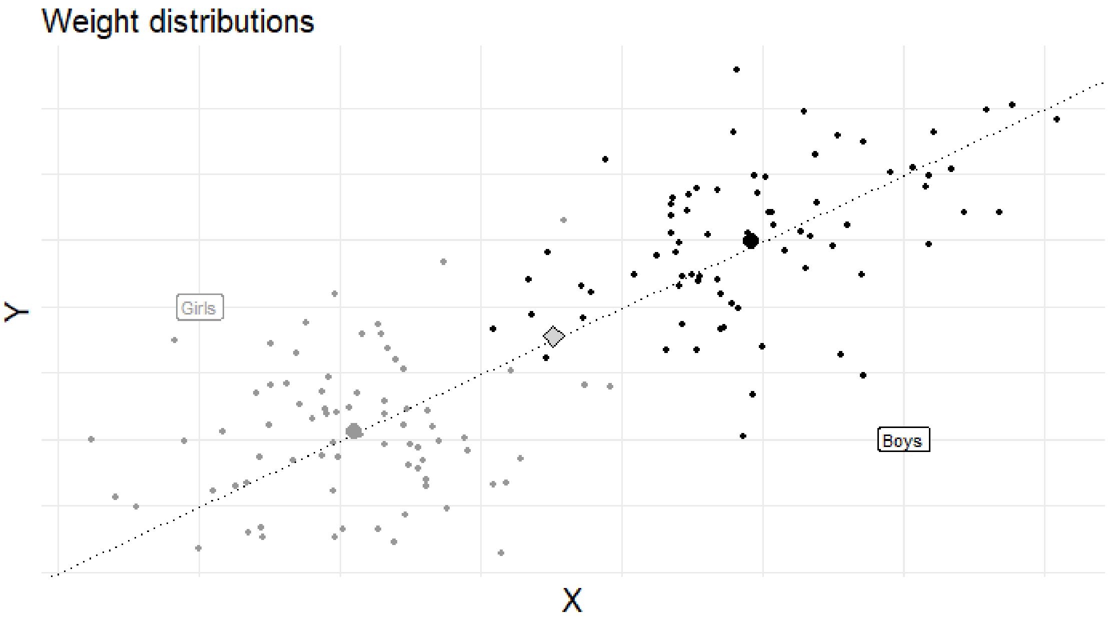

Figure 1.1: The weight in September (Y) in relation to the weight in June (X) for Boys (black) and Girls (grey)–group means are large dots and overall mean is diamond

[17] The code is omitted for clarity, but available for download on Github. As a matter of fact, you can find the code used to generate every single Figure and Table from the book.

[18] [Lord, 1967, p.304]

[19] The change scores would almost never be exactly 0, but omitting this consideration simplifies the analysis without Lord's arguments losing any force.

1.3.5 Duelling Statisticians and Their Competing Analyses

Statistician 1

As stated earlier, the Statistician 1, presumably after plotting the figure above, would compute change scores in both groups and presumably used a paired t-test to discern if the effect is significantly different from 0, as shown in Table 1.1.

group	estimate	statistic	p.value	CI
Girls	0.11	0.16	0.87	[-1.23 ; 1.45]
Boys	0.30	0.49	0.63	[-0.91 ; 1.51]

Table 1.1: Pairwise Two-Sided T-Tests Comparing Weigth in September (Y) with Weigth in June (X), for Girls and Boys

Hurray! The results of the t-test confirm that there are no significant weight differences on average in both groups.

Statistician 2

The second statistician, perhaps feeling a bit more sophisticated, is of the opinion that the proper way to analyze this is through ANCOVA, so she first plots the ANCOVA regression lines in Figure 1.2.

Then, seeing the magnitude of the difference between intercepts, she proceeds to compute the proper estimates to test this group difference. She indeed finds that both the group coefficient (Table 1.2) and the F-test (Table 1.3) on the group parameter in the ANCOVA are highly significant, contradicting statistician 1! [[Difference between beta parameter and F-test]]

term	estimate	std.error	statistic	p.value
(Intercept)	10.85	2.22	4.90	<0.01
X	0.48	0.10	4.57	<0.01
group	7.59	1.69	4.48	<0.01

Table 1.2: ANCOVA Model Coefficients for the Weight of Students in June (Y) regressed on the Weight of Students in September (X) and Group

What gives–how can data like this exist!? As the concrete simulated data shows, this is not simply a theoretical idea[20]. This problem has been at the core of decades of debate and discussions by statisticians, with the most recent breakthroughs coming from developments in

[20] Real-life examples can be found in [Ross, 2004], [Charig et al., 1986], and [Bickel et al., 1975].

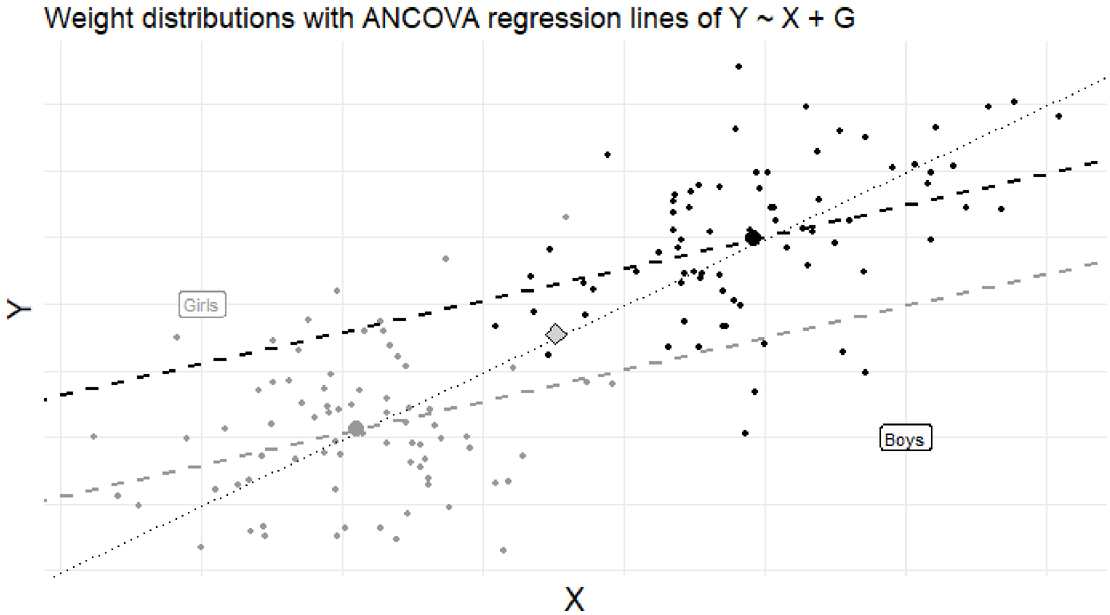

Figure 1.2: The weight in September (Y) in relation to the weight in June (X), with ANCOVA regression lines for both Boys (black) and Girls (grey)

term	sumsq	df	statistic	p.value
X	551.98	1.00	20.85	<0.01
group	531.51	1.00	20.08	<0.01
Residuals	3891.93	147.00		

Table 1.3: ANCOVA Model F-Tests for the Weight of Students in June (Y) regressed on the Weight of Students in September (X) and Group

causal inference. Can we solve this paradox and crown victorious the statistician that emerges unscathed from this statistical duel?

1.3.6 Potential Outcomes Approach

In Holland and Rubin [21], the problem is attacked using the Neyman-Rubin potential outcomes framework, which will be discussed in depth in Chapter 2. Concretely, they define

- Population (P) : The students at the university in the specified school
- Treatment (T) : The dining hall diet
- Control (C) : ??? (No control!)

[21] [[Holland Rubin citation]]

- Treatment indicator (S) : $S = t$ for all units in P
- Gender (G) : Student Gender (A = Girls, B = boys)
- Weight of Student in December (X)
- Weight of Student in June (Y)

The first problem raised is that of the absence of a control group. It is a fundamental idea in experimental science that to compare the effect of a treatment, or an experimental manipulation, there must an alternative called a control treatment, presumably identical in all aspects to the so-called active treatment, but inactive. Indeed, the question "Did a certain diet lead to weight gain?" is ill-posed in this framework. The question can only be rigorously asked by qualifying the statement like such : "Did a certain diet (T) lead to weight gain, when compared to another diet (C)?". This leads to defining two alternative potential outcomes for Y, one under the treatment $Y(T)$ and one under the control $Y(C)$, which we wish to compare. Ideally, we would then randomize a sample of units either to T or C. But Lord never mentioned an alternative diet. Rather, he was explicit that the two groups to be compared were based on G, the gender:

> It is widely recognized that ideally the research worker should assign cases or individuals at random to the groups that are to be studied by analysis of covariance. In behavioral research and in many other areas, such random assignment is usually difficult or impossible[22].

He proceeds to give the example of comparing racial groups, but he clearly places Gender (G) in the same category as that of a variable representing preexisting groups to be compared. In Lord's pessimistic conclusion, he says

> The researcher wants to know how the groups would have compared *if there had been no preexisting uncontrolled differences*[23].

The role of "if" in the previous statement, is what philosophers call "counterfactuals", or "counterfactual language". Counterfactuals describe "alternative worlds" so-to-speak,

[22] [Lord, 1967, p.304]

[23] [Lord, 1967, p.305, emphasis mine]

where things have gone differently than in our world, which is assumed to be real and unique[24].

This counterfactual language, if placed in an experimental framework, would presuppose that either we could randomize students to a given initial weight (X) at the experiment's onset, or that we could assign them to a given gender (G), which would in turn partially determine their initial weight (X). Of course, this is not possible in a true experiment.

Therefore, Rubin and Holland famously quipped "no causation without manipulation", although Holland later softened his position [25]. There is debate around what counterfactuals mean, if anything, due to their abstract, or even metaphysical nature. Indeed, if it's possible to imagine alternative worlds where some events happen differently, why wouldn't it be possible, at least in an abstract sense, or even in a thought experiment, to imagine what would have been different had a participant had a different initial weight (X)? Or even a different gender (G) altogether?

We will discuss Rubin and Holland's paper in more depth once we've put in place the proper notation and formal language required to discuss potential outcomes.

1.3.7 Directed Acyclic Graph (DAG) Approach

In 2016, Pearl published a paper in which he takes the position of having decisively put Lord's Paradox to bed[26]. Essentially, he stands as having solved this longstanding problem, a vantage point from which he is purportedly able to point out the shortcomings and missteps in all previous discussions around the issue. In his own words :

> The purpose of this paper is to trace back Lord's paradox from its original formulation, resolve it using modern tools of causal analysis, explain why it resisted prior attempts at resolution and, finally, address the general methodological issue of whether adjustments for preexisting conditions is justified in group comparison applications[27].

The first step in any Pearlian approach to causal problems is establishing a graphical model representing the causal relationships between the variables under study. We will for-

[24] We will not go in depth in philosophical debates surrounding counterfactuals and causality, but will rather refer the reader to appropriate literature when warranted. We will assume a somewhat naïve view of causality, which we will describe a bit later.

[25] [Holland, 2003]

[26] [Pearl, 2016]

[27] [Pearl, 2016, *Abstract*]

mally denote and define all these concepts later in Chapter 3, but one of the main advantages of DAGs as conceptual tools is their (superficially) simple and intuitive nature.

The first step is to represent all variables under study as nodes in a graph. Then, using unidirectional arrows, causal links between variables can be drawn. Indeed, if changes in A cause changes in B, we draw an arrow from A to B. The real challenge lies in establishing where to place arrows - if we knew all the causal links between variables with certainty, our study would be of less relevance. Still, for sake of exposition, we will assume the DAG Pearl posits as being appropriate for the situation at hand. We draw it in Figure 1.3.

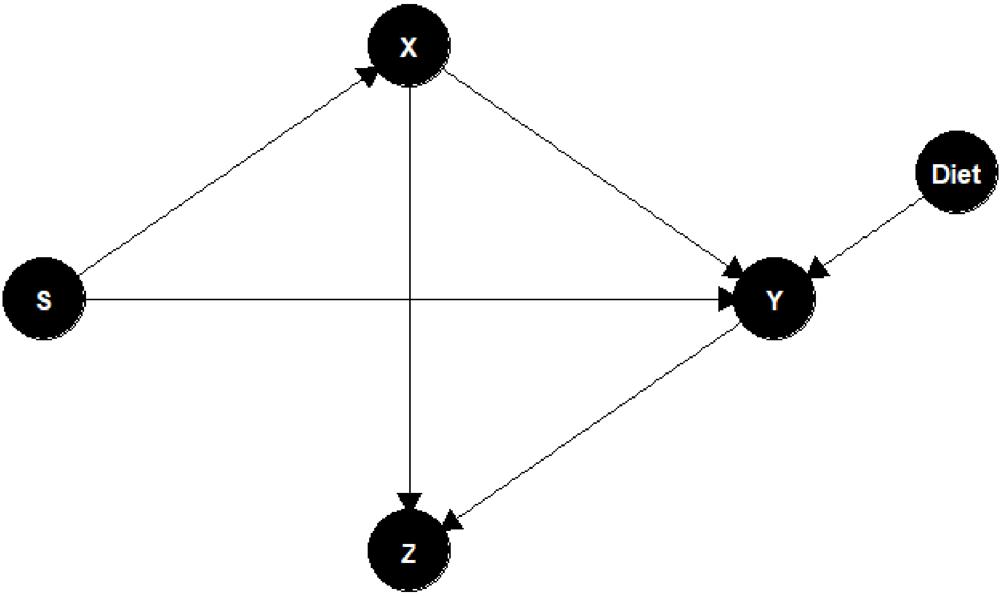

Figure 1.3: A Directed Acyclic Graph (DAG) Of Lord's Paradox

Pearl introduces a new variable, Z which is simply the change score $Z = Y - X$ and points two arrows into Z, one from X and one from Y. This is obvious, since Z is defined precisely equal to the difference of Y and X. Then, he posits that gender is a direct cause of both X and Y, a sensible assumption. Finally, he affirms that the initial weight X is a cause of the final weight Y[28]. His final assumption is that all the relationships in his diagram are **unconfounded**, which for now simply means that the estimates of causal effects are unbiased by

[28] Some readers might already feel uneasy of treating Gender as a cause of a person's weight, yet this is often done in research, based on solid pragmatic arguments. Pearl simply shrugs off the problem by reassuring us that if we wanted, the Gender variable could be replaced by a nexus of hormone variables that distinguish between biological sexes. This opens up a philosophical can-of-worms on the ontological status of social-construct study variables. We will not go down this rabbit-hole for now.

unobserved variables not included in the model. This is obviously a very strong assumption, as there could always be **hidden confounders**[29].

Recall that Pearl is interested in the following counterfactual statement, or at least something close to it : *"What would the subject's weight be in September if their Gender had been different?."* Obviously, this is abstract language but it shouldn't be too concerning as humans routinely make and use these kinds of counterfactual statements to understand and navigate the world[30].

1.3.8 DAGs, SCMs, SEMs and Mediation Analysis

At this stage, a few notes are in order. As Pearl insists in multiple different places throughout his work, a DAG causal model is **non-parametric**, i.e. the postulated causal relationships are not assumed to follow any specific functional form that could be reduced to a handful of parameters. These non-parametric models are sometimes referred to as **Structural Causal Models** (SCM). More informally, by non-parametric we mean that an arrow in a DAG is indeed a function mapping between two variables, but that the function is not necessarily described using a few symbols (i.e. parameters)–the function can be any complex function, even one that doesn't have neat equations to describe it[31]. On the contrary, a **parametric** function is usually defined by a small collection of values, called parameters. In statistics, we often used parametric models to simplify analyses. For example, the function describing the gaussian distribution can be boiled down to two parameters, the mean μ and variance σ^2.

The next step taken to make SCMs more useful, is imposing an additional fully **parametric** model structure. This essentially has the effect of restricting the functions linking variables together to have predefined forms which can then be estimated by estimating a handle of

[29] In this book, we will discuss what confounding means for causal estimation and how this phenomenon is formally represented in both the potential outcomes approach and the DAG approach to causal inference. We will get back to the notion of confounding especially in Chapters 2 and 3 but it will be discussed throughout. It is a central idea in modern science that has been defined and re-defined many times. In my opinion, this notion only makes sense when viewed through the lens of causal models, which is why it has had a tortuous existence.

[30] [Shpitser, 2012] says :

> Human beings reach an intuitive consensus on the meaning of many causal utterances and there have been numerous attempts to formalize causality in a way that is faithful to this consensus.

See also the references cited therein.

[31] A reminder that a function is a concept that tells us how to transform, or map, an input value to an output value. For example, $y = f(x)$ means that given input value x, the function f returns the output value y. A function that is not defined by an equation is sometimes said to have no *analytical form*. This idea is abstract, but an easier way to grasp the idea is to think of such a function as a huge lookup table. Given an input, say x again, we know how to find the output, here y, by looking in the table and finding x. Yet, this table has no clear, concise equation to describe it.

parameters. This tool is known as the **Structural Equation Model (SEM)** and is widely used in many areas of research, most notably in social sciences and biology[32].

Conceptually, the easiest fully parametric model to work with is one where we impose functions to be linear, in addition to giving every variable it's own independent gaussian error term. Linear, gaussian SEMs enjoy the property of being simply multivariate gaussian distributions with a given correlational structure. This essentially conceptualizes a DAG as a series of interwoven linear models. The mathematical simplicity afforded by this (strong) set of assumptions even allows the analyst to work with **latent variables**, that is, unobserved variables that we know are part of the causal story we wish to unravel. We will discuss this in much detail later in Chapter 7. Still, without taking the extra-step of moving from a DAG to and SEM, we can still say interesting things about the relationships between the variables in Lord's paradox, when represented using Pearl's DAG. This is one of the major strengths of the DAG approach–taken at its core, it relies on few assumptions yet still uncovers important truths about the system of variables being studied.

When following arrows in a DAG in their proper direction, we get **directed paths**. On the other hand, **undirected paths**, where we follow arrows without concern for their direction, will also be important when discussing confounding and the **back-door criterion** in Chapter 3. For example, $G \to X \to Y \to Z$ is a directed path from G to the outcome of interest, $Z = Y - X$. We thus say that X is a **mediator** between G and Z; it is affected by Gender and in turn affects the change score in Weight. It is also a mediator between G and Y. Stated in words, we say that the weight in September mediates the relationship between Gender and the outcome (either weight in June or change score in weights).

We also have a path $G \to Y$, which is of interest if we don't look at the change score Z, but rather the final weight in September Y. This path is unmediated, and represents a **direct effect.** It might already becoming clear that DAGs give us a neat conceptual tool to look at effects of variables and how they travel along networks of cause-effect relationships. We will see in Chapter 7 that what was traditionally considered simply as an *effect* of a variable on another, can in fact be looked at more specifically. This is sometimes called "mediation analysis". We will define notions of **total effect**, **controlled direct effect**, **natural direct effect**, and **natural**

[32] Social sciences and Biology encompass so many disciplines, that these alone account for a huge chunk of empirical research being done today.

indirect effect. We will then discuss the extremely important concept of **identification**: given our Structural Causal Model and the data on hand, what kind of effects can be estimated from the data? **Path analysis,** that is estimating effects along paths and using sets of rules for combining them, is a very old technique dating back to [Wright, 1921] that has seen a resurgence in applications. It will also be discussed in Chapter 7. Now, back to Lord's Paradox and its interpretation by Pearl.

Total Effects, Direct Effects, and Indirect Effects

The **total effect** of G on Z is akin to varying G and letting the effect of this variation propagate through the graph and looking how this affects the change score Z. Formally, this amounts to estimating

$$\begin{aligned}\text{Total Effect} &= \mathbb{E}(Z|G=\text{Boy}) - \mathbb{E}(Z|G=\text{Girl})\\ &= \mathbb{E}(Y-X|G=\text{Boy}) - \mathbb{E}(Y-X|G=\text{Girl})\\ &= \underbrace{\mathbb{E}(Y|G=\text{Boy}) - \mathbb{E}(X|G=\text{Boy})}_{\text{average change in weight for boys}} - \underbrace{\mathbb{E}(Y|G=\text{Girl}) - \mathbb{E}(X|G=\text{Girl}}_{\text{average change in weight for girls}},\end{aligned}$$

where the last equation is given by linearity of expectation. In the equation above, the expression we called *average change in weight for boys* can be estimated by taking the boys and computing $\bar{Y} - \bar{X}$. Equivalently, the *average change in weight for girls* can be estimated by doing the same computation using the group of girls[33]. This is of course what Statistician 1 did in Section 1.3.2 when he computed $\bar{Y} - \bar{X}$ in both groups and found them both to be 0.

[33] In this book, we will focus on causality and identification of causal effects of interest. That is, we will be more interested in determining if an estimation procedure can be reasonably assumed to estimate the true causal effect of interest. Obviously, estimating a causal effect using a plausible causal modeling strategy will not absolve us of the traditional duty that statisticians must observe. That is, by adding a layer of causal inference to an analysis, a statistician is not suddenly absolved of all traditional issues that statistics have been developed to tackle : sampling variability, research design, measurement error, inferring from a sample to a population, maximizing use of information in choosing estimators, etc. If anything, causal inference makes the duty of the statistician greater when conducting an analysis.

On the other hand, the **direct effect**[34] is rather

$$\begin{aligned}\text{Direct Effect} &= \int_x \left[\mathbb{E}(Z|G = \text{Boy}, X = x) - \mathbb{E}(Z|G = \text{Girl}, X = x)\right] \mathbb{P}(X = x|G = \text{Girl})\text{dx} \\ &= \int_x \left[\mathbb{E}(Y - X|G = \text{Boy}, X = x) - \mathbb{E}(Y - X|G = \text{Girl}, X = x)\right] \mathbb{P}(X = x|G = \text{Girl})\text{dx} \\ &= \int_x \underbrace{\left[\mathbb{E}(Y|G = \text{Boy}, X = x) - \mathbb{E}(Y|G = \text{Girl}, X = x)\right]}_{(*)} \mathbb{P}(X = x|G = \text{Girl})\text{dx}.\end{aligned}$$

To jump from the second line to the third in the above derivation, we simply used the linearity of expectation along with that fact that by conditioning on $X = x$, there is no more randomness involved in X. This yields

$$\mathbb{E}(X|G = \text{Boy}, X = x) = x$$

and

$$\mathbb{E}(X|G = \text{Girl}, X = x) = x,$$

so both of them cancel out.

Recall, Statistician 2 using ANCOVA in Section 1.3.3. Indeed, ANCOVA proceeds by estimating a linear model for the conditional expected value of the final weight Y as such

$$\mathbb{E}(Y|G, X) = \beta_0 + \beta_1 G + \beta_2 X.$$

Therefore, the expression $(*)$ becomes

$$\beta_0 + \beta_2 X - (\beta_0 + \beta_1 + \beta_2 X) = \beta_1, \tag{1.1}$$

if we set the value of G for boys to be 1 and for girls to be 0, as is traditionally done for binary variables in regression models. Putting Equation 1.1 back into the expression for the direct

[34] This expression is given in [Pearl, 2016]. We will also rely heavily on [Pearl, 2014] for Chapters 7 and 8 where we will formally define all these concepts and discuss when we can identify such effects. That is, when the effects can be parsed out from the data we have on hand.

effect, we get

$$\begin{aligned}\text{Direct Effect} &= \int_x \beta_1 \mathbb{P}(X = x|G = \text{Girl})\text{dx} \\ &= \beta_1,\end{aligned}$$

since β_1 does not depend on X and can be moved out of the integral and $\int_x \mathbb{P}(X = x|G = \text{Girl})\text{dx} = 1$, since we are integrating a probability density function. Thus, Statistician 2, by looking at the beta coefficient associated with the Gender variable in an ANCOVA, is estimating the direct effect under the assumption of a linear model for the expected value–which was found to be significantly different from 0. No wonder Statistician 1 and Statistician 2 do not agree, they are not estimating the same thing!

The conceptual clarity provided by the concepts, tools, and language associated with causal inference allows us to move beyond the nebulous so-called paradox illustrated by Lord in his famous paper. When everybody speaks the same language, the paradox dissolves. The idea of dissolving paradoxes is exciting! How about we attack other longstanding so-called paradoxes that have caused much confusion and debate in statistics?

1.4 To Adjust Or Not To Adjust?

It is a basic scientific principle that is repeated ad nauseam among many scientists that proper adjustment, or control, must be incorporated in statistical models. This is usually done by including variables such as Age, Sex, etc. that we believe are related somehow to the causal effect we are trying to estimate. Usually, there is an incentive to include all available variables in the adjustment–the more the better. Yet, without proper causal formalism, this can turn out to be ill-advised and even further bias the effect of interest. We will also learn in Chapter 3 that it is possible to eliminate bias resulting from confounding without necessarily including all confounders! We will also add a simple and intuitive online tool to our toolkit that, given a causal DAG, selects the smallest set of confounders needed to properly control for when estimating causal effects. But first, some paradoxical examples!

1.4.1 Simpson's Paradox

In [Simpson, 1951], popularized a curious and apparently paradoxical aspect of data analysis. Indeed, Simpson presents a data table in which a hypothetical experiment involving a treatment given to male and female patients. The death status of each patient is then classified according to the patient's sex, and whether the patient received the treatment or not. The Table 1.4 is a scaled-up version of the one given in [Simpson, 1951, p.241][35].

Male	Untreated	Treated	Female	Untreated	Treated
Alive	800	1600	Alive	400	2400
Dead	600	1000	Dead	600	3000

Table 1.4: Patients' Survival Rates in Simpson's Original Example

An elementary procedure learned in most Statistics 101 classes is that this sort of 2 × 2 contingency tables involving counts can be used to test the probabilistic independence of the row and column variables through the chi-square test, as in Table 1.5.

Sex	statistic	p.value	parameter	method
Male	7.33	<0.01	1	Pearson's Chi-squared test
Female	6.77	<0.01	1	Pearson's Chi-squared test

Table 1.5: Chi-Square Tests For Simpson's Original Example

The results of the chi-squared tests applied to the male and female tables separately are both highly significant, which strongly suggests that there is a significant positive effect of the treatment on death–for both males and females, the treatment reduces the chance of death.

Yet, if we look at the data on aggregate, as in Table 1.6 the chi-square statistic is perfectly equal to 0 - the data fits *exactly* what we would expect to observe if treatment status and death were completely independent of one another, as we see in Table 1.7.

Aggreagated	Untreated	Treated
Alive	1200	4000
Dead	1200	4000

Table 1.6: Aggregated Patients' Survival Rates in Simpson's Original Example

We could go even further. It is easy to construct similar examples where the aggregate

[35] We scale up the values from Simpson's original table to increase the power of the chi-square tests of independence as to render the observed differences statistically significant, to further reinforce the paradoxical element of this data.

statistic	p.value	parameter	method
0.00	1.00	1	Pearson's Chi-squared test

Table 1.7: Chi-Square Test For Simpson's Original Example (Aggregated)

data does not collapse to a null effect but is rather *reversed* as compared to the sign of the effect in the subgroups. This means, for example, that one could observe that a treatment is positive for both males and females, but negative for the overall population. [Pearl et al., 2016, p.2] eloquently state how disconcerting this paradoxical result should be :

> The data seem to say that if we know the patient's gender–male or female–we can prescribe the drug, but if the gender is unknown we should not! Obviously, that conclusion is ridiculous. If the drug helps men and women, it must help *anyone;* our lack of knowledge of the patient's gender cannot make the drug harmful.

What is the analyst left to do? Should their be an adjustment for sex of should the data be considered in aggregate?

Answering this question satisfactorily cannot be done by statistical arguments alone. The only satisfactory answer relies on **causal inference** concepts, namely understanding what kind of **causal relationship** holds between our Sex, Treatment, and Death variables. More precisely, the adjustment for Sex should only be considered if Sex is a **confounder** of the causal relationship between treatment and death.

The most basic confounding that sex S can have on the relationship between a treatment D and an outcome Y can be represented by the following DAG.

The (undirected) path $X \leftarrow S \rightarrow Y$ is called a **backdoor path.** As we will see in Chapter 3, we can properly identify and estimate causal effects on DAGs by using a criteria known as **d-separation**, that entails **blocking** all backdoor paths. Here, we would block the path, or d-separate the cause from the effect, by conditioning (or adjusting for, or controlling for) sex S when we estimate the association between D and Y. It is a common belief that adjusting for each and every variable is a valid strategy, perhaps after pruning those which don't pass significance tests in our model[36]

[36] We will discuss variable selection in a causal context, where we will look at when and how to include variables that could potentially confound our causal effect of interest. We will also learn how to use the **propensity score** estimates in Chapter 5, a very useful technique when the number of confounders is large, which can cause practical challenges.

In any case, it is strongly advised NOT to use stepwise (forward or backward) selection algorithms. Unfortunately,

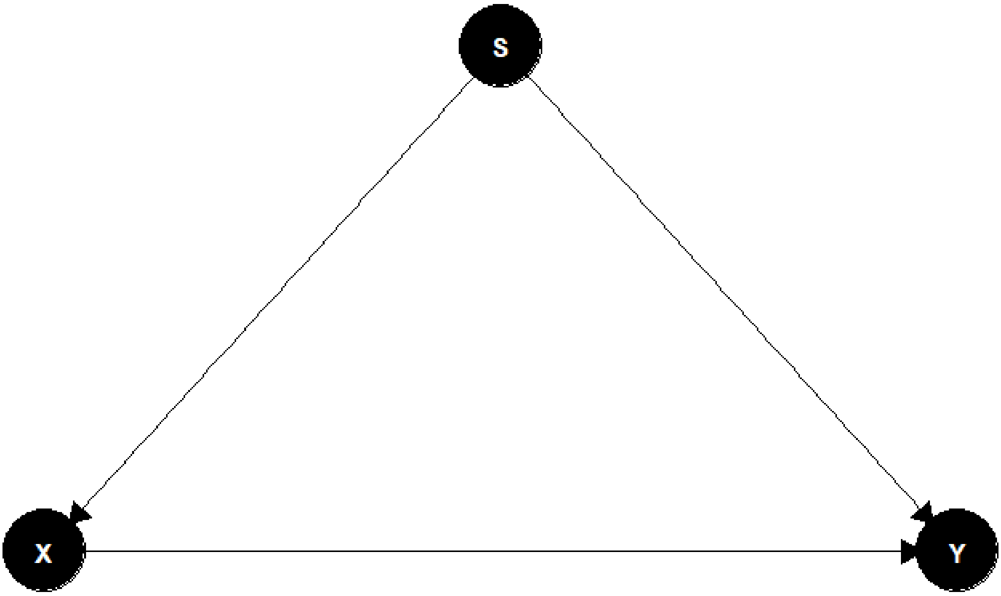

Figure 1.4: A DAG of Basic Confounding of X on Y by S

If the Sex variable turns out to be a **collider** in our causal model, adjusting for it would be catastrophic. Lets consider another example. You guessed it, it is also labelled a paradox.

1.4.2 Berkson's Paradox, Selection Bias, or Collider Effects

In [Berkson, 1946], numerical examples for 2×2 contingency tables are given, much in the same spirit than [Simpson, 1951], as discussed in Section 1.4.1. Berkson was worried that in case-control studies[37], selecting hospitalized patients could lead to a spurious association between exposure and disease (or between disease and disease). Indeed, as he shows, if the disease and exposure of interest causally affect the risk of being hospitalized, selecting

these tools are common, as they can be implemented automatically without much knowledge. This is precisely why they are wrong. I suggest reading Chapter 4, most notably Section 4.3 of [Harrell, 2015] for an excellent discussion of this topic. A free version is available online at [Harrell, 2024]. Harrell also gives workshops related to the book.

[37] A case-control study is an important epidemiological design used in observational studies that is usually learned in Epidemiology 101. As we will see in Section 1.5, Berkson thought this argument weakened the possibility that smoking tobacco had a **causal** effect on development on lung cancer. The paradox was contentious for a while in the research community, due to its peculiar nature. Indeed, the case-control design consists in finding all the diseased, so-called cases, in a certain population, called the study base. To ascertain whether there is an association between the case disease and an exposure of interest, another pool of disease-free units is selected from the study base as controls. Then, the exposure-disease association is measured. This may seem like a hopelessly biased approach, as we oversampled all the diseased from the population–the strategy seems to violate a fundamental intuition of selecting representative samples by, say, using simple random sampling from a well-defined population. Yet, as it turns out, from this design we can estimate odds-ratios without bias. Still, the design is extremely tricky. To my knowledge, there is no better discussion of case-control studies than in [Rothman et al., 2008], a must-read textbook for anyone interested in epidemiology.

hospitalized patients can induce a spurious exposure-disease association. This came to be known as **Berkson's Paradox**, **Berkson's Fallacy** or **Berkson's Bias**. It is a form of **selection bias**, a common term that represents many different forms of bias that result when a certain selection mechanism for units of observation misrepresents the true causal effect of interest.

Consider the following famous example, where we assume that intellectual ability and athletic ability are uncorrelated. The scatter-plot represents a population of units for which we have some measure of both these variables. It is apparent in the data-cloud that the two variables are not associated. Indeed, the regression line has slope 0, and no other patterns are detected.

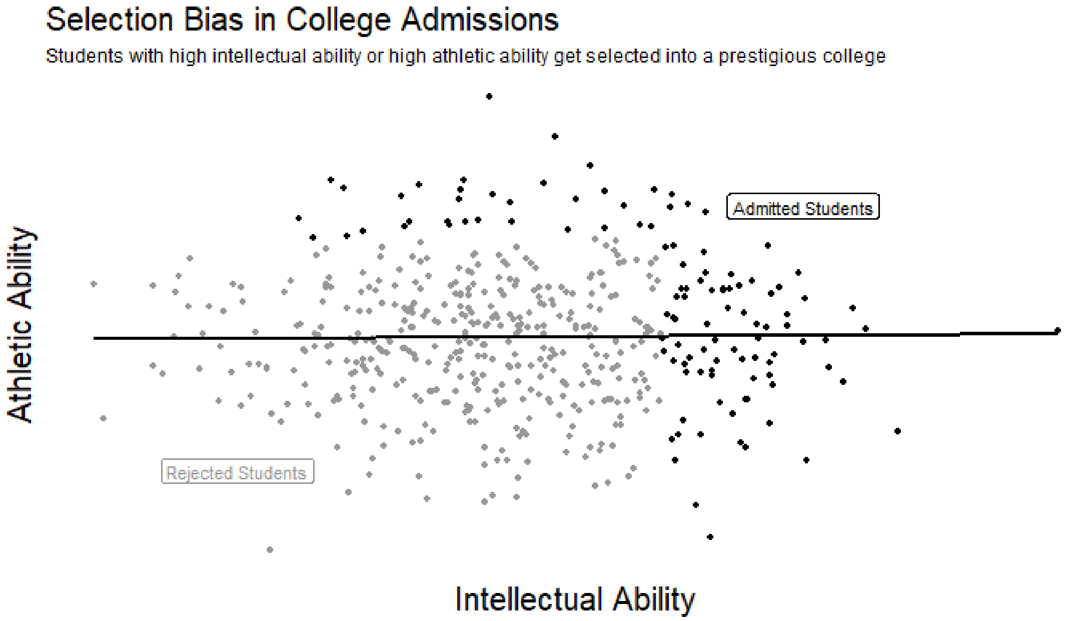

Figure 1.5: Intellectual ability and athletic ability are completely independent in the population, as attested by the flat regression slope

Now, consider we select a portion of the units for study, say graduates from the most prestigious university in the world[38]. Assume further that prestigious universities select their students based on their intellectual ability and/or their athletic ability[39] This can be represented by the following DAG.

This procedure will tend to select students in the upper-right corner of the data cloud :

[38] I will let the reader imagine the most prestigious university without committing to any!

[39] Especially in North America, college sports teams are vital to branding efforts by the universities, and thus there are complex selection processes used to find the best athletes possible.

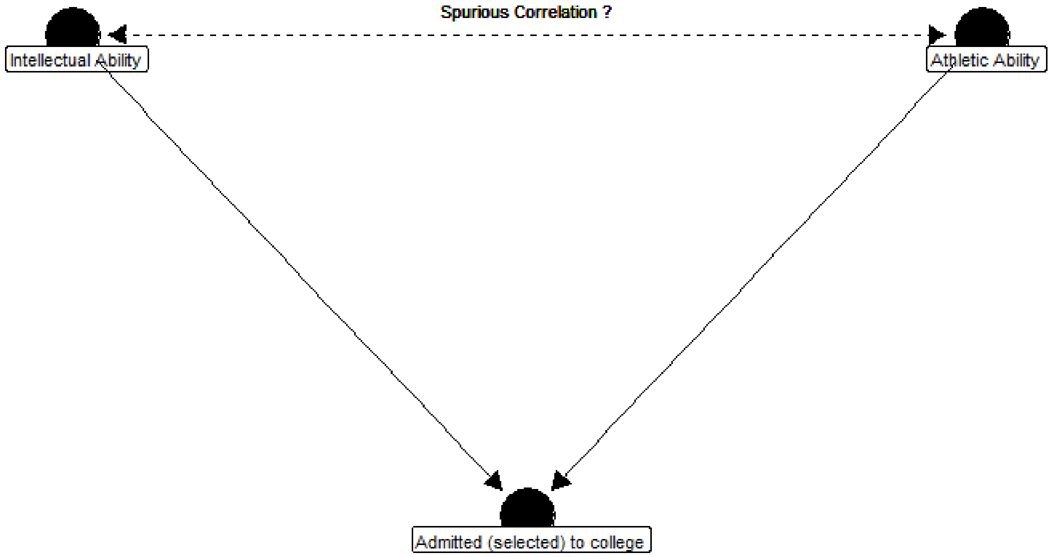

Figure 1.6: A DAG showing collider bias - Controlling or adjusting for college admission status, a collider, can create a spurious correlation between intellectual and athletic abilities

students with high intellectual ability are further on the right of the $X-$axis, and students with high athletic ability are further up on the $Y-$axis.

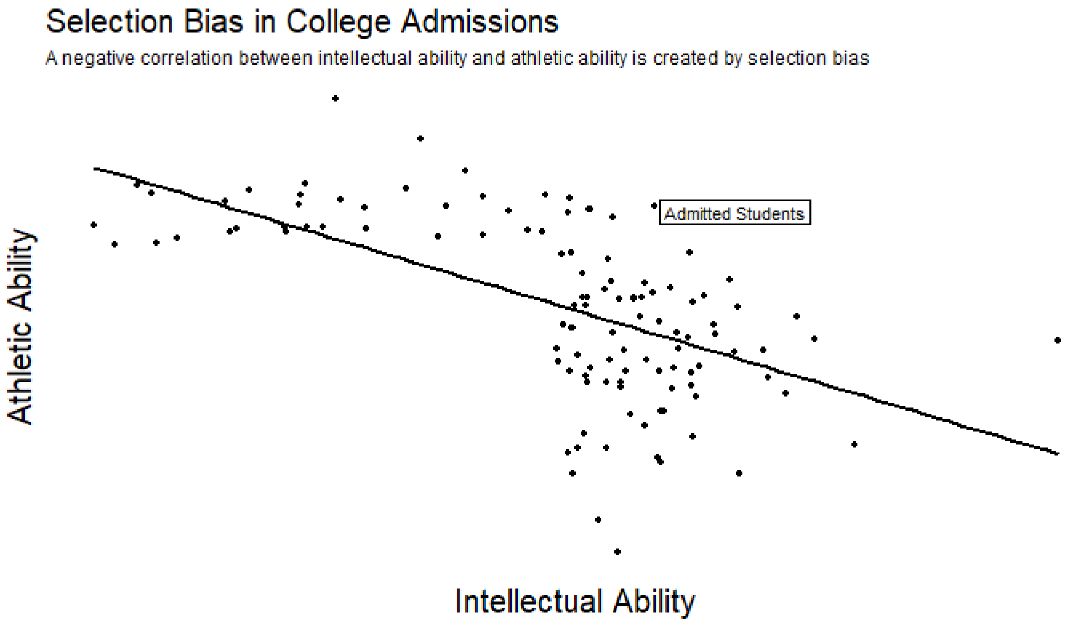

Figure 1.7: When selected into college, a (biased) negative correlation between intellectual ability and athletic ability is created in the selected group

When we isolate these students, we create a spurious negative association between intellectual and athletic abilities. Intuitively, this is simply because a student selected on high intellectual ability will, on average, have lower athletic ability, and vice-versa. Thus, when looking only at the units selected, the regression slope is negative, an artificial association that we've created through selection bias, or conditioning on a **collider**. That is, when two arrows in a DAG point to a common variable, this variable is called a collider–the causes *collide* at a common effect. It might first seem counter-intuitive that conditioning on a variable can create a spurious association between two of its causes, but hopefully the university graduates example gives you a good intuition of why this bias occurs.

We will learn to correctly identify confounders and colliders in DAGs, while taking care to not mix them up and properly address their related issues in Chapter 3. We will also learn about a slew of so-called **selection biases** and represent them with DAGs. As we will see, all forms of selection bias are induced by conditioning on colliders.

1.5 When Genius Errs and Observational Studies Prevail–The Story of Smoking and Lung Cancer

In 1959, Sir Ronald A. Fisher, a British statistician and geneticist widely considered to be one of the most influential scientists of all-time, wrote the following in a Letter to the Editors of Nature, a prestigious scientific journal :

> The *association* observable between the practice of cigarette smoking and the incidence of cancer of the lung [...] has been interpreted [...] almost as though it demonstrated a *causal connection* between these variables[Fisher, 1958d, emphasis mine].

He added in another paper :

> Unfortunately, considerable propaganda is now being developed to convince the public that cigarette smoking is dangerous [...] [Fisher, 1958b]

Yet, in an ironic twist of fate, it has been well established that the propaganda machine was working in exactly the opposite camp. Tobacco companies were found to be running vast and secretive campaigns to influence public opinion in favor of tobacco smoke, despite having themselves conducted much research on the topic, that all pointed towards tobacco smoke being dangerous to human health[40]. Was Fisher a victim of this propaganda?

The overwhelming evidence for the *association* between smoking and lung cancer was not considered controversial to Fisher. Rather, he refused to acknowledge the straightforward causal explanation for this association. He knew very well that once an association between two variables (e.g., X : smoking, and Y : lung cancer) has been established with a high degree of certainty, we find ourselves in one of three positions as pertains to their causal connection:

1. X causes Y;
2. Y causes X, sometimes called *reverse causation*;
3. Common unmeasured variables cause both X and Y, a phenomenon known as *confounding*.

In the quote above, it is clear that Fisher excluded the first possibility listed above. He even suggested the second possibility, that lung cancer could indeed cause smoking habits, a rather odd point of view:

> Is it possible, then, that lung cancer—that is to say, the precancerous condition which must exist and is known to exist for years in those who are going to show overt lung cancer—is one of the causes of smoking cigarettes? I don't think it can be excluded [Fisher, 1958a, p.162][41].

[40] There is now mounting evidence that the *playbook* used by tobacco companies to obscure scientific facts by sewing seeds of doubt in the population is being replayed when a large industry feels threatened by scientific evidence of the harms said industry is causing. The most distressing case comes from climate change deniers and greenwashing, but the same story is found over and over, in debates around the health concerns of asbestos, the impacts of insecticides on bee pollination and biodiversity, and more. See [Oreskes and Conway, 2019] for an excellent discussion on so-called *Merchants of Doubt*.

[41] Fisher goes on to further justify this point of view with curious anecdotal evidence, if we can even call it that :

> I don't think we know enough to say that it is such a cause. But the pre-cancerous condition is one involving a certain amount of slight chronic inflammation. The causes of smoking cigarettes may be studied among your friends, to some extent, and I think you will agree that a slight cause of irritation—a slight disappointment, an unexpected delay, some sort of a mild rebuff, a frustration—are commonly accompanied by pulling out a cigarette and getting a little compensation for life's minor ills in that way. And so, anyone suffering from a chronic inflammation in part of the body (something that does not give rise to conscious pain) is not unlikely to be associated with smoking more frequently, or smoking rather than not smoking. It is the kind of comfort that might be a real solace to anyone in the fifteen approaching lung cancer. And to take the poor chap's cigarettes away from him would be rather like taking away his white stick from a blind man. It would make an already unhappy person a little more unhappy than he need be.

Of course, he had no data to back this claim up, as it is quite ludicrous. He did in fact provide data to back up his suggestion of the third explanation, that is that there are unmeasured confounders that could explain the causal effect, specifically, genetic factors. We illustrate his argument in the following very simple DAG.

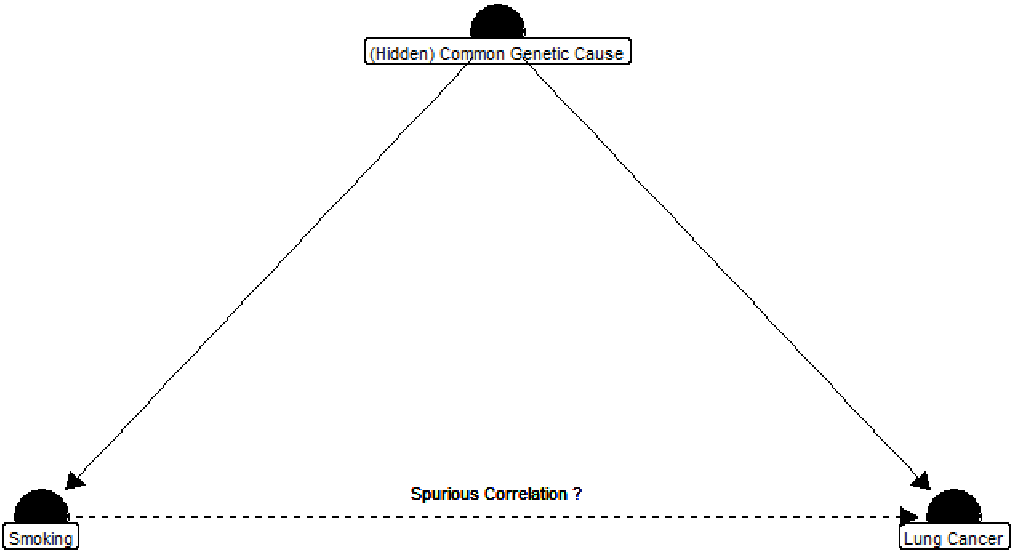

Figure 1.8: A DAG showing Fisher's argument for a hidden confounder - According to Fisher, a common cause for both tendency to smoke and lung cancer could explain the observed association between the two

The gist of the argument is that there could be genetic factors that influenced both smoking habits and the probability of developing lung cancer, creating a confounding bias in the estimates of the effect. While it was less controversial to assume that genetic factors were a direct cause of cancer, the argument for the cause-and-effect relationship between the same genetic factor and the proclivity to become a tobacco smoker were less evident. His argument was based on data that showed that in a few dozen pairs of monozygotic twins, their smoking behaviour was much more likely to be identical than in dizygotic twins, [Fisher, 1958c][42]. Let's look again at Fisher's argument using our DAG, this time highlighting the fact of the uncertainty of genetic factors directly causing tendency to smoke.

[42] A simplistic version of his argument says that people with the same genes are more likely to have similar behaviors than people with different genes. The use of twins helps to remove confounding from other factors, namely that people with different genes are also more likely to live in different environments, which in turn impacts their behavior.

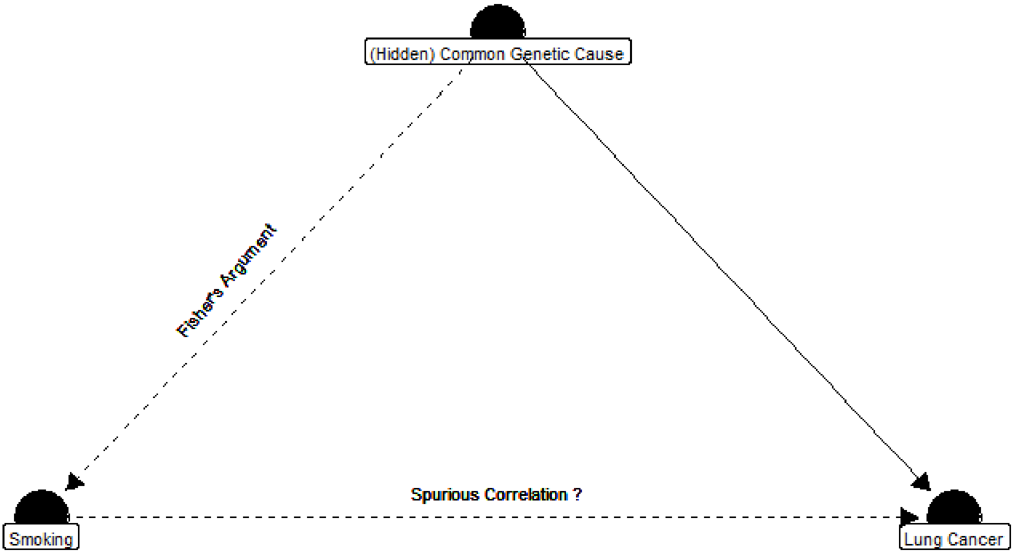

Figure 1.9: A DAG showing Fisher's argument for a hidden confounder - with dashed lines representing unconclusive evidence

In a very limited sense, Fisher was right, as his argument was based on the absence of randomization. By randomizing X, we automatically exclude the second and third possibilities listed above. Why? Because, by definition, the *only* cause of X is the randomization scheme! Fisher understood this better than anybody, since he was the first to truly develop the theory of the randomized experiment. Yet, if we believed Fisher had the last word in causal inference, many scientific problems would forever remain insoluble. The causal effect of cigarette smoking on lung cancer is a case in point: it is both unethical and unfeasible to conduct a randomized experiment to settle this matter. Fisher said it himself :

> [The proponents of the causal effect of cigarette smoking on lung cancer] cannot produce evidence in which a thousand children of teen age have been laid under a ban that they shall never smoke, and a thousand more chosen at random from the same age group have been under compulsion to smoke at least thirty cigarettes a day. If that type of experiment could be done, there would be no difficulty, [Fisher, 1958a, p.155].

There are other cases where we can never uncover causal mechanisms through randomization. Take as established that smoking causes lung cancer, which we will symbolize by

$X \to Y$. A more complete description would involve a nexus of physiological processes that we choose to leave unexplained, as a black-box. We could write this $X \to \boxed{B} \to Y$, where $\boxed{B}$ denotes the unexplained black-box phenomena by which physiological mechanisms turn cigarette smoking into lung cancer. The issue is that many sciences focus specifically on these sort of mechanisms, sometimes at cellular levels. Obviously, there are many cases where it is impossible to conduct randomized studies at this level of detail. What are we to do?

At the time of publication of these criticisms, Fisher was closing in on 70 years of age, and his best scientific days were behind him. It took a new generation of researchers to establish the idea that even without randomization, it is possible to make causal claims. Aren't we glad they did! Among other factors, these debates spurred researchers to discover *when and why observational evidence could be trusted.* The use of observational data to discover causal effects is often what is meant by *causal inference*, as experiments were widely considered the only way to detect causality.

Today, it is one of the most well-established scientific facts in Public Health and Epidemiology that cigarette is not only harmful to health, but it is one of the worst possible behaviors one can have in terms of impact on health outcomes. It not only *causes* lung cancer, but also a great deal of deadly health conditions[43]. Of course, there were also many prominent scientists, such as Sir Austin Bradford Hill, who held the position that the *association* found over-and-over again between tobacco smoke and lung cancer was indeed *causal.*

How did we go from a debate about hidden confounders to having scientifically established a clear scientific consensus on the *causal effect* of smoking on lung cancer through strictly observational means? A major step was the introduction of **sensitivity analysis**, a framework which allows to measure the amount of bias hidden confounding must bring in order to eliminate the effects measured[44].

[43] See [Doll, 1998], who claims that among nearly 40 diseases or causes of death which are positively associated with cigarette smoke, a great majority have been found to be causal. The paper also serves as a fascinating overview of the history of changing social and political attitudes towards the mounting scientific evidence of the harms caused by tobacco.

[44] Paul Rosenbaum, a student and collaborator of Rubin who has made tremendous contributions to causal inference, says in [Rosenbaum, 2023, p.64]

> There are always covariates that were not measured. Always.

1.6 Summary

I hope this Chapter has given you a nice overview of what we're about to learn together! We end with some exercises. This is the only Chapter where we forego a case-study - expect one at the end of every other Chapter.

Good luck!

1.7 Exercises

1. Go to Github and set up a free account. Star the repository (*repo*) associated with the book, if you like it!

2. For each of the following scenarios, identify the cause and the effect. Explain your reasoning.

   (a) A farmer applies fertilizer to a crop and observes an increase in yield.

   (b) A child plays outside in the rain without a coat and catches a cold.

   (c) A new law is enacted that increases the minimum wage, and unemployment rates decrease.

3. Explain why the causal conclusions stated below are wrong.

   (a) On a daily basis, there is a strong positive correlation between ice cream sales and temperature. Therefore, to help with global warming, we should prohibit ice cream sales.

   (b) There is a strong association between money spent on lawyer fees and probability of divorce. Therefore, to avoid divorce, couples should be mindful of their lawyer expenses and try to keep them at minimum.

4. For each situation, write a counterfactual statement that explores an alternative outcome.

   (a) "I studied adequately more for the exam, and I have passed."

(b) “Since the city has built the new park, many families have been moving into the area.”

(c) “I took the early bus at arrived on time for my meeting.”

5. Describe a real-world scenario (not related to the examples in the chapter) where Berkson’s Paradox might occur. Explain how the selection of a specific population could create a spurious association between two variables.

6. Select a recent study or news article that discusses a cause-and-effect relationship (e.g., the impact of a public health intervention). Analyze the following:

   - Identify the causal relationship being studied.
   - Discuss the potential outcomes framework as it applies to this study.
   - Suggest how a DAG could visually represent the causal relationships in this scenario.

7. For each scenario below, identify potential confounding variables and discuss how they might affect the causal interpretation:

   - A study finds that students who participate in after-school tutoring perform better on standardized tests.
   - A new traffic law is implemented, and accidents decrease at intersections.
   - A company notices that employees who attend wellness programs report higher job satisfaction.

8. What key lessons did you learn about the importance of understanding the causal relationships between variables in statistical analysis?

9. How do the concepts of confounding and selection bias impact the interpretation of observational data in public health research?

Chapter2

Fundamentals of Potential Outcomes Theory

2.1 Historical Context, Development, and Motivation

It is now well-known and almost universally applied in practice that experiments with randomization of treatment assignment can help uncover causality in empirical science, industrial research and development (R&D), data science, engineering, pharmaceutical research and drug development, online experimentation (e.g. A/B testing), and more. Scientists have always been experimenting, but it wasn't until relatively recently that the role of randomization was properly understood–a role we will cut and dissect in many different ways, as it is truly one of the greatest achievements of humanity's scientific enterprise. [Fisher, 1925] first made the explicit case for randomization for uncovering causality, although use of chance mechanisms by scientists can be found earlier in historical records[1]. Most of the traditional textbooks on design of experiments focused on industrial applications in manufacturing settings and quality control, for example [Box et al., 2005], a masterpiece we will discuss in this Chapter. With the advent of digital technology, there is now a large and growing demand for expertise in experimentation and analysis of online randomized experiments that is now added to the demand that already existed in more traditional sectors (i.e. pharmaceuticals, manufacturing, engineering, etc.). These experiments are done in marketing, web development, application optimization, etc. and are often called **A/B tests**. According to

[1] See for example [Peirce and Jastrow, 1884].