

Causal Inference: General Introduction

Ashley I Naimi

January 2024

Contents

1	Introduction	2
2	Correlation and Causation	2
3	Inference: Statistical and Machine Learning	3
4	Inference: Causal	4
5	Potential Outcomes Notation	5
6	Fundamental Problem of Causal Inference	6

1 Introduction

2 Correlation and Causation

In the *The Grammar of Science*, Karl [Pearson \(1911\)](#) wrote: “[b]eyond such discarded fundamentals as ‘matter’ and ‘force’ lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect.” He suggested that rather than pursue an understanding of cause-effect relations, scientists would be best served by measuring correlations through tables that classify individuals into specific categories. “Such a table is termed a contingency table, and the ultimate scientific statement of description of the relation between two things can always be thrown back upon such a contingency table.”

Over a century later, a majority of statistics courses tend to treat causal inference by simply stating that “correlation is not causation.” This treatment is hardly sufficient, for at least two reasons: 1) As scientists, our primary interest is (should be) in cause-effect relations; 2) People continue to conflate correlation with causation¹. For both of these reasons, we very much need to **clarify the conditions that would allow us to understand causality better**. This is what “causal inference” is all about.

Generally, I adopt the view that **the causal and statistical aspects of a scientific study should be kept as separate as possible**. The objective is to first define the effect and articulate the conditions under which causal inference is possible for this effect, and then to understand what statistical tools will enable us to answer the causal question.² Causal inference tells us what we should estimate, and whether we can. Statistics tells us how to estimate it. By implication, we should avoid the commonplace practice of treating statistical models as if they were causal.³ For example, the practice of reading the risk ratio, odds ratio, or risk difference for an exposure of interest from a generalized linear (statistical) model⁴ will sometimes work under very specific conditions, but is not the best approach for quantifying exposure effects ([Naimi and Whitcomb, 2020](#)).

¹ Daniel Westreich and I reviewed a book whose authors were so caught up in the allure of “Big Data”, they thoroughly forgot that correlation \neq causation. See [Naimi and Westreich \(2014\)](#)

² Loosely speaking: Causal inference is the “what?” Statistics is the “how?”

³ See the section on Inference below

⁴ or the hazard ratio from a Cox model, or the mean ratio from a Poisson model, or host of other types of regression models

3 Inference: Statistical and Machine Learning

As scientists and researchers, we encounter the word “inference” quite frequently. Yet it is often used in different ways, and sometimes used to convey different things. This is particularly true if we contrast “inference” in, say, the machine learning literature, to “inference” in statistics. It’s hard to see, but practitioners in machine learning disciplines use the word inference in a very different way than those in statistics (Breiman, 2001).

The key to understanding the difference between what we mean by “inference” in different disciplines is provided in a paper by Galit Shmueli (?). In this paper, she distinguishes between two fundamental actions in science: explanation and prediction.

In simple terms, explanation is the act of building a theoretical model of some aspect of the world we are studying. This model becomes our best representation of how the world works. This is typically a **causal exercise**, in that data are used to understand how variables are causally related to each other.

In contrast, prediction does not involve building theoretical models, or understanding cause-effect relations. The objective is to construct an algorithm that can be used to find any kind of relationship between variables that can be exploited to predict a dependent variable with independent variables.

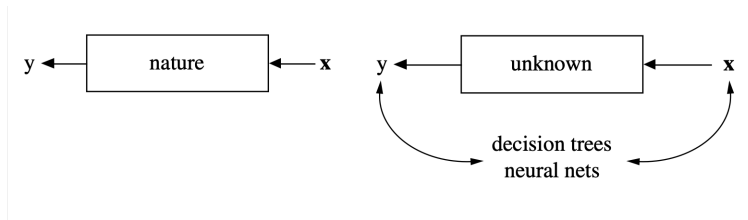


Figure 1: Demonstration of Leo Breiman’s ‘Data Modeling’ and ‘Algorithmic Modeling’ perspectives. In the Data Modeling approach, scientists use data and statistics to build theoretical models of ‘nature’, which provides insight about the world. In the Algorithmic Modeling approach, scientists are not particularly interested in nature, but more so in constructing an algorithm that enables us to use x to predict y .

In algorithmic modeling areas such as machine learning, “inference” is usually meant to connote a prediction from the algorithm on an out-of-sample observation. Consider, for example, the footnote in Chapter 4 (page 103) of Murphy (2022): “In the deep learning community, the term ‘inference’ refers to what we will call ‘prediction’, namely computing $p(y \mid x, \hat{\theta})$.” That’s typically the extent of what is meant by “inference” in machine learning settings.

On the other hand, in statistics, “inference” is usually formalized as a measure of the uncertainty that exists between the results that we get in a particu-

lar study, and the underlying model of the world (nature in Figure 1). Statistical inference allows us to provide an answer to the question: how confident should we be that our data support the model?⁵ There are several tools available for us to do this, including Frequentist tools focused on error control, and Bayesian tools focused on updating beliefs based on new evidence. However, both share a common goal which is to use data to attempt to quantify how wrong we might be about a statement about how the world works.

⁵ There are a lot of subtleties here that we do not have time to discuss. These subtleties have a longstanding history, originating in some of the earliest works in probability theory and statistics (e.g., Jacob Bernoulli's *Ars Conjectandi*, or the *Art of Conjecturing*). I refer the interested reader to the recent book by Clayton (2021), or chapter 4 of the book by Diaconis and Skyrms (2019).

4 Inference: Causal

"Causal inference" is not merely about error control or updating beliefs, but rather deals with the **formal mechanisms by which we can combine data, assumptions, and models to interpret correlations (or associations) causally**.⁶ Thus, causal inference shares similar objectives to statistical inference when "explanation" (in Galit Shmueli's sense) is the goal. However, statistical inference is usually insufficiently suited to the task. Causal inference allows us to fill in the gaps of understanding when we use statistics to explain causal effects.

⁶ There are a number of introductory books and articles on causal inference in the empirical sciences. Here are some excellent options: Hernán and Robins (2020), Pearl et al. (2016), Imbens and Rubin (2015), Cunningham (2021).

The framework in which we define what we mean by "causal relation" or "causal effect" is the **potential outcomes framework**. A central notion in the potential outcomes framework is the counterfactual. This notion stems from the intuitive and informal practice of interpreting cause-effect relations as **circumstances (e.g., health outcomes) that would have arisen had things (e.g., exposures) been different**.

While this intuition serves an important purpose, it is not sufficient for doing rigorous science. Suppose we ask: "what is the effect of smoking on the 5-year risk of CVD, irrespective of smoking's effect on body weight?" This question may seem clear and intuitive. To answer this question, we would do a study in which we collect data, enter these into a computer, perform some calculations, and obtain a number (we'd usually like to interpret as the "effect").

But there is a problem.⁷ The calculations performed by the computer are **rigorously defined (i.e., unambiguous) mathematical objects**. On the other hand, **English language sentences about cause effect relations are ambiguous**. For example, the "effect of smoking" can mean many different things:

⁷ This problem was articulated by Robins 1987, and I am using a version of the example from his paper.

- All people smoke any tobacco ever versus no people smoke tobacco ever.

- All people smoke 3 cigarettes per day versus all people smoke 2 cigarettes per day.
- All people who have smoked any tobacco in the last 15 years cease to smoke any tobacco whatsoever.

Similarly, “irrespective of” can mean a number of things:

- The effect of smoking on CVD risk that would be observed in a hypothetical world where smoking did not affect body mass?
- The effect of smoking on CVD risk if everyone were set to “normal” body mass?
- The effect of smoking on CVD risk if everyone were held at the body mass they had in the month prior to study entry?

But the numerical strings of data and the computer algorithms applied to these data are well defined mathematical objects, which do not admit such ambiguity. Depending on several choices, including the data collected, how variables are coded, the modeling strategy, and other aspects of the project, the computer is being told which question to answer. There is a lot of potential uncertainty in the space between the English language sentences we use to ask causal questions, and the computer algorithms we use to answer those questions. Causal inference is about clarifying this uncertainty.

5 Potential Outcomes Notation

The building blocks for causal inference are **potential outcomes** (Rubin, 2005).

Importantly, these are conceptually distinct from **observed outcomes**. That is, the outcome that one might observe in a dataset is not the same as the potential outcome.

Potential outcomes are functions of exposures. For a given exposure x , we will write the potential outcome as Y^x .⁸ **This is interpreted as “the outcome (Y) that would be observed if X were set to some value x ”.** For example, if X is binary [denoted $X \in \{0, 1\}$], then Y^x is the outcome that would be observed if $X = 0$ or $X = 1$. If we wanted to be specific about the value of x , we could write $Y^{x=0}$ or $Y^{x=1}$ (or, more succinctly, Y^0 or Y^1).

Similarly, when the exposure and/or outcome are measured repeatedly over follow-up, notation must account for this. We thus use subscripts to denote

⁸ Alternate notation includes: Y_x , $Y(x)$, $Y \mid \text{Set}(X = x)$, and $Y \mid \text{do}(X = x)$.

when the variable was measured. For example, if the exposure is measured three times, we can denote the first measurement X_0 , the second X_1 , and the third X_2 . Additionally, we use **overbars** to denote the past history of a variable over follow-up time. For example, for the three time-point scenario, \overline{X}_1 denotes the set $\{X_0, X_1\}$ (i.e., the exposure values for the first time-point, the second time-point, but NOT the third time-point.) Finally, we sometimes need to index future exposures, which we do using **underbars**. For example, \underline{X}_1 denotes the set $\{X_1, X_2\}$ (i.e., the exposure values for the second time-point and the third time-point, but NOT the first time-point.)

More generally, for some arbitrary point over follow-up j , \overline{X}_j denotes $\{X_0, X_1, X_2, \dots, X_j\}$.⁹ In contrast, \underline{X}_j denotes $\{X_j, X_{j+1}, X_{j+2}, \dots, X_J\}$. Note here that we use capital “ J ” to denote the last follow-up time in a study.

We can then define potential outcomes as a function of these exposure histories and/or futures. For example:

$$\begin{aligned} Y^{\overline{x}_j=1} &= Y^{x_0=1, x_1=1, \dots, x_j=1} \\ Y^{\underline{x}_j=1} &= Y^{x_j=1, x_{j+1}=1, \dots, x_J=1} \\ Y^{\overline{x}_{j-1}=1, \underline{x}_j=0} &= Y^{x_0=1, x_1=1, \dots, x_{j-1}=1, x_j=0, x_{j+1}=0, \dots, x_J=0} \end{aligned}$$

⁹ Note that this notation presumes a discrete time framework, where study time is broken up into distinct chunks that can be indexed by 0, 1, There are some important distinctions in the technical / theoretical literature between discrete versus continuous time analyses, particularly in the survival setting. We will not consi

6 Fundamental Problem of Causal Inference

The main task of causal inference is to quantify effects, which are usually defined as contrasts of potential outcomes. For example, for a binary exposure $X \in [0, 1]$ and binary or continuous outcome Y , the commonly targeted average treatment effect is defined as:

$$\psi = E(Y^1 - Y^0)$$

One of the key challenges behind causal inference is that, for a given individual, we can only observe their outcomes under one particular exposure setting. An individual cannot be simultaneously exposed and unexposed, and we can thus never take the difference between Y^1 and Y^0 for the same person.

This is the **fundamental problem of causal inference**, and the only way we can bypass this fundamental problem is by making (identifiability) assump-

tions.

Often, the causal inference literature suggests that this problem is easier to solve than it actually is. For example, in chapter 4 of his (really excellent) book, *Causal Inference: The Mixtape*, Scott Cunningham provides a table of made up data for eight people with missing counterfactuals:

Name	X	Y	Y^1	Y^0
Andy	1	10	10	.
Ben	1	5	5	.
Chad	1	16	16	.
Daniel	1	3	3	.
Edith	0	5	.	5
Frank	0	7	.	7
George	0	8	.	8
Hank	0	10	.	10

he notes that “If you look closely at Table 15, you will see that for each unit, we only observe one potential outcome.” This formulation is common in the causal inference literature (Brumback, 2021, p 38; de Mesquita and Fowler, 2021, p 41; Diggle et al., 2002, p 270). The fundamental problem then becomes one of combining assumptions with these data to recover the (single) missing potential outcome for each individual, which then allows us to compute the exposure effects of interest (e.g., ψ).

However, we noted above that potential outcomes are distinct from observed outcomes: the outcome that one might observe in a dataset is not the same as the potential outcome. But when the literature states that we can “only observe one potential outcome,” this implies that the data we collect on individuals can, at least partially, provide us with direct access to the potential outcomes. By contrast, I take the position that **we can never observe any potential outcomes**. We are only ever able to obtain information on potential outcomes by way of unverifiable assumptions. How can we reconcile these contradictory positions?

As we will see in the next section, formulations that argue that we can, indeed, observe only one potential outcome are already making assumptions about their system under study.

References

- Leo Breiman. Statistical modeling: The two cultures. *Stat Sci*, 16(3):199–215, 2001.
- B. A. Brumback. *Fundamentals of Causal Inference: With R*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2021.
- A. Clayton. *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. Columbia University Press, 2021.
- Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, New Haven, CT, 2021.
- E. B. de Mesquita and A. Fowler. *Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis*. Princeton University Press, 2021.
- P. Diaconis and B. Skyrms. *Ten Great Ideas about Chance*. Princeton University Press, 2019.
- Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L. Zeger. *Analysis of longitudinal data*. Oxford statistical science series ; 25. Oxford University Press, Oxford ;New York, 2002.
- M. A. Hernán and JM Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, 2020.
- Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, 2015.
- K. P. Murphy. *Probabilistic Machine Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 2022.
- Ashley I. Naimi and Daniel J. Westreich. Big data: A revolution that will transform how we live, work, and think. *American Journal of Epidemiology*, 179(9): 1143–1144, 2014.
- Ashley I Naimi and Brian W Whitcomb. Estimating risk ratios and risk differences using regression. *American Journal of Epidemiology*, 189(6):508–510, 2020.

Judea Pearl, Madelyn R Glymour, and Nicholas Jewell. *Causal Inference in Statistics: A Primer*. Wiley, United Kingdom, 2016.

Karl Pearson. *The Grammar of Science*. London, J.M. Dent & sons Ltd, 3rd edition, 1911.

Donald B Rubin. Causal inference using potential outcomes. *J Am Stat Assoc*, 100(469):322–331, 2005.