An Introduction to Causal Inference

Fabian Dablander[1]

[1] Department of Psychological Methods, University of Amsterdam

Causal inference goes beyond prediction by modeling the outcome of interventions and formalizing counterfactual reasoning. Instead of restricting causal conclusions to experiments, causal inference explicates the conditions under which it is possible to draw causal conclusions even from observational data. In this paper, I provide a concise introduction to the graphical approach to causal inference, which uses Directed Acyclic Graphs (DAGs) to visualize, and Structural Causal Models (SCMs) to relate probabilistic and causal relationships. Successively, we climb what Judea Pearl calls the "causal hierarchy" — moving from association to intervention to counterfactuals. I explain how DAGs can help us reason about associations between variables as well as interventions; how the *do*-calculus leads to a satisfactory definition of confounding, thereby clarifying, among other things, Simpson's paradox; and how SCMs enable us to reason about what could have been. Lastly, I discuss a number of challenges in applying causal inference in practice.

Keywords: DAGs, *d*-separation, *do*-calculus, Simpson's paradox, SCMs, counterfactuals
Word count: 8103

Introduction

Although skeptical of induction and causation, David Hume gave a definition of a *cause* that is widely used today (Hume, 1748, p. 115; see also Greenland, 2011):

> "We may define a cause to be an object, followed by another, [...] where, if the first object had not been, the second had never existed."

Karl Pearson, a foundational figure in mathematical statistics, would have none of it; for him, correlation was central to science, causality being merely a special case of correlation. He abhorred the counterfactual element inherent in Hume's definition, yet sought to classify correlations into "genuine" and "spurious" (Aldrich, 1995). While Pearson lacked the formal framework to do this rigorously, modern causal inference provides such a framework. Going beyond Pearson, causal inference takes the counterfactual element in Hume's definition as the key building block; yet it also lays bare its "fundamental problem": the fact that we, per definition, cannot observe counterfactuals. For example, a patient cannot receive and at the same time not receive the treatment. The *potential outcome framework*, formalized for randomized experiments by Neyman (1923/1990) and developed for observational settings by Rubin (1974), defines for all individuals such *potential outcomes*, only some of which are subsequently observed. This framework dominates applications in epidemiology, medical statistics, and economics, stating the conditions under which causal effects can be estimated in rigorous mathematical language (e.g., Rosenbaum & Rubin, 1983; Hernán & Robins, 2020, ch. 3). Another approach to causal inference has an equally long tradition, going back to the path diagrams of Wright (1921). Most of the subsequent developments of this approach to causal inference came from artificial intelligence and are associated with Judea Pearl, who proposed using directed acyclic graphs (DAGs) to depict causal relations (Pearl, 1995, 2009). Here, the fundamental building block from which interventional as well as counterfactual statements follow are Structural Causal Models (SCM); instead of starting from potential outcomes, this approach to causal inference sees them as being derived from SCMs.

In this paper, I focus on the graphical approach to causal inference. Following Pearl (2019b), I distinguish three "levels" of causal inference. At the most basic level is association, which corresponds to the activity of *seeing*. At this level, we merely observe that a set of variables are statistically related. Directed acyclic graphs allow us to describe these relations in the form of conditional independencies between variables, without endowing them with causal information. Only on the next level — intervention — do we interpret DAGs causally. At the intervention level, we can answer population-level questions such as "what would happen if we force every patient to take the treatment?". The activity of *doing* corresponds to this level. At the highest level are counterfactuals, which correspond to the activity of *imagining*. These require the strongest assumptions but allow us to answer individual-level

Correspondence concerning this article should be addressed to Fabian Dablander, Department of Psychological Methods, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 VZ Amsterdam, The Netherlands. E-mail: dablander.fabian@gmail.com

questions such as "would the patient have recovered if we had given him the treatment, even though he did not recover and has not received the treatment?"

In the following sections, we successively climb this causal hierarchy. At each level, we discuss the central concepts and illustrate them with examples. This paper is by no means exhaustive; instead, it should provide you with a first appreciation of the concepts that surround the graphical approach to causal inference. The goal is that, after reading the paper, you will be better equipped than Karl Pearson was in the quest to understand when correlation does imply causation — and when it does not.

Correlation Alone Does Not Imply Causation

It is a truth universally echoed by scientists that correlation does not imply causation. In daily life, however, the former is frequently mistaken for the latter. Messerli (2012), for example, showed a strong positive relationship between chocolate consumption and the number of Nobel Laureates per country. Using more recent data, I have found an even stronger relationship, which is visualized in Figure 1.[1] Although it is difficult to assess whether Messerli (2012) is facetious in his writing or not, he is careful not to mistake this correlation for causation. In reporting on the study, the chocolate industry was less careful, stating that "eating chocolate produces Nobel prize winners" (Nieburg, 2012).

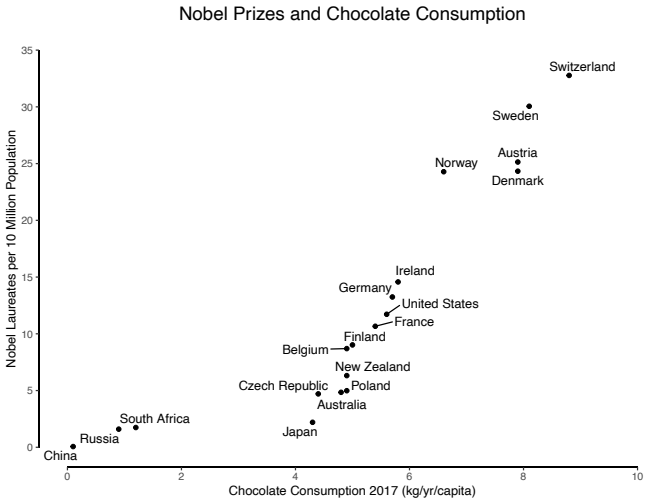

Figure 1. Shows the relationship between chocolate consumption and the number of Nobel Laureates per country.

Correlation by itself does not imply causation because statistical relations do not uniquely constrain causal relations. In particular, while chocolate consumption *could* cause an increase in Nobel Laureates, an increase in Nobel Laureates could likewise underlie an increase in chocolate consumption — possibly due to the resulting festivities, as Messerli (2012) conjectures. More plausibly, unobserved variables such as socio-economic status or quality of the education system might cause an increase in both chocolate consumption and Nobel Laureates, thus rendering their correlation spurious, that is, non-causal. The *common cause principle* states these three possibilities formally (Reichenbach, 1956):

> If two random variables X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then either (a) X causes Y, (b) Y causes X, or (c) there exists a third variable Z that causes both X and Y. Further, X and Y become independent given Z, i.e., $X \perp\!\!\!\perp Y \mid Z$.

An in principle straightforward way to break this uncertainty is to conduct an intervention: we could, for example, force the citizens of Austria to consume more chocolate and study whether this increases the number of Nobel laureates in the following years. Such interventions are clearly unfeasible; yet even in less extreme settings it is frequently unethical, impractical, or impossible — think of smoking and lung cancer — to intervene by for example conducting a randomized controlled trial.

Causal inference provides us with tools that allow us to draw causal conclusions even in the absence of a true experiment, given that certain assumptions are fulfilled. These assumptions increase in strength as we move up the levels of the causal hierarchy. In the remainder of this paper, I discuss the levels *association*, *intervention*, and *counterfactuals*, as well as the prototypical actions corresponding to each level — *seeing*, *doing*, and *imagining*.

Seeing

Association is on the most basic level, allowing us to see that two or more things are somehow related. Importantly, we need to distinguish between *marginal* associations, which look at the assocation between two variables without taking into account other variables, and *conditional* associations, which do take other variables into account. The latter are a key element of causal inference.

Figure 2 illustrates the difference between marginal and conditional assocations. The left panel shows the whole, aggregated data. Here, we see that the variables X and Y are positively correlated: an increase in values for X co-occurs with an increase in values for Y. This relation describes the *marginal* association of X and Y because we do not care whether $Z = 0$ or $Z = 1$. On the other hand, as shown in the right panel, if we condition on the binary variable Z, we find that there is no

[1] You can download the data from https://fabiandablander.com/assets/data/nobel-chocolate.csv. It includes Nobel Laureates up to 2019 and the 2017 chocolate consumption data as reported by https://www.statista.com/statistics/819288/worldwide-chocolate-consumption-by-country/.

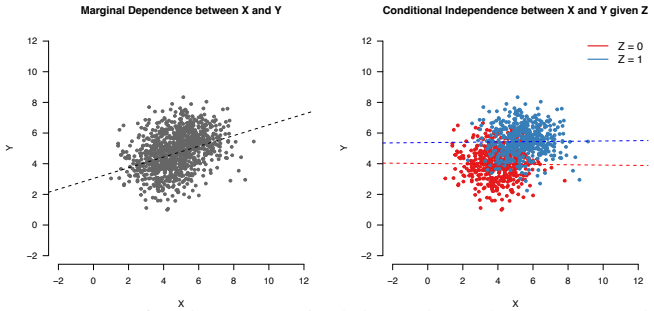

Figure 2. Left: Shows marginal dependence between X and Y. Right: Shows conditional independence between X and Y given Z.

relation: $X \perp\!\!\!\perp Y \mid Z$.[2] In some cases, the relationship between two variables can even become *reversed* in sub-populations compared to the relationship in the whole population. We will discuss the ramifications of this in a later section in some detail. For now, we focus on the simple fact that such a pattern of (conditional) (in)dependencies can exist. In the next section, we discuss a powerful tool that allows us to visualize such dependencies.

Directed Acyclic Graphs

We can visualize the statistical dependencies between the three variables X, Y, and Z using a graph. A graph $\mathcal{G}$ is a mathematical object that consists of nodes and edges. In the case of *Directed Acyclic Graphs* (DAGs), these edges are directed. We take our variables (X, Y, Z) to be nodes in such a DAG and we draw (or omit) edges between these nodes so that the conditional (in)dependence structure in the data is reflected in the graph. We will explain this more formally shortly. For now, let's focus on the relationship between the three variables. We have seen that X and Y are marginally dependent but conditionally independent given Z. It turns out that we can draw *three* DAGs that encode this fact; these are the first three DAGs in Figure 3. X and Y are dependent through Z in these graphs, and conditioning on Z *blocks* the path between X and Y. (We state this more formally shortly). While it is natural to interpret the arrows causally, at this first level of the causal hierarchy, we refrain from doing so. For now, the arrows are merely tools that help us describe associations between variables.

The rightmost DAG in Figure 3 encodes a different set of conditional (in)dependence relations between X, Y, and Z than the first three DAGs. Figure 4 illustrates this: looking at the aggregated data we do not find a relation between X and Y — they are *marginally independent* — but we do find one when looking at the disaggregated data — X and Y are *conditionally dependent* given Z.

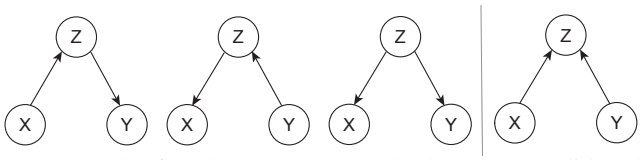

Figure 3. The first three DAGs encode the same conditional independence structure, $X \perp\!\!\!\perp Y \mid Z$. In the fourth DAG, Z is a collider such that $X \not\perp\!\!\!\perp Y \mid Z$.

A toy example might help build intuition: Assume that in the whole population — which includes singles as well as people in a relationship — being attractive (X) and being intelligent (Y) are two independent traits. This is what is illustrated in the left panel in Figure 4. Let's make the assumption that both being attractive and being intelligent are positively related with being in a relationship. What does this imply? First, it implies that, on average, single people are less attractive and less intelligent. This can be seen in the right panel in Figure 4, where singles ($Z = 0$) have a lower average value for X and Y compared to the people in a relationship ($Z = 1$). Second, and perhaps counter-intuitively, it implies that in the population of single people (and people in a relationship, respectively), being attractive and being intelligent are *negatively correlated*, as can also be seen in Figure 4.

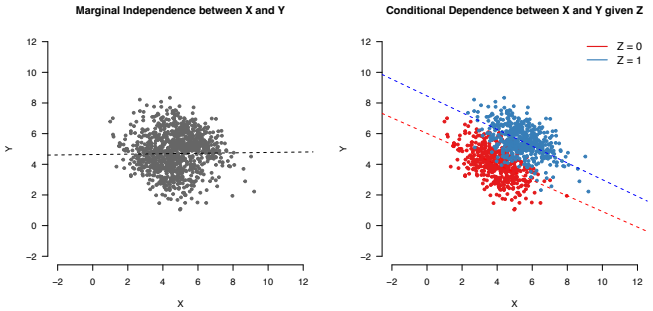

Figure 4. Left: Shows marginal independence between X and Y. Right: Shows conditional dependence between X and Y given Z

In the above example, visualized in the rightmost DAG in Figure 3, Z is commonly called a *collider*. Suppose we want to estimate the association between X and Y in the whole population. Conditioning on a collider (for example, by only analyzing data from people who are not in a relationship) and then computing the association between X and Y will lead to a different estimate, and the induced bias is known as *collider bias*. It is a serious issue not only in dating, but also for example in medicine, where it is known as Berkson's bias

[2] Instead of having Z only enter the regression as a main effect, we also include the interaction between Z and X, resulting in the two separate slopes (red and blue) in Figure 2 (and Figure 4) instead of one averaged slope. As long as Z enters the regression as a main effect, we say that we have *adjusted* for Z.

(Berkson, 1946; Cole et al., 2010).

The simple graphs shown in Figure 3 are the building blocks of more complicated graphs. In the next section, we describe a tool that can help us find (conditional) independencies between sets of variables. This becomes very important later when we introduce Structural Causal Models (SCMs), which relate causal to probabilistic statements. The resulting probabilistic statements, which include conditional (in)dependencies, can then be tested using data.

d-separation

For large graphs, it is not obvious how to conclude that two nodes are (conditionally) independent. *d*-separation is a tool that allows us to check this algorithmically (Geiger, Verma, & Pearl, 1990). To be able to use this tool, we need to define the following concepts:

- A *path* from X to Y is a sequence of nodes and edges such that the start and end nodes are X and Y, respectively.
- A conditioning set $\mathcal{L}$ is the set of nodes we condition on (it can be empty).
- Conditioning on a non-collider along a path *blocks* that path.
- A collider along a path blocks that path. However, conditioning on a collider (or any of its descendants) unblocks that path.

With these definitions out of the way, we call two nodes X and Y d-separated by $\mathcal{L}$ if conditioning on all members in $\mathcal{L}$ blocks all paths between the two nodes. To illustrate how d-separation works in practice, we apply it to the DAG shown in Figure 5. First, note that there are no *marginal* independencies; this means that without blocking nodes by conditioning on them, any two nodes are connected by a path. For example, there is a path going from X to Y through Z, and there is a path from V to U going through Y and W.

However, there are a number of *conditional* independencies. For example, X and Y are conditionally independent given Z. Why? There are two paths from X to Y: one through Z and one through W. However, since W is a collider on the path from X to Y, the path is already blocked. The only unblocked path from X to Y is through Z, and conditioning on it therefore blocks all remaining open paths. Additionally conditioning on W would unblock one path, and X and Y would again be associated.

So far, we have implicitly assumed that conditional (in)dependencies in the graph correspond to conditional (in)dependencies between variables. We make this assumption explicit now. In particular, note that d-separation provides us with an independence model $\perp\!\!\!\perp_{\mathcal{G}}$ defined on graphs. To

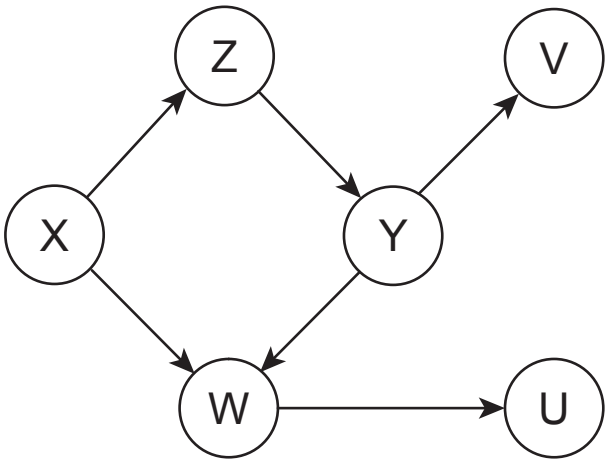

Figure 5. DAG to practice d-separation on, see main text.

connect this to our standard probabilistic independence model $\perp\!\!\!\perp_{\mathcal{P}}$ defined on random variables, we assume the following *Markov property*:

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z \implies X \perp\!\!\!\perp_{\mathcal{P}} Y \mid Z \ . \tag{1}$$

In words, we assume that if the nodes X and Y are d-separated by Z in the graph $\mathcal{G}$, the corresponding random variables X and Y are conditionally independent given Z. This implies that all conditional independencies in the data are represented in the graph. For example, the graph $X \to Y \to Z$ combined with the Markov property implies that the variables X, Y, and Z are all marginally dependent, but that X is conditionally independent of Y given Z. Moreover, Equation (1) implies (and is implied by) the following factorization of the joint probability distribution over all variables:

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathrm{pa}^{\mathcal{G}}(X_i)) \ , \tag{2}$$

where $\mathrm{pa}^{\mathcal{G}}(X_i)$ denotes the parents of the node X_i in graph $\mathcal{G}$ (see Peters, Janzing, & Schölkopf, 2017, p. 101). A node X is a parent of a node Y if there is an arrow from X to Y; for example, X is a parent of W in the graph shown in Figure 5. A node Y is a descendant of a node X if there exists a directed path from node X to Y; for example, V, W, and U are descendants of Y in the graph shown in Figure 5, but Z and X are not. The above factorization implies that a node X is independent of its non-descendants given its parents.

d-separation is an extremely powerful tool. Until now, however, we have used DAGs only to visualize (conditional) independencies. We do not merely want to see the world, but also change it; this requires a notion of *intervention*. In the next section, we go beyond *seeing* to *doing*.

Doing

From this section on, we are willing to interpret DAGs causally. As Dawid (2010) warns, this is a serious step. In merely describing conditional independencies — *seeing* — the arrows in the DAG played a somewhat minor role, being nothing but "incidental construction features supporting the d-separation semantics" (Dawid, 2010, p. 66). In this section, we endow the DAG with a causal meaning and interpret the arrows as denoting *direct causal effects*. What is a causal effect? Following Pearl and others, we take an *interventionist* position and say that a variable X has a causal influence on Y if changing X leads to changes in (the distribution of) Y. This position is a very useful one in practice, but not everybody agrees with it (Cartwright, 2007, ch. 6).

There are two principal ways how one might arrive at a DAG. First, one could try to learn it from data; this is known as *causal discovery* (e.g., Spirtes & Zhang, 2016). Second, one might posit a Structural Causal Model based on theory and an understanding of the problem one is modeling. From a SCM, a DAG follows; we will touch on this in a later section. Here, we assume that we have arrived at a causal DAG, and show what this enables us to do. Specifically, Figure 6 shows the observational DAGs from ealier (top row) as well as the manipulated DAGs (bottom row) where we have intervened on the variable X, that is, set the value of the random variable X to a constant x. Setting the value of $X = x$ cuts all incoming causal arrows. This is because the value of X is determined only by the intervention, not by any other factors.

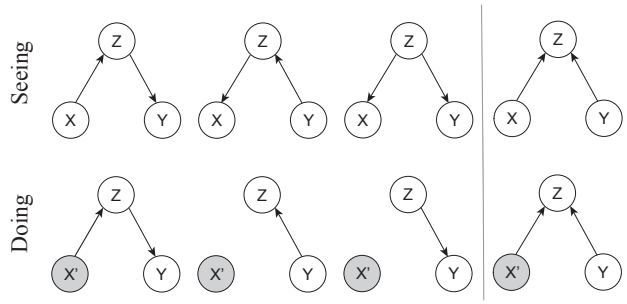

Figure 6. *Seeing*: DAGs are used to encode conditional independencies. The first three DAGs encode the same associations. *Doing*: DAGs are causal. All of them encode distinct causal assumptions.

As is easily verified with d-separation, the first three graphs in the top row encode the same conditional independence structure. This implies that we cannot distinguish them using only observational data. Interpreting the edges causally, however, we see that the DAGs have a starkly different interpretation. The bottom row makes this apparent by showing the result of an intervention on X. In the leftmost causal DAG, Z is on the causal path from X to Y, and intervening on X therefore influences Y through Z. In the DAG next to it, Z is on the causal path from Y to X, and so intervening on X does not influence Y. In the third DAG, Z is a common cause and — since there is no other path from X to Y — intervening on X does not influence Y. For the collider structure in the rightmost DAG, intervening on X does not influence Y because there is no unblocked path from X to Y. Note that we assume that the DAG adequately captures *all causal relations*, which implies that there is no unobserved confounding.

To make the distinction between seeing and doing, Pearl introduced the *do*-operator. While $p(Y \mid X = x)$ denotes the *observational* distribution, which corresponds to the process of seeing, $p(Y \mid do(X = x))$ corresponds to the *interventional* distribution, which corresponds to the process of doing. The former describes which values Y would likely take on when X *happened to be* x, while the latter describes which values Y would likely take on when X *would be set to* x.

Computing Causal Effects

$P(Y \mid do(X = x))$ describes the causal effect of X on Y, but how do we compute it? Actually *doing* the intervention might be unfeasible or unethical; side-stepping actual interventions and still getting at causal effects is the whole point of causal inference. We want to learn causal effects from observational data, and so all we have is the observational DAG. The causal quantity, however, is defined on the manipulated DAG. Consequently, we need to build a bridge between the observational DAG and the manipulated DAG, and we do this by making two assumptions.

First, we assume that *interventions are local*. This means that if I set $X = x$, then this only influences the variable X, with no other direct influence on any other variable. Of course, intervening on X will influence other variables, but only through X, as a side-effect of the intervention itself. In colloquial terms, we do not have a "fat hand" (e.g., Scheines, 2005), but act like a surgeon precisely targeting only a very specific part of the DAG.

Second, we assume that the mechanism by which variables interact do not change through interventions; that is, the mechanism by which a cause brings about its effects does not change whether this occurs naturally or by intervention (e.g., Pearl, Glymour, & Jewell, 2016, p. 56).

With these two assumptions in hand, further note that $p(Y \mid do(X = x))$ can be understood as the *observational* distribution in the manipulated DAG — which we denote as $p_m(Y \mid X = x)$ — that is, in the DAG where we set $X = x$. This is because after *doing* the intervention (which catapults us into the manipulated DAG, where all arrows pointing to the node we intervened on are cut), all that is left for us to do is to *see* its effect. Observe that the leftmost and rightmost

DAG in Figure 6 remain the same under intervention on X, and so the interventional distribution $p(Y \mid do(X = x))$ is just the conditional distribution $p(Y \mid X = x)$. The middle DAGs require a bit more work:

$$\begin{aligned} p(Y = y \mid do(X = x)) &= p_m(Y = y \mid X = x) \\ &= \sum_z p_m(Y = y, Z = z \mid X = x) \\ &= \sum_z p_m(Y = y \mid X = x, Z = z)\, p_m(Z = z) \\ &= \sum_z p(Y = y \mid X = x, Z = z)\, p(Z = z) \ . \end{aligned}$$

The first equality follows by definition. The second and third equality follow from the *sum* and *product* rule of probability. The last line follows from the assumption that the mechanism through which X influences Y is independent of whether we set X or whether X naturally occurs, that is, $p_m(Y = y \mid X = x, Z = z) = p(Y = y \mid X = x, Z = z)$, and the assumption that interventions are local, that is, $p_m(Z = z) = p(Z = z)$. Thus, the interventional distribution we care about is equal to the conditional distribution of Y given X when we adjust for Z. Graphically speaking, this blocks the path $X \leftarrow Z \leftarrow Y$ in the left middle graph and the path $X \leftarrow Z \rightarrow Y$ in the right middle graph in Figure 6. If there were a path $X \rightarrow Y$ in these two latter graphs, and if we would not adjust for Z, then the causal effect of X on Y would be *confounded*. For these simple DAGs, however, it is already clear from the fact that X is independent of Y given Z that X cannot have a causal effect on Y. In the next section, we study a more complicated graph and look at confounding more closely.

Confounding

Confounding has been given various definitions over the decades, but usually denotes the situation where a (possibly unobserved) common cause obscures the causal relationship between two or more variables. Using the framework of causal inference, we can be more precise and call a causal effect of X on Y confounded if $p(Y \mid X = x) \neq p(Y \mid do(X = x))$, which also implies that collider bias is a type of confounding. Confounding occured in the middle two DAGs in Figure 6, as well as in the chocolate consumption and Nobel Laureates example. Confounding is the bane of observational data analysis. Helpfully, causal DAGs provide us with a tool to state multivariate causal relations between variables, and the *do*-calculus subsequently provides us with a means to know what variables we need to adjust for so that causal effects are unconfounded.

A very useful tool to see whether a causal effect is confounded or not is the *backdoor criterion* (Pearl et al., 2016, p. 61), which states:

> Given two nodes X and Y, an adjustment set $\mathcal{L}$ fulfills the backdoor criterion if no member in $\mathcal{L}$ is a descendant of X and members in $\mathcal{L}$ block all paths between X and Y that contain an arrow into X. Adjusting for $\mathcal{L}$ thus results in the unconfounded causal effect of X on Y.

Assume that $\mathcal{L}$ consists of a set of variables Z. Formally, the backdoor criterion states that:

$$P(Y = y \mid do(X = x)) = \sum_z P(Y = y \mid X = x, Z = z)\, P(Z = z) \tag{3}$$

The key observation is that this *adjustment formula* (a) blocks all spurious, that is, non-causal paths between X and Y, (b) leaves all directed paths from X to Y unblocked, and (c) creates no spurious paths. This means that, if the backdoor criterion is satisfied, and we condition on $\mathcal{L}$, then the causal effect of X on Y is unconfounded. To see this action, let's again look at the DAG in Figure 5. The causal effect of Z on U is confounded by X, because in addition to the legitimate causal path $Z \rightarrow Y \rightarrow W \rightarrow U$, there is also an unblocked path $Z \leftarrow X \rightarrow W \rightarrow U$ which confounds the causal effect. The backdoor criterion would have us condition on X, which blocks the spurious path and renders the causal effect of Z on U unconfounded. Note that conditioning on W would also block this spurious path; however, it would also block the causal path $Z \rightarrow Y \rightarrow W \rightarrow U$. The backdoor criterion is very useful, but it does not exhaust all the ways in which one can arrive at the causal effect; specifically, there are causal effects which it fails to identify, but situations in which this is the case occur less frequently (see for example the *front-door criterion*, Pearl et al., 2016, p. 66). Beyond such criteria, the *do*-calculus provides a full account as to whether a particular causal effect can be estimated (Pearl, 2009, pp. 85–86)

Let's recap what we have discussed so far. At the lowest level of the causal hierarchy — association — we have discovered DAGs and d-separation as a powerful tool to reason about conditional (in)dependencies between variables. Moving to intervention, the second level of the causal hierarchy, we have satisfied our need to interpret the arrows in a DAG causally. Doing so required strong assumptions, but it allowed us to go beyond *seeing* and model the outcome of interventions. We used the *do*-calculus to clarify the notion of confounding. In particular, collider bias is a type of confounding, which has important practical implications: we should not blindly enter all variables into a regression in order to "control" for them, but think carefully about what the underlying causal DAG could look like. Otherwise, we might induce spurious associations.

The concepts from causal inference can help us understand methodological phenomena that have been discussed for decades. In the next section, we apply the concepts we have

seen so far to make sense of one such phenomenon: *Simpson's Paradox*.

Simpson's Paradox

Suppose two doctors, Dr. Hibert and Dr. Nick, perform a number of heart surgeries and band-aid removals; Table 1 records their respective performance (taken from Blitzstein & Hwang, 2014, p. 67). Strikingly, while Dr. Hibert has a higher success rate than Dr. Nick in surgery (77.8% vs 20%) as well as band-aid removal (100% vs 90%), his overall sucess rate is lower (80% vs 83%). While Karl Pearson has been aware of similar effects already in 1899 (Aldrich, 1995), it was the article by Simpson (1951) which drew renewed attention to this fact; Blyth (1972) was the first to call it a "paradox". Formally, such a reversal means that:

$$P(E \mid D) < P(E \mid \neg D) \tag{4}$$

$$P(E \mid D, S) > P(E \mid \neg D, S) \tag{5}$$

$$P(E \mid D, \neg S) > P(E \mid \neg D, \neg S) \ . \tag{6}$$

In our case, E denotes success, D denotes whether Dr. Hibert performed the procedure, and S denotes whether the procedure was heart surgery. The symbol $\neg$ denotes negation; for example, $\neg D$ denotes "not Dr. Hibert", thus referring to Dr. Nick.

This reversal can be explained by referring to base rates. In particular, heart surgery is clearly a more difficult procedure than removing band-aids. Since Dr. Hibert conducts considerably more heart surgeries than Dr. Nick, his overall performance suffers. To see how these base rates enter formally, observe that:

$$\begin{aligned} P(E \mid D) &= P(E \mid D, S)\, P(S \mid D) + P(E \mid D, \neg S)\, P(\neg S \mid D) \\ &= 0.778 \times 0.90 + 1.00 \times 0.10 \\ &= 0.80 \ . \end{aligned}$$

The weights $P(S \mid D)$ and $P(\neg S \mid D)$ constitute the base rates; a similar calculation can be done for Dr. Nick. Observe that his weights are reversed, that is, Dr. Nick performs considerably more band-aid removals — $P(\neg S \mid \neg D) = 0.90$ — than heart surgeries — $P(S \mid \neg D) = 0.10$. By computing the overall performance we lose information about which doctor performs which medical procedure how frequently, opening up the possibility for a reversal.

As demonstrated above, such a reversal can be intuitively explained by referring to base rates. Why call this a "paradox"? Following Pearl (2014), I believe it is useful to distinguish between Simpson's *reversal* and Simpson's *paradox*. The former refers to situations such as the one explained above. The latter refers to a "psychological phenomenon that evokes surprise and disbelief" (Pearl, 2014, p. 9). Such surprise and disbelief is easily evoked when reading Lindley and Novick (1981), who studied contingency tables not about doctors and surgery, but about treatment and sex. I provide a different, but similarly illustrative example here (see also Pearl et al., 2016, ch. 1). Suppose you observe $N = 700$ patients who either *choose* to take the treatment drug or not; note that this is not a randomized control trial. Table 2 shows the number of recovered patients split across sex (taken from Pearl et al., 2016, p. 2). Observe that more men as well as more women recover when taking the treatment (93% and 73%) compared to when not taking the treatment (87% and 69%). And yet, when taken together, *fewer* patients who took the treatment recovered (78%) compared to patients who did not take the treatment (83%). This is puzzling — should a doctor prescribe the treatment or not? Clearly, the answer has important real-world ramifications. Yet Lindley and Novick (1981) showed that there is no purely statistical criterion which allows us to decide whether to prescribe or not prescribe the treatment. While the authors suggested that *exchangeability*, a technical condition referring to sequences of random variables, provides an answer, subsequent literature showed that one instead needs to rely on explicit causal knowledge (e.g., Hernán, Clayton, & Keiding, 2011).

Table 1
Dr. Hibert outperforms Dr. Nick both in surgery and band-aid, yet his overall performance is worse.

	Dr. Hibert		Dr. Nick	
	Surgery	Band-Aid	Surgery	Band-Aid
Successes	70	10	2	81
Failures	20	0	8	9
Success Rate	77.8%	100%	20%	90%
Overall Success Rate	80%		83%	

In particular, to decide whether to prescribe treatment or not based on the data in Table 2, we need to compute the causal effect that the treatment has on recovery. As a first step, we draw the causal DAG. Suppose we know that women are more likely to take the treatment, that being a woman has an effect on recovery more generally, and that the treatment has an effect on recovery. Moreover, we know that the *treatment cannot cause sex*. This is a trivial yet crucial observation — it is impossible to express this in purely statistical language. One of the reasons why causal DAGs are such powerful tools is because they allow us to formalize such assumptions; the graph in Figure 7 makes explicit that sex (S) is a common cause of both treatment (T) and recovery (R). We denote $S = 1$ as being female, $T = 1$ as having chosen treatment, and $R = 1$ as having recovered. The left DAG in Figure 7 is observational while the right DAG indicates the intervention $do(T)$, that is, forcing every patient to either take the treatment

Table 2
Data used in the Simpson's paradox example, see main text.

	Treatment	No Treatment
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Men & Women	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

($T = 1$) or to not take the treatment ($T = 0$).

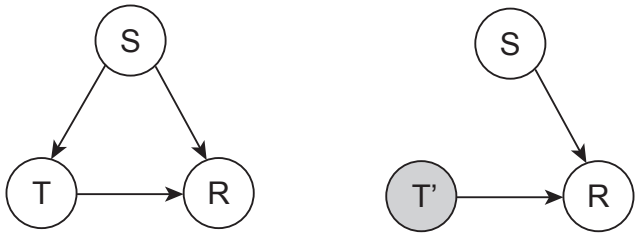

Figure 7. Underlying causal DAG of the example with treatment (T), biological sex (S), and recovery (R).

We are interested in the probability of recovery if we would force everybody to take, or not take, the treatment drug; we call the difference between these two probabilities the *average causal effect* in the population. This is key: the *do*-operator is about populations, not individuals. Using it, we cannot make statements that pertain to the recovery of an individual patient; we can only refer to the probability of recovery as defined on populations of patients. We will discuss *individual causal effects* once we have ascended to the third, and final level of the causal hierarchy. Computing the average causal effect requires knowledge about the interventional distributions $p(R \mid do(T = 0))$ and $p(R \mid do(T = 1))$. As discussed above, these correspond to the conditional distribution in the manipulated DAG which is shown in Figure 7 on the right. The backdoor criterion tells us that the conditional distribution in the observational DAG will correspond to the interventional distribution when blocking the spurious path $T \leftarrow S \rightarrow R$. Using the adjustment formula given in (3), we write:

$$
\begin{aligned}
p(R = 1 \mid do(T = 1)) &= \sum_s p(R = 1 \mid T = 1, S = s)\, p(S = s) \\
&= p(R = 1 \mid T = 1, S = 0)\, p(S = 0) \\
&+ p(R = 1 \mid T = 1, S = 1)\, p(S = 1) \\
&= \frac{81}{87} \times \frac{87 + 270}{700} + \frac{192}{263} \times \frac{263 + 80}{700} \\
&\approx 0.83 \ .
\end{aligned}
$$

In words, we first compute the benefit of the treatment separately for men and women, and then we average the result by weighting it with the fraction of men and women in the population. This tells us that, if we force everybody to take the treatment, about 83% of people will recover. Similarly, we can compute the probability of recovery given we force all patients to not take the treatment:

$$
\begin{aligned}
p(R = 1 \mid do(T = 0)) &= \sum_s p(R = 1 \mid T = 0, S = s)\, p(S = s) \\
&= p(R = 1 \mid T = 0, S = 0)\, p(S = 0) \\
&+ p(R = 1 \mid T = 0, S = 1)\, p(S = 1) \\
&= \frac{234}{270} \times \frac{87 + 270}{700} + \frac{55}{80} \times \frac{263 + 80}{700} \\
&\approx 0.78 \ .
\end{aligned}
$$

To compute the average causal effect (*ACE*) of treatment on recovery, we compute:

$$
\begin{aligned}
\text{ACE}(T \rightarrow R) &= \mathbb{E}\left[R \mid do(T = 1)\right] - \mathbb{E}\left[R \mid do(T = 0)\right] \\
&= 0.83 - 0.78 \\
&= 0.05 \ .
\end{aligned}
$$

On average, 5% more patients would recover if they were given the treatment; note that this is the exact opposite to the conclusion we had drawn when looking at the aggregated data in Table 2. The treatment does indeed have a positive effect on recovery on average, and the doctor should prescribe it.

Note that this conclusion heavily depended on the causal graph. While graphs are wonderful tools in that they make our assumptions explicit, these assumptions are — of course — not at all guaranteed to hold. These assumptions are strong, stating that the graph must encode all causal relations between variables, and that there is no unmeasured confounding, something we can only ever approximate in observational settings.

Let's look at another example in which we have the same data, but other causal relations are plausible. In particular, instead of the variable sex we look at the *post-treatment* variable blood pressure; see Table 3. Here we have measured blood pressure after the patients have taken the treatment drug. The question we wish to answer remains the same: Should a doctor prescribe the treatment or not? Since blood pressure is a post-treatment variable, it cannot influence a patient's decision to choose the treatment or not. Suppose the true causal DAG is as shown in Figure 8, asserting that the treatment has an indirect effect on recovery through blood pressure, in addition

Table 3
Data used in the Simpson's paradox example, see main text.

	Treatment	No Treatment
Low Blood pressure	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High Blood pressure	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Low & High Blood pressure	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

to having a direct causal effect. Note that a causal effect is *direct* only at at particular level of abstraction. The treatment drug works by inducing certain biochemical reactions that might themselves be described by DAGs. On a finer scale, then, the direct effect ceases to be direct.

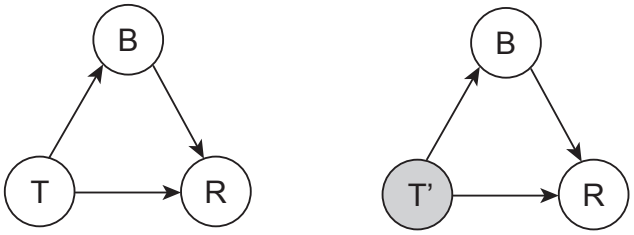

Figure 8. Underlying causal DAG of the example with treatment (T), blood pressure (B), and recovery (R).

From this DAG, we find that the causal effect of T on R is unconfounded. Therefore, the two causal quantities of interest are given by:

$$p(R = 1 \mid do(T = 1)) = p(R = 1 \mid T = 1) = 0.78$$
$$p(R = 1 \mid do(T = 0)) = p(R = 1 \mid T = 0) = 0.83 \ .$$

This means that the treatment is indeed harmful, since:

$$\begin{aligned} \text{ACE}(T \to R) &= \mathbb{E}\left[R \mid do(T = 1)\right] - \mathbb{E}\left[R \mid do(T = 0)\right] \\ &= 0.78 - 0.83 \\ &= -0.05 \ . \end{aligned}$$

Thus, the treatment has a negative effect in the general population (combined data). Suppose that the treatment has a direct positive effect on recovery, but an indirect negative effect through blood pressure. If we look only at patients with a particular blood pressure, then only the treatment's positive effect on recovery remains. However, since the treatment does influence recovery negatively through blood pressure, it would be misleading to take the association between T and R conditional on B as our estimate for the causal effect. In contrast to the previous example, using the aggregate data is the correct way to analyze these data in order to estimate the average causal effect — assuming that the underlying DAG is true.

So far, our discussion has been entirely model-agnostic, that is, we have not assumed a data-generating model. In the next section, we discuss Structural Causal Models (SCM) as the fundamental building block of this approach to causal inference. This will unify the previous two levels of the causal hierarchy — *seeing* and *doing* — as well as open up the third and final level: *imagining*.

Structural Causal Models

Structural Causal Models (SCM) relate causal and probabilistic statements. As an example, consider the following SCM:

$$\begin{aligned} X &:= \varepsilon_X \\ Y &:= f(X, \varepsilon_Y) \ . \end{aligned}$$

X directly causes Y in a manner specified by the function f; the noise variables ε_X and ε_Y are assumed to be independent. In a SCM, we take each equation to be a causal statement, and we stress this by using the assignment symbol := instead of the equality sign =. Note that this is in stark contrast to standard regression models; here, we explicitly state our causal assumptions.

As we will see below, Structural Causal Models imply observational distributions (*seeing*), interventional distributions (*doing*), as well as counterfactuals (*imagining*). Thus, they can be seen as the unifying element of this approach to causal inference. In what follows, we restrict the class of Structural Causal Models by allowing only linear relationships between variables and assuming independent Gaussian error terms, but note that more complicated models are possible (e.g., Hoyer, Janzing, Mooij, Peters, & Schölkopf, 2009). Structural Causal Models are closely related to Structural Equation Models, whose causal content has been debated throughout the last century. For more information, see for example Bollen and Pearl (2013).

As an example of an SCM, assume the following:

$$\begin{aligned} X &:= \varepsilon_X \\ Y &:= X + \varepsilon_Y \\ Z &:= Y + \varepsilon_Z \ , \end{aligned}$$

where $\varepsilon_X, \varepsilon_Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and $\varepsilon_Z \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.10)$ (see also Peters et al., 2017, p. 90). Again, each line states how variables are causally related. For example, we assume that X has a

direct causal effect on Y, that this effect is linear, and that it is obscured by independent Gaussian noise.

The assumption of Gaussian errors induces a multivariate Gaussian distribution on (X, Y, Z) whose independence structure is encoded in the leftmost DAG in Figure 9. The middle DAG shows an intervention on Z, while the rightmost DAG shows an intervention on X. Recall that, as discussed above, intervening on a variable cuts all incoming arrows. In the following, we illustrate an important fact: while a variable Z can be an excellent predictor of a variable Y, it need not have a causal effect on Y.

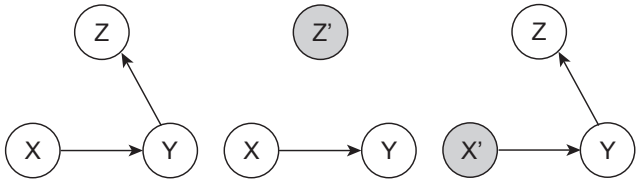

Figure 9. An excellent predictor (Z) need not be causally effective.

At the first level of the causal hierarchy — association — we might ask ourselves: does X or Z predict Y better? To illustrate the answer for our example, we simulate $n = 1000$ observations from the Structural Causal model using R:

```
set.seed(1)

n <- 1000
x <- rnorm(n, 0, 1)
y <- x + rnorm(n, 0, 1)
z <- y + rnorm(n, 0, 0.10)
```

Figure 10 shows that Y has a much weaker association with X (left panel) than with Z (right panel); this is because the standard deviation of the error ε_X is only a tenth of the standard deviation of the error ε_Z. For prediction, therefore, Z is the more relevant variable.

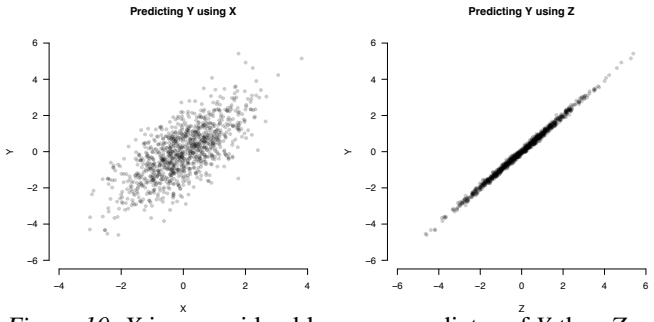

Figure 10. X is a considerably worse predictor of Y than Z.

But does Z actually have a causal effect on Y? This is a question about intervention, which is squarely located at the second level of the causal hierarchy. Assuming an underlying Structural Causal Model, we can easily simulate interventions in R and visualize their outcomes:

```
intervene_z <- function(z, n = 1000) {
  x <- rnorm(n, 0, 1)
  y <- x + rnorm(n, 0, 1)
  cbind(x, y, z)
}

intervene_x <- function(x, n = 1000) {
  y <- x + rnorm(n, 0, 1)
  z <- y + rnorm(n, 0, 0.1)
  cbind(x, y, z)
}

set.seed(1)
datz <- intervene_z(z = 2)
datx <- intervene_x(x = 2)
```

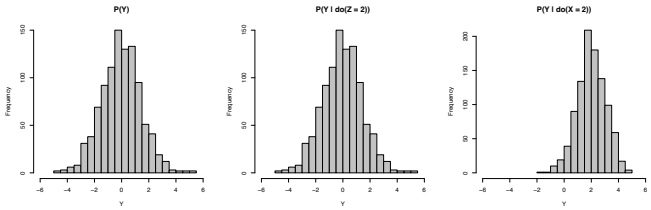

Figure 11. Shows marginal distribution of Y (left), interventional distribution of $P(Y \mid \text{do}(Z = 2))$ (middle), and interventional distribution of $P(Y \mid \text{do}(X = 2))$ (right).

The leftmost histogram in Figure 11 below shows the marginal distribution of Y when no intervention takes place. The histogram in the middle shows the marginal distribution of Y in the manipulated DAG where we set $Z = 2$. Observe that, as indicated by the causal graph, Z does not have a causal effect on Y such that $p(Y \mid do(Z = 2)) = p(Y)$. The histogram on the right shows the marginal distribution of Y in the manipulated DAG where we set $X = 2$. Clearly, then, X has a causal effect on Y. More precisely, we can again compute the average causal effect:

$$\text{ACE}(X \to Y) = \mathbb{E}\left[Y \mid do(X = x + 1)\right] - \mathbb{E}\left[Y \mid do(X = x)\right] \quad ,$$

which in our case equals one, as can also be seen from the structural assignments in the SCM above. Thus, SCMs allow us to model the outcome of interventions. However, note again that this is strictly about populations, not individuals. In the next section, we see how SCMs allow us to reach the final level of the causal hierarchy, moving beyond the average to define individual causal effects.

Imagining

In the *Unbearable Lightness of Being*, Milan Kundera has Tomáš ask himself:

> "Was it better to be with Tereza or to remain alone?"

To which he answers:

> "There is no means of testing which decision is better, because there is no basis for comparison. We live everything as it comes, without warning, like an actor going on cold. And what can life be worth if the first rehearsal for life is life itself?"

Kundera is describing, as (Holland, 1986, p. 947) put it, the "fundamental problem of causal inference", namely that we simply cannot observe counterfactuals. If Tomáš chooses to stay with Tereza, then he cannot not choose to stay with Tereza. He cannot go back in time and revert his decision, living instead "everything as it comes, without warning". This does not mean, however, that Tomáš cannot assess afterwards whether his choice has been wise. As a matter of fact, humans constantly evaluate mutually exclusive options, only one of which ever comes true; that is, humans reason *counterfactually*.

To do this formally requires strong assumptions. The *do*-operator, introduced above, is too weak to model counterfactuals. This is because it operates on distributions that are defined on populations, not on individuals. We can define an average causal effect using the *do*-operator, but — unsurprisingly — it only ever refers to averages. However, Structural Causal Models also allow counterfactual reasoning on the level of the individual; we illustrate this with the following example.

Suppose we want to study the causal effect of grandma's treatment for the common cold (T) on the speed of recovery (R). Usually, people recover from the common cold in seven to ten days, but grandma swears she can do better with a simple intervention — we agree on doing an experiment. Assume we have the following SCM:

$$\begin{aligned} T &:= \varepsilon_T \\ R &:= \mu + \beta T + \varepsilon \ , \end{aligned}$$

where μ is an intercept, $\varepsilon_T \sim$ Bernoulli(0.50) indicates random assignment to either receive the treatment ($T = 1$) or not receive it ($T = 0$), and $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma)$. The SCM tells us that the direct causal effect of the treatment on how quickly patients recover from the common cold is β. This causal effect is obscured by individual error terms for each patient $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$, which can be interpreted as aggregate terms for all the things left unmodelled. In particular, ε_k summarizes all the things that have an effect on the speed of recovery for patient k.

Table 4
Data simulated from the SCM concerning grandma's treatment of the common cold.

Patient	Treatment	Recovery	ε_k
1	0	5.80	0.80
2	0	3.78	-1.22
3	1	3.68	0.68
4	1	0.74	-2.26
5	0	7.87	2.87

Once we have collected the data, suppose we find that $\mu = 5$, $\beta = -2$, and $\sigma = 2$. This does speak for grandma's treatment, since it shortens the recovery time by 2 days on average:

$$\begin{aligned} \text{ACE}(T \to R) &= \mathbb{E}\left[R \mid do(T = 1)\right] - \mathbb{E}\left[R \mid do(T = 0)\right] \\ &= \mathbb{E}\left[\mu + \beta + \varepsilon\right] - \mathbb{E}\left[\mu + \varepsilon\right] \\ &= (\mu + \beta) - \mu \\ &= \beta \ . \end{aligned}$$

Given the value for ε_k, the Structural Causal Model is fully determined, and we may write $R(\varepsilon_k)$ for the speed of recovery for patient k. To make this example more concrete, Table 4 shows data for five patients, simulated from the SCM.

We see that the first patient did not receive the treatment ($T = 0$), took about $R = 5.80$ days to recover from the common cold, and has a unique value $\varepsilon_1 = 0.80$. Would this particular patient have recovered more quickly if we had given him grandma's treatment even though we did not? We denote this quantity of interest as $R_{T=1}(\varepsilon_1)$ to contrast it with the actually observed $R_{T=0}(\varepsilon_1)$. To compute this seemingly otherworldly quantity, we simply plug the value $T = 1$ and $\varepsilon_1 = 0.80$ into our Structural Causal Model, which yields:

$$R_{T=1}(\varepsilon_1) = 5 - 2 + 0.80 = 3.80 \ .$$

Using this, we can define the *individual causal effect* of treatment T on recovery R for the first patient as:

$$\begin{aligned} \text{ICE}(T \to R) &= R_{T=1}(\varepsilon_1) - R_{T=0}(\varepsilon_1) \\ &= 3.80 - 5.80 \\ &= -2 \ , \end{aligned}$$

which in this example is equal to the average causal effect due to the linearity of the underlying SCM (Pearl et al., 2016, p. 106). In general, individual causal effects are not identified, and we have to resort to average causal effects.

Answering the question of whether a particular patient would have recovered more quickly had we given him the treatment

even though we did not give him the treatment seems almost fantastical. It is a *cross-world* statement: given what we have observed, we ask about what would have been if things had turned out different. It may strike you as a bit eerie to speak about different worlds. Peters, Janzing, & Schölkopf (2017, p. 106) state that it is "debatable whether this additional [counterfactual] information [encoded in the SCM] is useful." It certainly requires strong assumptions. More broadly, Dawid (2000) argues in favour of causal inference without counterfactuals. Yet if we want to design machines that can achieve human level reasoning, it is likely that we need to endow them with counterfactual thinking (Pearl, 2019b). Moreover, many concepts that a relevant in legal and ethical domains, such as fairness (Kusner, Loftus, Russell, & Silva, 2017), require counterfactuals.

Challenges of Applying Causal Inference

While causal inference provides a powerful tool to reason about interventions and to formalize counterfactual reasoning, it comes with a number of challenges. In the following, I categorize the challenges into *statistical* and *conceptual* ones, but note that this is not a strict separation. First, as is the case with all modeling, the model is likely misspecified. If one derives causal quantities based on linear models, then the causal quantity will be accurate in so far as the relationship between variables is indeed linear. Linearity is a strong assumption, so caution is well-adviced.

Computing causal effects usually requires one to condition on covariates to adjust for confounding. Selecting the appropriate covariates is a difficult statistical problem, however. A large number of strategies to select covariates in order to adjust for confounding exist, yet Witte and Didelez (2019) show — as may be expected — that no such method uniformly performs best in all situation. Instead, they suggest that the choice of method should be informed by the hypothesized confounding structure.

In the social and behavioural sciences, item-level responses are usually taken as a stand-in for underlying constructs. However, as Westfall and Yarkoni (2016) show, adjusting for confounding using item-level responses that are a noisy proxy for the underlying construct can lead to incorrect causal conclusions. The authors give the classic example of observing that the number of drownings as well as the sale of ice-cream increases on sunny days. Clearly, the resulting correlation between ice-cream sales and drownings is spurious — sunny weather causes people to go swimming and buy ice-cream — and controlling for temperature would expose it as such. However, in the social and behavioural sciences we generally do not have direct access to the constructs we are interested in. To illustrate this on the ice-cream example, assume that we cannot measure temperature but that we have to take self-reported feelings of heat as a proxy for temperature. If this stand-in is noisy, then controlling for it does not remove the spurious correlation between ice-cream sales and drownings (Westfall & Yarkoni, 2016). To avoid this, Westfall and Yarkoni (2016) suggest latent variable modeling, which reduces noise by combining various indicators in a measurement model, and generally a much larger sample of participants.

As we have seen throughout the paper, statistics is not enough to constrain causal inference. For example, the two DAGs in the example of Simpson's paradox are observationally equivalent — they imply the same conditional independencies — yet provide different causal conclusions. This problem is immediate in *causal discovery*, where one tries to infer causal relations from data, which outputs not one DAG but an equivalence class (e.g., Kalisch & Bühlmann, 2007). Similarly, in the field of structural equation modeling, this is known as the "model equivalence" problem (e.g., Raykov & Penev, 1999). Thus, unless one has a strong theory that undergirds one's causal model, thereby excluding all other equivalent models, drawing strong conclusions based only on statistical model fit is perilous.

The fact that directed acylic graphs do not allow feedback loops may seem to run counter to real-life experience. In psychology, for example, it is obvious that variables can reinforce each other, resulting in a cycle. One may argue, however, that at a sufficiently high temporal resolution, no system has cycles. Instead, the system can be modeled as a DAG at each time unit, where $X \rightarrow Y$ at time t and $Y \rightarrow X$ at time $t + 1$. Under some conditions, Structural Causal Models can allow cyclic assignments, which leads to *directed cyclic graphs* (e.g., Spirtes, 1995). In the Structural Equation Modeling literature, such models are known as *nonrecursive* Structural Equation Models (e.g., Bollen, 1989, p. 83).

Correctly interpreting DAGs causally requires two crucial assumptions. First, we assumed that the DAG includes all relevant variables and their causal relations. In practical applications, this is hard to verify, and likely always violated to some extent. Second, we assumed that interventions are local, that is, that intervening on one variable does not influence other variables that are causally unrelated to the intervened variable. This is likely violated in many applications. In psychology, for example, we usually do have a "fat hand"; for example, an intervention geared towards decreasing suicidal thoughts most certainly influences other variables as well. The extent to which this is problematic must be assessed on a case by case basis.

Lastly, I want to touch upon a conceptual issue which concerns the interpretation of the *do*-operator for non-manipulable causes. Recall that the *do*-operator allows us to answer questions such as "how does the recovery rate change if we force every patient to take the treatment?". How-

ever, nothing in the formalism prevents us from applying it to variables that are (generally believed to be) non-manipulable; for example, we might ask what would happen if we force every patient to be a woman. Researchers on causal inference disagree as to the interpretation of such interventions. Pearl (2019a), for example, argues that one need not to worry about the distinction between manipulable and non-manipulable causes, stating that one could interpret the causal effect as an upper bound of an ideal intervention which might be possible in the future. M. A. Hernán (2016), on the other hand, calls for the need of "well-defined interventions". The more difficult it is to define how one would conduct a randomized trial in which one would manipulate the quantity of interest, the harder it becomes to interpret the estimated causal effect. I believe that this is a very sensible position, and researchers should carefully think about how they would implement the interventions whose causal effects they estimate (see also M. A. Hernán & Robins, 2016).

Conclusion

In this introductory paper, we have touched on several key concepts of causal inference. We have started with the puzzling observation that chocolate consumption and the number of Nobel Laureates are strongly positively related. At the lowest level of the causal hierarchy — association — we have seen how directed acyclic graphs can help us visualize conditional independencies, and how d-separation provides us with an algorithmic tool to check such independencies.

Moving up to the second level — intervention — we have seen how the *do*-operator models populations under interventions. This helped us define *confounding* — the bane of observational data analysis — as occuring when $p(Y \mid X = x) \neq p(Y \mid do(X = x))$. This comes with the important observation that entering all variables into a regression in order to "control" for them is misguided; rather, we need to carefully think about the underlying causal relations lest we want to introduce bias by for example conditioning on a collider. The *backdoor criterion* provided us with a graphical way to assess whether an effect is confounded or not.

At the top level of the causal hierarchy, we have seen that Structural Causal Models (SCMs) provide the building block from which observational and interventional distributions follow. SCMs further imply counterfactual statements, which allow us to move beyond the *do*-operator and average causal effects: they enable us to answer questions about what would have been if things had been different. Lastly, we have discussed a number of challenges researchers face when applying tools from causal inference in practice.

Suggested Readings

For further reading, I recommend the excellent textbooks by Pearl et al. (2016) and Peters et al. (2017). For less technical reading, which also provides a historical perspective, see Pearl and Mackenzie (2018). Miguel Hernán teaches an introductory online course on causal diagrams, freely available from https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your. This manuscript is based on a blog post on causal inference (Dablander, 2019).

Acknowledgements

I want to thank Oisín Ryan, Jonas Haslbeck, Sacha Epskamp, and Peter Edelsbrunner for valuable comments on this manuscript.

References

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, *10*(4), 364–376.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, *2*(3), 47–53.

Blitzstein, J. K., & Hwang, J. (2014). *Introduction to Probability*. Chapman; Hall/CRC.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of Causal Analysis for Social Research* (pp. 301–328). Springer.

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press.

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., & Poole, C. (2010). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, *39*(2), 417–420.

Dablander, F. (2019). An introduction to Causal inference. Retrieved from https://fabiandablander.com/r/Causal-Inference.html

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, *95*(450), 407–424.

Dawid, A. P. (2010). Beware of the DAG! In *Causality: Objectives and Assessment* (pp. 59–86).

Geiger, D., Verma, T., & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, *20*(5), 507–534.

Greenland, S. (2011). Causation and Causal Inference. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 216–221). Springer.

Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, *26*(10), 674–680.

Hernán, M. A., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, *40*(3), 780–785.

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, *183*(8), 758–764.

Hernán, M., & Robins, J. (2020). Causal inference: What if. *Boca Raton: Chapman & Hill/CRC*.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems* (pp. 689–696).

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. LaSalle, IL.

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, *8*(Mar), 613–636.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066–4076).

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *The Annals of Statistics*, *9*(1), 45–58.

Messerli, F. (2012). Chocolate consumption, cognitive function, and nobel laureates. *The New England Journal of Medicine*, *367*(16), 1562–1564.

Neyman, J. (1990). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (1923). English translations excerpts by D. Dabrowska and T. Speed. *Statistical Science*, *4*(5), 465–480. (Original work published 1923)

Nieburg, O. (2012). Eating chocolate produces Nobel prize winners, says study. Retrieved from https://www.confectionerynews.com/Article/2012/10/11/Chocolate-creates-Nobel-prize-winners-says-study

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2009). *Causality (2nd Edition)*. Cambridge University Press.

Pearl, J. (2014). Comment: Understanding Simpson's Paradox. *The American Statistician*, *68*(1), 8–13.

Pearl, J. (2019a). On the interpretation of do(x). *Journal of Causal Inference*, *7*(1).

Pearl, J. (2019b). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT press.

Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research*, *34*(2), 199–244.

Reichenbach, H. (1956). *The direction of time*. University of California Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science*, *72*(5), 927–940.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238–241.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the 11th*

Conference on Uncertainty in Artificial Intelligence (pp. 491–498).

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. In *Applied Informatics* (Vol. 3, p. 3). Springer.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS One*, *11*(3).

Witte, J., & Didelez, V. (2019). Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*, *61*(5), 1270–1289.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–580.