Notes on causal inference and directed acyclic graphs

Ben Lansdell

1 But what is causality?

Say we observe a negative relationship between number of apples eaten per day and heart disease. Does this relationship mean that apples are protective against disease? Maybe. It is well known that correlation does not imply causation. Perhaps in this case number of apples eaten per day correlates with general diet, or general fitness, which instead are the cause of lower heart disease. Such factors are a source of confounding. How do we distinguish between these possibilities? A statistician answers these causal inference questions in two ways: by considering counterfactuals and interventions.

A counterfactual is simply a potential event that did not occur. A given patient either does or does not receive the treatment on a given trial. Whichever event does not occur is the counterfactual. Under a counterfactual account of causality to claim that a proposed treatment causes disease remission is to claim that had the patient not received the treatment then the disease outcome would be different (greater).

Interventionist accounts are similar but focus on the notion of manipulability. Here to claim that one variable causes another is to claim that if through intervention one variable is forced to a given state then a change in the other variable will be observed. Here the notion of intervention is treated as a primitive and causal relationships are derived from that.

Thus, had a given patient *not* had so many apples per day, would their health be worse? And, if a patient was *forced* to eat many apples per day, would their health be better? Here we focus on frameworks that attempt to answer these questions in the presence of confounding. The basic idea is that if we do observe all the factors we reasonably consider to be confounding the estimate then we can correct for this.

2 Learning causal relationships

Randomized controlled trials (RCTs) are the gold standard for causal inference. The idea simply being that if assignment to a treatment group is randomized then the distribution of covariates in the control and treatment groups will be identical, and therefore any difference in outcome between the control and treatment groups can then only be attributed to the fact that one group received a treatment while the other did not.

However, sometimes RCTs are difficult, expensive, or unethical to perform. This motivates considering when causal relationships can be inferred from observational data alone. In the absence of randomization, receiving treatment may be correlated with many other factors which could also impact the outcome. What are conditions in which the effects of confounding can be mitigated?

Counterfactual outcomes are not observed for individual patients – they either receive a treatment or do not. This is known as the fundamental problem of causal inference. As a result often we need to (or in fact want to) consider aggregate causal effects estimated over a population. This has two consequences for analysis.

The first is that in considering causal relationships in this aggregate sense, the timing of pairs of events is often unspecified, vaguely defined, or implicit in how data is collected. By losing this timing information it is harder to analyze cases of mutual causation. Thus the assumption made here is that one variable is the cause of another, or vice versa, or not at all – there is a directedness to the relationship over the time window in which observations are made. It addition to excluding mutual causation from consideration, it is simplest to further exclude cycles, or causal chains (e.g. $A \to B \to C \to A$). The second consequence is that, by considering a population of subjects/events, it becomes more necessary to allow for probabilistic causal relationships, in which one variable's occurrence affects another's probability of occurring and the relationship need in no way be deterministic. These considerations motivate summarizing causal relationships between a set of variables using directed acyclic graphs (DAGs), and using a probabilistic framework.

3 Counterfactuals: the causal effect as difference in potential outcomes

Measuring causal effects in terms of counterfactuals is a relatively old idea (as far as statistics goes), dating back to 1923 from work of Neyman. The Neyman-Rubin causal model provides a framework for reasoning about causal effects with counterfactuals. In a simple setting, the model considers two *potential outcomes*: an outcome when a subject does receive a treatment, $Y(1)$, and an outcome when a subject does not receive a treatment, $Y(0)$ (i.e. a control subject). For a given subject, i, the *causal effect* is the difference in potential outcomes:

$$E_i = Y_i(1) - Y_i(0). \tag{1}$$

If we let W_i be a treatment random variable:

$$W_i = \begin{cases} 1, & \text{subject } i \text{ receives treatment;} \\ 0, & \text{subject } i \text{ assigned control;} \end{cases} \tag{2}$$

then assuming consistency between potential and observed outcome, Y_i, we have:

$$Y_i = W_i Y_i(0) + (1 - W_i) Y_i(1). \tag{3}$$

As an aside, note that the potential outcomes $Y(i)$ are treated as kinds of hypothetical random variables. In a sense neither is observed, and they are only related to observation through the assumption that (3) holds. This is a somewhat subtle point that is perhaps not well reflected in the notation. Equations in causal models can have quite different interpretations to standard statistical models, despite having similar notation, which is important to be aware of.

Per the *fundamental problem of causal inference*, only one of these potential outcomes is ever observed. To get around this, causal effects can be measured over a population of subjects, some of which receive the treatment and some of which do not. Over a population we can consider the *average causal effect*:

$$\tau = \mathbb{E}(Y_I(1) - Y_I(0)). \tag{4}$$

If W_i is assigned to each subject at random then τ can be computed directly from the treatment and control subpopulation means. In randomized cases, W_i is independent from the potential outcomes. If W_i were not independent from the potential outcomes then the measured causal effect (difference in means) could simply be a result of this correlation.

3.1 Causal assumptions for identifiability

Being able to measure a causal effect in an unbiased (unconfounded) way means the effect is *identifiable*. Within this counterfactual framework, this linking of potential outcomes to causal effects relies on four *causal assumptions*. Some of these have been alluded to above. They are:

1. (**SUTVA**) Stable Unit Treatment Value Assumption. This means:
   - There is no interference in treatments – one subject receiving treatment does not affect others' treatment.
   - There is only one form of treatment.
2. (**Consistency**) This assumption links the hypothetical potential outcomes to observed data. If we assume consistency then we are assuming:
$$Y_i = W_i Y_i(0) + (1 - W_i) Y_i(1), \tag{5}$$
as discussed above.
3. (**No unmeasured confounders/ignorability**) The treatment assignment is independent of the potential outcomes:
$$Y(1), Y(0) \perp\!\!\!\perp W. \tag{6}$$
In most cases of interest both the outcome and treatment variable are related to a set of observed covariates, X. Causal inference then requires:
$$Y(1), Y(0) \perp\!\!\!\perp W | X. \tag{7}$$
In RCTs this assumption may be reasonable. This says that the distribution of potential outcomes $(Y(1), Y(0))$ is the same across treatment levels W, conditioned on X. In observational settings often this is the primary assumption that is a road block to identifiability.
Another way to understand this is as follows. We want to relate observed quantities to hypothetical potential outcomes. We can do this if we assume ignorability:
$$\begin{aligned}
\mathbb{E}(Y|W=1) - \mathbb{E}(Y|W=0) &= \mathbb{E}(WY(1) + (1-W)Y(0)|W=1) - \mathbb{E}(WY(1) + (1-W)Y(0)|W=0) \\
&= \mathbb{E}(Y(1)|W=1) - \mathbb{E}(Y(0)|W=0) \\
\text{(ignorability)} \quad &= \mathbb{E}(Y(1) - Y(0))) \\
&= \tau
\end{aligned}$$
4. (**Positivity**) Additionally, causal inference requires a non-zero probability of assignment to a treatment group for all subjects:
$$0 < \mathbb{P}(W_i = 1 | X_i = x) < 1, \quad \forall x. \tag{8}$$
This is known as the *positivity*, or overlap, assumption.
Simply, a causal effect cannot be measured if no subjects receive the treatment, or they all do.

4 Directed acyclic graphs and probability distributions

In a sense the conditional independence between treatment and potential outcome is the main assumption that requires analysis in the above set of assumptions. This analysis can be aided by encoding our assumptions about the relations between different variables in a graph. This section defines and describes the behavior of these graphs. The following section contains criteria that can be used to identify sets of variables that are sufficient to act as controls, that remove the effect of confounding and hence that satisfy ignorability. These models are types of graphical models, sometimes known as *Bayesian networks*, and were first developed by Pearl in the 1980s.

Here we will consider a set of random variables $\mathcal{X}$ as nodes on a directed acyclic graph $\mathcal{G}$. Let this graph have edges $\mathcal{E}$ that represent relations between the variables. Ignorability requires conditional independence of the outcome from the treatment variable, so here we will let the directed edges encode conditional independence assumptions. (A *causal Bayesian network* has additional semantics that are discussed in Section 7. For the moment the directed edges only encode information about conditional independence.)

First note that the DAG imposes an ordering on the variables $\mathcal{X}$, from which we can talk about a node's parents, children, ancestors or descendants. Note also that any multivariate distribution can be decomposed into a product of conditional probabilities for any ordering of the variables:

$$P(X) = \prod_{j=1}^{N} P(X_j | \{X_k\}_{k>j}).$$

Given this, if we assume that the variables are ordered in a way that respects the ordering of the DAG, then we will say $\mathcal{X}$ is a Bayesian network with respect to $\mathcal{G}$ if the joint distribution over variables $\mathcal{X}$ factors according to:

$$P(X) = \prod_{j=1}^{N} P(X_j | \text{Pa}(X_j)),$$

where $\text{Pa}(X_j)$ is the parents of node X_j. That is, each node X_j is conditionally independent of its non-descendants given its parents:

$$P(X_j | \{X_k\}_{k>j}) = P(X_j | \text{Pa}(X_j)).$$

This is the *Markov condition*, or Markov assumption, for a Bayesian network. A node is conditionally independent of the entire network given its *Markov blanket* – its parents, its children, and its children's other parents.

Often also invoked is the *faithfulness condition*, which is the condition that the conditional independencies implied by the graph are the only conditional independencies in the distribution. E.g. assuming faithfulness in the graph $A \to B$ says that there is in fact a dependence between A and B.

4.1 Some types of graphs

Some properties of a Bayesian network can be inferred graphically. For instance three basic components of DAGs are:

1. Chain: $A \to B \to C$
2. Fork: $A \leftarrow B \to C$
3. Collider (inverted fork): $A \to B \leftarrow C$

These graphs behave differently when conditioning on parts of them. Compare the fork and the inverted fork.

- For the fork, A and C are dependent. Yet when conditioned on B, A and C become independent.
- The converse is true for the inverted fork. Without conditioning, A and B are independent. Yet when conditioned on B, A and C become dependent. This may seem a little counter-intuitive. An example of this phenomenon is if B is determined through tossing two independent coins, A and C. If B is determined as

$$B = \begin{cases} 1, & A = H, C = H; \\ 0, & \text{else} \end{cases}$$

  By itself, knowing A tells you nothing about C. But knowing B and A now tells you something about C.

Note that the fork and the chain have the same behavior:

- For the fork, A and C are dependent. Yet when conditioned on B, A and C become independent.
- For the chain, A and C are dependent. Yet when conditioned on B, A and C become independent.

4.2 d-separation

For more complicated graphs, are a given set of variables sufficient controls to render two nodes conditionally independent? Here the notion of *d-separation* is useful.

The d stands for dependence. Let P be a path from node u to v. A path is a loop-free, undirected (i.e. all edge directions are ignored) path between two nodes. Then P is said to be d-separated by a set of nodes Z if any of the following conditions holds:

- P contains a directed chain such that the middle node m is in Z, or
- P contains a fork, $u \cdots \leftarrow m \rightarrow \cdots v$, such that the middle node m is in Z, or
- P contains an inverted fork (or collider), $u \cdots \rightarrow m \leftarrow \cdots v$, such that the middle node m is *not* in Z and no descendant of m is in Z.

Nodes u and v are said to be d-separated by Z if all paths between them are d-separated. If u and v are not d-separated, they are called d-connected.

We have the result that X_u and X_v being d-separated by Z tells us that X_u and X_v are conditionally independent given Z.

4.3 Markov equivalence classes

Note that a DAG may prescribe a factorization of the probability distribution, but the converse is not true. That is, knowing a factorization of the joint distribution does not always imply a unique DAG. Instead it prescribes a *Markov equivalence class* of DAGs. This means that if we want to think of the directed edges as representing causal relationships then knowing a joint distribution factorization does not always provide a unique graph of causal relationships. This limits what we can learn about causal relationships from a joint (observational) distribution alone.

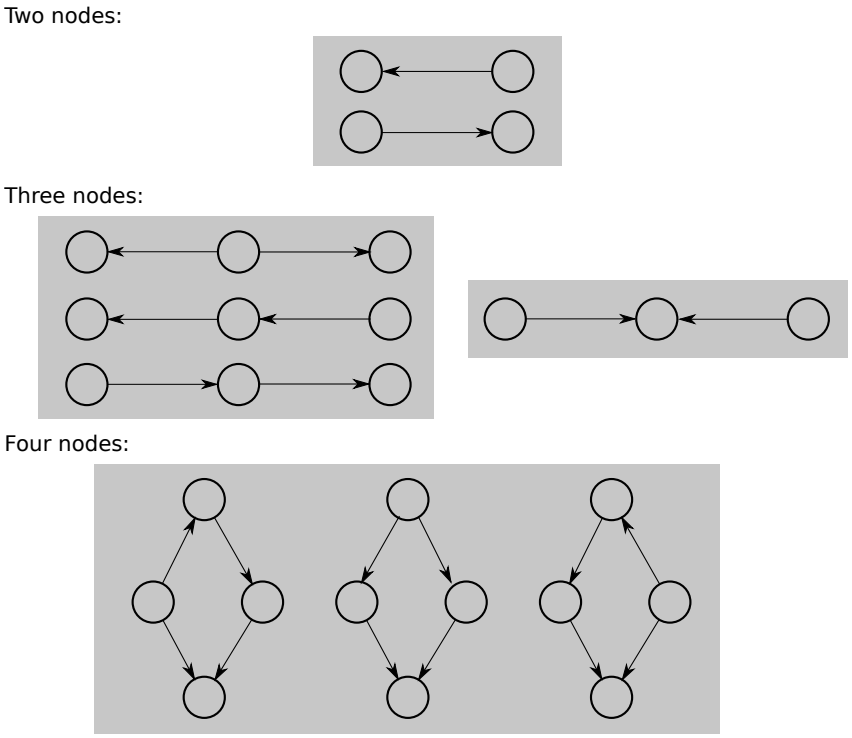

Figure 1: Examples of DAGs in the same Markov equivalence class.

Two graphs are Markov equivalent iff they share the same conditional independencies. Equally, they are Markov equivalent iff they have the same d-separations. That is, if u and v are d-separated by C in $\mathcal{G}_1$ then they are d-separated by C in $\mathcal{G}_2$, and vice versa. Some examples of DAGs that are Markov equivalent are shown in Figure 1.

In fact a simple graphical rule tells us if two DAGs are in the same Markov equivalence class. The *skeleton* of a network is the undirected graph. Two DAGs are in the same equivalence class (observationally equivalent) if they have the same skeleton and the same set of 'v-structures' – the same set of two converging arrows whose tails are not connected by an arrow.

5 Controlling for confounders

Now we know some of the behavior of Bayesian networks we can return to the question of identifying variables that can be controlled for to remove confounding. This means we want to identify variables X such that ignorability holds:

$$Y(1), Y(0) \perp\!\!\!\perp W | X.$$

Note that the observed outcome is of the form $Y = WY(1) + (1 - W)Y(0)$, which induces a conditional dependence between W and Y – the corresponding DAG will have a directed edge from W to Y. Ignorability requires essentially that any *other* paths from W to Y are blocked (i.e. controlled for, conditioned on). Which

choices of X achieve this? Three such criteria are identified below, stated without proof. An example of each is shown in Figure 2.

5.1 Backdoor criterion

If a set of variables X satisfy the following conditions:

1. X blocks every path from W to Y that has an arrow into W (blocks the back door), and
2. No node in X is a descendant of W.

then X satisfies the backdoor criterion with respect to nodes W and Y.

5.2 Disjunctive cause criterion

Sometimes simpler than using the backdoor criterion, which can involve analyzing the entire DAG is the disjunctive cause criterion. It is simply:

- Control for all parents of the treatment variable, the effect variable (that are not descendants of the treatment), or both.

Sometimes this is an easier set to identify than other (potentially smaller) sets that satisfy the backdoor criterion.

5.3 Frontdoor criterion

If a set of variables Z satisfy the following conditions:

1. Z blocks all directed paths from X_i to X_j, and
2. there is no backdoor path from X_i to Z, and
3. all backdoor paths from Z to X_j are blocked by X_i

then Z satisfies the frontdoor criterion with respect to nodes X_i and X_j.

6 Some common methods

Once a set of variables to control for has been identified, how do we actually use this knowledge to identify causal effects? In theory, if we observe controls X then we can measure the causal effect from:

$$\tau = \mathbb{E}(\mathbb{E}(Y|W=1,X) - \mathbb{E}(Y|W=0,X)).$$

In practice however this requires a lot of data to get reliable estimates of each conditional expectation. In biomedical/social science settings this is often an issue. Generally each conditional expectation has to be estimated parametrically to capture the dependence on X. This introduces bias through choice of model, etc. Thus methods that can estimate causal effects without this modeling are attractive. A way of doing this is to effectively match the confound distribution X between the control and treatment groups. Thereby making treatment independent of the covariates, and the data more like what is produced in a randomized control trial. This balancing of distributions among control and treatment groups is achieved through sampling subjects in different ways.

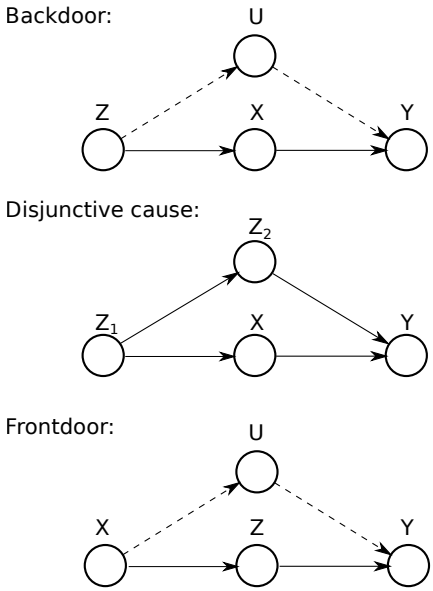

Figure 2: Three criteria through which conditioning on Z will render the effect of X on Y identifiable.

6.1 Matching

The basic idea of matching is as follows. For each condition $W = 1$ and $W = 0$ there are only a finite number of samples:

$$\{y_i^{w=0}, x_i^{w=0}\}_{i=1}^{I_0} \text{ and } \{y_i^{w=1}, x_i^{w=1}\}_{i=1}^{I_1}.$$

Matching simply pairs one sample in the treatment group with one sample in the control group whose control covariates are close:

$$(y_i^{w=0}, x_i^{w=0}) \leftrightarrow (y_j^{w=1}, x_j^{w=1}), \quad x_i^{w=0} \sim x_j^{w=1}.$$

Since between treatment groups X have roughly the same distribution, this dependence does not need to be modeled. This allows the above causal effect expectation to be approximated.

Choices must be made about the metric that is used to decide when two points are similar. And choices must be made about how to deal with different treatment and control population sizes. One possibility is to discard all samples for which no match is made. Another possibility is to match one sample in the treatment group to more than one sample in the control group.

A common way is to match on the treatment group. This then estimates what is known as the *causal effect of treatment on the treated*, often a quantity of interest. If we let $C(i)$ represent the sample index in the control population that is matched to sample i in the treatment population then the causal effect is estimated from:

$$\tau \approx \frac{1}{I_1} \sum_{i=1}^{I_1} y_i^{w=1} - y_{C(i)}^{w=0}.$$

Matching can be performed on all covariates, or just covariates that are identified as confounders, according to the backdoor or other criterion. Note that matching does not remove the need for ignorability – unmeasured confounders can still affect the analysis, thus X still must satisfy the backdoor criteria.

6.2 Propensity score matching

Matching directly on controls X can be difficult if X is high-dimensional. Instead, we can match on what is called the propensity score, which is the probability of being treated given a set of controls:

$$\pi(X) = P(W = 1|X).$$

Matching on $\pi(X)$ has the same effect as matching on X directly. This is because subjects at the same propensity level have, by definition, the same probability of being assigned to the treatment group. Thus, for these subjects, treatment assignment is randomized (independent of X). In this way the distribution of X in treatment and control groups are made to be the same.

The propensity score is known, by definition, in randomized control trials. It has to be estimated in observational studies. But since it only involves observed data X and W this is straightforward. For example, one can use logistic regression.

Again, propensity score matching still requires the ignorability assumption with controls X. Without it, even if the distribution of X is balanced between control and treatment groups, unobserved confounders can still be different amongst control and treatment.

6.3 Inverse probability of treatment weighting

Instead of matching on propensity score, which may discard some samples, we can simply reweight each subject by the inverse of its probability of receiving treatment – known as the inverse probability of treatment weighting (IPTW). This matches one unit in a treatment group with a certain number of 'pseudo-units' in the control group at a rate proportional to the relative probability of receiving treatment at a given level in X. In this way balance is achieved across levels.

This is a type of importance sampling.

7 Causal Bayesian networks

This is the framework developed most significantly by Pearl [1]. A causal model is a Bayesian network along with a mechanism to determine how the model will respond to intervention. Now, rather than using the notion of potential outcomes and counterfactuals, causal effects are measured as the result of intervention. In addition to parents/children, we also think of the directed edges in the DAG as representing causal relationships, meaning a node's parents and children are also its causes and effects.

The *causal Markov condition* is the condition that all nodes are independent of their non-effects, given their direct causes. In the event that the structure of a Bayesian network accurately depicts causality, this is equivalent to the Markov condition. However, a network may accurately embody the Markov condition without depicting causality, in which case it should not be assumed to embody the causal Markov condition.

7.1 Interventions and causal effects

An intervention on a single variable is denoted $\text{do}(X_i = y)$. Intervening on a variable removes the edges to that variable from its parents and forces the variable to take on a specific value: $P(x_i|\text{Pa}_{X_i} = \mathbf{x_i}) = \delta(x_i = y)$. The interventional joint distribution, $P_{X_i=y}$, is then defined as:

$$P_{X_i=y}(\mathbf{x}) \coloneqq \prod_{j \neq i}^{N} P(x_j|\text{Pa}_{X_j} = \mathbf{x}_j)\delta(x_i = y),$$

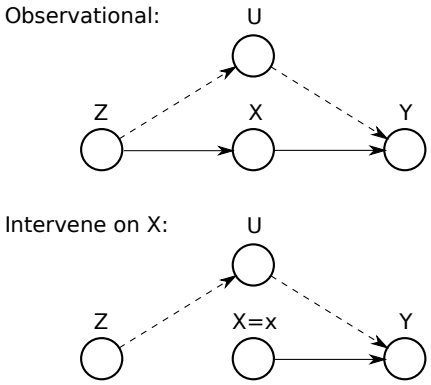

Figure 3: Intervening on X changes the graph and underlying distribution.

also abbreviated to P_{X_i}. Expectations under interventions then take the form:

$$\mathbb{E}(X_j|\text{do}(X_i = y)) \coloneqq \int x_j P_{X_i=y}(x_j)\, dx_j \coloneqq \mathbb{E}_{X_i=y}(X_j).$$

The idea of intervention is shown in Figure 3.

Now, given the ability to intervene, the average causal effect between an outcome variable X_j and a binary variable X_i can be defined as:

$$\tau \coloneqq \mathbb{E}(X_j|\text{do}(X_i = 1)) - \mathbb{E}(X_j|\text{do}(X_i = 0)).$$

In general, the 'do' conditional is different to standard probabilistic conditioning. However criteria exist under which the interventional conditional distribution coincides with the probabilistic conditional distribution. The causal effect from node X_i to X_j can be inferred for conditional distributions that satisfy these criteria. These are actually the same criteria identified above in the counterfactual framework when searching for controls that provide ignorability. The interventional and counterfactual frameworks thus are compatible with one another. Pearl argues the interventional framework subsumes the older counterfactual framework.

For instance, if S_{ij} satisfy the backdoor criteria with respect to $X_i \to X_j$ then we can relate the interventional and observational expectations as follows:

$$\begin{aligned}
\mathbb{E}(X_j|\text{do}(X_i = y)) &= \int x_j P_{X_i=y}(x_j)\, dx_j \\
&= \int\int x_j P_{X_i=y}(x_j|\mathbf{s}_{ij}) P_{X_i=y}(\mathbf{s}_{ij})\, dx_j d\mathbf{s}_{ij} \\
&= \int\int x_j P(x_j|\mathbf{s}_{ij}, X_i = y) P(\mathbf{s}_{ij})\, dx_j d\mathbf{s}_{ij} \\
&= \mathbb{E}\left(\mathbb{E}(X_j|\mathbf{S}_{ij}, X_i = y)\right), \qquad (9)
\end{aligned}$$

from which a causal effect can be measured.

8 Structural equation models

The above frameworks are non-parametric, dealing simply with factorizations of joint distributions. The parametric form of a causal Bayesian network is the structural equation model (SEM). Each node is described by:

$$X_j = f_j(\mathrm{Pa}(X_j), \epsilon_j; \theta_j),$$

for some independent noise variable ϵ_j, and parameters θ_j.

Note that the equality here is of a different nature to an algebraic equality. It conveys assignment rather than comparison. (Similar to the difference between = and == in programming languages.) Some authors use $\leftarrow$ instead of = to communicate this difference. This means that structural equation models have an invariance property that standard statistical models do not: the SEM is robust to intervention. The model should describe the data equally well regardless of whether it comes from observation or interventional experiments.

9 Some further reading

An overview of the counterfactual framework can be found in the short Coursera course. The interventionist framework of Pearl is described in his influential 2000 book. A more modern treatment, based on structural equation models, which in some sense subsume the above two frameworks can be found in Peters et al 2017.

- "A Crash Course in Causality: Inferring Causal Effects from Observational Data" Coursera course. Jason Roy (here at Penn Medicine). `www.coursera.org/learn/crash-course-in-causality/`
- "Causality: Models, Reasoning and Inference" Judea Pearl, 2000.
- "Elements of Causal Inference: Foundations and Learning Algorithms" Jonas Peters, Dominik Janzing and Bernhard Schölkopf, 2017.

References

[1] Judea Pearl. *Causality: models, reasoning and inference.* Cambridge Univ Press, 2000.