1. Introduction

1.1. Causation versus Association

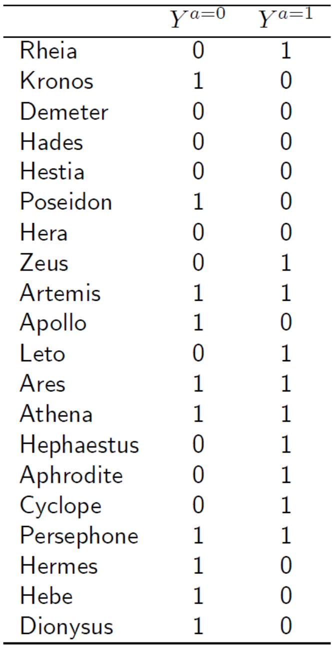
Causation

Consider a dichotomous treatment variable A (1: treated, 0: untreated) and a dichotomous outcome variable Y (1: death, 0: survival), both of which are random variables. Let $Y^{a=1}$ be the outcome variable that would have been observed under the treatment value $a = 1$, and $Y^{a=0}$ the outcome variable that would have been observed under the treatment value $a = 0$, which are also random variables and referred to as *potential outcomes* or *counterfactual outcomes*.

We can now provide a formal definition of ***causal effect for an individual***: The treatment A has a causal effect on an individual's outcome Y if $Y^{a=1} \neq Y^{a=0}$ for the individual.

Therefore, we are now ready to provide a formal definition of the ***average causal effect*** in the population: An average causal effect of treatment A on outcome Y is present if $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$ (dichotomous outcomes) or $\mathrm{E}[Y^{a=1} = 1] \neq \mathrm{E}[Y^{a=0} = 1]$ (non-dichotomous outcomes) in the population of interest.

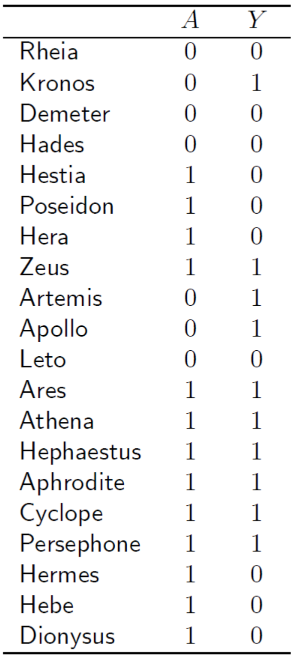
Under this definition, treatment A does not have an average causal effect on outcome Y in our population because both the risk of death under treatment $\Pr[Y^{a=1} = 1]$ and the risk of death under no treatment $\Pr[Y^{a=0} = 1]$ are 0.5.



	$Y^{a=0}$	$Y^{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Leto	0	1
Ares	1	1
Athena	1	1
Hephaestus	0	1
Aphrodite	0	1
Cyclope	0	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

Association

Obviously, the data available from actual studies look different from those shown in the table above. For example, we would not usually expect to learn Zeus's outcome if treated $Y^{a=1} = 1$ and also Zeus's outcome if untreated $Y^{a=0} = 0$. In the real world, we only get to observe one of those outcomes because Zeus is either treated or untreated. We referred to the observed outcome as Y. Thus, for each individual, we know the observed treatment level A and the outcome Y as in the table below.



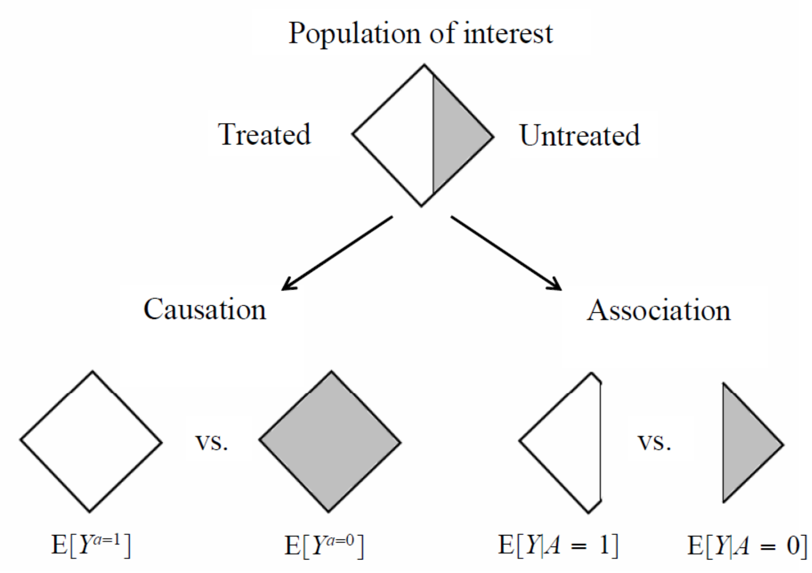
	A	Y
Rheia	0	0
Kronos	0	1
Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Leto	0	0
Ares	1	1
Athena	1	1
Hephaestus	1	1
Aphrodite	1	1
Cyclope	1	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

When the proportion of individuals who develop the outcome in the treated $\Pr[Y = 1|A = 1]$ equals the proportion of individuals who develop the outcome in the untreated $\Pr[Y = 1|A = 0]$, we say that treatment A and outcome Y are *independent*, that A is *not associated with* Y, or that A *does not predict* Y. We say that treatment A and outcome Y are ***dependent*** or ***associated*** when $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$. In our population, treatment and outcome are indeed associated because $\Pr[Y = 1|A = 1] = 7/13$ and $\Pr[Y = 1|A = 0] = 3/7$.

For dichotomous or continuous outcomes Y, we can also rewrite the definition of ***association*** in the population as $\mathrm{E}[Y = 1|A = 1] \neq \mathrm{E}[Y = 1|A = 0]$.

Causation-Association Difference

The figure below depicts the causation-association difference. The population (represented by a diamond) is divided into a white area (the treated) and a smaller grey area (the untreated). The definition of ***causation*** implies a contrast between the whole white diamond (all individuals treated) and the whole grey diamond (all individuals untreated), whereas ***association*** implies a contrast between the white (the treated) and the grey (the untreated) areas of the original diamond. That is, inferences about causation are concerned with ***what if*** questions in *counterfactual worlds*, such as "what would be the risk if everybody had been treated?" and "what would be the risk if everybody had been untreated?", whereas inferences about association are concerned with questions in the *actual world*, such as "what is the risk in the treated?" and "what is the risk in the untreated?"



We can use the notation we have developed thus far to formalize this distinction between causation and association. The risk $\Pr[Y=1|A=1]$ is a conditional probability: the risk of Y in the subset of the population that meet the condition 'having actually received treatment value a' (i.e., $A=a$). In contrast the risk $\Pr[Y^a=1]$ is an unconditional (also known as marginal probability), the risk of Y^a in the entire population. Therefore, ***association*** is defined by a different risk in two disjoint subsets of the population determined by the individuals' actual treatment value ($A=1$ or $A=0$), whereas ***causation*** is defined by a different risk in the same population under two different treatment values ($a=1$ or $a=0$).

These radically different definitions explain the well-known adage "***association is not causation***." In our population, there was association because the mortality risk in the treated ($7/13$) was greater than that in the untreated ($3/7$). However, there was no causation because the risk if everybody had been treated ($10/20$) was the same as the risk if everybody had been untreated. This discrepancy, which is referred to as ***confounding***, would not be surprising if those who received heart transplants were, on average, sicker than those who did not receive a transplant.

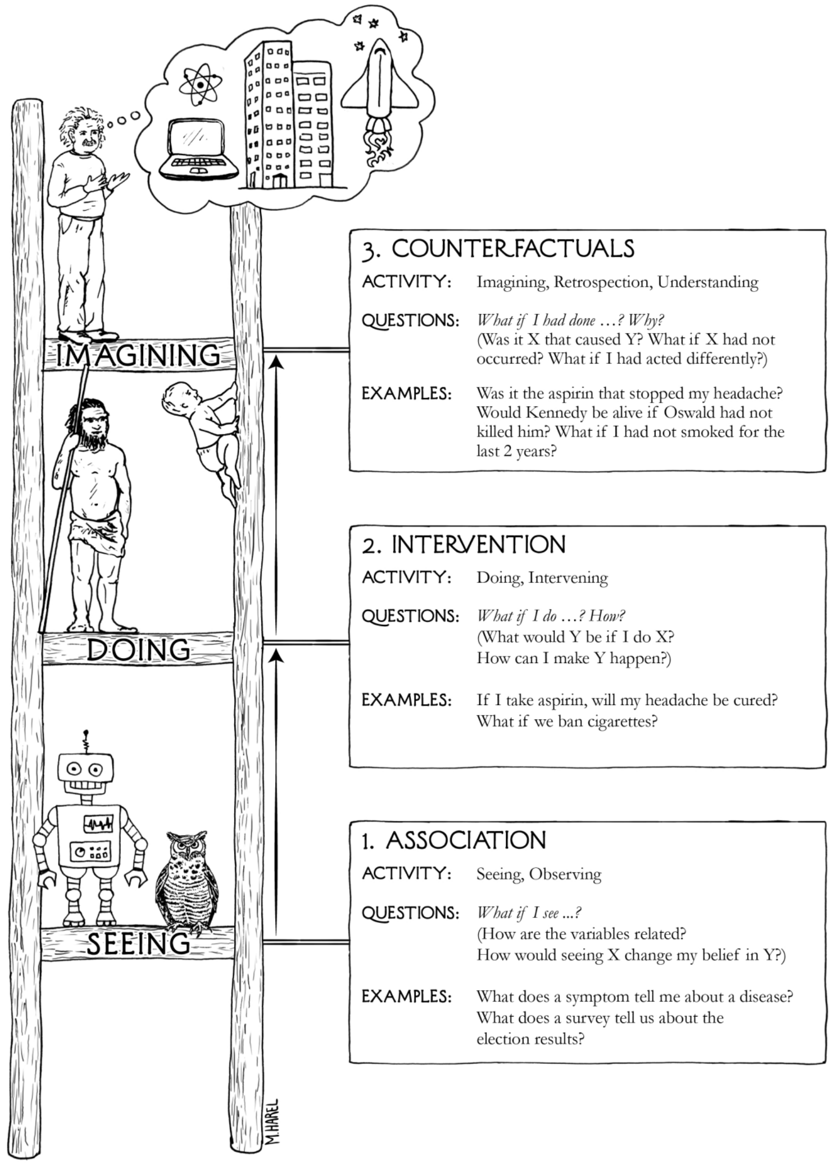
1.2. Why Study Causation?

The questions that motivate most studies in the health, social and behavioral sciences are ***not associational but causal*** in nature. For example,

- What is the efficacy of a given drug in a given population?
- Whether data can prove an employer guilty of hiring discrimination?
- What fraction of past crimes could have been avoided by a given policy?
- What was the cause of death of a given individual, in a specific incident?

These are causal questions because they require **some knowledge of the data-generating process**; they cannot be computed from the data alone, nor from the distributions that govern the data.

1.3. The Ladder of Causation



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?* (What would Y be if I do X? How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?* (How are the variables related? How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

The Ladder of Causation, with representative organisms at each level.

- Most animals, as well as present-day learning machines, are on the first rung, learning from association.
- Tool users, such as early humans, are on the second rung if they act by planning and not merely by imitation.
- We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge. Counterfactual learners, on the top rung, can imagine worlds that do not exist and infer reasons for observed phenomena.

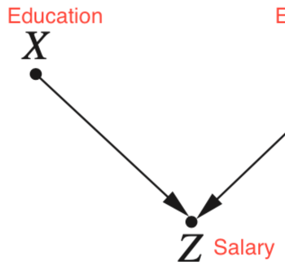
While rung one deals with the seen world and rung two deals with a brave new world that is seeable, rung three deals with a world that cannot be seen (because it contradicts what is seen).

2. Structural Causal Models

Structural causal models (SCMs) provide a comprehensive theory of causality.

2.1. Causal Graph

A causal graph $G = (V, E)$ is a directed graph that describes the causal effects between variables, where V is the node set and E the edge set. In a causal graph, each node represents a random variable including the treatment, the outcome, other observed and unobserved variables. A directed edge $x \rightarrow y$ denotes a causal effect of x on y.



2.2. Product Decomposition

For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(child|parents)$ over all the “families” in the graph. Formally, we write this rule as

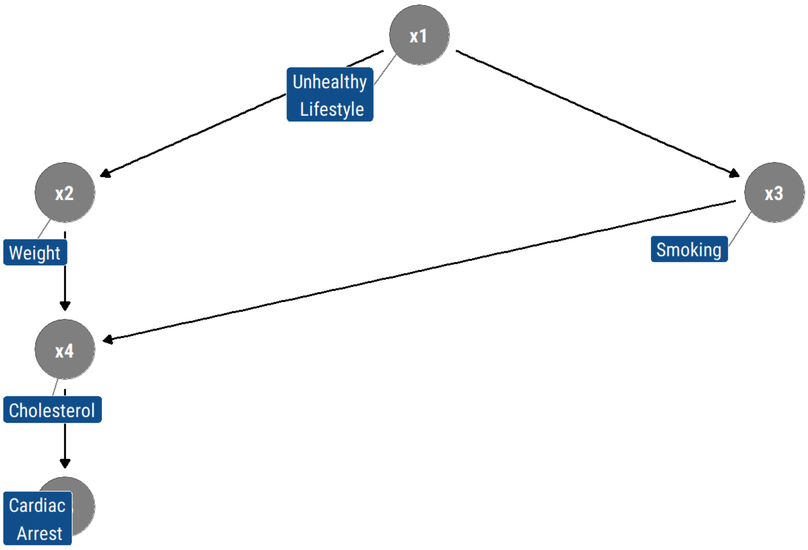
$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i | pa_i)$$

where pa_i stands for the values of the parents of variable X_i.

For example, in a simple chain graph $X \rightarrow Y \rightarrow Z$, we can write directly:

$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y | X = x)P(Z = z | Y = y)$$

This knowledge allows us to save an enormous amount of space when laying out a joint distribution. To estimate the joint distribution from a data set generated by the above model, we need not count the frequency of every triple (x, y, z); we can instead count the frequencies of each x, $(y|x)$, and $(z|y)$ and multiply.



$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4)$$

2.3. $\boldsymbol{d}$-Separation

Given a SCM, the conditional independence embedded in its causal graph provides sufficient information to confirm whether it satisfies the criteria such that we can apply certain causal inference methods. To understand the **conditional independence**, we need the concept of **dependency-separation** (d-separation) based on the definition of **blocked** path.

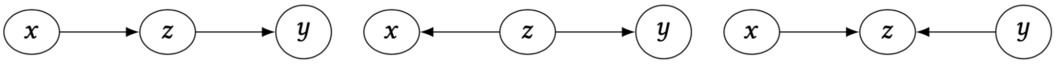
Two events A and B are conditionally independent given a third event C if $P(A|B, C) = P(A|C)$ and $P(B|A, C) = P(B|C)$.

There are three typical DAGs for conditional independence:

- **Chain**: x causally affects y through its influence on z

- **Fork**: z is the common cause of both x and y, that is, x is associated with y but there is no causation between them

- **Collider**: both x and y cause z but there is no causal effect or association between x and y



(a) Chain (b) Fork (c) Collider

In the chain and fork, the path between x and y is blocked, that is, the variables become independent, if we condition on z. Contrarily, in a collider, conditioning on z introduces an association between x and y. Generally, we say conditioning on a set of nodes $\mathcal{Z}$ blocks a path p iff there exists at least one node $z \in \mathcal{Z}$ in the path p.

Definition 1 ($\boldsymbol{d}$-Separation)

A path p is blocked (or d-separated) by a set of nodes $\mathcal{Z}$ if and only if

1. p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in $\mathcal{Z}$ (i.e., B is conditioned on), or

2. p contains a collider $A \leftarrow B \rightarrow C$ such that the collision node B is not in $\mathcal{Z}$, and no descendant of B is in $\mathcal{Z}$.

The probabilistic implications of the d-separation criterion:

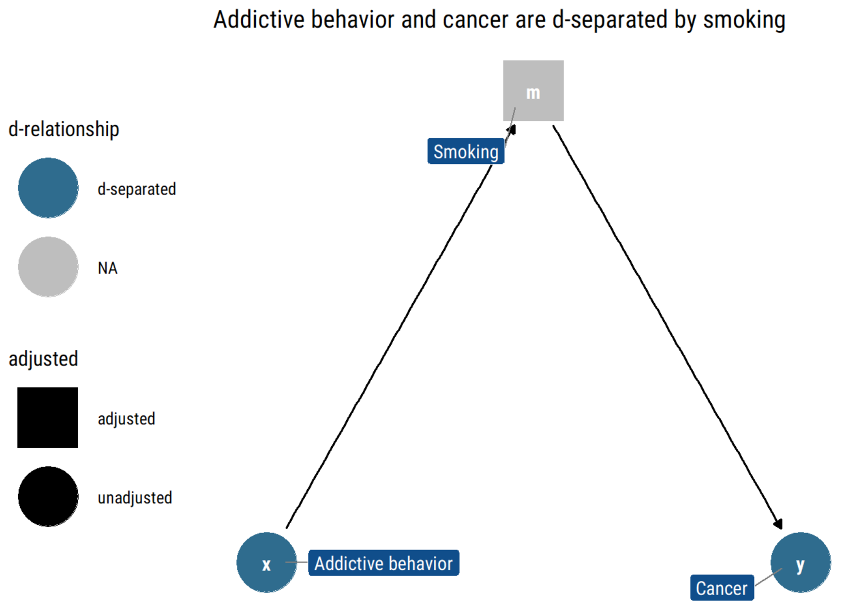
- If X and Y are d-separated by Z in causal graph G, then X is independent of Y conditional on Z in every distribution compatible with G

- Conversely, if X and Y are not d-separated by Z in causal graph G, then X and Y are dependent conditional on Z

Chain

Rule 1 (Conditional Independence in Chains)

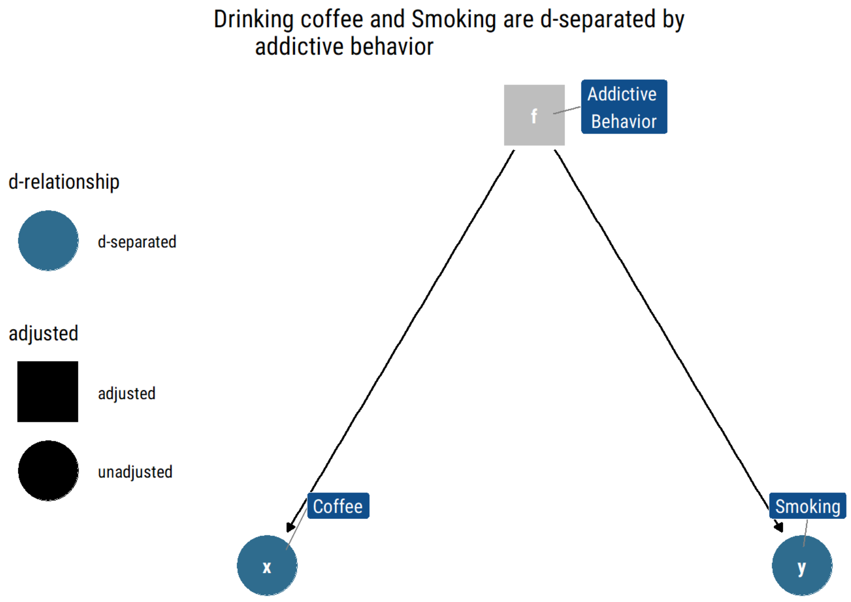
Two variables, X and Y, are conditionally independent given Z, if there is only one unidirectional path between X and Y, and Z is any set of variables that intercepts that path.



Fork

Rule 2 (Conditional Independence in Forks)

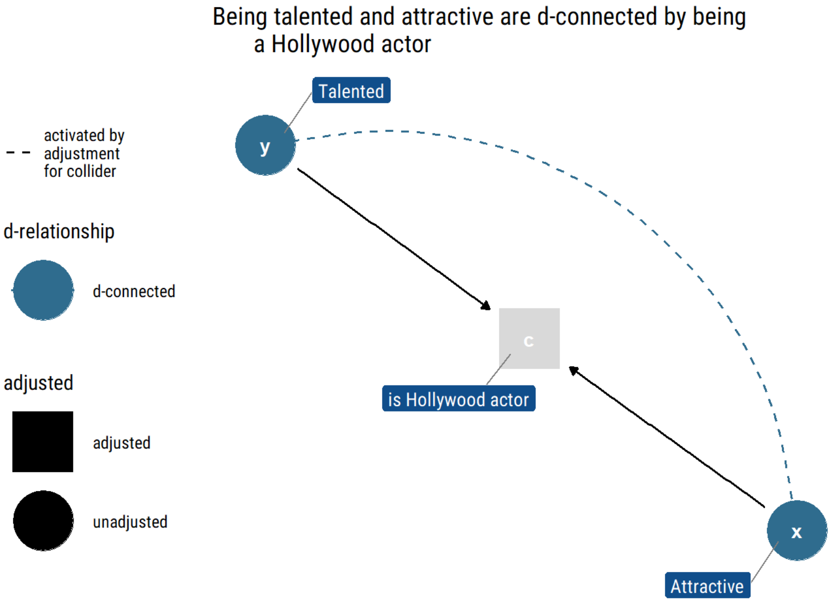
If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.



Collider

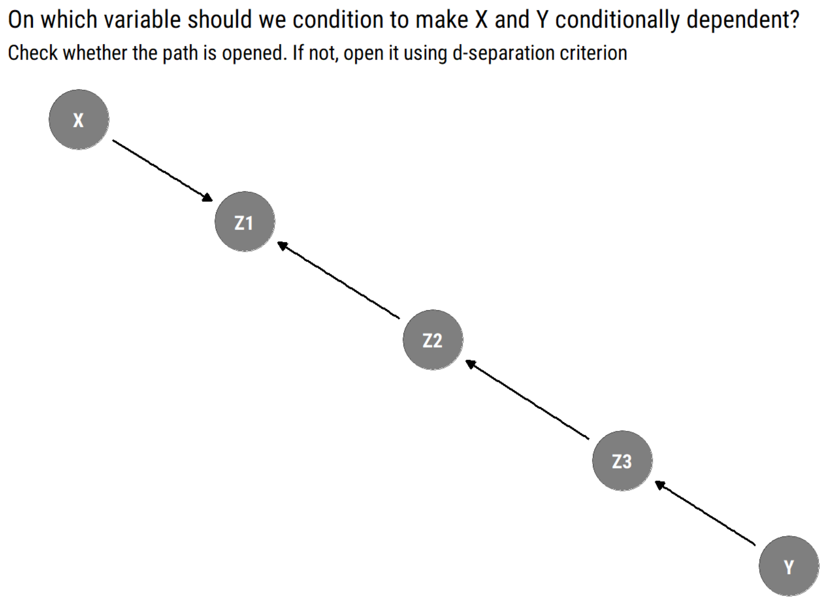
Rule 3 (Conditional Independence in Colliders)

If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendants of Z.


Being talented and attractive are d-connected by being a Hollywood actor
Talented
y
activated by adjustment for collider
d-relationship
d-connected
adjusted
adjusted
unadjusted
c
is Hollywood actor
x
Attractive

Example

Armed with the tool of d-separation, we can now look at some more complex graphical models and determine which variables in them are independent and dependent, both marginally and conditional on other variables.



On which variable should we condition to make X and Y conditionally dependent?

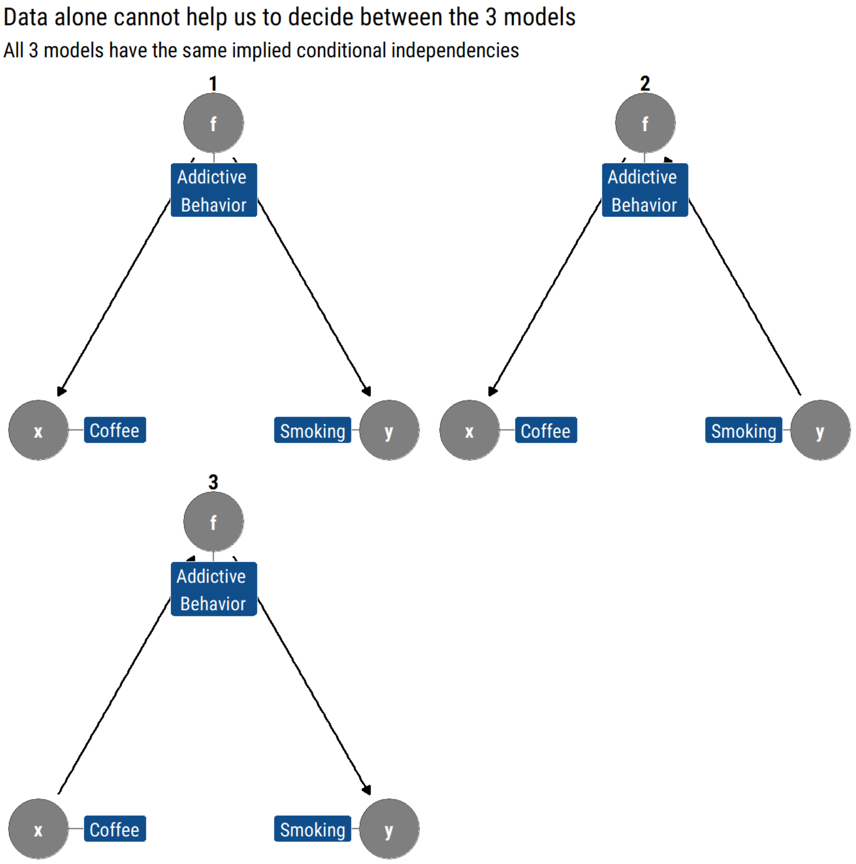
Check whether the path is opened. If not, open it using d-separation criterion

Can we distinguish models from data alone?

The preceding sections demonstrate that causal models have **testable implications** in the data sets they generate. For instance, if we have a graph G that we believe might have generated a data set S, d-separation will tell us which variables in G must be independent conditional on which other variables. Conditional independence is something we can test for using a data set. Suppose we list the d-separation conditions in G, and note that variables A and B must be independent conditional on C. Then, suppose we estimate the probabilities based on S, and discover that the data suggests that A and B are not independent conditional on C. We can then reject G as a possible causal model for S.

But there are other models that imply the same conditional independencies. Two graphical models G_1 and G_2 are observationally equivalent when they imply the same conditional independencies. The set of all the models with indistinguishable implications is called an **equivalence class**.

If two models are observationally equivalent, we **cannot use data alone to distinguish from them**. We must use our structural knowledge about the problem at hand to decide which model is the right one.


Data alone cannot help us to decide between the 3 models
All 3 models have the same implied conditional independencies
1
f
Addictive Behavior
x
Coffee
Smoking
y
2
f
Addictive Behavior
x
Coffee
Smoking
y
3
f
Addictive Behavior
x
Coffee
Smoking
y

As Pearl says, **data are fundamentally dumb**: if we rely only in data to inform our models, we are extremely limited on what we can learn from them. We cannot predict the consequences of intervening (i.e., causal effects) in one of the variables with only observational data. Therefore, we must extend our theory beyond conditional probabilities gain causal understanding with only observational data.

3. The Effects of Interventions

Causal effects cannot be estimated from the data itself without a causal story, but this dost not mean that we cannot gain any causal understanding, far from which it just means that we must have a causal model. With the combination of causal models and observational data, causal effects can be estimated by leveraging the invariant information from the pre-intervention distribution.

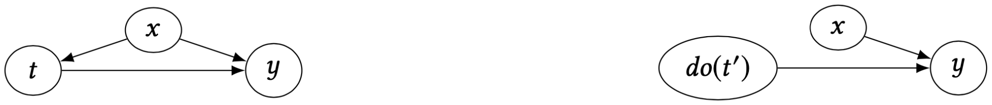
3.1. Interventions

The **do-calculus** of Pearl was proposed to define **intervention** in SCMs. Specifically, the do-calculus introduces a new operator $do(t)$, which denotes the intervention of setting the value of the variable t to t'. The notation of $do(t)$ leads to a formal expression of the interventional distributions:

Definition 2 (Interventional Distribution)

The interventional distribution $P(y|do(x'))$ denotes the distribution of the variable y when we rerun the modified data-generation process where the value of variable x is set to x'.

Consider a study of how Yelp ratings influence potential restaurant customers. Below are SCMs without and under intervention $do(t')$ for the study, where x, t and y denote restaurant category, Yelp rating and customer flow. For the causal graph, the interventional distribution $P(y|do(t'))$ refers to the distribution of variable y as if the rating t is set to t' by intervention, **where all the arrows into t are removed**.



(a) An SCM without intervention.

(b) An SCM under the intervention $do(t)$.

The difference between **intervening** on a variable and **conditioning** on that variable should, hopefully, be obvious. When we intervene on a variable in a model, we fix its value. We change the system, and the values of other variables often change as a result. When we condition on a variable, we change nothing; we merely narrow our focus to the subset of cases in which the variable takes the value we are interested in. What changes, then, is our perception about the world, not the world itself.

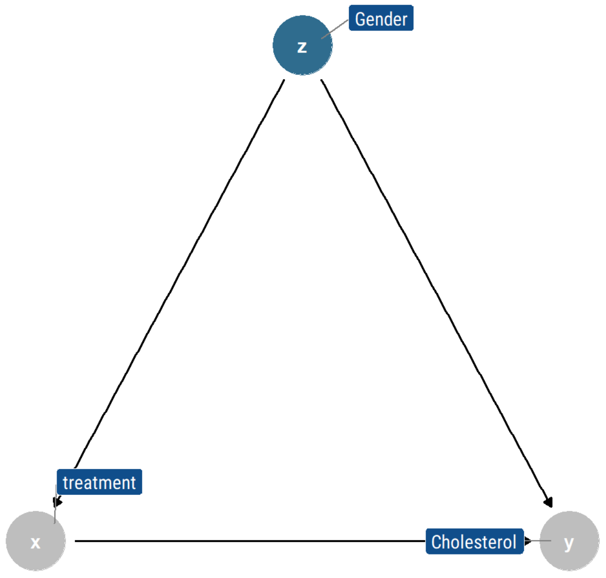
In notation, we distinguish between cases where a variable X takes a value x naturally and cases where we fix $X = x$ by denoting the latter $do(X = x)$. So $P(Y = y|X = x)$ is the probability that $Y = y$ conditional on finding $X = x$, while $P(Y = y|do(X = x))$ is the probability that $Y = y$ when we intervene to make $X = x$. In the distributional terminology, $P(Y = y|X = x)$ reflects the population distribution of Y among individuals whose X value is x. On the other hand, $P(Y = y|do(X = x))$ represents the population distribution of Y if **everyone in the population had their X value fixed at x**. We similarly write $P(Y = y|do(X = x), Z = z)$ to denote the conditional probability of $Y = y$, given $Z = z$, in the distribution created by the intervention $do(X = x)$.

3.2. The Adjustment Formula

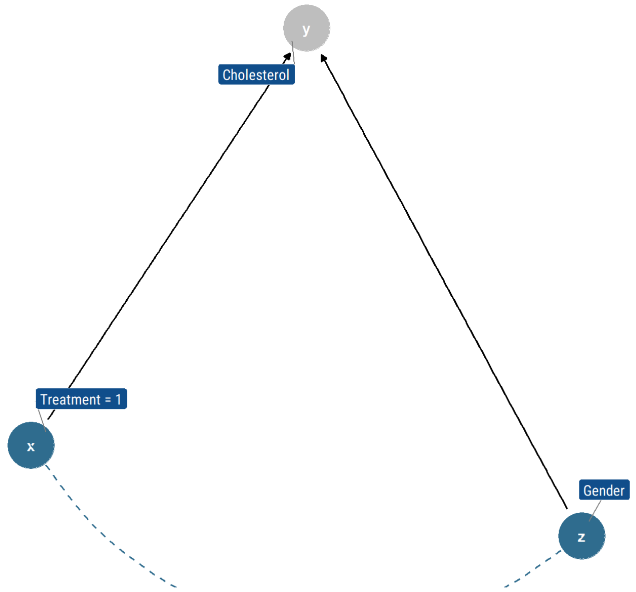
In the classical example of Simpson's paradox, where X stands for drug usage, Y stands for recovery, and Z stands for gender. To find out how effective the drug is in the population, we imagine a hypothetical intervention by which we administer the drug uniformly to the entire population and compare the recovery rate to what would obtain under the complementary intervention, where we prevent everyone from using the drug. Denoting the first intervention by $do(X = 1)$ and the second by $do(X = 0)$, our task is to estimate the difference

$$P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0))$$

which is known as the “causal effect difference,” or “***average causal effect***” (***ACE***).



We know from first principles that causal effects cannot be estimated from the data set itself without a causal story. That was the lesson of Simpson’s paradox: The data itself was not sufficient even for determining whether the effect of the drug was positive or negative. But with the aid of the graph G above, we can compute the magnitude of the causal effect from the data. To do so, we simulate the intervention in the form of a graph surgery (showed in the graph G_m below), then the causal effect $P\left(Y = y\middle|do(X = x)\right)$ is equal to the conditional probability $P_m(Y = y|X = x)$ that prevails in the *manipulated* model, where P is the pre-intervention distribution probability and P_m is the post-intervention distribution probability.



The key to computing the causal effect lies in the observation that P_m, the manipulated probability, shares two essential properties with P (the original probability function that prevails in the pre-intervention model):

- First, the marginal probability $P(Z = z)$ is invariant under the intervention, because the process determining Z is not affected by removing the arrow from Z to X

- Second, the conditional probability $P(Y = y|Z = z, X = x)$ is invariant, because the process by which Y responds to X and Z remains the same, regardless of whether X changes spontaneously or by deliberate manipulation

Therefore, two of the equations of invariance are $P_m(Y = y|Z = z, X = x) = P(Y = y|Z = z, X = x)$ and $P_m(Z = z) = P(Z = z)$.

We can also use the fact that Z and X are d-separated in the modified model and are, therefore, independent under the intervention distribution, which tells us that $P_m(Z = z|X = x) = P_m(Z = z) = P(Z = z)$. Putting these considerations together, we have

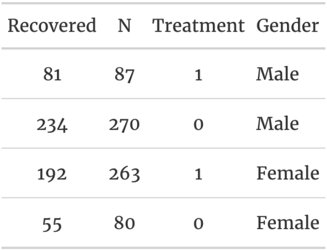
$$\begin{aligned} P\big(Y = y \mid do(X = x)\big) &= P_m(Y = y \mid X = x) \quad \text{(by definition)} \\ &= \sum_z P_m(Y = y \mid X = x, Z = z)\, P_m(Z = z \mid X = x) \\ &= \sum_z P_m(Y = y \mid X = x, Z = z)\, P_m(Z = z) \end{aligned}$$

Finally, using the invariance relations, we obtain a formula for the causal effect, in terms of pre-intervention probabilities:

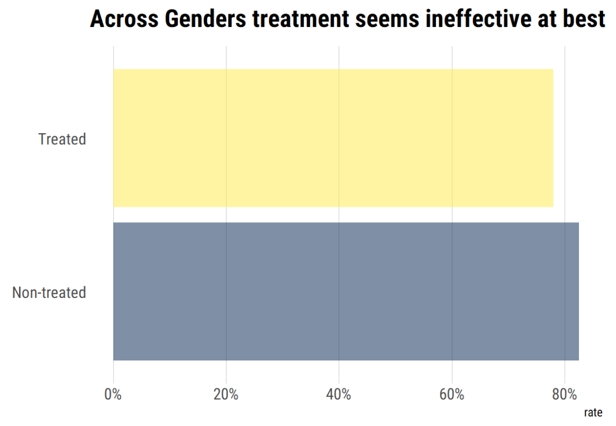
$$P\big(Y = y \mid do(X = x)\big) = \sum_z P(Y = y \mid X = x, Z = z)\, P(Z = z)$$

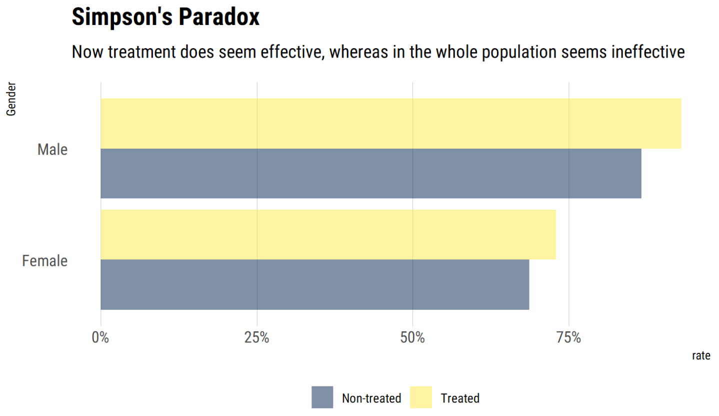
which is called ***adjustment formula***, and it computes the association between X and Y for each value z of Z, then averages over those values. This procedure is referred as “adjusting for Z” or “controlling for Z.”

Example: Simpson’s Paradox



Recovered	N	Treatment	Gender
81	87	1	Male
234	270	0	Male
192	263	1	Female
55	80	0	Female





To demonstrate the working of the adjustment formula, let us apply it numerically to the example, with $X = 1$ standing for the patient treated, $Z = 1$ standing for the patient being male, and $Y = 1$ standing for the patient recovering. We have

$$P\big(Y = 1 \mid do(X = 1)\big) = P(Y = 1 \mid X = 1, Z = 1)P(Z = 1) + P(Y = 1 \mid X = 1, Z = 0)P(Z = 0)$$

Substituting the given figures, we obtain

$$P\big(Y = 1 \mid do(X = 1)\big) = \frac{0.93(87 + 270)}{700} + \frac{0.73(263 + 80)}{700} = 0.832$$

Similarly,

$$P\big(Y = 1 \mid do(X = 0)\big) = \frac{0.87(87 + 270)}{700} + \frac{0.69(263 + 80)}{700} = 0.7818$$

Thus, comparing the effect of treatment ($X = 1$) to the effect of non-treatment ($X = 0$), we obtain

$$ACE = P\big(Y = 1 \mid do(X = 1)\big) - P\big(Y = 1 \mid do(X = 0)\big) = 0.832 - 0.7818 = 0.0502$$

giving a clear positive advantage to treatment. A more informal interpretation of ACE here is that it is simply the difference in the fraction of the population that would recover *if everyone took the drug compared to when no one takes the drug*.

We are now in a position to understand what variable, or set of variables, Z can legitimately be included in the adjustment formula. The intervention procedure, which led to the adjustment formula, dictates that Z should coincide with the parents of X, because it is the influence of these parents that we neutralize when we fix X by external manipulation. Denoting the parents of X by $PA(X)$, we can therefore write a general adjustment formula and summarize it in a rule:

Rule 4 (The Causal Effect Rule)

Given a graph G in which a set of variables PA are designated as the parents of X, the causal effect of X on Y is given by

$$P\big(Y = y \mid do(X = x)\big) = \sum_{z} P(Y = y \mid X = x, PA = z)\, P(PA = z)$$

where z ranges over all the combinations of values that the variables in PA can take.

However, what happens when we do not measure the parents of X and thus cannot adjust for them? Can we adjust for some other observed variable(s)? The question boils down to finding a set of variables that satisfy the ***backdoor criterion***.

Back-door Criterion → Front-door Criterion → Confounding Component