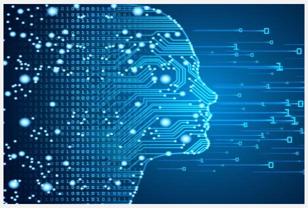
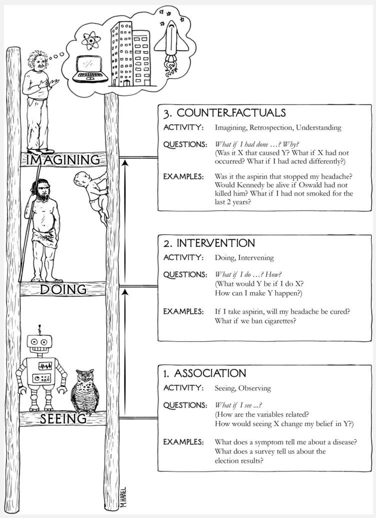


Principles of Causal Inference

Vasant G. Honavar

Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Center for Artificial Intelligence Foundations and Scientific Applications
Institute for Computational and Data Sciences
Huck Institutes of the Life Sciences
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Ladder of Causation



- Seeing: Most animals, learning machines populate the first rung. They learn from association.
- Doing: Tool users, including early humanoids, and perhaps some animals, populate the second rung. They can reason about and learn from interventions.
- Imagining: Humans populate the top rung. They can imagine worlds that do not exist and reason about, and learn from, counterfactuals.

Motivation for Counterfactuals

- Suppose you are driving to NYC
  - You have two options
    - Take freeway $do(X = 1)$
    - Take side road $do(X = 0)$.
  - You take side road
  - You reach NYC, after 6 hours
  - You tell yourself "I should have taken the freeway"
  - What does this really mean?
    - If you had taken the freeway, you would have reached NYC earlier
  - If the If condition (antecedent) is unrealized, we call it a counterfactual
  - Your assertion suggests that whatever slowed down your trip to NYC when you took the side road might not have slowed you down had you taken the freeway

Motivation for Counterfactuals

- After reaching NYC, you tell yourself "I should have taken the freeway"
- You are thinking If you had taken the freeway, you would have reached NYC earlier
- Your assertion is informed by your experience – that it took you 6 hours to reach NYC using side roads (not freeway)
- But when you decided to take the side road, had you anticipated that it would take you 6 hours to get to NYC using the side road, you would have taken the freeway instead if you thought that taking a freeway would get you to NYC in less than 6 hours!
- Suppose we try to express this using a do-expression
$$E(t \mid do(X = 1), t = 6\ hours)$$
- Does this make sense?

Motivation for Counterfactuals

- When you decided to take the side road, had you anticipated that it would take you 6 hours to get to NYC using the side road, you would have taken the freeway instead if you thought that taking a freeway would get you to NYC in less than 6 hours!
- Suppose we try to express this using a do-expression
$$E(t \mid do(X = 1), t = 6 \; hours)$$
- Does this make sense?
- No! We need to distinguish between the hypothetical driving time on freeway when driving on the side road takes 6 hours and the driving time on the side road!

Note on notation

- We use $Y^{a=1}$, $Y_{a=1}$ or Y_1 to denote the value of Y under the intervention $do(A=1)$

Motivating the counterfactuals

- The *do* operator lets us distinguish between $P(t|do(X=0)$ and $P(t|do(X=1)$
- But the *do* operator is too crude to distinguish between the hypothetical time on freeway conditioned on relevant factors and the actual time on side road
- We need a notation to distinguish between
  - Freeway driving time to NYC from SC: $Y_{X=1}$ or Y_1
  - Sideroad driving time to NYC from SC: $Y_{X=0}$ or Y_0
- We need to estimate $E(Y_{X=1} \,|X=0, Y= \; Y_0=1)$
- The expression contains a hypothetical event $Y_{X=1}$ predicated on the event $X=1$, conditioned on a conflicting event $X=0$ that actually occurred (and hence observed)!
- That is, $Y=Y_{X=1}$ and $X=0$ (and $Y=Y_{X=0}$) occur in different worlds!

Motivating the counterfactuals

- We need a notation to distinguish between
  - Freeway driving time to NYC from SC: $Y_{X=1}$ or Y_1
  - Sideroad driving time to NYC from SC: $Y_{X=0}$ or Y_0
- We need to estimate $E(Y_{X=1} \mid X = 0, Y = Y_0 = 1)$
- $Y = Y_{X=1}$ and $X = 0$ (and $Y = Y_{X=0}$) occur in different worlds!
- $E(Y_{X=1} \mid X=0, Y= Y_0 = 1)$ is very different from $E(Y|do(X = 0))$
  - The first is about estimation of a quantity in one world conditioned on observations in another world.
  - The second is about estimation of a quantity in one world conditioned on intervention in the same world.
- We can't reduce the first expression to a do expression
- We can't estimate it from an intervention experiment

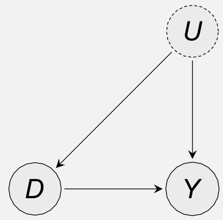
Motivating the counterfactuals

- We can't reduce $E(Y_{X=1} \mid X=0, Y=Y_0=1)$ to a do expression
- Hence, we cannot apply do-calculus!
  - You can only
    - *do* an intervention on everyone in the population (or everyone with the same covariates *X*)
  - However, as the preceding example shows, there are interesting causal questions having to do with individual level counterfactuals that cannot be operationalized using the *do*-operator
- What does it say about the completeness of do-calculus?
- Nothing!
- Why? *do*-calculus is about causal effects in populations, NOT individuals!

Motivating the counterfactuals

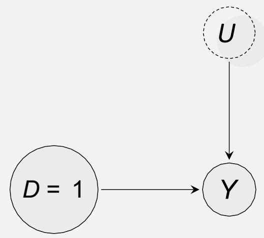
- Can we use an RCT to get at $E(Y_{X=1} \mid X=0, Y = Y_0 = 1)$?
  - An RCT will get us $E(Y \quad \mid do(X=0))$ and $E(Y \quad \mid do(X=1))$
  - An RCT will NOT get us $E(Y_{X=1} \mid X=0, Y = Y_0 = 1)$!
  - Why not?
  - Because X cannot simultaneously be both 1 and 0!
- If we cannot estimate $E(Y_{X=1} \mid X=0, Y = Y_0 = 1)$ from an RCT, there is no hope of estimating it from observational data!
- What if we estimate the freeway driving time for another driver or at another time of the day as a surrogate for your driving time from SC to NYC had you taken the freeway?
  - That would be an approximation
  - The quality of the approximation depends on many factors

Recap: Causal Effects as Interventions



- This model corresponds to the following structural equations
- $D = f_D(U)$
- $Y = f_Y(D, U)$
- What do the graph and the equations look like when we intervene and "do" $D = 1$?

Recap: Causal Effects as Interventions



- This model corresponds to
- $D = 1$
- $Y = f_Y(1, U)$

Recap: Causal Effects as Interventions

- If we $do(D = 1)$, then $D = 1$, and $Y = f_Y(1, U)$
- This Y under $do(D = 1)$ is a function of U and hence differs across individuals
- The mean of Y under the intervention $do(D = 1)$ is:

$$E[Y \mid do(D = 1)] = \sum_u f_Y(1, u) P(U = u)$$

- $f_Y(1, u)$ is Y if D is set to 1 for a unit with infinitely many features u
- This value $f_Y(1, u)$ is in fact a (unit-level) counterfactual
- "What would Y be if D were set to 1 in a unit with feature values u"?

Structural Interpretation of Counterfactuals

- If we $do(D = d)$ in an SCM M,
  - We get the SCM M_d where $D = f_D(u)$ is replaced by $D = d$ The counterfactual value of Y in unit u in model M when D is set to d is $Y_{M_d}(u)$, or $Y_d(u)$ or just Y_d
- The variable Y is passively observed, and the variable Y_d denotes the result of an intervention $D = d$
- This definition of counterfactuals relies on a causal model

Fundamental law of counterfactuals

Counterfactuals obey:

- Consistency: if $D = d$ then $Y_d = Y$
- If D is binary, $Y = D\,Y_1 + (1 - D)\,Y_0$

- Y_1 is the observed value of Y when X is set to 1

Causal Effects using Counterfactuals

- What is then the average causal effect of binary *D* on *Y* using not the *do*-operator, but counterfactuals?
  - $E[Y_1] - E[Y_0]$
- In the literature that uses only counterfactuals but no graphs, this is often called the average treatment effect (of D on Y)
- Nothing new:
  - $E[Y_1] - E[Y_0] = E[Y|do(D = 1)] - E[Y|do(D = 0)]$
- However, we can now also think about causal effects for individuals: $Y_1(u) - Y_0(u)$
- This individual treatment effect will vary across individuals as a function of u

Structural Causal Models Recap

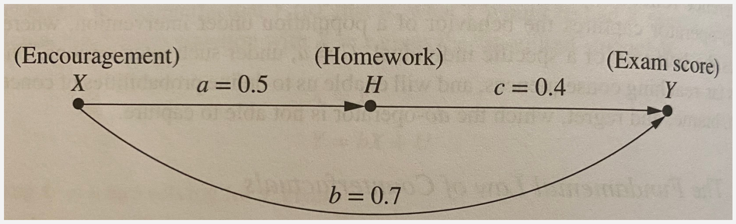
A structural causal model $M = (V, U, F, \mathrm{P}(u))$ where:

- V is a set of endogenous (observed) variables.
- U is a set of exogenous (unobserved) variables.
- F is a set of functions $f : D \to V_i$ where $D \subseteq V \cup U$ and $V_i \in V$.
- $P(u)$ is a probability distribution on U.

Counterfactuals defined

- Suppose M is a structural causal model (V, U, F), exogenous variables U (latent) with known domains.
- $U = u$ implies an individual in the population (e.g., a person, a situation in Nature)
- $X(u)$ denotes the characteristics of an individual with $U = u$
- Counterfactual sentence: Y would be y had D been d
- $Y_d(u) = y$ where Y and D are any variables in V
- We can interpret "had D been d" as an instruction to the causal reasoner to make a minimal modification in the current model so as to establish the antecedent condition $D = d$

From population data to individual behavior



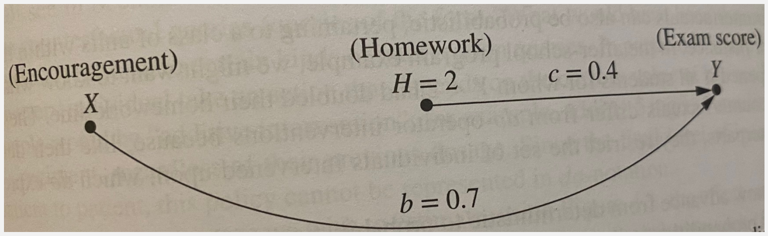
$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

- Suppose Joe has $X = 0.5$, $H = 1$, and $Y = 1.5$

$$0.5 = U_X$$
$$1 = (0.5)(0.5) + U_H$$
$$1.5 = (0.7)\,(0.5) + (0.4)\,(1) + U_Y$$

- We find that
  - $U_X = 0.5$
  - $U_H = 0.75$ and
  - $U_Y = 0.75$

From population data to individual behavior



What happens to Joe's score when we double the homework?
$H = 2, U_X = 0.5, U_H = 0.75$, and $U_Y = 0.75$

$$X = U_X \qquad\qquad Y_{H=2} = (0.7)(0.5) + (2)(0.4) + 0.75$$
$$H = 2 \qquad\qquad = 0.35 + 0.8 + 0.75$$
$$Y = b\,X + c\,H + U_Y \qquad\qquad = 1.90$$

Counterfactuals in Linear Systems

- Structural model $Y = \alpha + \beta D + E$
- This model claims that for every unit u, $Y_d(u) = \alpha + \beta d + E$ so that for every u, $Y_1(u) - Y_0(u) = \beta$
- β is one structural coefficient (identifiable from observational data under certain conditions)
- Given the causal assumptions embodied in this structural causal model, β, the causal effect of D on Y the same for every individual.
- This is almost always wrong
  - If motherhood M affects wages W differently among women
  - We couldn't possibly assert that $W = \alpha + \beta M + E$
- Structural models are not regressions, but the structural coefficients, under certain conditions (which we went over in previous lectures), can be identified from observational data

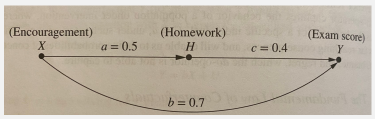
Computing Deterministic Counterfactuals Given a Causal Model

- Abduction[1]: Use evidence $E = e$ to determine the value of (past) U
- Action: Modify the model M, by removing the structural equations for the variables in D by replacing them with $D = d$, to get the modified model M_d
- Prediction: Use the modified Model M_d and the (past) value of U to compute the value of Y_d, the consequence of the counterfactual based on our understanding of the past and the imagined intervention $D = d$
- Counterfactuals, which are taken as primitives in Rubin's potential outcomes framework, are derived properties of structural equation models

- [1] Peirce, C. S. *Collected Papers of Charles Sanders Peirce*, C. Hartshorne, P. Weiss, and A. Burks (ed), 1931–1958, Cambridge MA: Harvard University Press.

Probabilistic counterfactuals

- What if counterfactuals pertain to a subset of individuals in a population?
- It is unlikely that their counterfactual outcomes are identical
  - Example: Effect of motherhood on income of women
- Suppose we wanted to know what would have happened if all students with $Y < 2$ had their homework H doubled?

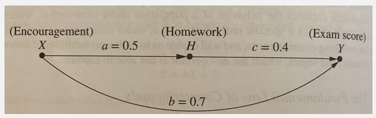


$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

- Can we use do expressions to express such counterfactuals?
- No. Because do expressions cannot restrict the set of individuals intervened in the manner specified

Probabilistic counterfactuals

- Suppose we wanted to know what would have happened if all students with $Y < 2$ had their homework H doubled?



$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

- Suppose Joe's score was less than 2. We want to know what his score have been had his homework been doubled?
- Unlike in the deterministic case, we don't know everything (X, Y, H) about Joe. All we know is that he is in the group with $Y < 2$
- We cannot determine the precise value of $U = \{U_X, U_H, U_Y\}$ for Joe
- $P(U)$ induces a distribution over the observables $\{X, Y, H\}$
- This presents us with the problem of answering probabilistic counterfactual queries

Probabilistic Counterfactual Given a Causal Model

- Given that we observe the feature $E = e$ for a given individual, what is the expected outcome Y for that individual had D been d?
  - That is, we want to know: $E[Y_{D=d}|E = e]$
- Computing the probabilistic counterfactual given a causal model M involves 3 steps:
  - Abduction: Use evidence $E = e$ to update $P(U)$ to $P(U|E = e)$
  - Action: Modify the model M, by removing the structural equations by replacing the variables in D by replacing them with $D = d$, to get the modified model M_d
  - Prediction: Use the modified Model M_d and $P(U|E = e)$ to compute the expectation of Y, the consequence of the counterfactual
- Counterfactuals, which are taken as primitives in Rubin's potential outcomes framework, are derived properties of structural equation models

Example

$X = aU$

$Y = bX + U$

Suppose $a = b = 1$

$U = \{1,2,3\}$

$P(u=1) = \frac{1}{2}$, $P(u=2) = \frac{1}{3}$ and $P(u=3) = \frac{1}{6}$

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	$\boxed{1}$	$\boxed{2}$	2	3	4	1	1	1
2	2	4	$\boxed{3}$	4	5	2	2	2
3	3	6	4	5	6	3	3	3

- $X(1) = (1)(1) = 1$.
- $Y(1) = (1)X(1) + 1 = (1)(1) + 1 = 2$
- How do we compute $Y_1(2)$
- $Y_1(2)$ is the result of intervention setting $X = 1$ on Y with $U = 2$
  - Drop the first Structural equation and set $X = 1$.
  - Use second structural equation to calculate $Y_1(2)$=(1)(1)+2=3

Example

$X = aU$

$Y = bX + U$

Suppose $a = b = 1$

$U = \{1,2,3\}$

$P(u = 1) = \frac{1}{2}$, $P(u = 2) = \frac{1}{3}$ and $P(u = 3) = \frac{1}{6}$

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
$\boxed{1}$	1	2	2	$\boxed{3}$	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

- We can compute the probability that Y would be 3 had X been 2
  - $P(Y_2 = 3)$
  - $Y_2(u) = 3$ occurs only in the first row, when $U = 1$ which occurs with probability $P(1) = 1/2$

Example

$X = aU$

$Y = bX + U$

Suppose $a = b = 1$

$U = \{1,2,3\}$

$P(u=1) = \frac{1}{2}$, $P(u=2) = \frac{1}{3}$ and $P(u=3) = \frac{1}{6}$

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

- We can compute any counterfactual probability
  - $P(Y_2 = 4) = P(U = 2) = 1/3$
- We can compute any joint probability
  - $P(Y_1 < 4, Y_2 > 3) = 1/3$
  - Note that this is a cross-world event spanning $X = 1$ and $X = 2$ which intersect at $U = 2$

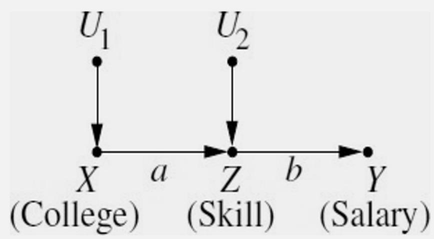
The power of probabilistic counterfactuals

- Given an SCM, we can compute any counterfactual probability
- Given an SCM, we can compute any joint probability over combinations of counterfactuals
  - E.g. $P(Y_1 = y_1 \quad , Y_2 = y_2)$
- This allows us to compute conditional probabilities over counterfactuals and define independence among counterfactuals just as we did over observables
- This is something we cannot do using the $do(X = x)$ notation

Limitation of the do-expressions

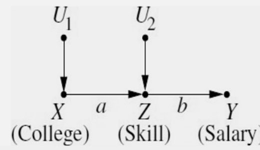
Example model:

- $X = 1$ denotes college educated
- U_2 =1 denotes having work experience
- Z denotes skill level
- Y denotes salary



$$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$$

Limitation of the do-expressions



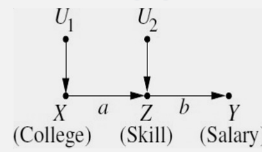
$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$

Suppose we want to compute $\mathbb{E}[Y_{X=1} | Z = 1]$ the expected salary of individuals with skill level $Z = 1$, had they received a college education

- Can we use $\mathbb{E}[Y \mid do(X = 1), Z = 1]$?
- The do-expression stands for the expected salary of individuals who all finished college and have since acquired skill level $Z = 1$.
- The salaries of these individuals, as the SCM shows, depend only on their skill, and are not affected by whether they obtained the skill through college or work experience.
- Conditioning on $Z = 1$, in this case, cuts off the effect of the intervention that we're interested in.

Limitation of the do-expressions

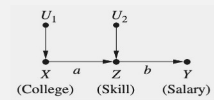
Suppose we want to compute $E[Y_{X=1}|Z = 1]$ the expected salary of individuals with skill level Z = 1, had they received a college education



$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$

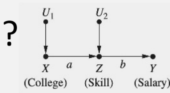
- Can we use $\mathbb{E}[Y \,|do(X = 1), Z = 1]$?
- Conditioning on Z = 1cuts off the effect of the intervention $Z = 1$
- Some of those with skill, i.e., Z = 1 might not have gone to college yet might have attained higher skill (and salary) had they received college education.
- The individuals that are relevant for computing $\mathbb{E}[Y_{X=1}|Z = 1]$ are excluded by the do-expression $\mathbb{E}[Y \,|do(X = 1), Z = 1]$
- Here,
  - $\mathbb{E}[Y \,|do(X = 1), Z = 1] = \mathbb{E}[Y \,|do(X = 0), Z = 1]$ (Z d-separates X from Y)
  - $\mathbb{E}[Y_{X=1}|Z = 1] \neq \mathbb{E}[Y_{X=0}|Z = 1]$

Counterfactual versus do-expression



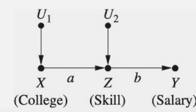
- $\mathbb{E}[Y \mid do(X = 1), Z = 1] = \mathbb{E}[Y \mid do(X = 0), Z = 1]$?
  - Yes, Y only depends on Z
  - Conditioning on Z d-separates X from Y
  - $Z = 1$ refers to current skills; intervention $do(X = 1)$ is about the effect of hypothetical education in an unrealized past, given current skills
- $\mathbb{E}[Y_{X=1} \mid Z = 1] \neq \mathbb{E}\,[Y_{X=0} \mid Z = 1]$?
  - No, $Z = 1$ selects a subset of the population in which we examine the effect of intervening on X
  - $Z = 1$ and $X = 1$ refer to different worlds (pre and post-intervention)

Can counterfactual encode a do-expression?



- Yes. $\mathbb{E}[Y \mid do(X = 1), Z = 1] = \mathbb{E}[Y_{X=1} \mid Z_{X=1} = 1]$
- That is, we condition on the post-intervention value of Z
- $P[Y = y \mid do(X = 1), Z = z] = \frac{P(Y=y, Z=z \mid do(X=1))}{P(Z=z \mid do(X=1))}$

Counterfactual and do Calculations



$X = u_1 \quad Z = aX + u_2 \quad Y = bZ$

u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a+1)b$	1	$a+1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a+1$	$(a+1)b$	b	$(a+1)b$	1	$a+1$

Suppose $a \neq 1, a \neq 0 , ab \neq 0$

$\mathbb{E}[Y_1|Z = 1] = (a+1)b$

$\mathbb{E}[Y_0|Z = 1] = b$

$\mathbb{E}[Y|do(X = 1), Z = 1] = b$

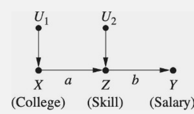
$\mathbb{E}[Y|do(X = 0), Z = 1] = b$

$\mathbb{E}[Y_1 - Y_0 \mid Z = 1] = ab$

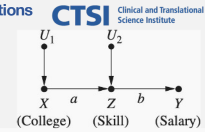
- Even though Z d-separates X from Y, X has a causal effect on Y among those with $Z = 1$
- While the salary of those at skill level $Z = 1$ depends only on their skill and not on education

Counterfactual and do Calculations

- Even though Z d-separates X from Y, X has a causal effect on Y among those with $Z = 1$
- While the salary of those at skill level $Z = 1$ depends only on their skill and not on education X, the salary of individuals currently at skill level $Z = 1$ could have been different had they had a different past
- Dependencies of this sort needed for retrospective reasoning about an unrealized past are not represented in standard structural causal models and cannot be expressed using do expressions
- Performing such reasoning requires augmenting causal graphs with counterfactual variables



Counterfactual and do Calculations



$X = u_1 \quad Z = aX + u_2 \quad Y = bZ$

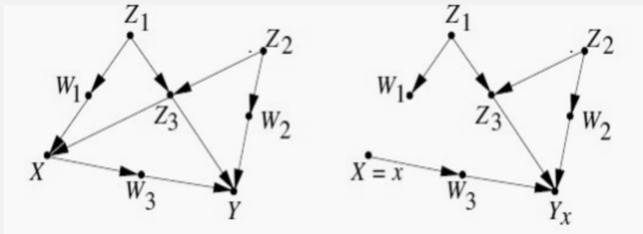
u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a+1)b$	1	$a+1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a+1$	$(a+1)b$	b	$(a+1)b$	1	$a+1$

- With $a \neq 0, a \neq 1$, $P(U_1)$ and $P(U_2)$ do not appear in the calculations because the condition $Z = 1$ occurs only for $u_1 = 0$ and $u_2 = 1$ forcing Y, Y_1 and Y_2 to take a definite value.
- But with $a = 1$, $Z = 1$ occurs when $u_1 = 0$ and $u_2 = 1$ as well as when $u_1 = 1$ and $u_2 = 0$
- $\mathbb{E}[Y_{X=1}|Z = 1] = b\left(1 + \frac{P(u_1=0)P(u_2=1)}{P(u_1=0)P(u_2=1)+P(u_1=1)P(u_2=0)}\right)$
- $\mathbb{E}[Y_{X=0}|Z = 1] = b\left(\frac{P(u_1=0)P(u_2=0)}{P(u_1=0)P(u_2=1)+P(u_1=1)P(u_2=0)}\right)$

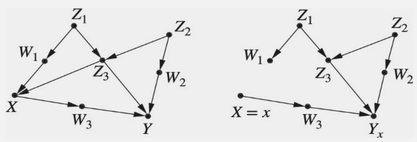
Graphical Representation of Counterfactuals

Can we see counterfactual in our causal model's graph?

Yes. Based on the fundamental law of counterfactuals $Y_x(u) = Y_{M_x}(u)$

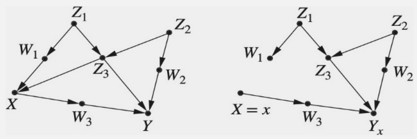


Graphical representation of counterfactuals



- How can we visualize counterfactual Y_{x}?
- Remove arrows going into X yielding M_{x} in which Y is now Y_{x}
- Remember that conditioning Y_{x} on W_3 is a pre-interventional conditioning
- In M_{x}, which variables cause Y to vary?
  - Z_3, W_2, U_{W_3} and U_{y} (conditioning on W_3 couples X and U_{W_3})
- How do we hold X constant?
- We simply remove effect of arrows going into X?
- Condition on variables satisfying the backdoor criterion

The Graphical Representation of Counterfactuals



- What would cause Y_x to vary?
- All exogenous variables capable of influencing Y in M_x
- Observed and unobserved parents of Y, and parents of nodes on the pathways between X and Y
  - Z_3, W_2, U_{W_3} and U_y (the unobserved variables are not shown in the fig)
- If we can identify a set of variables Z in M_x that satisfy the back door criterion with respect to (X, Y_x), we render X independent of Y_x given Z

Counterfactual Interpretation of Backdoor Criterion

- If a set Z of variables satisfies the backdoor condition relative to (X, Y), then the counterfactual Y_{x} is conditionally independent of X given Z

$$P(Y_{\mathrm{x}} | X, Z) = P(Y_{\mathrm{x}} | Z)$$

- How can we calculate $P(y_{\mathrm{x}})$ from data?

$$P(y_{\mathrm{x}}) = \sum_z P(y_{\mathrm{x}} | Z = z)\ P(Z = z)$$

LoT

$$= \sum_z P(y_{\mathrm{x}} | x, Z = z)\ P(Z = z)$$

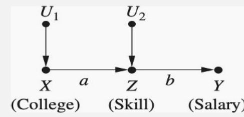
BDC

$$= \sum_z P(y | x, Z = z)\ P(Z = z)$$

Consistency

This is just backdoor adjustment in the counterfactual setting!

Counterfactual Independence



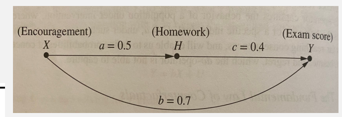
- Does the effect of education on salary (Y_{x}) depend on education (X), given skill Z?

  $$Y_{\mathrm{x}} \perp\!\!\!\perp X \mid Z? \text{ or } E[Y_{\mathrm{x}}|X,Z] = E[Y_{\mathrm{x}}|Z]?$$

- We know $E[Y|X,Z] = E[Y|Z]$
- $\because$ Z blocks all paths from X to Y
- Is the situation different for Y_{x}?
  - Yes!
  - Remove arrows into X to get M_x in which Y is Y_x
  - Which variables cause Y_x to vary when conditioned on Z?
  - U_2 - Why? Because U_2 and X become d-connected when conditioned on Z
  - Hence, $E[Y_{\mathrm{x}}|X,Z] \neq E[Y_{\mathrm{x}}|Z]$
  - In this case, Education matters in estimating the causal effect of Skill (Z) on Salary (Y)!

Counterfactual in Experimental Settings

- We saw that counterfactuals can be answered from a fully specified structural model
- Consider data for 10 students



Participant	Participant characteristics U_X	U_H	U_Y	Observed behavior X	Y	H	Predicted potential outcomes Y_0	Y_1	H_0	H_1	Y_{00} ...
1	0.5	0.75	0.75	0.5	1.50	1.0	1.05	1.95	0.75	1.25	0.75
2	0.3	0.1	0.4	0.3	0.71	0.25	0.44	1.34	0.1	0.6	0.4
3	0.5	0.9	0.2	0.5	1.01	1.15	0.56	1.46	0.9	1.4	0.2
4	0.6	0.5	0.3	0.6	1.04	0.8	0.50	1.40	0.5	1.0	0.3
5	0.5	0.8	0.9	0.5	1.67	1.05	1.22	2.12	0.8	1.3	0.9
6	0.7	0.9	0.3	0.7	1.29	1.25	0.66	1.56	0.9	1.4	0.3
7	0.2	0.3	0.8	0.2	1.10	0.4	0.92	1.82	0.3	0.8	0.8
8	0.4	0.6	0.2	0.4	0.80	0.8	0.44	1.34	0.6	1.1	0.2
9	0.6	0.4	0.3	0.6	1.00	0.7	0.46	1.36	0.4	0.9	0.3
10	0.3	0.8	0.3	0.3	0.89	0.95	0.62	1.52	0.8	1.3	0.3

$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$
$$\sigma_{u_H u_Y} = 0$$

- We used the model to predict the potential outcomes
- In reality, we never can get such data (why?)
- Nevertheless, we can use the model to compute the $\mathbb{E}[Y_{X=1} - Y_{X=0}]$

Counterfactual in Experimental Settings

- Suppose we do not have the model
- But we have data from an experiment in which X is assigned at random to members of the population
- The observed data correspond to the last two columns

Participant	Predicted potential outcomes Y_0	Y_1	Observed outcomes Y_0	Y_1
1	1.05	1.95	1.05	■
2	0.44	1.34	■	1.34
3	0.56	1.46	■	1.46
4	0.50	1.40	■	1.40
5	1.22	2.12	1.22	■
6	0.66	1.56	0.66	■
7	0.92	1.82	■	1.82
8	0.44	1.34	0.44	■
9	0.46	1.36	■	1.36
10	0.62	1.52	0.62	■

True average treatment effect: 0.90 — Study average treatment effect: 0.68

- Now, because X is randomly assigned, the backdoor adjustment formula applies in the counterfactual setting with $Z = \{\ \}$
- $\mathbb{E}[Y_x]$=$\mathbb{E}[Y \ \ |X = x]$

- Because $\mathbb{E}[Y_x]$=$\mathbb{E}[Y \ \ |X = x]$, we can estimate $\mathbb{E}[Y_{X=1} - Y_{X=0}]$ $= \mathbb{E}[Y_{X=1}] - \mathbb{E}[Y_{X=0}]$ from the observed data!
- Note that the quality of the estimate depends on sample size etc.

Applications of Counterfactuals

- Recruitment program
- Additive Interventions
- Personal decision making
- Gender discrimination in hiring
- Mediation and path disabling

Effectiveness of a Job Training Program

- The government funds a job training program aimed at getting jobless people back into the workforce.
- A pilot randomized experiment shows that the program is effective; a higher percentage of people were hired among those who finished the the program than among those who did not enroll in the program.
- The program is approved, and the training is offered to any unemployed person who wants to enroll.
- The hiring rate among those who complete the program turns out even higher than in the randomized pilot study.

Effectiveness of a Job Training Program

- The hiring rate among those who complete the program turns out even higher than in the randomized pilot study.
- Critics: Those who self-enroll, may be more intelligent, more resourceful, and more socially connected than those who were eligible but did not enroll and hence were more likely to have found a job regardless of the training.
- What we really need to estimate is the differential benefit of the program on those enrolled: the extent to which hiring rate has increased among the enrolled, compared to what it would have been had they not enrolled.

Effectiveness of a Job Training Program

- What we really need to estimate is the differential benefit of the program on those enrolled: the extent to which hiring rate has increased among the enrolled, compared to what it would have been had they not enrolled.
- How? Counterfactuals to the rescue!
- Let $X = 1$ represent training and $Y = 1$ represent hiring
- The effect of training on the trained is $\mathbb{E}[Y_{X=1} - Y_{X=0} \mid X = 1]$
- This is called the effect of treatment on the treated

Personal Decision Making

- Cancer patients must decide between two treatments:
  - lumpectomy alone, or
  - lumpectomy plus irradiation
- Ms. Jones, in consultation with her oncologist, decides on the second option. Ten years later, Ms. Jones is alive, and the tumor has not recurred. She speculates: Do I owe my life to irradiation?
- Mrs. Smith, on the other hand, chooses the first option and her tumor recurred after a year. She regrets: I should have gone through irradiation.

Can these speculations ever be substantiated using data?

Discriminatory hiring practices

- Mary files a lawsuit against the TechGigs, alleging discriminatory hiring practices.
- She claims she applied for a job with TechGigs and despite having all the credentials for the job, she was not hired, allegedly because she mentioned, during the course of her interview, that she is gay.
- Moreover, she claims, the hiring record of TechGig shows consistent preferences for straight employees.
- Does she have a case? Can hiring records prove whether TechGig was discriminating when declining her job application?

Mediation path disabling

- A policy maker wishes to assess the extent to which gender disparity in hiring can be reduced by making hiring decisions gender-blind, rather than eliminating gender inequality in education or job training.
- The former concerns the "direct effect" of gender on hiring, whereas the latter concerns the "indirect effect," or the effect mediated via job qualification.

Freedom of choice and causal effects

- In most modern nations, some people may attend university, but are not forced to.
- For those who attend university, does it pay off in terms of lifetime earnings?
- Y lifetime earnings, $D = 1$ if you go to university
- We can model the above using a very simple SCM would be:
- $Y = f_Y(D, U_Y)$ $\quad D = f_D(U_D)$
- U_Y and U_D may be correlated
- Why?
  - You love learning, do well in school; so you attend university, but of course your learning skills also have an effect on your earnings regardless of whether you attend university

Freedom of choice and causal effects

- Perhaps (some) people can (roughly) predict their lifetime earnings depending on their decision to attend or not attend university
- Counterfactuals as potential outcomes: You deliberate an action and its potential outcome; you take an action and experience an actual outcome
    - For those who choose $D = d$, the potential outcome Y_d becomes an actual outcome Y (**consistency**) $\mathbb{E}[Y_d | D = d] = \mathbb{E}[Y | D = d]$:
- You want to know whether "universities pay off".
- You have data on **actual** earnings for people with $D = 1$ and $D = 0$.
- Can you somehow get from this the average causal effect (ATE) of D?
- We know $\mathbb{E}[Y | do(D = 1)] \neq \mathbb{E}[Y | D = 1]$
- Hence $\mathbb{E}[Y_1] = \mathbb{E}[Y | do(D = 1)]$

$$\neq \mathbb{E}[Y | D = 1] = \mathbb{E}[Y_1 | D = 1]$$

Freedom of choice and causal effects

- $\mathbb{E}[Y_1] \neq \mathbb{E}[Y_1 | D = 1]$
  - Problem is dependency between D and potential outcomes (NOT observed outcomes).
  - Is there a substantive interpretation to this dependency?
- Perhaps (some) people can (roughly) predict their lifetime earnings depending on their decision to attend or not attend university
- Extreme case: People know exactly their potential outcomes and choose $D = 1$ if $Y_1 \geq Y_0$
- If people behave this way, we cannot identify the ATE of attending university using only measures of D and Y
- Because treatment D is a function of $Y_d = f_Y(d, U)$, it's a function of $U \Rightarrow$ back-door path from D to Y

$\mathbb{E}$

- Extreme case: People know exactly their potential outcomes and choose $D = 1$ if $Y_1 \geq Y_0$ → unobserved confounding
- Perhaps, in some situations, one may ask people about their potential outcomes. Then no fancy analysis needed, just directly calculate $E[Y_1 - Y_0]$.
- But this is in general, unlikely to be the case
- If individuals have the freedom to choose D and their choice is influenced by their knowledge of potential outcomes, you have unobserved confounding between D and Y
- Thinking in terms of who is choosing treatment and how is extremely useful for analyzing causal problems

Freedom to choose and causal effects: ATT

- "For those who attend university, does it pay off in terms of lifetime earnings?"
- First clause is conditioning filter: People for whom $D = 1$
- But $\mathbb{E}[Y \mid do(D = 1), D = 1] - \mathbb{E}[Y \mid do(D = 0), D = 1]$ does not address this question
  - *do* requires you to **force** people to attend university
  - *do* is not an option when people have the freedom to choose whether to attend university
  - Also the state of affairs $do(D = 0), D = 1$ just does not make sense
- Every person has a counterfactual Y_1 if they go to university, and Y_0 if they don't. Try to write down an expression that captures above question using counterfactuals!
- $\mathbb{E}[Y_1 - Y_0 \mid D = 1]$

Freedom to choose and causal effects: ATT

- $\mathbb{E}[Y_1 - Y_0 | D = 1]$: The average treatment effect on the treated (ATT)
- $\mathbb{E}[Y_1 - Y_0 | D = 1] = \mathbb{E}[Y_1 | D = 1] - \mathbb{E}[Y_0 | D = 1]$
- ATE is relevant if you can/want/must force all people to have
  - $D = 1$ *or* $D = 0$
- ATT is relevant if people can **choose** D!

Relevance of ATT

- In general, if ATT = ATE, individuals do not make choices depending on their potential outcomes/no confounding
- If ATT > ATE, individuals (or some other force) optimizes the choice of treatment with respect to the outcome Y
- If ATT < ATE, people (or someone else) chooses the treatment as to "hurt" people (with respect to Y)!
- Identification of ATT and ATE gives some insight into decision-making processes

Estimation of ATT

- If X satisfies BDC wrt (D, Y), then

$ATT = \mathbb{E}[Y_1 - Y_0 | D = 1] =$

$$\sum_x (\mathbb{E}[Y | D = 1, X = x] - E[Y | D = 0, X = x]) \cdot P(X = x | D = 1)$$

Estimation strategy:

- Bin/discretize continuous X cleverly via coarsened exact matching (CEM, King et al.)
- In each stratum of X, estimate X –specific effect $E[Y | D = 1, X = x] - E[Y | D = 0, X = x]$ via regression
- Estimate $P(X = x | D = 1)$ nonparametrically:

$$P(X = x | D = 1) = \frac{P(X = x, D = 1)}{P(D = 1)} \quad \text{(Bayes' law)}$$

Direct and Indirect Effects

Substantive Examples for (in)direct effects

- Do macroeconomic conditions affect the vote for the incumbent mostly through individual evaluations of the economy?
- Does the incumbency effect exist because strong incumbents scare off high-quality challengers?
- Are hiring processes discriminatory? Is there a direct effect of socio-economic background/gender/race on the probability of being hired?
- Do some genes cause lung cancer only through their effect on smoking behaviour?
- Does Cognitive Behavioral Therapy only work because it leads people to use anti-depressants more often?

History of (in)direct effects

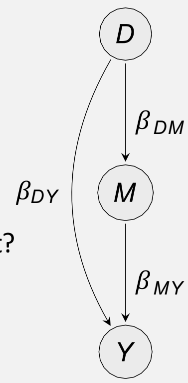
- Being clear about the theoretical causal mechanism is a precondition for a good theory
- Very often, disagreement is not about direction of some causal effect, but about the mechanism
- The methodological literature on how to learn about causal mechanisms from data only started around 2000!
- While this is clearly of interest to **all** sciences
  - Pearl in 1st ed. of "Causality" (2000): "Indirect effects lack intrinsic operational meaning"
  - Rubin (2004): Indirect effects are "ill-defined" and "more deceptive than helpful"
  - Pearl changed his opinion in 2001 and gave general definitions, identification results, and policy implications of indirect effects

Direct and indirect effects in linear models

- What is the direct, what is the indirect effect of *D* on *Y* in this model?
- Direct: β_{DY} , indirect: $\beta_{DM}\beta_{MY}$
- Can you think of the direct effect as a *do*-effect?
- Yes

$$\mathbb{E}[Y \mid do(D = 1), do(M = m)] - \mathbb{E}[Y \mid do(D = 0), do(M = m)]$$

  - Can you think of the indirect effect as a *do*-effect?
  - No: not possible to *do* D so as to isolate $\beta_{DM}\beta_{MY}$
  - As we have seen, linear causal models permit causal estimation using series of linear regressions



From linear to general case

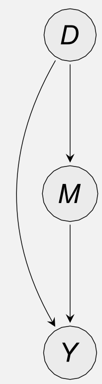
- In linear structural equation models, it is "clear" what direct and indirect effects are: Just look at path coefficients
- However, it is not clear how to generalize this notion to nonlinear models, or unknown functions *f* where we do not even know what "coefficients" they may have
  - Pearl in 1st ed. of "Causality" (2000): "Indirect effects lack intrinsic **operational** meaning" - not possible to write them using *do*-operator; not clear to which **action** or **policy** they refer
  - Pearl 2001, also 2nd ed. Of "Causality (2009): "An indirect effect is the effect of a variable when its direct effect is disabled"

Substantive Examples for path-disabling interventions

- Women $G = 1$ acquire musical skills M. They play in an audition for an orchestra, and then are hired $Y = 1$ with some probability p_1
- Women $G = 1$ acquire musical skills M. They play in an audition **behind a curtain** for an orchestra, and then are hired $Y = 1$ with a different probability p_2
- Playing behind a veil makes sure committee does not know gender, thus disabling the direct effect of gender on hiring
- Goldin 2001 found that introduction of gender-blind audition in various professional US orchestras substantially increased the representation of women in orchestra
- Using only *do*-interventions, we would worry about how to $do(M)$ (increase musical skills, perhaps conditional on gender) which is harder and perhaps more costly

Substantive Examples for path-disabling interventions

- A drug D helps to cure some illness Y , but also results in headache which leads patients to take aspirin M. Aspirin possibly also affects Y
- What happens if drug company manipulates the drug so that it does not cause headache, but direct effect stays the same?
- Disable path $D \rightarrow M$
- Some people may still naturally take aspirin for other reasons, which may moderate the remaining direct effect of D
- If we can identify the direct effect of D not going through M from observational data, we can predict the causal effect of the drug D on illness Y



Path disabling interventions

- Because we are dealing with disabling the influence of one variable on another, we cannot express the effect of such interventions using the do operator
- However, we can express it using the language of counterfactuals
- Suppose we want to assess hiring disparity after putting in place gender-blind hiring procedures
- We require that all female $(X = 0)$ applicants be treated like males $(X = 1)$ with comparable qualifications $(Q = q)$ and proceed to estimate the hiring rate under this counterfactual condition $Y_{X=1,Q=q}$
- But because q varies among applicants, we need to average this according to distribution of q

Path disabling interventions

- Suppose we want to assess hiring disparity after putting in place gender-blind hiring procedures
- We require that all female $(X = 0)$ applicants be treated like males $(X = 1)$ with comparable qualifications $(Q = q)$ and proceed to estimate the hiring rate under this counterfactual condition $Y_{X=1,Q=q}$
- But because q varies among applicants, we need to average this according to distribution of q among females

$$\sum_{q} \mathbb{E}\left[Y_{X=1,Q=q}\right] P(Q = q|X = 0)$$

Path disabling interventions

- Males should have a similar chance of getting hired except that the average is with respect to distribution of q among males

$$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] P(Q=q|X=1)$$

- The indirect effect of gender on hiring as mediated by qualification is given by

$$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] [P(Q=q|X=0) - P(Q=q|X=1)]$$

Path disabling interventions

- The indirect effect of gender on hiring mediated by qualification is given by

$$\sum_q \mathbb{E}\big[Y_{X=1,Q=q}\big]\,[P(Q=q|X=0)-P(Q=q|X=1)]$$

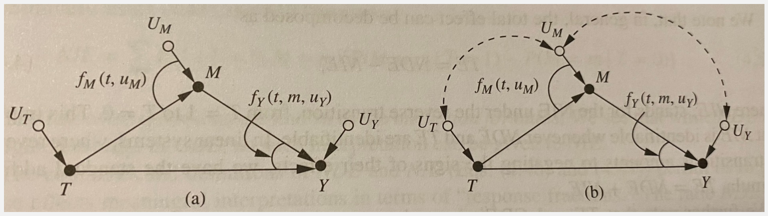
- This is the Natural indirect effect (NIE) of X on Y, mediated by q
- Can we estimate NIE from observational data?
- In the absence of confounding, we can show that

$$\sum_q \mathbb{E}\big[Y_{X=1,Q=q}\big]\,[P(Q=q|X=0)-P(Q=q|X=1)]$$

$$=\sum_q \mathbb{E}[Y|X=1,Q=q]\,[P(Q=q|X=0)-P(Q=q|X=1)]$$

We call this the mediation formula

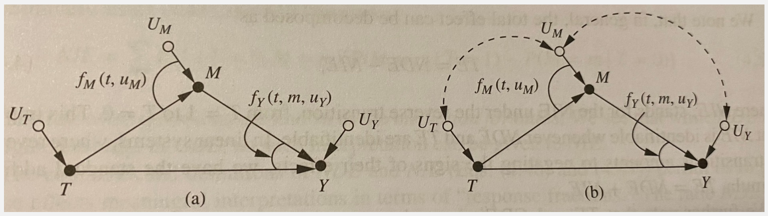
A toolkit for mediation



$t = f_T(u_T)$ $\quad m = f_M(t, u_M)$ $\quad y = f_Y(t, m, u_Y)$

- Total effect $TE = E[Y_1 - Y_0]$
  $= E[Y|do(T = 1) - E(Y|do(T = 0)]$
- TE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is allowed to track the change in T naturally as dictated by f_M

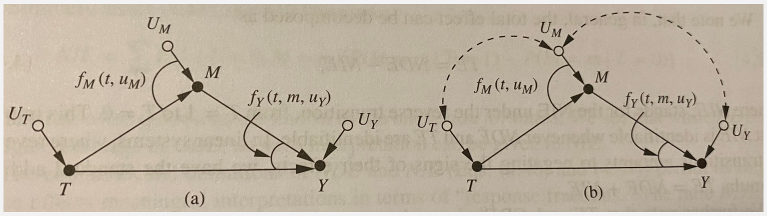
A toolkit for mediation



$t = f_T(u_T)$ $m = f_M(t, u_M)$ $y = f_Y(t, m, u_Y)$

- Controlled direct effect $CDE(m) = E[Y_{1,m} - Y_{0,m}]$
  $= E[Y|do(T = 1, M = m) - E(Y|do(T = 0, M = m)]$
- CDE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to m uniformly for the entire population

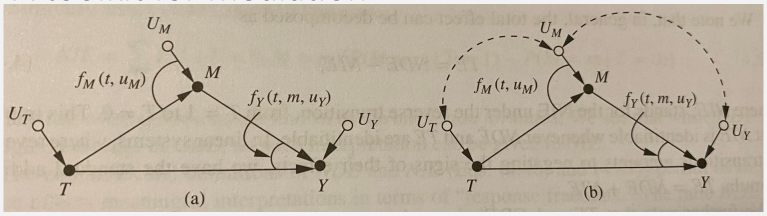
A toolkit for mediation



$t = f_T(u_T)$ $\quad$ $m = f_M(t, u_M)$ $\quad$ $y = f_Y(t, m, u_Y)$

- Natural direct effect $NDE = E[Y_{1,M_0} - Y_{0,M_0}]$
- NDE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to whatever value it would have attained, for each individual, prior to the change, i.e., under $T = 0$

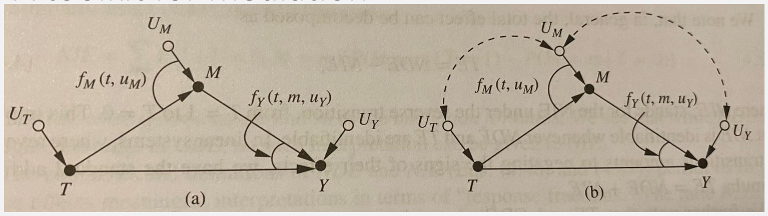
A toolkit for mediation



$t = f_T(u_T)$ $\quad m = f_M(t, u_M)$ $\quad y = f_Y(t, m, u_Y)$

- Natural indirect effect $NIE = E[Y_{0,M_1} - Y_{0,M_0}]$
- NIE measures the expected increase in Y when the treatment is held constant at $T = 0$ while the mediator M changes to whatever value it would have attained, for each individual, under $T = 1$.
- NIE captures, the portion of the effect that can be explained by mediation alone, while disabling the capacity of Y to respond to T

A toolkit for mediation



$t = f_T(u_T)$ $\quad m = f_M(t, u_M)$ $\quad y = f_Y(t, m, u_Y)$

- In general, $TE = NDE - NIE_{\,r}$
- Where $NIE_{\,r}$ denotes the NIE under the reverse change, i.e., $T = 1$ to $T = 0$
- In linear systems, $TE = NDE + NIE$
- Why?
- Because reversal of change flips the sign of the coefficient

Conditions for identifying natural effects

We can identify NDE and NIE provided there exists a set W of measured covariates such that

- No member of W is a descendent of T
- W blocks all backdoor paths from M to Y
- The W-specific effect of T on M is identifiable (possibly using experiments or adjustments)
- The W-specific joint effect of $\{T, M\}$ on Y is identifiable (possibly using experiments or adjustments)

Identification of NDE

- When the first two conditions hold,

$$NDE = \sum_m \sum_w [\mathbb{E}[Y|T=1, M=m, W=w] - \mathbb{E}[Y|T=0, M=m, W=w]].\ P(M=m|T=0, W=w)P(W=w)$$

- What about in the non-confounded case?

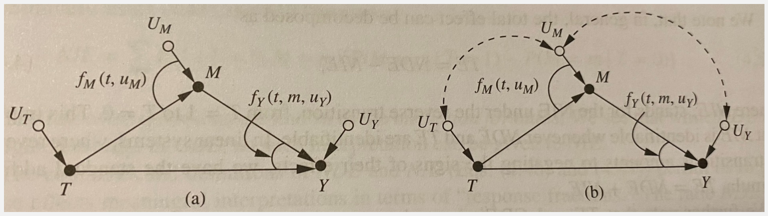
In the non-confounded case, this further reduces to

$$NDE = \sum_m \sum_w [\mathbb{E}[Y|T=1, M=m] - \mathbb{E}[Y|T=0, M=m]]P(M=m|T=0)$$

- Similarly in the non-confounded case,

$$NIE = \sum_m \mathbb{E}[Y|T=0, M=m][P(M=m|T=1) - P(M=m|T=0)]$$

Response fraction due to mediation



- NDE/TE measures the fraction of the response that is transmitted directly, with M frozen
- NIE/TE measures the fraction of the response that is transmitted through M, with Y blinded to T
- $(TE - NDE)/TE$ measures the fraction of the response that is necessarily due to M

Summary: Counterfactuals, path disabling interventions, Mediation

- Mediation is easy to analyze in the case of linear causal models; need to only worry about unobserved confounding
- But in general, causal models may not be linear – we considered the non-parametric setting
- Definition of direct/indirect effects relies on "path-disabling" interventions instead of "variable-setting" or "do" interventions
- Leads to
  - Nested counterfactuals $E[Y_{D=1,M_{D=0}}]$ and definition of "natural" direct (as distinct from controlled direct) and indirect effects.
  - Identification via blocking back-door paths

Summary: Counterfactuals, path disabling interventions, Mediation

- Definition of direct/indirect effects relies on "path-disabling" interventions instead of "variable-setting" or "do" interventions
- Leads to
  - Nested counterfactuals $E[Y_{D=1,M_{D=0}}]$ and definition of "natural" direct (as distinct from controlled direct) and indirect effects.
  - Identification via blocking back-door paths
- Estimand is very similar to ATT/ATE estimand:
  - For all values m of mediator M compute mean differences in Y for different values d of treatment;
  - Weight each difference with distribution of M under "baseline" value of D
- Nonparametric identification of NDE/NIE impossible in the presence of post-treatment confounders