Methods for Causal Inference
Lecture 16

Ava Khamseh
School of Informatics

2021-2022

Counterfactuals: Separate identification from estimation

Defining the counterfactual should not require approximation. Definitions should accurately capture what we wish to estimate precisely.
(How we then estimate it is a different problem).

Defining and estimating counterfactual allows us to address complex problems:

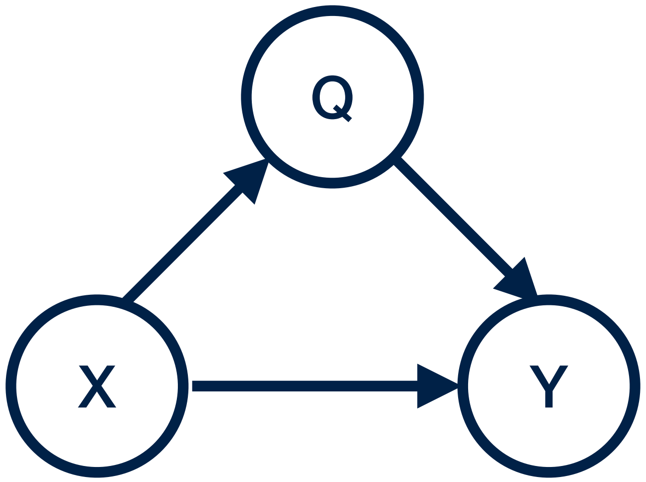
- efficacy of a job training programme by identifying how may enrolled would have gotten jobs had they not enrolled
- Predict the effect of an additive intervention (adding 5 mg/l of insulin to a group of patient with varying insulin levels), from experimental studies
- Obtain the likelihood that an individual cancer patient would have had a different outcome, had they chosen a different treatment

Mediation and Path-disabling Interventions

Example 4.4.5 *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.*

Aim: Which of the two causal effects is greater (i) the direct effect (gender on hiring), or (ii) the indirect effect (education on job qualification on hiring)?
—> Could inform policy where to invest resources to address disparity

X: gender
Q: job qualification
Y: hiring decision



Pearl's Primer book Chapter 4, page 114

Mediation: CDE

Controlled Direct Effect (CDE):

$$\boxed{p(Y=y|do(X=x),do(Q=q))-p(Y=y|do(X=x'),do(Q=q))}$$

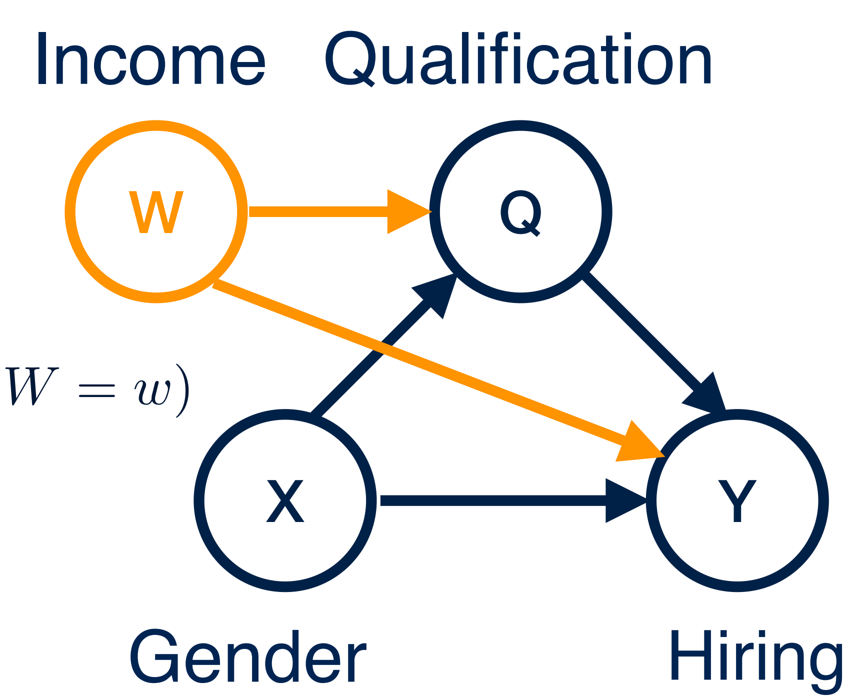
There are no backdoor paths from T to Y, hence the above is equal to:

eliminate the do(x)

$$p(Y=y|X=x,do(Q=q))-p(Y=y|X=x',do(Q=q))$$

There are 2 back-door paths from Q to Y in the original graph:

1) through gender X, which is blocked by X
2) through income W, so we condition on W



eliminate do(q)

$$\sum_w \Big(p(Y=y|X=x,Q=q,W=w)-p(Y=y|X=x',Q=q,W=w)\Big)p(W=w)$$

Mediation and Path-disabling Interventions

Example 4.4.5 *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.*

Aim: Which of the two causal effects is greater (i) the direct effect (gender on hiring), or (ii) the indirect effect (education on job qualification on hiring)?
—> Could inform policy where to invest resources to address disparity

This concerns enabling/disabling processes (e.g., educational reforms) rather than lowering/raising values of specific variables. Thus, the do-operator and the controlled direct effect (CDE) seen earlier do not suffice …

… as before, we phrase the problem mathematically via counterfactuals!

Pearl's Primer book Chapter 4, page 114

Mediation and Path-disabling Interventions

How do we phrase this in a counterfactual manner?

For example, we want to know how the gender disparity changes *after* successfully implementing gender-blind hiring procedures.

In words: We estimate gender disparity under the counterfactual condition that all female applicants be treated as males

Mediation and Path-disabling Interventions

How do we phrase this in a counterfactual manner?

For example, we want to know how the gender disparity changes *after* successfully implementing gender-blind hiring procedures.

In words: We estimate gender disparity under the counterfactual condition that all female applicants be treated as males

Hiring status (Y) of a female applicant with qualification Q = q, given that the employer treats her as though she is a male (X=1) is captured by the counterfactual $Y_{X=1,Q=q}$

Since Q varies over the population, we average this quantity according to the distribution of the qualification of female applicants, $p(Q = q|X = 0)$

Mediation and Path-disabling Interventions

Since Q varies over the population, we average this quantity according to the distribution of the qualification of female applicants, $p(Q = q | X = 0)$

The result is $$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] p(Q = q | X = 0)$$

Male applicants have similar chances, but averaging over $p(Q = q | X = 1)$

Mediation and Path-disabling Interventions

Since Q varies over the population, we average this quantity according to the distribution of the qualification of female applicants, $p(Q = q | X = 0)$

The result is $$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] p(Q = q | X = 0)$$

Male applicants have similar chances, but averaging over $p(Q = q | X = 1)$

Subtracting the two quantities yields the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q:

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

the difference between different gender

Allow Q to vary naturally between applicants, as opposed to the CDE. Here we disable the capacity of Y to respond to X but leave its response to Q unaltered.

$$\boxed{p(Y = y | do(X = x), do(Q = q)) - p(Y = y | do(X = x'), do(Q = q))}$$

Pearl's Primer book Chapter 4, page 115

Mediation and Path-disabling Interventions

It remains to identify the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q, in order to allow estimation:

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

The following result is known as Pearl's ***Mediation formula***

Mediation and Path-disabling Interventions

It remains to identify the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q, in order to allow estimation:

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

The following result is known as Pearl's ***Mediation formula***

Theorem (Pearl, 2001)

In the absence of confounding, the NIE can be identified as follows

$$\mathrm{NIE} = \sum_q \mathbb{E}\left[Y|X=1, Q=q\right]\left(p(Q=q|X=0) - p(Q=q|X=1)\right)$$

In words: It measures the extent to which the effect of X on Y is *explained* by its effect on the mediator Q. In the NIE we "freeze" the direct effect of X on Y, yet allow the mediator Q of each unit to react to X in a natural "unfrozen" way.

Pearl's Primer book Chapter 4, page 115

Mathematical toolkit for Attribution and Mediation

The various applications of counterfactuals we have seen share many features in their mathematical description. Examples are:

1. **ETT (Effect of Treatment on the Treated)**, i.e., $\mathbb{E}[Y_x \mid X = x']$
   Showed up in questions related to *recruitment to a programme* and *additive interventions*

Mathematical toolkit for Attribution and Mediation

The various applications of counterfactuals we have seen share many features in their mathematical description. Examples are:

1. **ETT (Effect of Treatment on the Treated)**, i.e., $\mathbb{E}[Y_x \mid X = x']$
   Showed up in questions related to *recruitment to a programme* and *additive interventions*
2. **Probability of necessity**, i.e., $\mathrm{PN} = P(Y_0 = 0 \mid X = 1, Y = 1)$
   In words: "Had Y not happened in case X was 0 (i.e., Y_0=0), i.e., was treatment (X=1) necessary to obtain Y=1?"
   Showed up in *the cancer treatment example* and *legal liability*

Mathematical toolkit for Attribution and Mediation

The various applications of counterfactuals we have seen share many features in their mathematical description. Examples are:

1. **ETT (Effect of Treatment on the Treated)**, i.e., $\mathbb{E}[Y_x \mid X = x']$
   Showed up in questions related to *recruitment to a programme* and *additive interventions*
2. **Probability of necessity**, i.e., $\mathrm{PN} = P(Y_0 = 0 \mid X = 1, Y = 1)$
   In words: "Had Y not happened in case X was 0 (i.e., Y_0=0), i.e., was treatment (X=1) necessary to obtain Y=1?"
   Showed up in *the cancer treatment example* and *legal liability*
3. **Nested counterfactual expression**, i.e., $\mathbb{E}\left[Y_{x,M_{x'}}\right]$
   In words: "The expected outcome (Y) had the treatment been X=x, and, simultaneously, had the mediator M attained the valued $M_{x'}$ it would have attained had X been x'. "
   This is the key quantity in *mediation*

Pearl's Primer book Section 4.5

Mathematical toolkit: **Attribution**

First, we consider the attribution of cause

To keep things clear yet precise we consider binary events:

- X = x and Y = y represent treatment and outcome respectively
- X = x' and Y = y' represent their negations (no treatment / negative outcome)

Our target quantity is the probability of necessity:

"Find the probability that if X had been x', Y would be y', given that, in reality, X is x and Y is y"

In these variables, the probability of necessity reads

$$\mathrm{PN}(x, y) = P(Y_{x'} = y' \mid X = x, Y = y)$$

Mathematical toolkit: **Attribution**

This counterfactual quantity captures the legal criterion of “but for”

Example

The probability that the damage would not have occurred had the action not been taken ($Y_0 = 0$) given that, in fact, the damage did occur ($Y = 1$) and the action was taken ($X = 1$).

In this case, the plaintiff has to argue that “it is more probable than not that the damage would not have occurred *but for* the actions of the defendant.”

We now consider conditions under which this “but for”, the probability of necessity, can be identified from empirical studies.

Attribution: Identification

The following identification result can be found in [Pearl, 2000, Chapter 9].

Theorem

If Y is monotonic relative to X, $Y_x(u) \geq Y_{x'}(u)$ for all u for x > x' (e.g., additional chemotherapy can remove the cancer but never prevent removal), and if the causal effect $p(Y = y|\mathrm{do}(X = x))$ is identifiable, then PN is identifiable and

$$\mathrm{PN} = \frac{p(y) - p(y|\mathrm{do}(x'))}{p(x, y)}$$

Attribution: Identification

The following identification result can be found in [Pearl, 2000, Chapter 9].

Theorem

If Y is monotonic relative to X, $Y_x(u) \geq Y_{x'}(u)$ for all u for x > x' (e.g., additional chemotherapy can remove the cancer but never prevent removal), and if the causal effect $p(Y = y|\mathrm{do}(X = x))$ is identifiable, then PN is identifiable and

$$\mathrm{PN} = \frac{p(y) - p(y|\mathrm{do}(x'))}{p(x, y)}$$

Equivalently, using the total law $p(y) = p(y|x)p(x) + p(y|x')(1 - p(x))$

$$\mathrm{PN} = \frac{p(y|x) - p(y|x')}{p(y|x)} + \frac{p(y|x') - p(y|\mathrm{do}(x'))}{p(x, y)}$$

Note: The required causal effect can be estimated from randomised trials or from observational data, e.g., using the backdoor criterion

Attribution: Identification

This second expression has a helpful interpretation

Example: Suppose there is a case brought against a car manufacturer, claiming that its car's faulty design led to a man's death in a car crash.

$$\mathrm{PN} = \boxed{\frac{p(y|x) - p(y|x')}{p(y|x)}} + \frac{p(y|x') - p(y|\mathrm{do}(x'))}{p(x,y)}$$

Excess Risk Ratio (ERR) or **Attributable Risk Fraction among the exposed**
It tells us how much more likely people are to die in crashes when driving one of the manufacturer's cars (X=x) than not (X=x')

Attribution: Identification

This second expression has a helpful interpretation

Example: Suppose there is a case brought against a car manufacturer, claiming that its car's faulty design led to a man's death in a car crash.

$$\mathrm{PN} = \boxed{\frac{p(y|x) - p(y|x')}{p(y|x)}} + \boxed{\frac{p(y|x') - p(y|\mathrm{do}(x'))}{p(x,y)}}$$

Excess Risk Ratio (ERR) or **Attributable Risk Fraction among the exposed**
It tells us how much more likely people are to die in crashes when driving one of the manufacturer's cars (X=x) than not (X=x')

Confounding Factor (CF): This factor corrects for confounding bias due to confounding of the causal effect of X on Y, i.e., when $p(y|x') \neq p(y|\mathrm{do}(x'))$
E.g. People buying the manufacturer's cars are more likely to drive too fast

Pearl's Primer book Section 4.5.1

Example: Attribution in Legal Setting

Lawsuit against: the manufacturer of a drug x

Charge: drug x is likely to have accuse the death of Mr A, who took it to relive back pains.

Manufacturer's defence: Experimental data for patients with back pains show conclusively that drug x has only minor effects on death.

Pearl's Primer book Section 4.5.1

Example: Attribution in Legal Setting

Lawsuit against: the manufacturer of a drug x
Charge: drug x is likely to have accuse the death of Mr A, who took it to relive back pains.
Manufacturer's defence: Experimental data for patients with back pains show conclusively that drug x has only minor effects on death.

Plaintiff argues: Experimental data is not relevant here because it represents average effects on patients in the study, not patients like Mr A, who did not participate in the study. In particular, Mr A used the drug of his own volition, unlike subject in the experimental study who took the drug to comply with the experimental protocols. The plaintiff then provides non-experimental (observational) data for patients similar to Mr A who chose drug x to relive back pains but were not part of such experiments, and experienced lower death rates than those who didn't take the drug.

Pearl's Primer book Section 4.5.1

Example: Attribution in Legal Setting

The court must now decide: Based on experimental and non-experimental data, is it "**more probable than not**" that drug x was in fact the cause of Mr A's death.

Experimental:

$$P(y|do(x)) = 16/1000 = 0.016$$
$$P(y|do(x')) = 14/1000 = 0.014$$

Non-experimental

$$P(y) = 30/2000 = 0.015$$
$$P(x, y) = 2/2000 = 0.001$$
$$P(y|x) = 2/1000 = 0.002$$
$$P(y|x') = 28/1000 = 0.028$$

Table 4.5 Experimental and nonexperimental data used to illustrate the estimation of PN, the probability that drug x was responsible for a person's death (y)

	Experimental		Nonexperimental	
	$do(x)$	$do(x')$	x	x'
Deaths (y)	16	14	2	28
Survivals (y')	984	986	998	972

Pearl's Primer book page 119

Example: Attribution in Legal Setting

The court must now decide: Based on experimental and non-experimental data, is it "**more probable than not**" that drug x was in fact the cause of Mr A's death.

Experimental:

$$P(y|do(x)) = 16/1000 = 0.016$$
$$P(y|do(x')) = 14/1000 = 0.014$$

Non-experimental

$$P(y) = 30/2000 = 0.015$$
$$P(x,y) = 2/2000 = 0.001$$
$$P(y|x) = 2/1000 = 0.002$$
$$P(y|x') = 28/1000 = 0.028$$

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)}$$
$$= \boxed{\frac{0.002 - 0.028}{0.002}} + \frac{0.028 - 0.014}{0.001} = \boxed{-13} + 14 = 1$$

Negative observational ERR: gives the impression that the drug is preventing death

Example: Attribution in Legal Setting

The court must now decide: Based on experimental and non-experimental data, is it "**more probable than not**" that drug x was in fact the cause of Mr A's death.

Experimental:

$$P(y|do(x)) = 16/1000 = 0.016$$
$$P(y|do(x')) = 14/1000 = 0.014$$

Non-experimental

$$P(y) = 30/2000 = 0.015$$
$$P(x, y) = 2/2000 = 0.001$$
$$P(y|x) = 2/1000 = 0.002$$
$$P(y|x') = 28/1000 = 0.028$$

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x, y)}$$
$$= \boxed{\frac{0.002 - 0.028}{0.002}} + \boxed{\frac{0.028 - 0.014}{0.001}} = \boxed{-13} + \boxed{14} = 1$$

Bias-correction term rectifies this impression!

Example: Attribution in Legal Setting

The court must now decide: Based on experimental and non-experimental data, is it "**more probable than not**" that drug x was in fact the cause of Mr A's death.

Experimental:

$$P(y|do(x)) = 16/1000 = 0.016$$
$$P(y|do(x')) = 14/1000 = 0.014$$

Non-experimental

$$\begin{aligned} P(y) &= 30/2000 = 0.015 \\ P(x, y) &= 2/2000 = 0.001 \\ P(y|x) &= 2/1000 = 0.002 \\ P(y|x') &= 28/1000 = 0.028 \end{aligned}$$

$$\begin{aligned} PN &= \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x, y)} \\ &= \boxed{\frac{0.002 - 0.028}{0.002}} + \boxed{\frac{0.028 - 0.014}{0.001}} = \boxed{-13} + \boxed{14} = 1 \end{aligned}$$

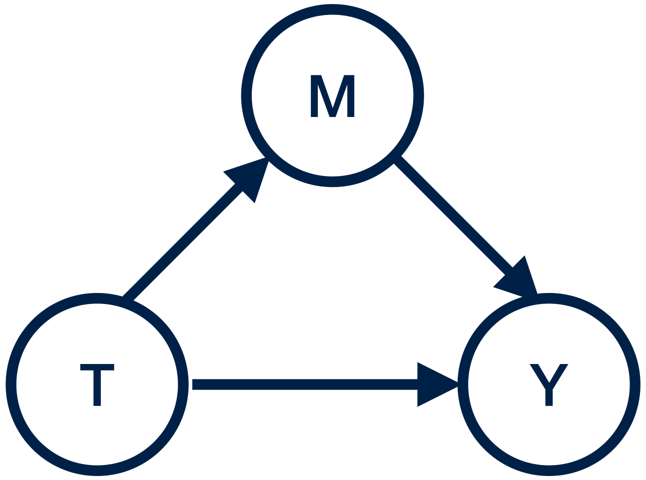
(Barring sampling errors)

full assurance that drug x was in fact responsible for death.

100% sure

Mathematical toolkit: **Mediation**

Next, we consider a typical mediation problem and the various associated causal effects



Treatment (T), mediator (M), and outcome (Y)

Structural causal model:

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

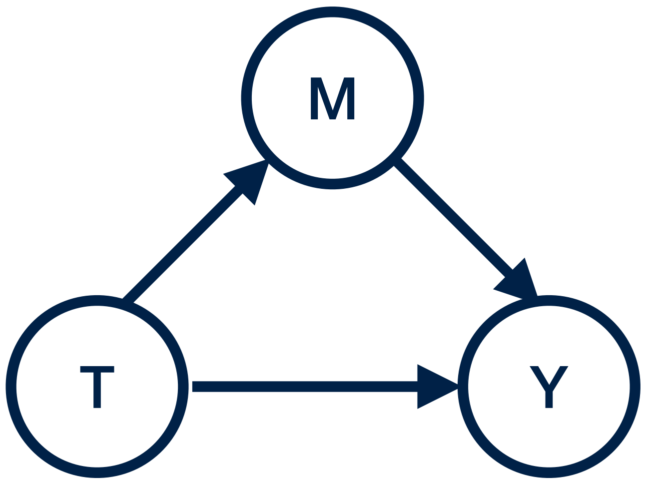
As always, the omitted factors $U = (U_T, U_M, U_Y)$ that influence treatment, mediator, and outcome may very well be dependent

(All expectations on the next slides are with respect to Y, U_M and U_Y.)

Pearl's Primer book Section 4.5.2

Mathematical toolkit: **Mediation**

Four types of effects when we go from T=0 to T=1:

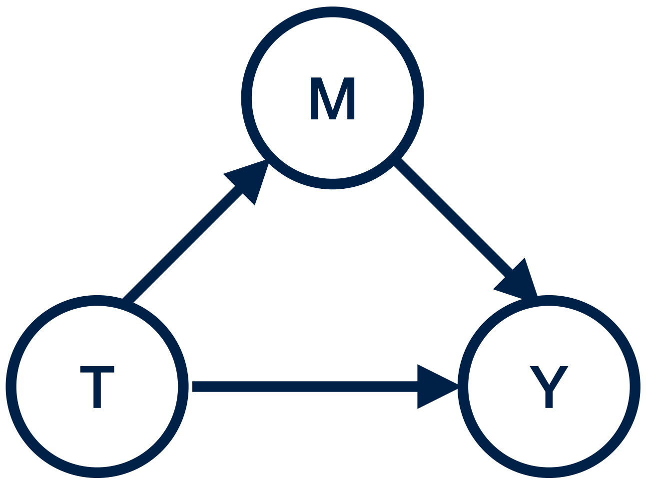


1. **Total effect:** Measures the increase in Y as treatment changes from T=0 to T=1 while mediator M changes freely as per the structural function f_M

$$\begin{aligned}\text{TE} &= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}[Y|\text{do}(T=1)] - \mathbb{E}[Y|\text{do}(T=0)]\end{aligned}$$

Mathematical toolkit: **Mediation**

Four types of effects when we go from T=0 to T=1:



1. **Total effect:** Measures the increase in Y as treatment changes from T=0 to T=1 while mediator M changes freely as per the structural function f_M

$$\begin{aligned}\mathrm{TE} &= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}[Y|\mathrm{do}(T=1)] - \mathbb{E}[Y|\mathrm{do}(T=0)]\end{aligned}$$

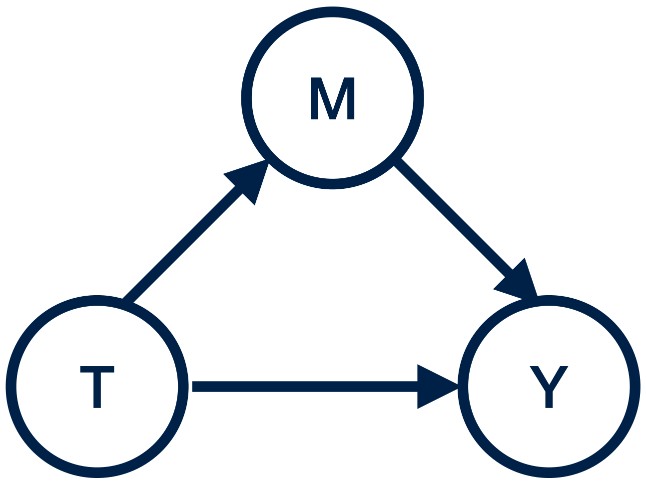
2. **Controlled direct effect (CDE(m)):** Measures the expected increase in Y as treatment changes from T=0 to T=1 while mediator is set to M = m uniformly

$$\begin{aligned}\mathrm{CDE} &= \mathbb{E}[Y_{1,m} - Y_{0,m}] \\ &= \mathbb{E}[Y|\mathrm{do}(T=1, M=m)] - \mathbb{E}[Y|\mathrm{do}(T=0, M=m)]\end{aligned}$$

Pearl's Primer book Section 4.5.2

Mathematical toolkit: **Mediation**

Four types of effects when we go from T=0 to T=1:

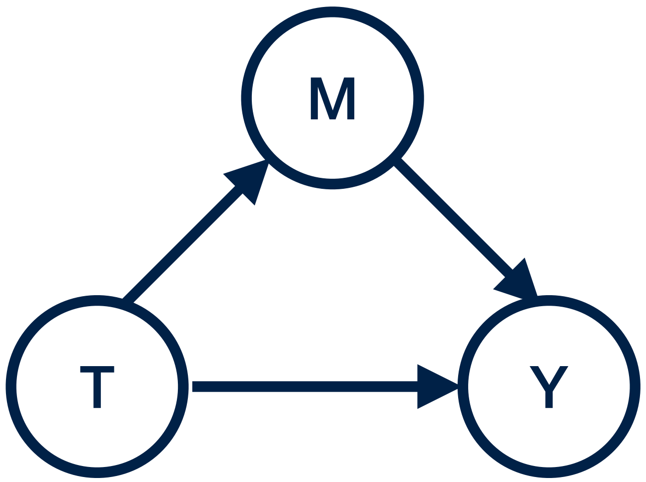


3. **Natural direct effect (NDE):** Measures expected increase in Y as treatment changes from T=0 to T=1 while mediator is set to whatever value it *would have attained* (for each individual) prior to change, that is, under T = 0.

$$\mathrm{NDE} = \mathbb{E}\left[Y_{1,M_0} - Y_{0,M_0}\right]$$

Mathematical toolkit: **Mediation**

Four types of effects when we go from T=0 to T=1:



3. **Natural direct effect (NDE)**: Measures expected increase in Y as treatment changes from T=0 to T=1 while mediator is set to whatever value it *would have attained* (for each individual) prior to change, that is, under T = 0.

$$\mathrm{NDE} = \mathbb{E}\left[Y_{1,M_0} - Y_{0,M_0}\right]$$

4. **Natural indirect effect (NIE)**: Measures the expected increase in Y when the treatment is held constant at T=0 and the mediator M changes to whatever value it *would have attained* (for each individual) under T=1

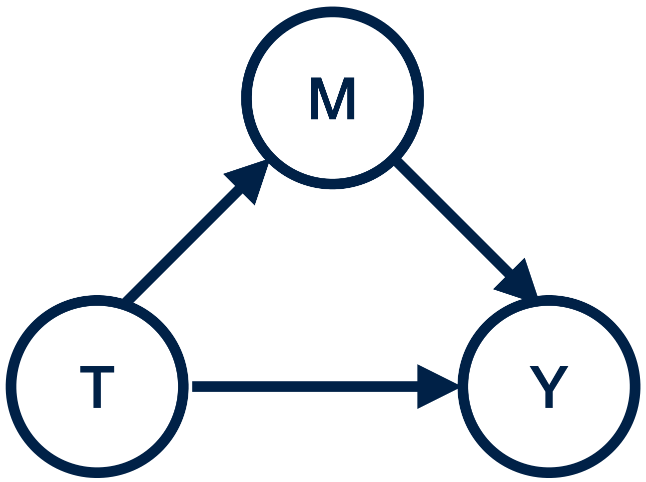
$$\mathrm{NIE} = \mathbb{E}\left[Y_{0,M_1} - Y_{0,M_0}\right]$$

It captures the portion of the effect that can be explained by mediation alone, while disabling (or “freezing”) the capacity of Y to respond to T

Pearl’s Primer book Section 4.5.2

Mathematical toolkit: **Mediation**

Some remarks on these four types of effects

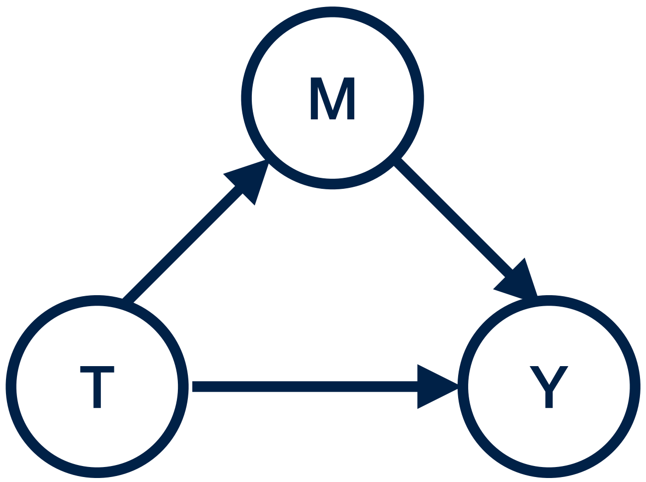


1. TE and CDE(m) are *do*-expressions so can be estimated from experimental data or observational studies using the backdoor and front-door criteria

2. NDE and NIE are **not** *do*-expressions, so their causal identifiability will require a new set of results and, possibly, further assumptions

Identifying NDE and NIE

There should exist a set W of measured covariates s.t.



A. No member of W is a descendant of T
B. W blocks all backdoor paths from M to Y (after removing the arrows T —> M and T —> Y)
C. W-specific effect of T on M is identifiable, possibly using experiments
D. W-specific joint effect of {T,M} on Y is identifiable, possibly using experiments

Theorem

When A and B hold, NDE is experimentally identifiable and is given by

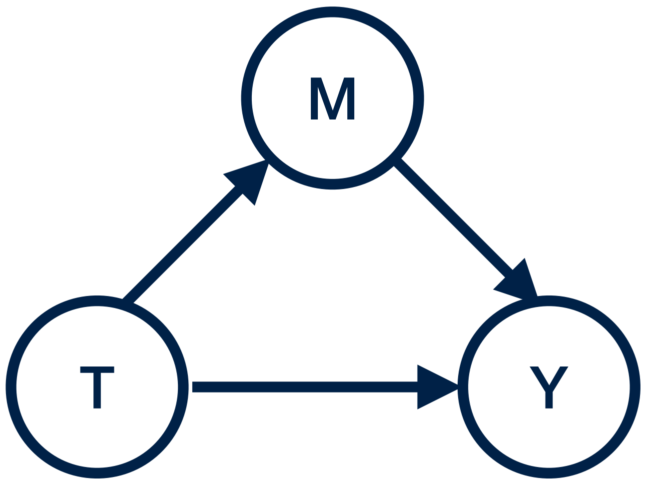
$$\mathrm{NDE} = \sum_m \sum_w \Big[\mathbb{E}[Y|\mathrm{do}(T=1, M=m), W=w] - \mathbb{E}[Y|\mathrm{do}(T=0, M=m), W=w]\Big] \times p(M=m|\mathrm{do}(T=0), W=w)p(W=w)$$

Identifiability of the *do*-expression is guaranteed by conditions C and D and can be determined using the backdoor or front-door criteria

Pearl's Primer book Section 4.5.2

Identifying NDE and NIE

There should exist a set W of measured covariates s.t.



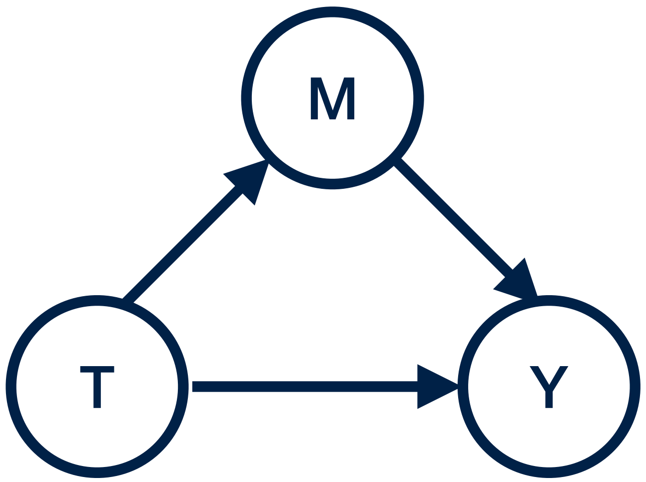
A. No member of W is a descendant of T
B. W blocks all backdoor paths from M to Y (after removing the arrows T —> M and T —> Y)
C. W-specific effect of T on M is identifiable, possibly using experiments
D. W-specific joint effect of {T,M} on Y is identifiable, possibly using experiments

Corollary
If A and B hold, and the W deconfound the relationships in C and D, then the *do*-expressions in the theorem reduce to conditional expectations, and we have

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

Identifying NDE and NIE

Finally, one can give simpler expression assuming that



E. The exogenous variables $U = (U_T, U_M, U_Y)$ are mutually independent

If all conditions A, B, C, D, and E hold then

$$\text{NDE} = \sum_m \left[\mathbb{E}[Y|T=1, M=m] - \mathbb{E}[Y|T=0, M=m]\right] p(M=m|T=0)$$

and, similarly,

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

These two expressions are known as the *mediation formulas*

Note that NDE is a weighted average of CDE(m), whereas NIE is not

Pearl's Primer book Section 4.5.2

Methods for Causal Inference
Lecture 16

Ava Khamseh
School of Informatics

2021-2022