Counterfactuals and Mediation

Brady Neal

causalcourse.com

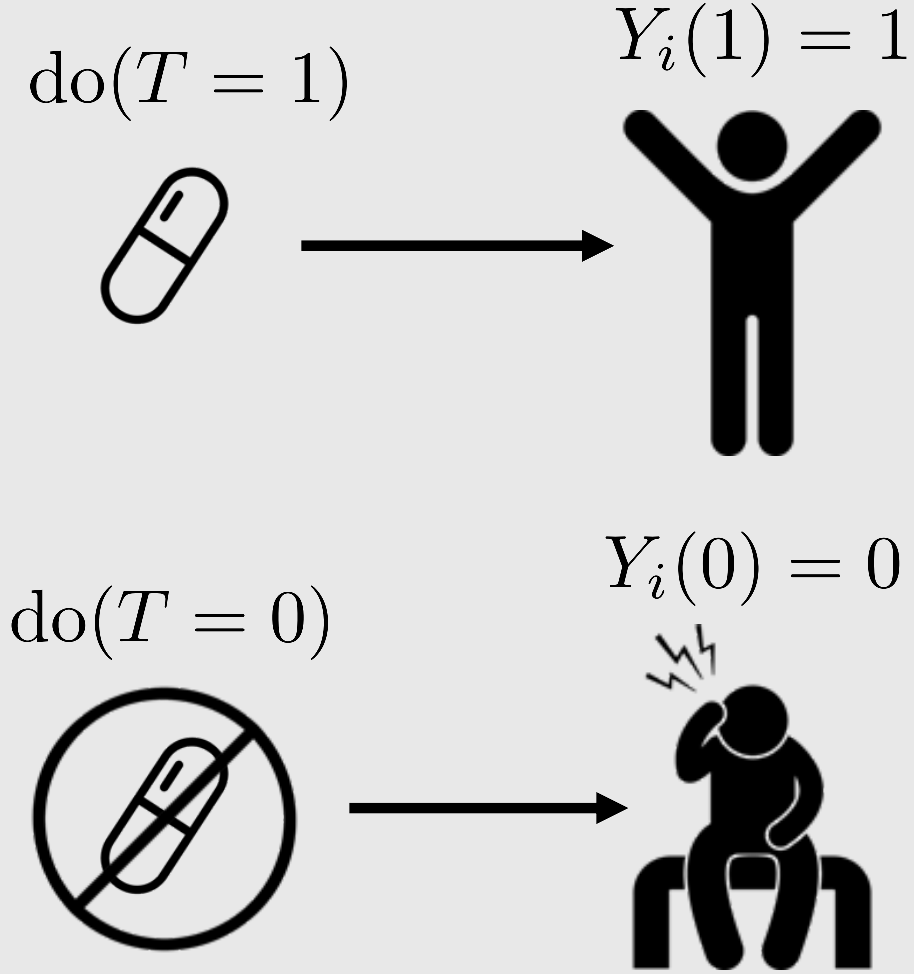
Counterfactuals Basics

Important Application: Mediation

Counterfactuals Basics

Important Application: Mediation

Fundamental Problem of Causal Inference



T	: observed treatment
Y	: observed outcome
i	: used in subscript to denote a specific unit/individual
$Y_i(1)$	: potential outcome under treatment
$Y_i(0)$	: potential outcome under no treatment

Causal effect

$$Y_i(1) - Y_i(0) = 1$$

Fundamental Problem of Causal Inference

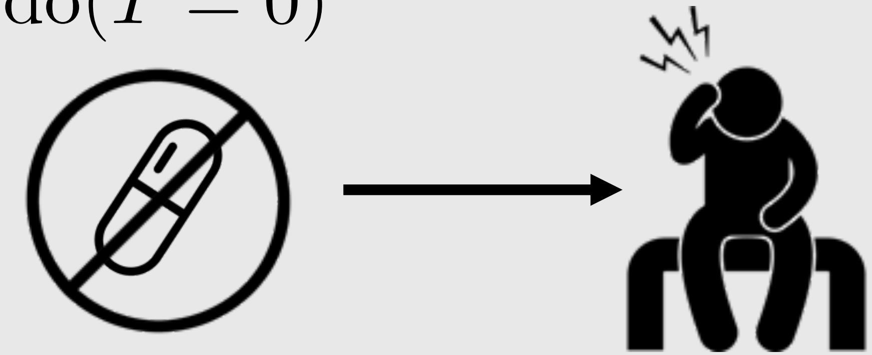
Counterfactual

$\mathrm{do}(T=1)$ $\quad$ $Y_i(1) = 1$

T	: observed treatment
Y	: observed outcome
i	: used in subscript to denote a specific unit/individual
$Y_i(1)$	: potential outcome under treatment
$Y_i(0)$	: potential outcome under no treatment

Factual

$\mathrm{do}(T=0)$ $\quad$ $Y_i(0) = 0$



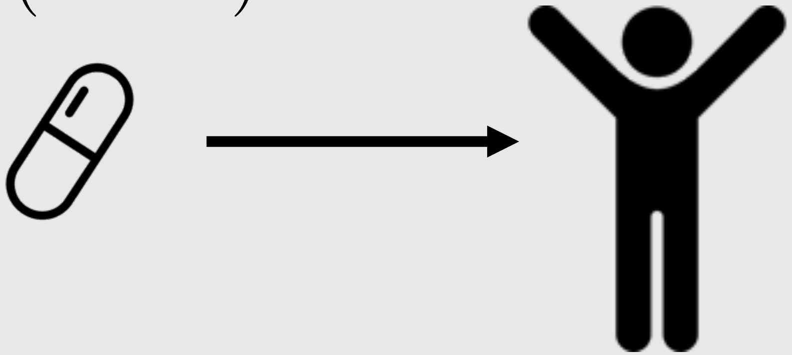
Causal effect

$$Y_i(1) - Y_i(0) = 1$$

Fundamental Problem of Causal Inference

Factual

$\mathrm{do}(T = 1)$ $\quad Y_i(1) = 1$



Counterfactual

$\mathrm{do}(T = 0)$ $\quad Y_i(0) = 0$

T : observed treatment
Y : observed outcome
i : used in subscript to denote a specific unit/individual
$Y_i(1)$: potential outcome under treatment
$Y_i(0)$: potential outcome under no treatment

Causal effect

$$Y_i(1) - Y_i(0) = 1$$

We can compute counterfactuals using a parametric SCM.

Counterfactuals

Counterfactual: $P(Y(t) \mid T = t', Y = y')$

Counterfactuals

Counterfactual: $P(Y(t) \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

Counterfactuals

Counterfactual: $P(Y(t) \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

hypothetical condition

Counterfactuals

Counterfactual: $P(Y(t) \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

hypothetical condition

Different from CATE: $\mathbb{E}[Y(t) \mid X = x]$

Counterfactuals

Counterfactual: $P(\underbrace{Y(t)}_{\text{hypothetical condition}} \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

Different from CATE: $\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y \mid do(t), X = x]$

Counterfactuals

Counterfactual: $P(\underbrace{Y(t)}_{\text{hypothetical condition}} \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

Different from CATE: $\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y \mid do(t), X = x]$

Cannot express counterfactuals using do-notation

Roadmap for Computing Counterfactuals

Roadmap for Computing Counterfactuals

Given: Observation of (T, Y) (observation of potential outcome $Y(t)$ where t is the observed value of T)

Roadmap for Computing Counterfactuals

Given: Observation of (T, Y) (observation of potential outcome Y(t) where t is the observed value of T)

Main ingredient necessary: correct parametric model for the structural equation for Y

Roadmap for Computing Counterfactuals

Given: Observation of (T, Y) (observation of potential outcome Y(t) where t is the observed value of T)

Main ingredient necessary: correct parametric model for the structural equation for Y

Result: access to counterfactuals Y(t') at the unit-level

Computing Counterfactuals: Simple Example

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $$\begin{aligned} T &:= \dots \\ Y &:= UT + (1 - U)(1 - T) \end{aligned}$$

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$$Y := UT + (1 - U)(1 - T)$$

Observation: T = 0 and Y = 0

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$Y := UT + (1 - U)(1 - T)$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$Y := UT + (1 - U)(1 - T)$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \ldots$

$\underline{Y := UT + (1 - U)(1 - T)}$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y_u(1)?$

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$Y := UT + (1 - U)(1 - T)$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y_u(1)$?

Step 1: Solve for U

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \ldots$

$\underline{Y := UT + (1-U)(1-T)}$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y = UT + (1-U)(1-T)$ $\quad Y_u(1)?$

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$Y := UT + (1 - U)(1 - T)$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$

$$0 = U(0) + (1 - U)(1 - 0)$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$\underline{Y := UT + (1 - U)(1 - T)}$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$

$$0 = U(0) + (1 - U)(1 - 0)$$

$$0 = 1 - U$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$\underline{Y := UT + (1-U)(1-T)}$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1-U)(1-T)$$

$$0 = U(0) + (1-U)(1-0)$$

$$0 = 1 - U$$

$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \ldots$

$$Y := UT + (1 - U)(1 - T)$$

Step 2: Individualized SCM

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$
$$0 = U(0) + (1 - U)(1 - 0)$$
$$0 = 1 - U$$
$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$$\underline{Y := UT + (1-U)(1-T)}$$

Step 2: Individualized SCM

$$T := \dots$$

$$Y := (1)T + (1-1)(1-T)$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1-U)(1-T)$$

$$0 = U(0) + (1-U)(1-0)$$

$$0 = 1 - U$$

$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$Y := UT + (1 - U)(1 - T)$$

Step 2: Individualized SCM
$$T := \dots$$
$$Y := T$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$
$$0 = U(0) + (1 - U)(1 - 0)$$
$$0 = 1 - U$$
$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$\underline{Y := UT + (1-U)(1-T)}$$

Step 2: Individualized SCM
$$T := 1$$
$$Y := T$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1-U)(1-T)$$
$$0 = U(0) + (1-U)(1-0)$$
$$0 = 1 - U$$
$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \dots$

$$Y := UT + (1 - U)(1 - T)$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y_u(1)$?

$$Y = UT + (1 - U)(1 - T)$$

$$0 = U(0) + (1 - U)(1 - 0)$$

$$0 = 1 - U$$

$$U = 1$$

Step 2: Individualized SCM

$$T := 1$$

$$Y := T$$

$$Y_u(1) = 1$$

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

U: unobserved variable describing the individual (1 if dog person; 0 if anti-dog person)

SCM: $T := \ldots$

$$Y := UT + (1 - U)(1 - T)$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$

$$0 = U(0) + (1 - U)(1 - 0)$$

$$0 = 1 - U$$

$$U = 1$$

$Y_u(1)$?

Step 2: Individualized SCM

$$T := 1$$

$$Y := T$$

$$Y_u(1) = 1$$

ITE: $Y_u(1) - Y_u(0) = 1 - 0 = 1$

General Steps for Deterministic Counterfactuals

General Steps for Deterministic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

General Steps for Deterministic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation to determine the value of U

General Steps for Deterministic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation to determine the value of U
2. Action: Modify the SCM, by replacing the structural equation for T with T := t

General Steps for Deterministic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation to determine the value of U
2. Action: Modify the SCM, by replacing the structural equation for T with T := t
3. Prediction: Use the value of U from step 1 and the modified SCM from step 2 to compute the value of Y(t)

Question:

Given the observation T = 1 and Y = 0, compute Y(0) for this individual given the following SCM:

$$T := \dots$$
$$Y := UT + (1 - U)(1 - T)$$

Can’t Always Determine Counterfactual

Even when we have the structural equation for Y, we can’t always determine counterfactuals with probability 1

Can’t Always Determine Counterfactual

Even when we have the structural equation for Y, we can’t always determine counterfactuals with probability 1

What if we can’t solve for U (function that maps U to Y for a fixed value of T isn’t invertible)?

Can't Always Determine Counterfactual

Even when we have the structural equation for Y, we can't always determine counterfactuals with probability 1

What if we can't solve for U (function that maps U to Y for a fixed value of T isn't invertible)?

Example:

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & U = \text{never happy} \\ T & U = \text{dog-needer} \\ 1 - T & U = \text{dog-hater} \end{cases}$$

Can't Always Determine Counterfactual

Even when we have the structural equation for Y, we can't always determine counterfactuals with probability 1

What if we can't solve for U (function that maps U to Y for a fixed value of T isn't invertible)?

Example:

Observation: T = 1 and Y = 0

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & U = \text{never happy} \\ T & U = \text{dog-needer} \\ 1 - T & U = \text{dog-hater} \end{cases}$$

Can’t Always Determine Counterfactual

Even when we have the structural equation for Y, we can’t always determine counterfactuals with probability 1

What if we can’t solve for U (function that maps U to Y for a fixed value of T isn’t invertible)?

Example:

Observation: T = 1 and Y = 0

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & U = \text{never happy} \\ T & U = \text{dog-needer} \\ 1 - T & U = \text{dog-hater} \end{cases}$$

Non-Invertible Example

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & \underline{U = \text{never happy}} \\ T & U = \text{dog-needer} \\ 1 - T & \underline{U = \text{dog-hater}} \end{cases}$$

Observation: T = 1 and Y = 0

Non-Invertible Example

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & \underline{U = \text{never happy}} \\ T & U = \text{dog-needer} \\ 1 - T & \underline{U = \text{dog-hater}} \end{cases}$$

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

Non-Invertible Example

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & U = \text{never happy} \\ T & U = \text{dog-needer} \\ 1 - T & U = \text{dog-hater} \end{cases}$$

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$Y_u(0) = ?$$

Non-Invertible Example

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & \underline{U = \text{never happy}} \\ T & U = \text{dog-needer} \\ 1 - T & \underline{U = \text{dog-hater}} \end{cases}$$

$$P(U = \text{always happy}) = 0.3$$
$$P(U = \text{never happy}) = 0.2$$
$$P(U = \text{dog-needer}) = 0.4$$
$$P(U = \text{dog-hater}) = 0.1$$

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$Y_u(0) = ?$$

Non-Invertible Example

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & \underline{U = \text{never happy}} \\ T & U = \text{dog-needer} \\ 1 - T & \underline{U = \text{dog-hater}} \end{cases}$$

$$P(U = \text{always happy}) = 0.3$$
$$\underline{P(U = \text{never happy}) = 0.2}$$
$$P(U = \text{dog-needer}) = 0.4$$
$$\underline{P(U = \text{dog-hater}) = 0.1}$$

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$Y_u(0) = ?$$

Non-Invertible Example

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & \underline{U = \text{never happy}} \\ T & U = \text{dog-needer} \\ 1 - T & \underline{U = \text{dog-hater}} \end{cases}$$

$$P(U = \text{always happy}) = 0.3$$
$$\underline{P(U = \text{never happy}) = 0.2}$$
$$P(U = \text{dog-needer}) = 0.4$$
$$\underline{P(U = \text{dog-hater}) = 0.1}$$

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$P(U = \text{never happy} \mid T = 1, Y = 0) = \frac{0.2}{0.2 + 0.1} = \frac{2}{3}$$

$$P(U = \text{dog-hater} \mid T = 1, Y = 0) = \frac{0.1}{0.2 + 0.1} = \frac{1}{3}$$

$$Y_u(0) = ?$$

Non-Invertible Example

Structural equation for Y:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & \underline{U = \text{never happy}} \\ T & U = \text{dog-needer} \\ 1 - T & \underline{U = \text{dog-hater}} \end{cases}$$

$$P(U = \text{always happy}) = 0.3$$
$$\underline{P(U = \text{never happy}) = 0.2}$$
$$P(U = \text{dog-needer}) = 0.4$$
$$\underline{P(U = \text{dog-hater}) = 0.1}$$

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$P(U = \text{never happy} \mid T = 1, Y = 0) = \frac{0.2}{0.2 + 0.1} = \frac{2}{3}$$

$$P(U = \text{dog-hater} \mid T = 1, Y = 0) = \frac{0.1}{0.2 + 0.1} = \frac{1}{3}$$

$$Y_u(0) = ?$$

$$P(Y_u(0) = 1) = \frac{1}{3}$$

General Steps for Probabilistic Counterfactuals

General Steps for Probabilistic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

General Steps for Probabilistic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation Z to update the distribution of U: $P(U \mid Z)$

General Steps for Probabilistic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation Z to update the distribution of U: P(U | Z)
2. Action: Modify the SCM, by replacing the structural equation for T with T := t

General Steps for Probabilistic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation Z to update the distribution of U: P(U | Z)
2. Action: Modify the SCM, by replacing the structural equation for T with T := t
3. Prediction: Use the the updated distribution of U step 1 and the modified SCM from step 2 to compute the distribution of Y(t)

No Unit-Level Counterfactuals without Parametric Model

Main ingredient necessary for computing counterfactuals:
parametric model for the structural equation for Y

No Unit-Level Counterfactuals without Parametric Model

Main ingredient necessary for computing counterfactuals: parametric model for the structural equation for Y

Strong assumption

No Unit-Level Counterfactuals without Parametric Model

Main ingredient necessary for computing counterfactuals: parametric model for the structural equation for Y

Strong assumption

Without it, we are stuck with the fundamental problem of causal inference.

Question:

Given the observation T = 1 and Y = 1, compute Y(0) for this individual given the following SCM and prior:

$$Y := \begin{cases} 1 & U = \text{always happy} \\ 0 & U = \text{never happy} \\ T & U = \text{dog-needer} \\ 1 - T & U = \text{dog-hater} \end{cases}$$

$$P(U = \text{always happy}) = 0.3$$
$$P(U = \text{never happy}) = 0.2$$
$$P(U = \text{dog-needer}) = 0.4$$
$$P(U = \text{dog-hater}) = 0.1$$

Population-Level Doesn't Require a Parametric Model

Population-level counterfactual: $\mathbb{E}[Y(t) \mid T = t']$

Population-Level Doesn't Require a Parametric Model

Population-level counterfactual: $\mathbb{E}[Y(t) \mid T = t']$

Just like we were able to identify the ATE $\mathbb{E}[Y(1) - Y(0)]$ nonparametrically (using just the causal graph), we can do the same with population-level counterfactual quantities, if they are identifiable

Population-Level Doesn't Require a Parametric Model

Population-level counterfactual: $\mathbb{E}[Y(t) \mid T = t']$

Just like we were able to identify the ATE $\mathbb{E}[Y(1) - Y(0)]$ nonparametrically (using just the causal graph), we can do the same with population-level counterfactual quantities, if they are identifiable

Same with CATEs: $\mathbb{E}[Y(1) - Y(0) \mid X = x]$

Population-Level Doesn't Require a Parametric Model

Population-level counterfactual: $\mathbb{E}[Y(t) \mid T = t']$

Just like we were able to identify the ATE $\mathbb{E}[Y(1) - Y(0)]$ nonparametrically (using just the causal graph), we can do the same with population-level counterfactual quantities, if they are identifiable

Same with CATEs: $\mathbb{E}[Y(1) - Y(0) \mid X = x]$

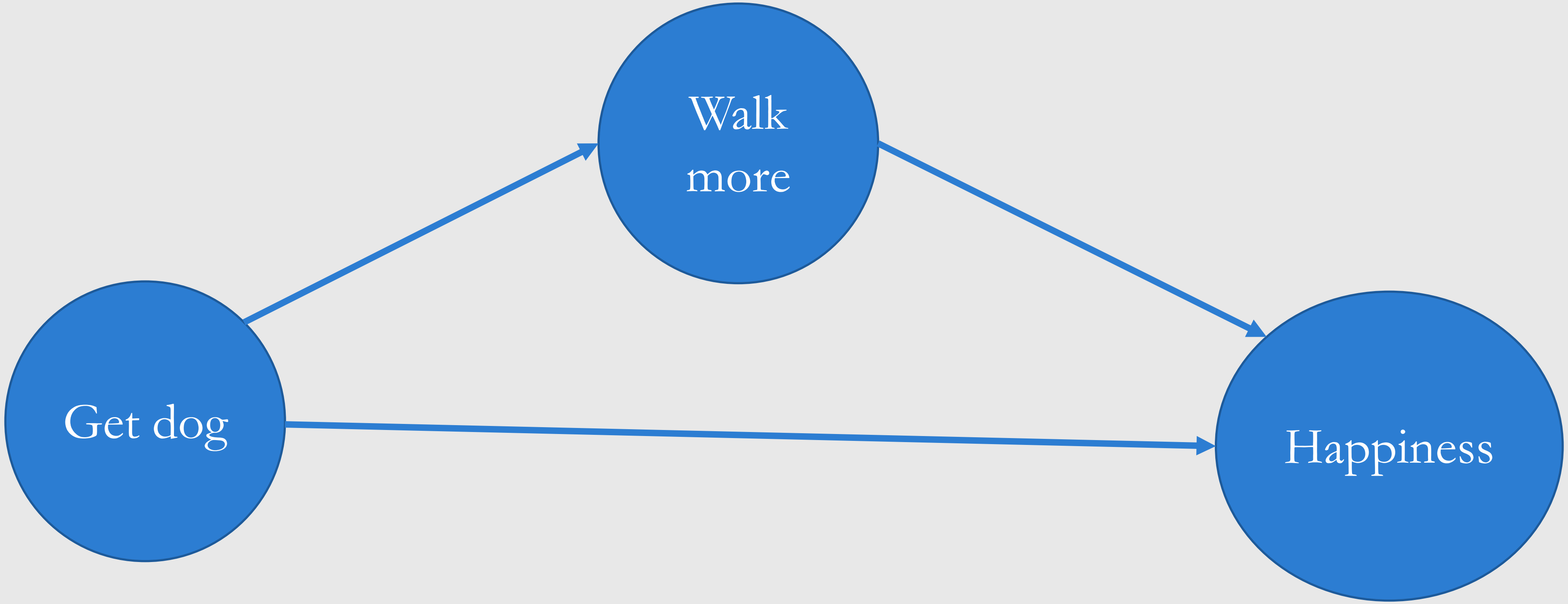
See Malinsky et al. (2019)'s potential outcome calculus (generalization of do-calculus) for general identification of counterfactual quantities

Counterfactuals Basics

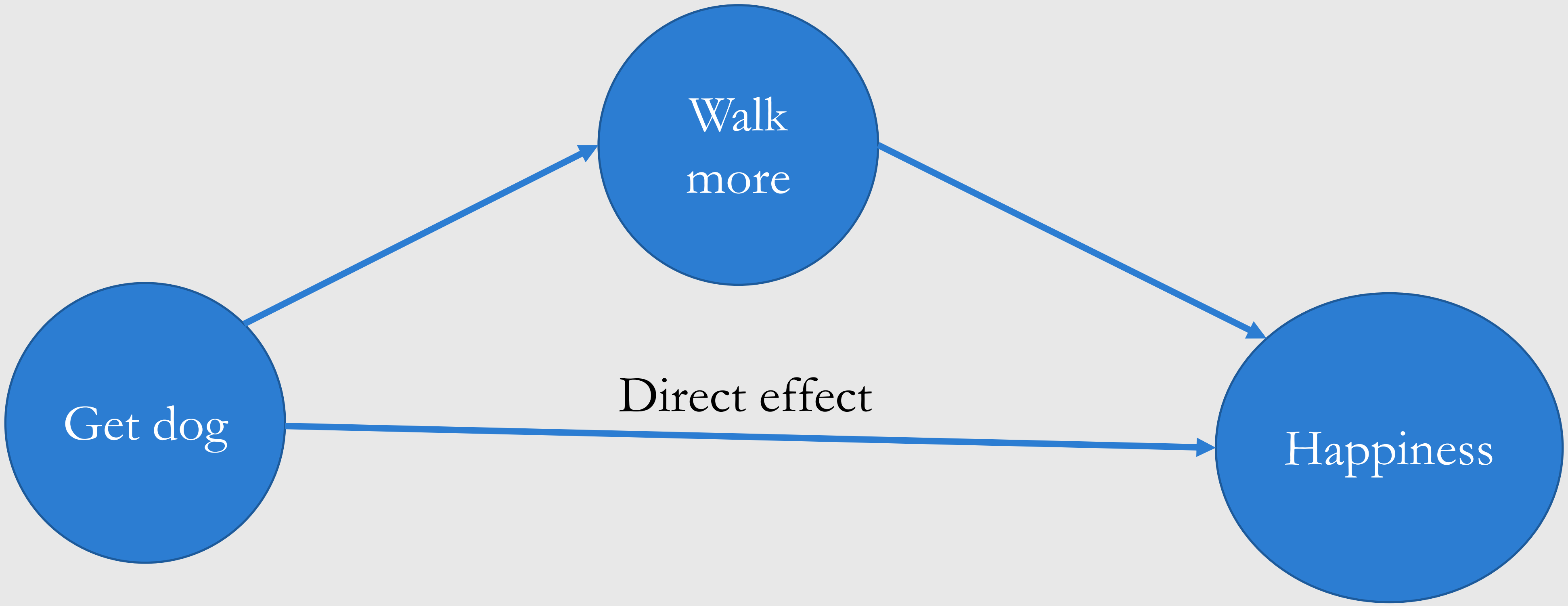
Important Application: Mediation

Mediation

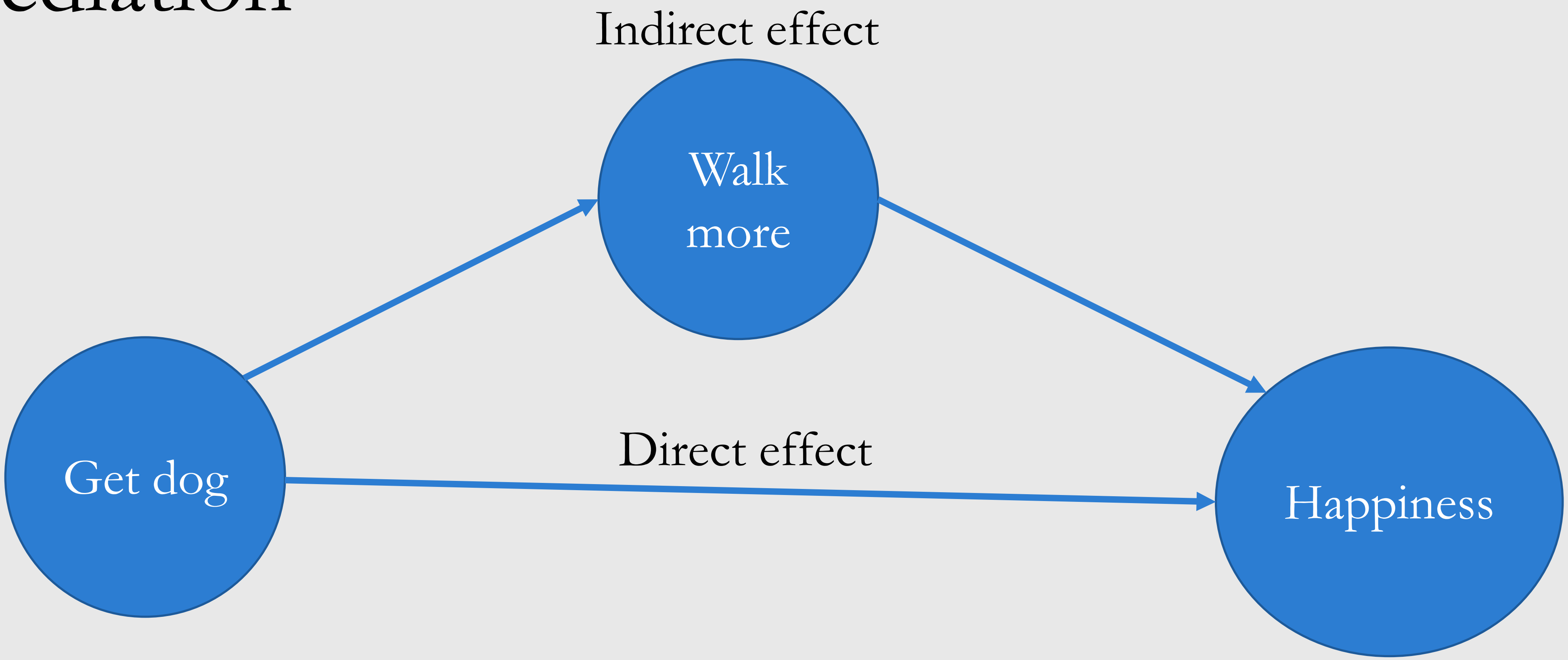
Mediation



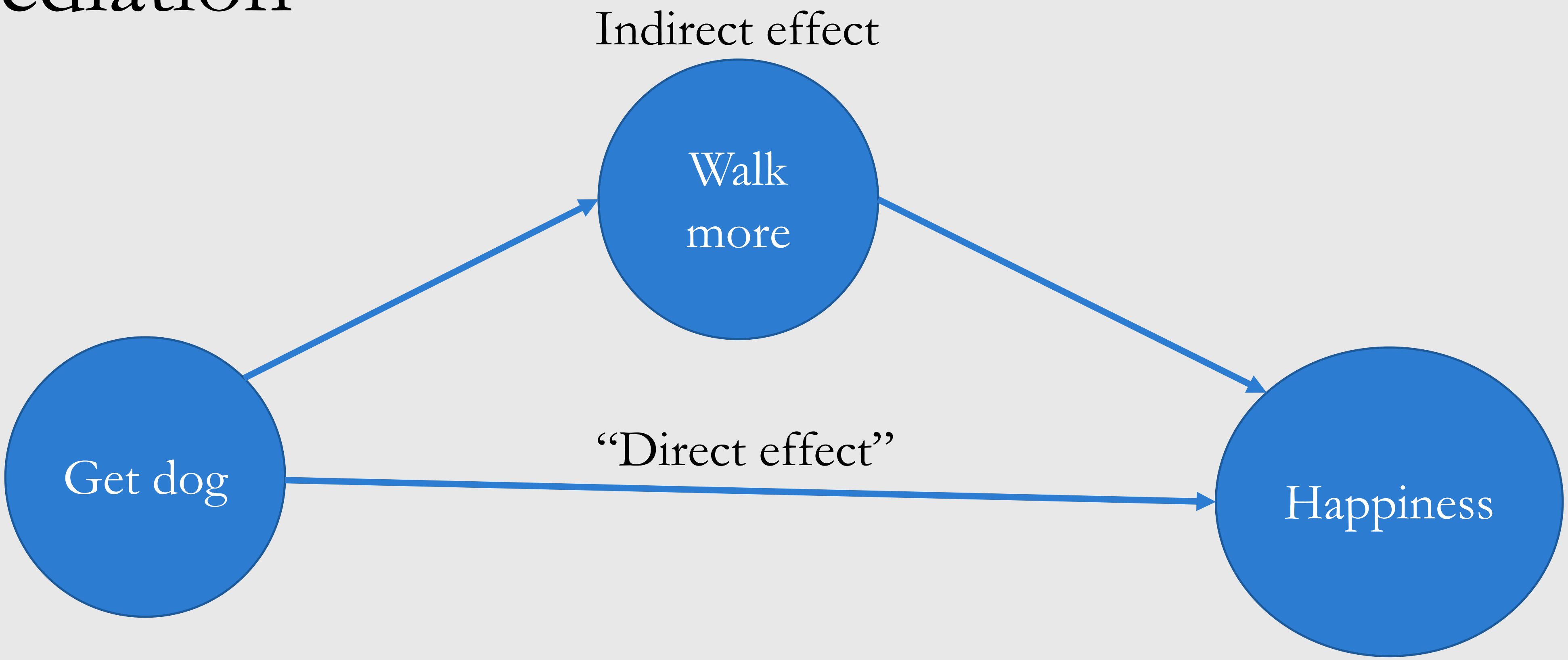
Mediation



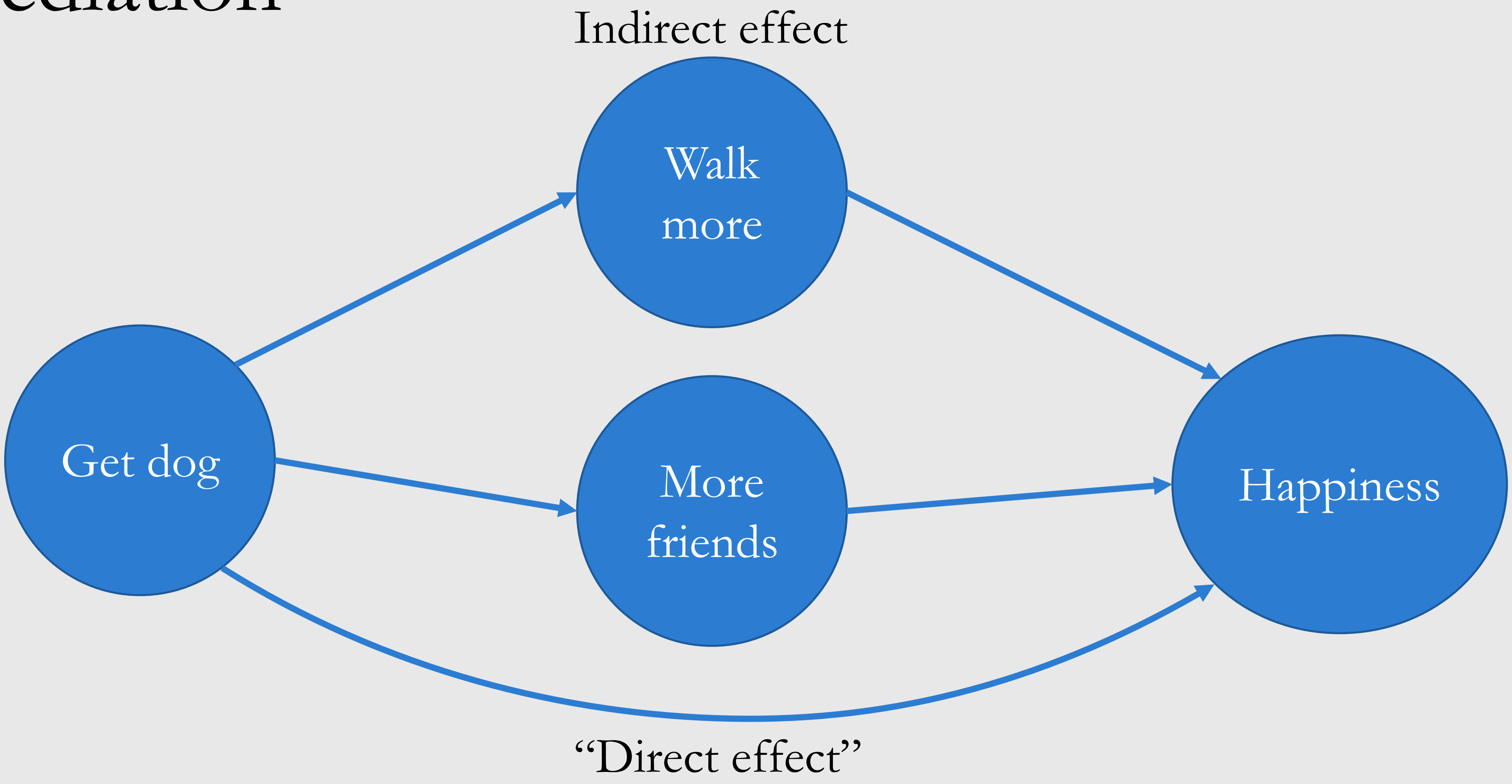
Mediation



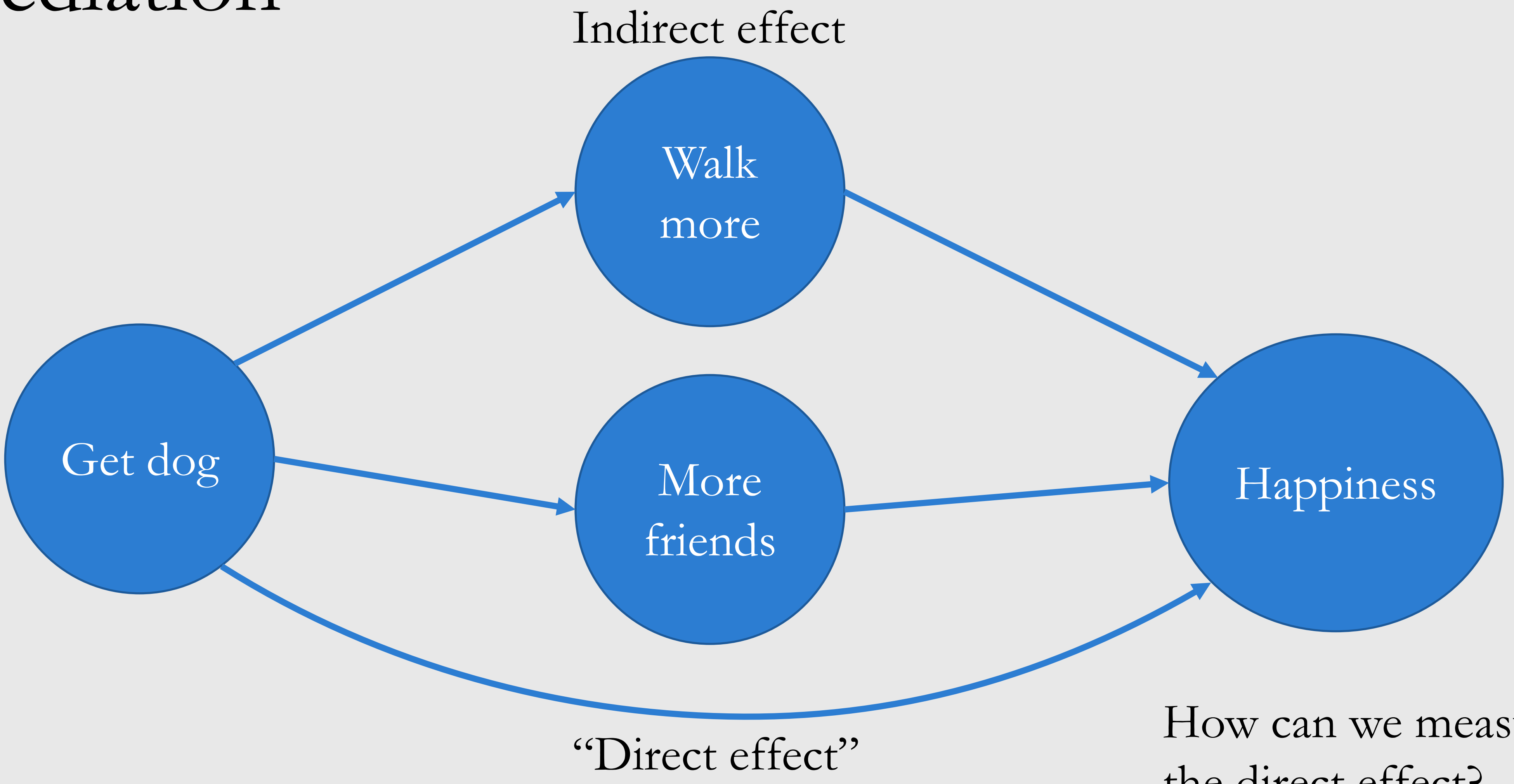
Mediation



Mediation

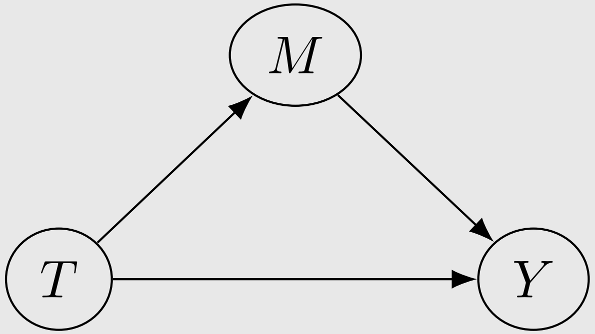

Indirect effect
Walk more
More friends
Get dog
Happiness
"Direct effect"

Mediation

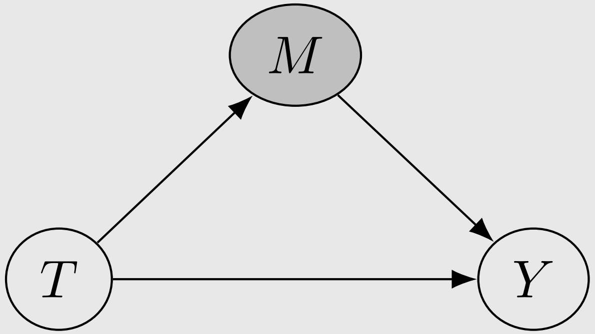


How can we measure the direct effect?

Controlled Direct Effect (CDE)

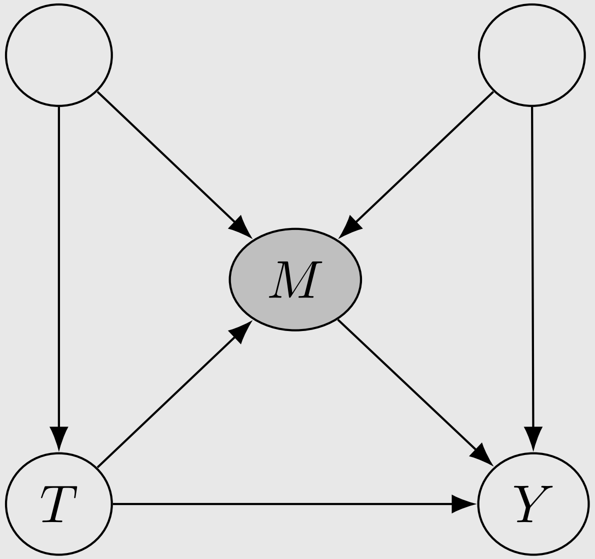


Controlled Direct Effect (CDE)



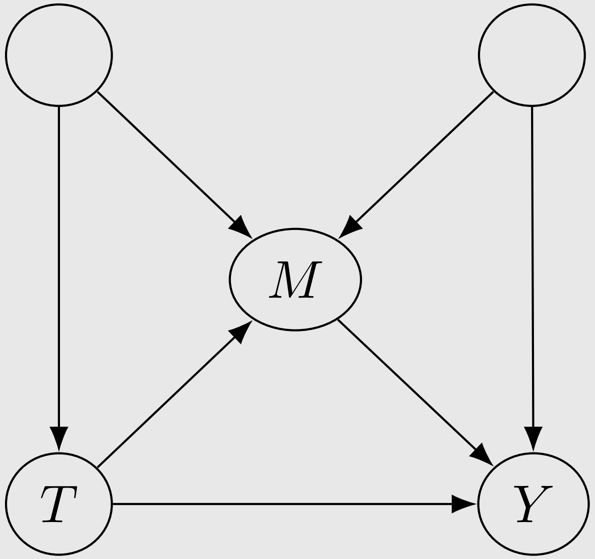
$$\mathbb{E}[Y \mid do(T=1), M=m] - \mathbb{E}[Y \mid do(T=0), M=m]$$

Controlled Direct Effect (CDE)



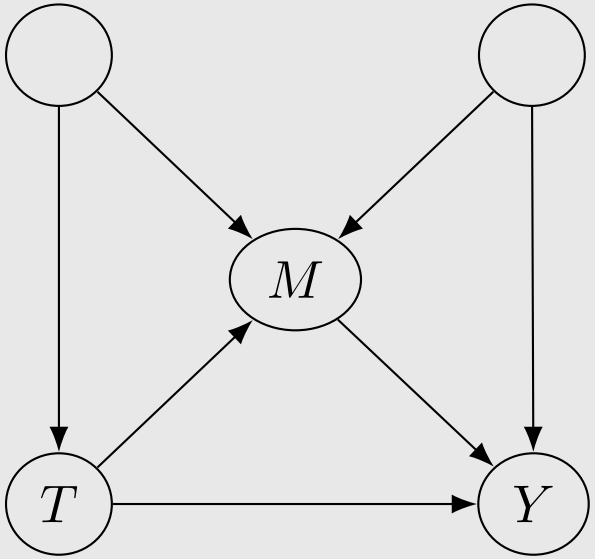
$$\mathbb{E}[Y \mid do(T = 1), M = m] - \mathbb{E}[Y \mid do(T = 0), M = m]$$

Controlled Direct Effect (CDE)



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

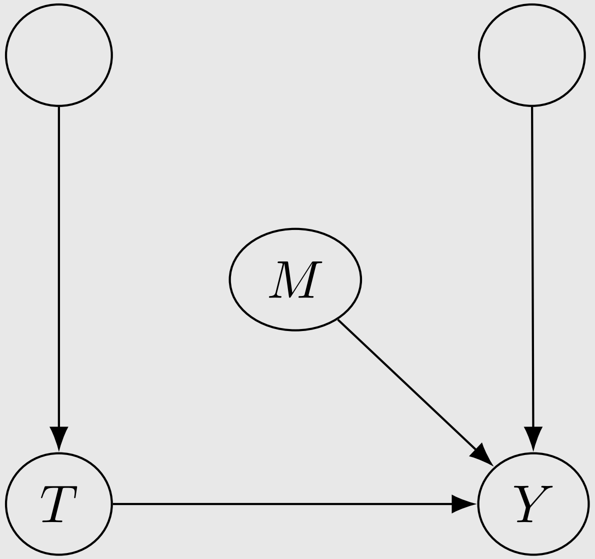
Controlled Direct Effect (CDE)



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

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

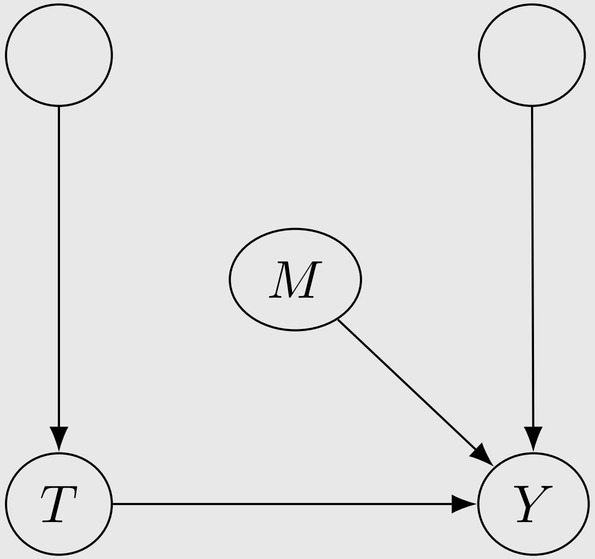
Controlled Direct Effect (CDE)



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

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

Controlled Direct Effect (CDE)

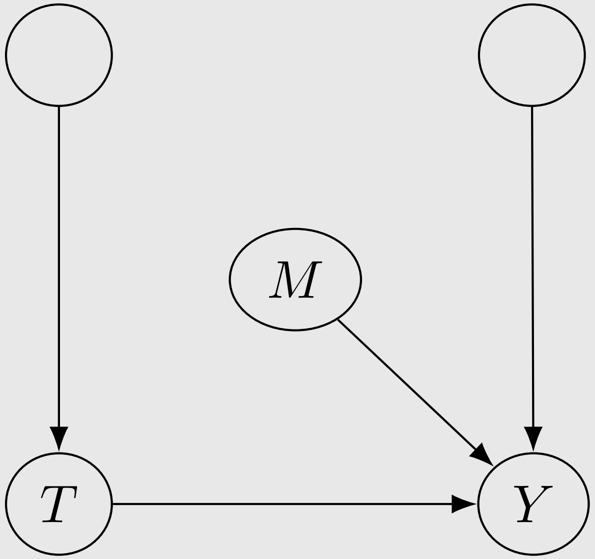


Problems:

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

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

Controlled Direct Effect (CDE)



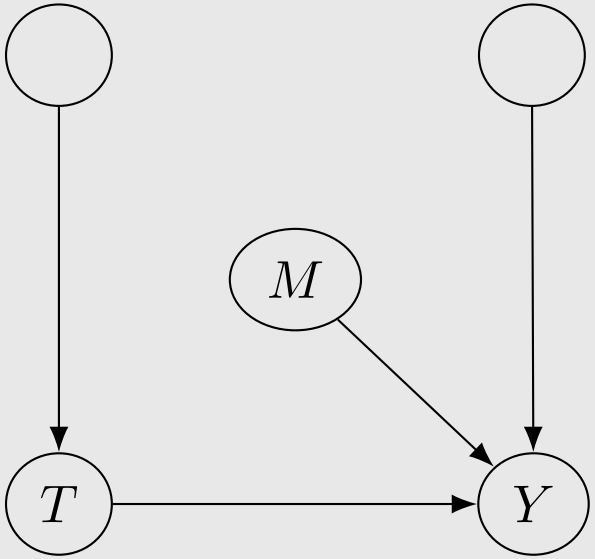
Problems:

- CDE is specific to the arbitrary choice of m

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

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

Controlled Direct Effect (CDE)



Problems:

- CDE is specific to the arbitrary choice of m
- How do we get the indirect effect? Can't just subtract the CDE from the total effect

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

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

Natural Direct and Indirect Effects

Natural Direct and Indirect Effects

Subscript notation:

$$\mathbb{E}[Y_{t,m}] \triangleq \mathbb{E}[Y \mid do(T = t, M = m)]$$

Natural Direct and Indirect Effects

Subscript notation:

$$\mathbb{E}[Y_{t,m}] \triangleq \mathbb{E}[Y \mid do(T = t, M = m)] \qquad \mathbb{E}[M_t] \triangleq \mathbb{E}[M \mid do(T = t)]$$

Natural Direct and Indirect Effects

Subscript notation:

$$\mathbb{E}[Y_{t,m}] \triangleq \mathbb{E}[Y \mid do(T = t, M = m)] \qquad \mathbb{E}[M_t] \triangleq \mathbb{E}[M \mid do(T = t)]$$

$$\text{CDE} \triangleq \mathbb{E}[Y_{1,m} - Y_{0,m}]$$

Natural Direct and Indirect Effects

Subscript notation:

$$\mathbb{E}[Y_{t,m}] \triangleq \mathbb{E}[Y \mid do(T = t, M = m)] \qquad \mathbb{E}[M_t] \triangleq \mathbb{E}[M \mid do(T = t)]$$

$$\text{CDE} \triangleq \mathbb{E}[Y_{1,m} - Y_{0,m}]$$

$$\text{NDE} \triangleq \mathbb{E}[Y_{1,M_0} - Y_{0,M_0}]$$

Natural Direct and Indirect Effects

Subscript notation:

$$\mathbb{E}[Y_{t,m}] \triangleq \mathbb{E}[Y \mid do(T=t, M=m)] \qquad \mathbb{E}[M_t] \triangleq \mathbb{E}[M \mid do(T=t)]$$

$$\text{CDE} \triangleq \mathbb{E}[Y_{1,m} - Y_{0,m}]$$

$$\text{NDE} \triangleq \mathbb{E}[Y_{1,M_0} - Y_{0,M_0}]$$

$$\text{NIE} \triangleq \mathbb{E}[Y_{0,M_1} - Y_{0,M_0}]$$

Natural Direct and Indirect Effects

Subscript notation:

$$\mathbb{E}[Y_{t,m}] \triangleq \mathbb{E}[Y \mid do(T=t, M=m)] \qquad \mathbb{E}[M_t] \triangleq \mathbb{E}[M \mid do(T=t)]$$

$$\text{CDE} \triangleq \mathbb{E}[Y_{1,m} - Y_{0,m}]$$

$$\text{NDE} \triangleq \mathbb{E}[Y_{1,M_0} - Y_{0,M_0}]$$

$$\text{NIE} \triangleq \mathbb{E}[Y_{0,M_1} - Y_{0,M_0}]$$

Recall problems with CDE:

- CDE is specific to the arbitrary choice of m
- How do we get the indirect effect? Can't just subtract the CDE from the total effect

Natural Direct and Indirect Effects

Subscript notation:

$$\mathbb{E}[Y_{t,m}] \triangleq \mathbb{E}[Y \mid do(T = t, M = m)] \qquad \mathbb{E}[M_t] \triangleq \mathbb{E}[M \mid do(T = t)]$$

$$\text{CDE} \triangleq \mathbb{E}[Y_{1,m} - Y_{0,m}]$$

$$\text{NDE} \triangleq \mathbb{E}[Y_{1,M_0} - Y_{0,M_0}]$$

$$\text{NIE} \triangleq \mathbb{E}[Y_{0,M_1} - Y_{0,M_0}]$$

$$\text{TE} = \text{NDE} - \text{NIE}_r$$

Recall problems with CDE:

- CDE is specific to the arbitrary choice of m
- How do we get the indirect effect? Can't just subtract the CDE from the total effect

Natural Direct and Indirect Effects

Subscript notation:

$$\mathbb{E}[Y_{t,m}] \triangleq \mathbb{E}[Y \mid do(T=t, M=m)] \qquad \mathbb{E}[M_t] \triangleq \mathbb{E}[M \mid do(T=t)]$$

$$\text{CDE} \triangleq \mathbb{E}[Y_{1,m} - Y_{0,m}]$$

$$\text{NDE} \triangleq \mathbb{E}[Y_{1,M_0} - Y_{0,M_0}]$$

$$\text{NIE} \triangleq \mathbb{E}[Y_{0,M_1} - Y_{0,M_0}]$$

$$\text{TE} = \text{NDE} - \text{NIE}_r$$

Recall problems with CDE:

- CDE is specific to the arbitrary choice of m
- How do we get the indirect effect? Can't just subtract the CDE from the total effect

For example in linear setting, $\text{TE} = \text{NDE} + \text{NIE}$

Question:

Show that TE = NDE – NIE_r,
where $\text{NIE}_r \triangleq \mathbb{E}[Y_{1,M_0} - Y_{1,M_1}]$.

Comparison of Controlled vs. Natural Mediation

Comparison of Controlled vs. Natural Mediation

CDE can always be measured via experiments (do-operator), but it has no clear undirect effect since there is no decomposition

Comparison of Controlled vs. Natural Mediation

CDE can always be measured via experiments (do-operator), but it has no clear undirect effect since there is no decomposition

NDE cannot always be measured via experiments since it is counterfactual, but it allows for the complete decomposition of the total effect into the NDE and NIE, which is what we'd like in mediation analysis

When We Can Measure NDE and NIE

When We Can Measure NDE and NIE

Adjustment set W

When We Can Measure NDE and NIE

Adjustment set W

Sufficient conditions for identifying NDE:

When We Can Measure NDE and NIE

Adjustment set W

Sufficient conditions for identifying NDE:

1. No member of W is a descendant of T

When We Can Measure NDE and NIE

Adjustment set W

Sufficient conditions for identifying NDE:

1. No member of W is a descendant of T
2. W blocks all backdoor paths from M to Y

When We Can Measure NDE and NIE

Adjustment set $\mathbb{W}$

Sufficient conditions for identifying NDE:

1. No member of $\mathbb{W}$ is a descendant of T
2. $\mathbb{W}$ blocks all backdoor paths from M to Y

$$\begin{aligned}\text{NDE} = \sum_m \sum_w & \left(\mathbb{E}[Y \mid do(T=1, M=m), W=w] - \mathbb{E}[Y \mid do(T=0, M=m), W=w]\right) \\ & \times P(M=m \mid do(T=0), W=w) P(W=w)\end{aligned}$$

When We Can Measure NDE and NIE

Adjustment set $\mathbb{W}$

Sufficient conditions for identifying NDE:

1. No member of $\mathbb{W}$ is a descendant of T
2. $\mathbb{W}$ blocks all backdoor paths from M to Y
3. $P(M = m \mid do(T = 0), W = w)$ is identifiable (e.g. no unblockable backdoor paths from T to M)
4.

$$\text{NDE} = \sum_m \sum_w \left(\mathbb{E}[Y \mid do(T = 1, M = m), W = w] - \mathbb{E}[Y \mid do(T = 0, M = m), W = w] \right)$$
$$\times P(M = m \mid do(T = 0), W = w) P(W = w)$$

When We Can Measure NDE and NIE

Adjustment set $\mathbb{W}$

Sufficient conditions for identifying NDE:

1. No member of $\mathbb{W}$ is a descendant of T
2. $\mathbb{W}$ blocks all backdoor paths from M to Y
3. $P(M = m \mid do(T = 0), W = w)$ is identifiable (e.g. no unblockable backdoor paths from T to M)
4. $\mathbb{E}[Y \mid do(T = t, M = m), W = w]$ is identifiable (e.g. no unblockable backdoors paths from T to Y)

$$\begin{aligned}\text{NDE} = \sum_m \sum_w & (\mathbb{E}[Y \mid do(T = 1, M = m), W = w] - \mathbb{E}[Y \mid do(T = 0, M = m), W = w]) \\ & \times P(M = m \mid do(T = 0), W = w) P(W = w)\end{aligned}$$

When We Can Measure NDE and NIE

Adjustment set W

Sufficient conditions for identifying NDE:

1. No member of W is a descendant of T
2. W blocks all backdoor paths from M to Y
3. $P(M = m \mid do(T = 0), W = w)$ is identifiable (e.g. no unblockable backdoor paths from T to M)
4. $\mathbb{E}[Y \mid do(T = t, M = m), W = w]$ is identifiable (e.g. no unblockable backdoors paths from T to Y)

$$\begin{aligned} \text{NDE} &= \sum_m \sum_w (\mathbb{E}[Y \mid do(T = 1, M = m), W = w] - \mathbb{E}[Y \mid do(T = 0, M = m), W = w]) \\ &\quad \times P(M = m \mid do(T = 0), W = w) P(W = w) \\ &= \sum_m \sum_w (\mathbb{E}[Y \mid T = 1, M = m, W = w] - \mathbb{E}[Y \mid T = 0, M = m, W = w]) \\ &\quad \times P(M = m \mid T = 0, W = w) P(W = w) \end{aligned}$$

When We Can Measure NDE and NIE

Adjustment set W

Sufficient conditions for identifying NDE:

$$\text{TE} = \text{NDE} - \text{NIE}_r$$

1. No member of W is a descendant of T
2. W blocks all backdoor paths from M to Y
3. $P(M = m \mid do(T = 0), W = w)$ is identifiable (e.g. no unblockable backdoor paths from T to M)
4. $\mathbb{E}[Y \mid do(T = t, M = m), W = w]$ is identifiable (e.g. no unblockable backdoors paths from T to Y)

$$\begin{aligned}
\text{NDE} = &\sum_m \sum_w \left(\mathbb{E}[Y \mid do(T = 1, M = m), W = w] - \mathbb{E}[Y \mid do(T = 0, M = m), W = w]\right) \\
&\quad \times P(M = m \mid do(T = 0), W = w) P(W = w) \\
= &\sum_m \sum_w \left(\mathbb{E}[Y \mid T = 1, M = m, W = w] - \mathbb{E}[Y \mid T = 0, M = m, W = w]\right) \\
&\quad \times P(M = m \mid T = 0, W = w) P(W = w)
\end{aligned}$$

Question:

Come up with your own example of mediation and the corresponding graph. Then, determine whether you can identify the NDE and NIE from observational data.

Path-Specific Effects

Measure causal effects along arbitrary path or set of paths in the causal graph

See [“Identifiability of Path-Specific Effects” (Avin et al., 2005)]