THE UNIVERSITY of EDINBURGH

Methods for Causal Inference
Lecture 14: Mediation

Ava Khamseh

School of Informatics
2025-2026

Conditional Interventions & Covariate-specific effects

So far: Interventions have been limited to actions that force a variable T to take on a specified value t.

More generally: Interventions can involve dynamic policies, i.e., T is made to respond in a specific way to another variable Z, via. $t = g(z)$ or
T=t with probability $P(t|z)$

Conditional Interventions & Covariate-specific effects

So far: Interventions have been limited to actions that force a variable T to take on a specified value t.

More generally: Interventions can involve dynamic policies, i.e., T is made to respond in a specific way to another variable Z, via. $t = g(z)$ or
T=t with probability $P(t|z)$

Example: A doctor administers a drug only to patients whose temperature Z exceed a certain level Z=z. The action of the doctor is **conditional** on the value of Z, $do(T = g(Z))$, where

$$g(Z) = \begin{cases} 1 \text{ when } Z > z & \text{Make stochastic (to avoid positivity violation, e.g. 75\% vs 25\%)} \\ 0 \text{ otherwise} & \text{50\%, 50\%} \end{cases}$$

The result of such a policy is: $p(Y = y|do(T = g(Z))$

Conditional Interventions & Covariate-specific effects

"z-specific effect" of T on Y: $p(Y = y|do(T = t), Z = z)$
Distribution of Y in a subset of the population for which Z=z
(Recall, ATE vs CATE)

Example: How does the treatment affect a specific age group, or individuals with blood sugar levels = z, etc.

We will use the adjustment formula, but modified for the conditional case above:
Paths need to remain blocked when we additionally condition on Z.

Conditional Interventions & Covariate-specific effects

"z-specific effect" of T on Y: $p(Y = y|do(T = t), Z = z)$
Distribution of Y in a subset of the population for which Z=z
(Recall, ATE vs CATE)

Example: How does the treatment affect a specific age group, or individuals with blood sugar levels = z, etc.

We will use the adjustment formula, but modified for the conditional case above: Paths need to remain blocked when we additionally condition on Z.

The z-specific effect $p(Y = y|do(T = t), Z = z)$ is **identified** whenever we can measure a set S of variable such that $S \cup Z$ satisfies the backdoor criterion. The z-specific effect is given by the **modified adjustment formula**:

$$p(Y = y|do(T = t), Z = z) = \sum_s p(Y = y|T = t, S = s, Z = z)P(S = s|Z = z)$$

Conditional Interventions & Covariate-specific effects

"z-specific effect" of T on Y: $p(Y = y|do(T = t), Z = z)$
Distribution of Y in a subset of the population for which Z=z
(Recall, ATE vs CATE)

Example: How does the treatment affect a specific age group, or individuals with blood sugar levels = z, etc.

We will use the adjustment formula, but modified for the conditional case above: Paths need to remain blocked when we additionally condition on Z.

The z-specific effect $p(Y = y|do(T = t), Z = z)$ is **identified** whenever we can measure a set S of variable such that $S \cup Z$ satisfies the backdoor criterion. The z-specific effect is given by the **modified adjustment formula**:

$$p(Y = y|do(T = t), \boxed{Z = z}) = \sum_s p(Y = y|T = t, S = s, \boxed{Z = z}) P(S = s|\boxed{Z = z})$$

Conditional Interventions & Covariate-specific effects

Need to ensure conditional on Z does not open back-doors, e.g.,
if Z is a collider, we need to make sure spurious paths created by it are blocked

Back to our z-dependent policy $p(Y = y|do(T = g(Z))$:

$$p(Y = y|do(T = g(Z)) = \sum_z p(Y = y|do(T = g(Z)), Z = z)p(Z = z|do(T = g(Z))$$

Since Z occurs before T

$$= \sum_z p(Y = y|do(T = g(Z)), Z = z)p(Z = z)$$

$$= \sum_z p(Y = y|do(T = t), Z = z)|_{t=g(z)}p(Z = z)$$

Conditional Interventions & Covariate-specific effects

Need to ensure conditional on Z does not open back-doors, e.g.,
if Z is a collider, we need to make sure spurious paths created by it are blocked

Back to our z-dependent policy $p(Y = y|do(T = g(Z))$:

$$p(Y = y|do(T = g(Z)) = \sum_z p(Y = y|do(T = g(Z)), Z = z)p(Z = z|do(T = g(Z))$$

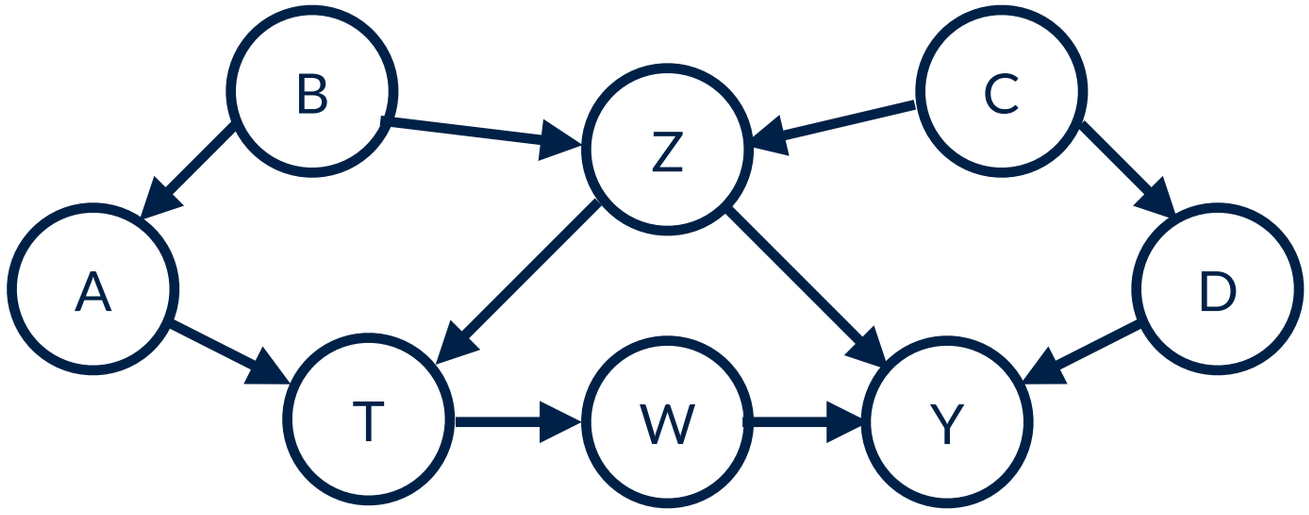
Since Z occurs before T

$$= \sum_z p(Y = y|do(T = g(Z)), Z = z)p(Z = z)$$

$$= \sum_z p(Y = y|do(T = t), Z = z)|_{t=g(z)}p(Z = z)$$

Suppose Z only takes one value in this sum z*, then only one term Z=z* occurs in the sum with probability one. More generally,
If multiple Zs can occur, the stochastic intervention is an average of the z-specific effects.

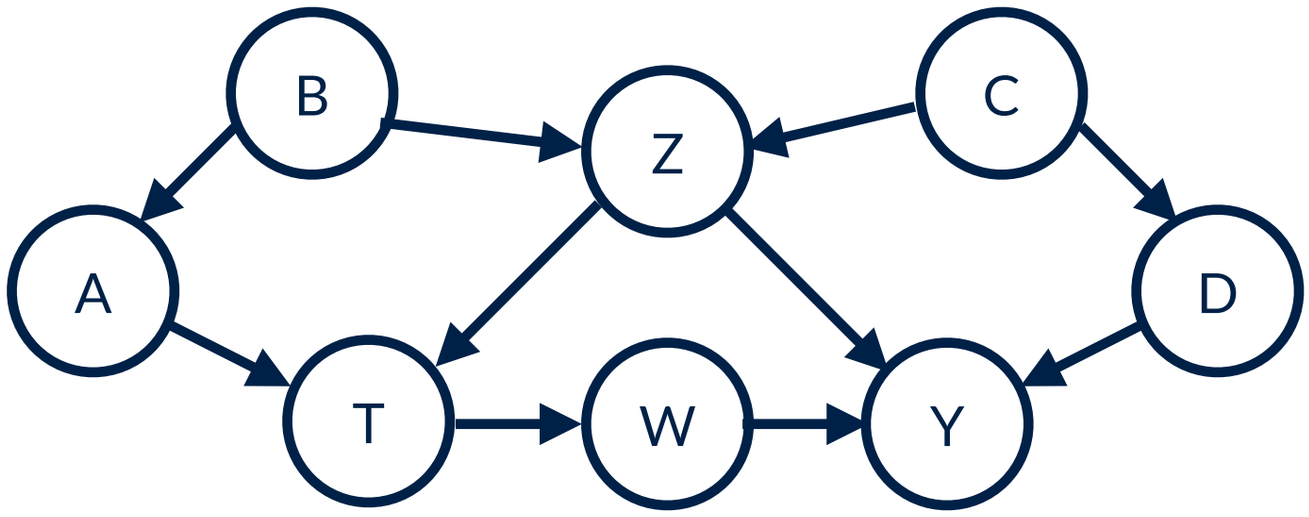
Example: c-specific effect of T on Y



Q: What is the causal effect of T on Y?

Pearl's Primer book, 3.5.1

Example: c-specific effect of T on Y



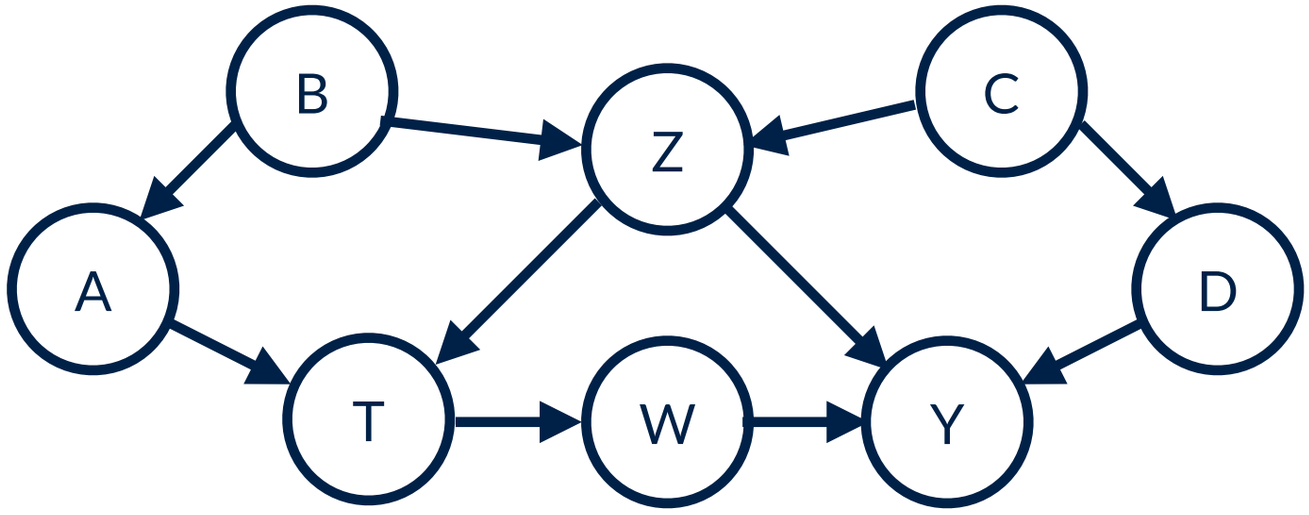
Q: What is the causal effect of T on Y?

Recalling the adjustment formula, we need to condition on Z, which is a collider node, so need to block the spurious path by e.g., condition on A (a parent of T)

$$p(Y = y|do(T = t)) = \sum_{z,a} p(Y = y|T = t, Z = z, A = a)p(Z = z, A = a)$$

Pearl's Primer book, 3.5.1

Example: c-specific effect of T on Y



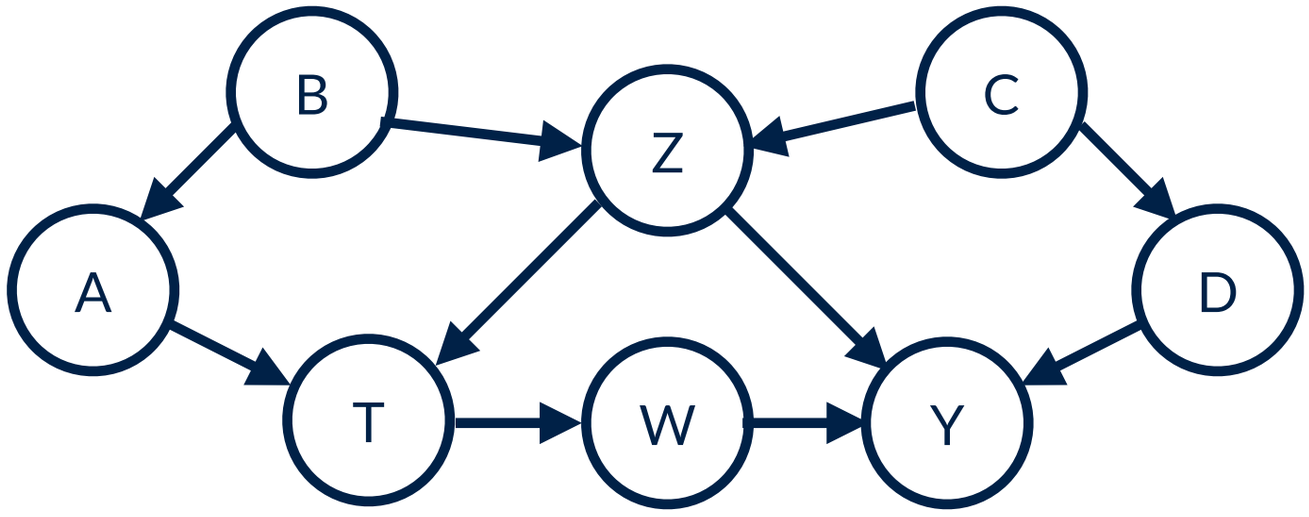
Q: What is the causal effect of T on Y?

Recalling the adjustment formula, we need to condition on Z, which is a collider node, so need to block the spurious path by e.g., condition on C will also work:

$$p(Y = y|do(T = t)) = \sum_{z,c} p(Y = y|T = t, Z = z, C = c)p(Z = z, C = c)$$

Pearl's Primer book, 3.5.1

Example: c-specific effect of T on Y



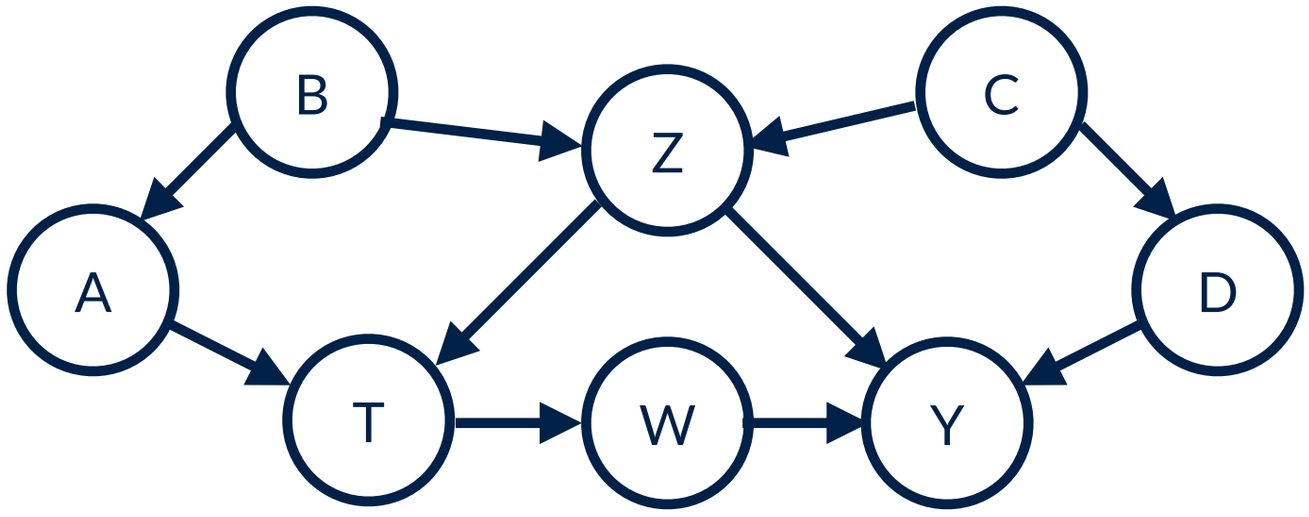
Q: What is the c-specific causal effect of T on Y?

$$p(Y = y|do(T = t), C = c) = \sum_z p(Y = y|T = t, Z = z, C = c)p(Z = z|C = c)$$

(From the rule on slide 6.)

Pearl's Primer book, 3.5.1

Example: c-specific effect of T on Y



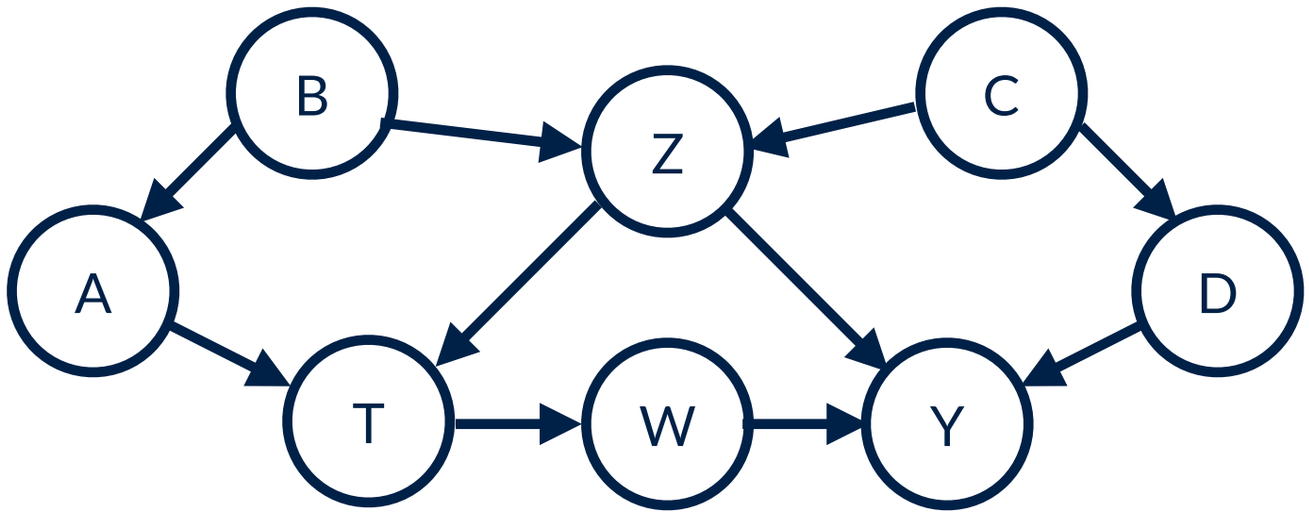
Q: What is the z-specific causal effect of T on Y?

$$p(Y = y|do(T = t), Z = z) = \sum_c p(Y = y|T = t, Z = z, C = c)p(C = c|Z = z)$$

(From the rule on slide 6.)

Pearl's Primer book, 3.5.1

Example: c-specific effect of T on Y



Q: What is the z-dependent causal effect of T on Y, under the strategy:

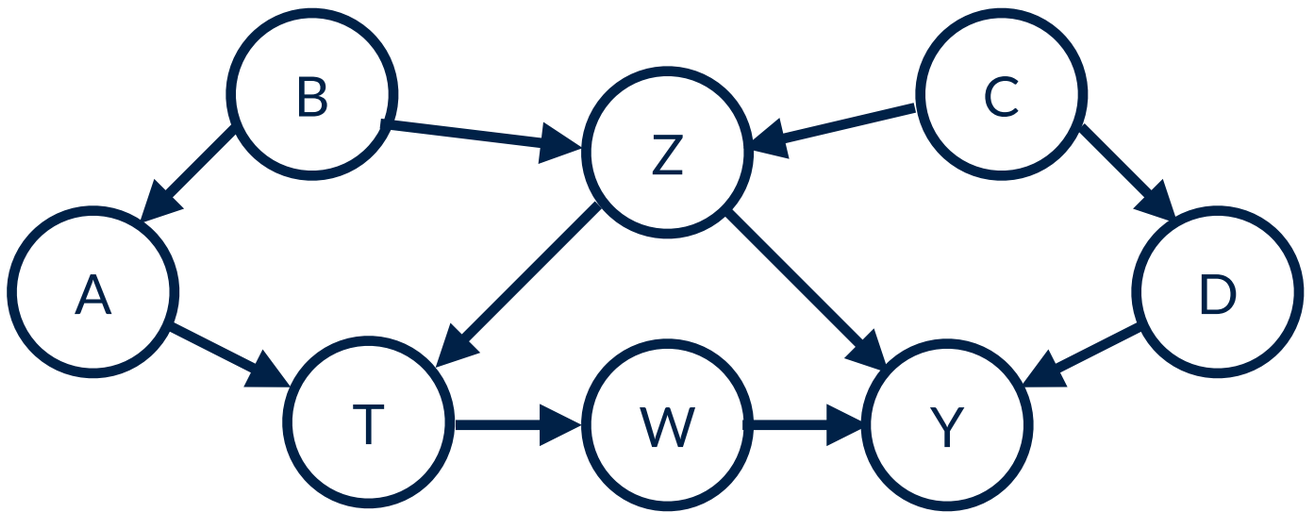
$$g(Z) = \begin{cases} 0 & Z \leq 2 \\ 1 & Z > 2 \end{cases} \quad \text{where} \quad Z \in \{1, 2, 3, 4, 5\}$$

$$p(Y = y|do(T = g(Z))) = \sum_z p(Y = y|do(T = g(Z)), Z = z)p(Z = z) \quad \text{product rule}$$

$$\begin{aligned} &= p(Y = y|do(T = 0), Z = 1)p(Z = 1) \\ &+ p(Y = y|do(T = 0), Z = 2)p(Z = 2) \\ &+ p(Y = y|do(T = 1), Z = 3)p(Z = 3) \\ &+ p(Y = y|do(T = 1), Z = 4)p(Z = 4) \\ &+ p(Y = y|do(T = 1), Z = 5)p(Z = 5) \end{aligned}$$

Pearl's Primer book, 3.5.1

Example: c-specific effect of T on Y



Q: What is the <u>z-dependent</u> causal effect of T on Y, under the strategy:

$$g(Z) = \begin{cases} 0 & Z \leq 2 \\ 1 & Z > 2 \end{cases} \quad \text{where} \quad Z \in \{1, 2, 3, 4, 5\}$$

$$p(Y = y|do(T = g(Z)) = \sum_z p(Y = y|do(T = g(Z)), Z = z)p(Z = z) \quad \text{product rule}$$

Use results from the z-specific effect

$$\begin{aligned} &= p(Y = y|do(T = 0), Z = 1)p(Z = 1) \\ &+ p(Y = y|do(T = 0), Z = 2)p(Z = 2) \\ &+ p(Y = y|do(T = 1), Z = 3)p(Z = 3) \\ &+ p(Y = y|do(T = 1), Z = 4)p(Z = 4) \\ &+ p(Y = y|do(T = 1), Z = 5)p(Z = 5) \end{aligned}$$

Pearl's Primer book, 3.5.1

Estimation via Inverse Probability Weighing (IPW)

Practical problem with intervention procedures: backdoor and front-door criteria tell us whether it is ***possible*** to predict the result of interventions, e.g.,

$$p(Y = y|\mathrm{do}(X = x))$$

can be expressed in terms of observed probabilities.

This requires conditioning on a set Z of covariates satisfying one of the criteria

… but in practice, this conditioning may be problematic!

For example:

1. Z may consist of many variables, each spanning many values
2. Number of samples with $Z = z$ may be small —> poor statistics

One approach (recall): **Inverse Probability Weighing (IPW)**

Estimation via Inverse Probability Weighing (IPW)

Approach: Inverse Probability Weighing (IPW)

Requires:

1. Adjustment set of variables Z to apply one of the criteria
2. Corresponding propensity score function: $g(x, z) = p(X = x|Z = z)$

Justification for Inverse Probability Weighing (IPW)

Comparison with <u>conditional probability</u>

Like filtering:

1. Omit all cases for which, e.g., X = x does *<u>not</u>* hold
2. Normalise the surviving cases so probabilities add up to one.

In practice, this is done by uniformly multiplying by a factor $1/p(X = x)$, i.e.,

$$p(Y = y, Z = z|X = x) = \frac{p(Y = y, Z = z, X = x)}{p(X = x)}$$

Can think of this as the probability of each surviving case being boosted by this 1/P(X=x) factor (**uniformly** so).

Justification for Inverse Probability Weighing (IPW)

Hypothetical probability

Examine the population created by the operation $\mathrm{do}(X = x)$, and see how each case is changed as a result of this operation.

This follows from the adjustment formula w.r.t. an adjustment set Z:

$$
\begin{aligned}
p(Y = y|\mathrm{do}(X = x)) &= \sum_z p(Y = y|X = x, Z = z)p(Z = z) \\
&= \sum_z \frac{p(Y = y|X = x, Z = z)p(X = x|Z = z)p(Z = z)}{p(X = x|Z = z)} \\
&= \sum_z \frac{p(Y = y, X = x, Z = z)}{p(X = x|Z = z)}
\end{aligned}
$$

So, each case $(Y = y, X = x, Z = z)$ in the population has its probability **(non-uniformly!)** boosted by the factor $1/p(X = x|Z = z)$.

Example on Inverse Probability Weighing (IPW)

Simpson's paradox: Drug that seems to help men and woman separately, but hurt the general population

X = took drug, Y = recovered, Z = sex

Table 3.3 Joint probability distribution $P(X, Y, Z)$ for the drug-gender-recovery story of Chapter 1 (Table 1.1)

X	Y	Z	% of population
Yes	Yes	Male	0.116
Yes	Yes	Female	0.274
Yes	No	Male	0.01
Yes	No	Female	0.101
No	Yes	Male	0.334
No	Yes	Female	0.079
No	No	Male	0.051
No	No	Female	0.036

Pearl's Primer book, p73-74

Example on Inverse Probability Weighing (IPW)

Condition on “X = Yes”:

Table 3.4 Conditional probability distribution $P(Y, Z|X)$ for drug users $(X = yes)$ in the population of Table 3.3

X	Y	Z	% of population
Yes	Yes	Male	0.232
Yes	Yes	Female	0.547
Yes	No	Male	0.02
Yes	No	Female	0.202

$$P(Y, Z|X) = \frac{P(Y, Z, X)}{P(X)}$$

Table is produced in two steps:

1. Those with “X = No” are removed
2. Weights in final row are renormalised (so they add up to one) by multiplying with the constant 1 / P(X = yes), computed by using “X = Yes” from 3.3:

$$P(X = yes) = 0.116 + 0.274 + 0.01 + 0.101 = 0.501$$

Pearl’s Primer book, p73-74

Example on Inverse Probability Weighing (IPW)

Next, consider the population $\mathrm{do}(X = \mathrm{Yes})$:

1. Calculate the distribution of weights, i.e., according to Table 3.3:

$$P(X = yes|Z = Male) = \frac{(0.116 + 0.01)}{(0.116 + 0.01 + 0.334 + 0.051)} = 0.247$$

$$P(X = yes|Z = Female) = \frac{(0.274 + 0.101)}{(0.274 + 0.101 + 0.079 + 0.036)} = 0.765$$

Be careful with rounding errors

2. Multiply weights by 1/0.247 (Z = Male) and 1/0.765 (Z = Female), obtain:

Table 3.5 Probability distribution for the population of Table 3.3 under the intervention $do(X = Yes)$, determined via the inverse probability method

X	Y	Z	% of population
Yes	Yes	Male	0.476
Yes	Yes	Female	0.357
Yes	No	Male	0.041
Yes	No	Female	0.132

We deduce: $P(Y = yes|do(X = yes)) = 0.476 + 0.357 = 0.833$

Pearl's Primer book, p73-74

Remarks on Inverse Probability Weighing (IPW)

Remarks:

1. Redistribution of the probabilities is not uniform (cf. 3.5: Rows 1 and 2)

2. May lead to significant computational savings: Only need to estimate the propensity score $g(x, z) = p(X = x|Z = z)$ for the values Z = z that are *actually observed in the data*, i.e., for at most as many Z as the sample size

   If Z has many more values than the sample size, this can be a great help

3. Caution: The method of IPW to compute

$$p(Y = y|\mathrm{do}(X = x))$$

   is only valid when the set of variables Z satisfies the backdoor criterion.

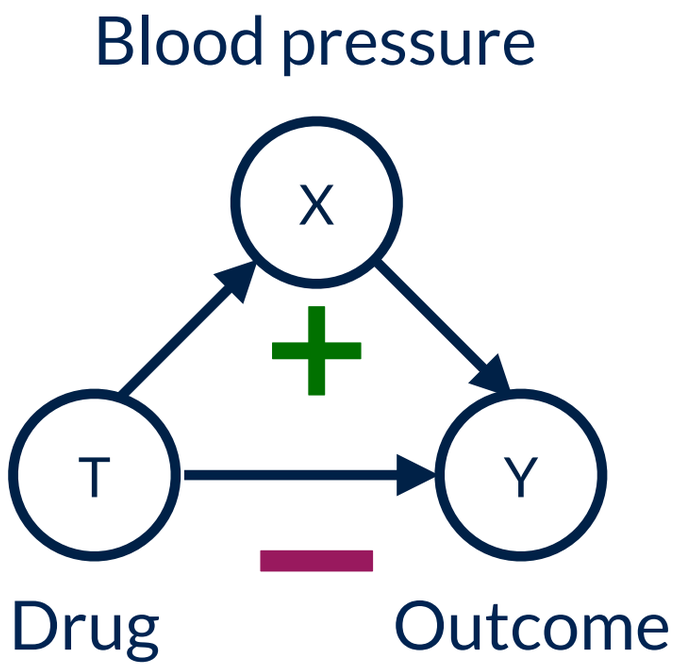
Mediation

Mediation

A variable may cause another, directly or indirectly through a set me mediating variables.

Example: Treatment decrease blood pressure, and through this process, increases recovery. But treatment has a direct negative effect on recovery. ‘Overall, is treatment good or bad?’ (We did this in lecture 10)

‘How much of the effect is direct and how much indirect?’ Non-trivial!



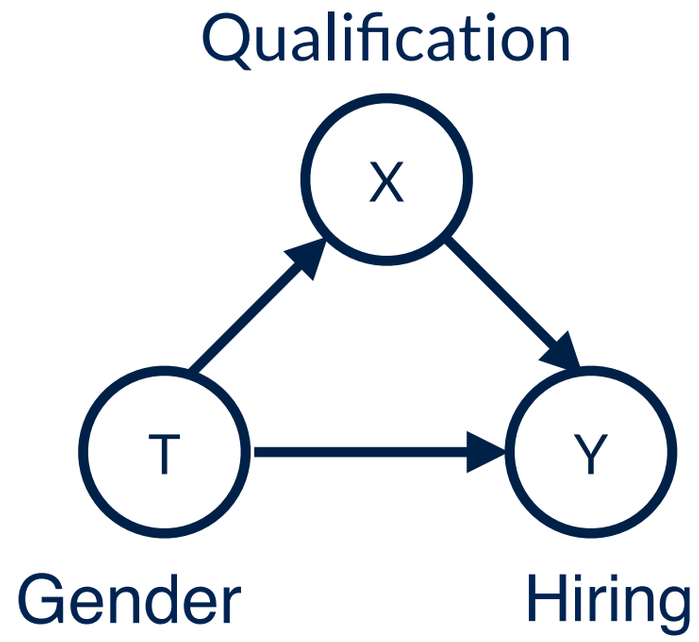
Mediation: Example

Q: If and to what degree a company discriminates by gender (T) in hiring (Y).

(i) Direct discrimination based on gender (illegal)
(ii) Indirect: Gender affects hiring practices, e.g., could be that women are more/ less likely to go in a particular field

To answer the question, we are focusing on (i), meaning we need to keep (ii) steady and measure the remaining relationship between gender and hiring

‘With qualifications held constant, any change hiring would have been due to gender alone.’



Mediation: Example

Q: If and to what degree a company discriminates by gender (T) in hiring (Y).

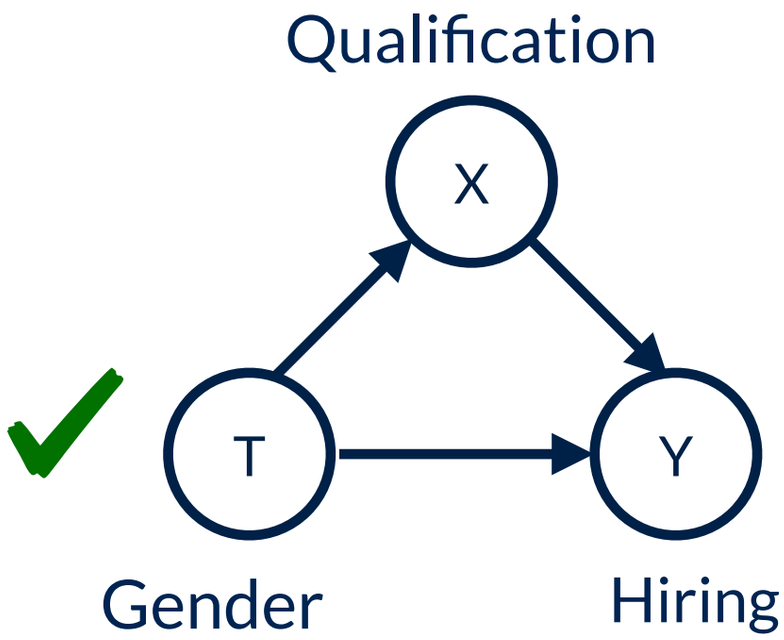
(i) Direct discrimination based on gender (illegal)
(ii) Indirect: Gender affects hiring practices, e.g., could be that women are more/ less likely to go in a particular field

To answer the question, we are focusing on (i), meaning we need to keep (ii) steady and measure the remaining relationship between gender and hiring

One way is to condition on the mediator, and see if:

p(Hired | Female, Highly qualified) = or !=
p(Hired | Male, Highly qualified)

If not equal, then there is a direct effect of gender on hiring



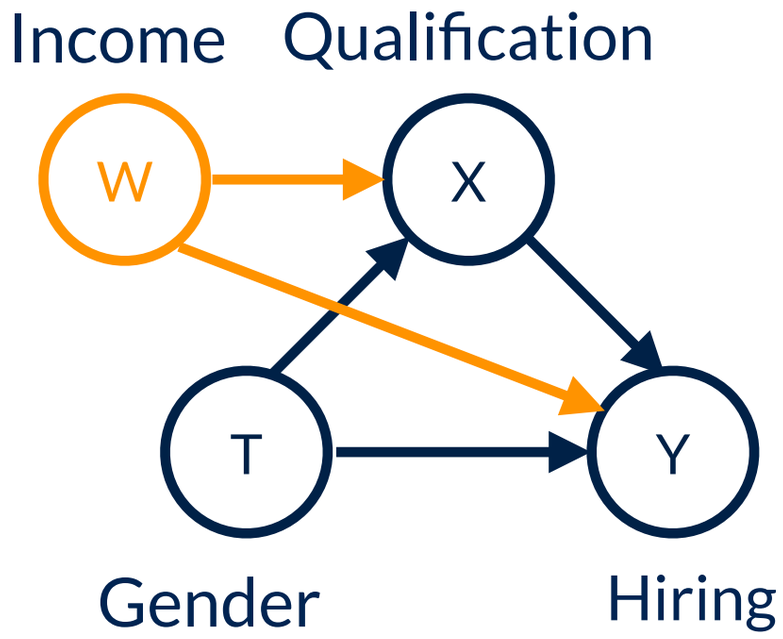
Mediation: Example

Complication: Suppose there is a confounder for the mediator X and outcome Y.

Individuals from higher income families are more likely to have gone to uni and/or have more connection that would help them get hired.

Now, if we condition on qualification to get the direct effect, we have a collider!

T -> X <- W -> Y



Mediation: Example

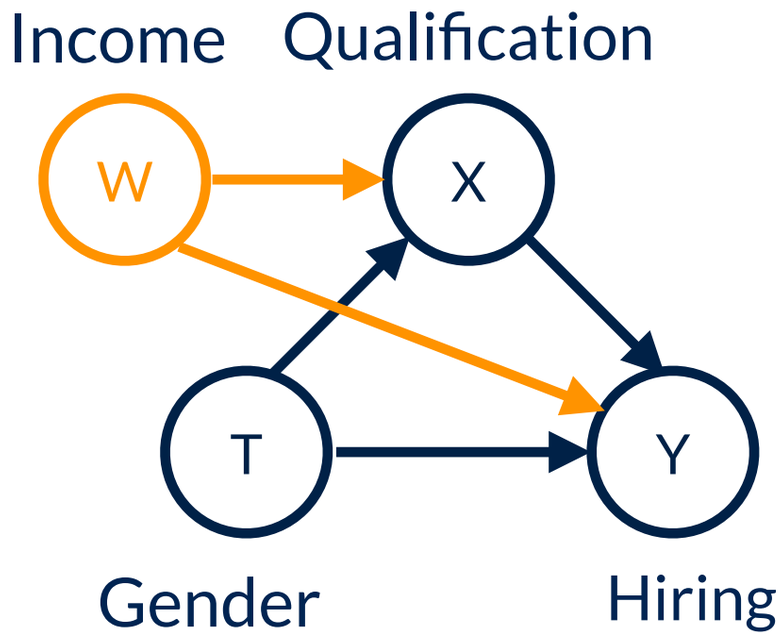
Complication: Suppose there is a confounder for the mediator X and outcome Y.

Individuals from higher income families are more likely to have gone to uni and/or have more connection that would help them get hired.

Now, if we condition on qualification to get the direct effect, we have a collider!

T -> X <- W -> Y

So we have to block that path …



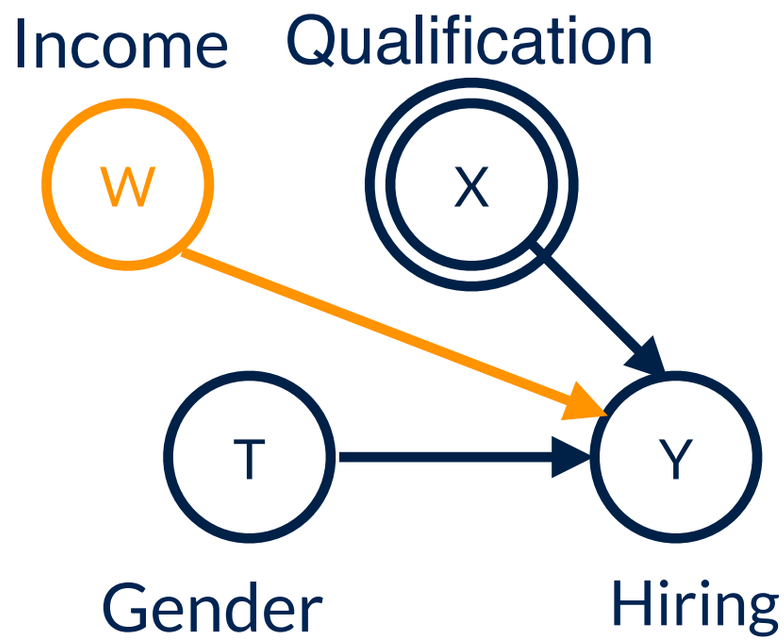
Mediation: Example

Apply the do-operator on qualification.

In the hypothetical graph below, there are no spurious paths and only the direct effect remains. Again, need to reduce the do-operators to usual expression in terms of probabilities. We have:

Controlled Direct Effect (CDE):

$$p(Y = y|do(T = t), do(X = x)) - p(Y = y|do(T = t'), do(X = x))$$



Mediation: Example

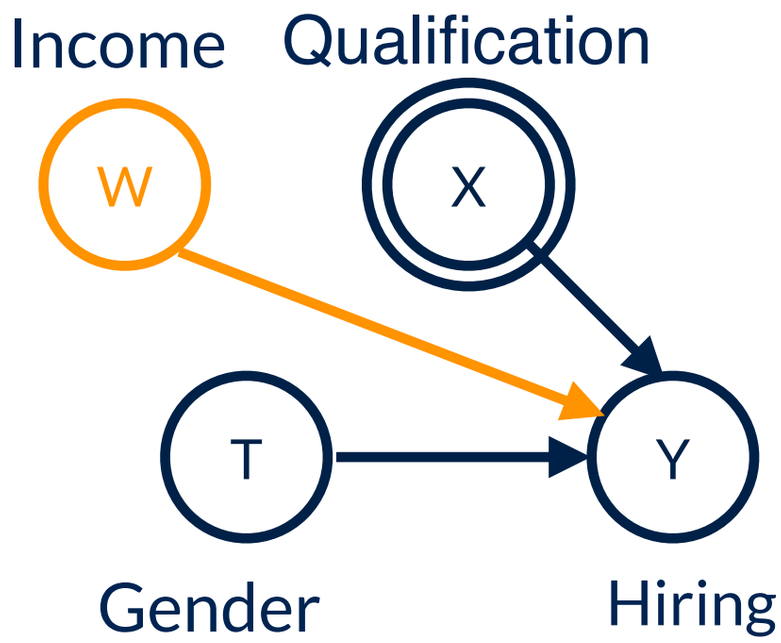
Apply the do-operator on qualification.

In the hypothetical graph below, there are no spurious paths and only the direct effect remains. Again, need to reduce the do-operators to usual expression in terms of probabilities. We have:

Controlled Direct Effect (CDE):

$$p(Y = y|do(T = t), do(X = x)) - p(Y = y|do(T = t'), do(X = x))$$

Notice: Direct effect may differ for different values of X, e.g., hiring practices may discriminate against women in jobs requiring higher qualifications etc.



Mediation: Example

Controlled Direct Effect (CDE):

$$p(Y = y|do(T = t), do(X = x)) - p(Y = y|do(T = t'), do(X = x))$$

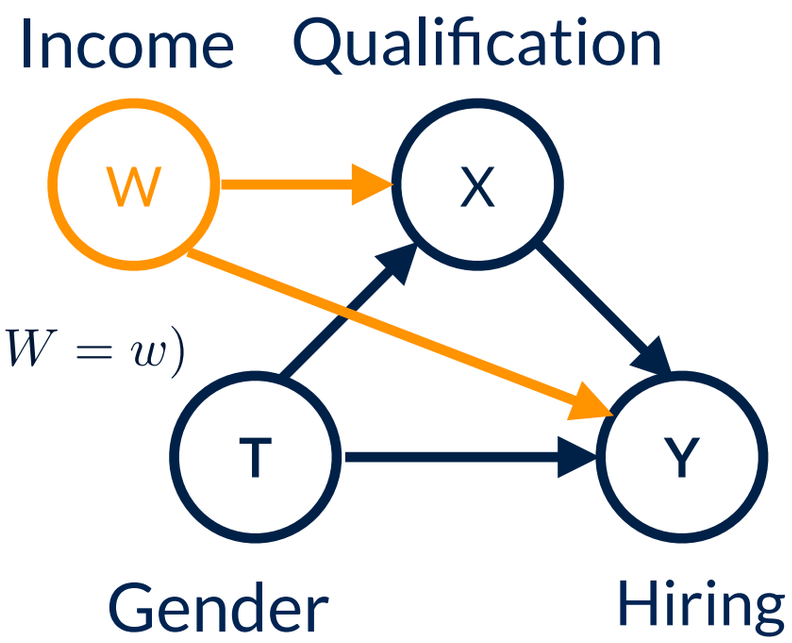
There are no backdoor paths from T to Y, hence the above is equal to:

$$p(Y = y|T = t, do(X = x)) - p(Y = y|T = t', do(X = x))$$

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

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

$$\sum_w \Big(p(Y = y|T = t, X = x, W = w) - p(Y = y|T = t', X = x, W = w)\Big)p(W = w)$$



Mediation

In general: the CDE of T on Y, meditated by X is identifiable if:

1) There exits a set S_1 of variables that blocks all back-door paths from X to Y
2) There exits a set S_2 of variables that blocks all back-door paths from T to Y, after deleting all arrows entering X.

Remark: (2) is not necessary in randomised control trials