Lecture 4: Linking the observed data to the SCM, The statistical model

A roadmap for causal inference

1. Specify **Causal Model** representing <u>real</u> background knowledge
2. Specify **Causal Question**
3. Specify **Observed Data** and link to causal model
4. **Identify** : Knowledge + data sufficient?
5. Commit to an **estimand** as close to question as possible, and a **statistical model** representing <u>real</u> knowledge.
6. **Estimate**
7. **Interpret** Results

Outline

1. The observed data and their link to the SCM
2. From the SCM to a Statistical Model
   - What causal structures result in statistical independence
     - The d-separation criteria
   - How to tell what restrictions, if any, the SMC places on the allowed distributions for the observed data

References

- TLB. Chapter 2
- Pearl. “An Introduction to Causal Inference” *Int J Biostat*, 6(2): Article 7, 2010.
- Greenland and Pearl. “Causal Diagrams” in S. Boslaugh, editor, Encyclopedia of Epidemiology. Sage Publications, Thousand Oaks, CA, 2007
- Greenland, Pearl and Robins. “Causal Diagrams for Epidemiologic Research” *Epidemiology*, 10(1): 37-48, 1999.

Specify the Observed Data

- Of course, you can do this without the benefit of a causal model
- Example: Observed data are
  - Baseline covariates W= (age, sex, SES)
  - Exposure A= Vitamin Use
  - Outcome Y= Breast cancer status at 5 years
  - $O=(W,A,Y)\sim P_0$
- Note: the observed data are a (multidimensional) random variable and have a distribution

Linking the observed data to the SCM

- We assume that the observed data were generated by a data generating system compatible with our SCM
  - O is a subset of X
- The distribution of U and the structural equations F identify the distribution of X, and thus the distribution of O
  - For example: For O=X

$$P_0(O=o) = \sum_u P_f(X=x|U=u)P(U=u) = \sum_u I(X(u)=x)P(U=u)$$

Linking the observed data to the SCM

- We observe a sample of size n of the random variable O
  - For now we will work with independent samples
  - The framework is not restricted to this
- This gives us n i.i.d. copies $O_1, O_2, \ldots, O_n$ drawn from probability distribution P_0
- In other words, we assume our observed data were generated by sampling *n* times from the data generating system specified in our causal model

What restrictions, if any, does the SCM put on the observed data distribution?

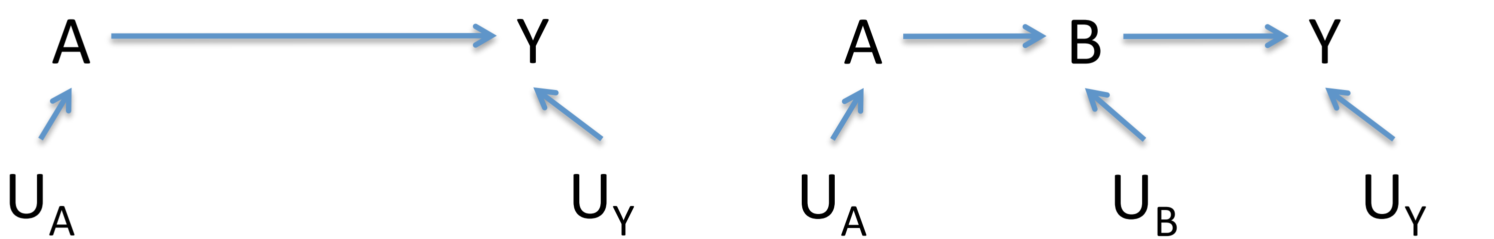
- How to tell what these are?
  - You can work it out based on functions F and allowed distributions for P_U
  - You can evaluate the causal graph using the d-separation criteria

- Why is this helpful? Any such restrictions

1. Give us our statistical model
   - Set of allowed distributions for P_0
2. Give us the testable implications (if any) of our SCM

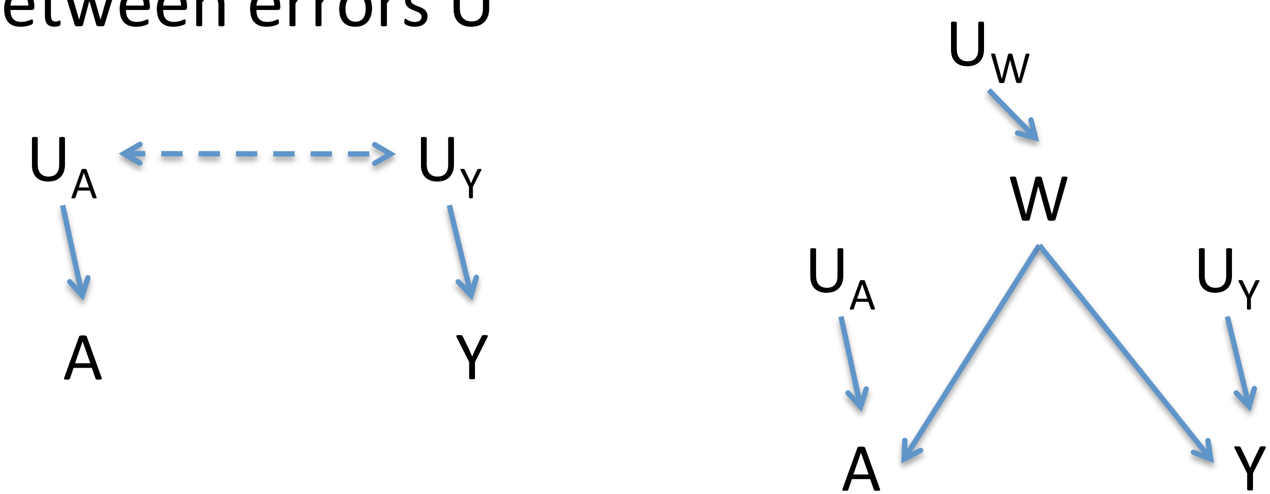
What causal structures can lead to dependence between two observed variables?

1. <u>Direct and Indirect Effects</u>
   – An effect of A on Y can result in an association



What causal structures can lead to dependence between two observed variables?

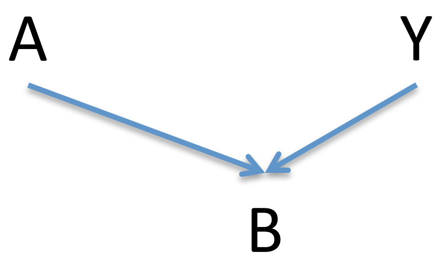
2. <u>Shared common cause</u>
   - Common cause (measured or unmeasured) of A and Y can result in an association
   - When the common cause is not included in X, it is represented through the dependence it induces between errors U

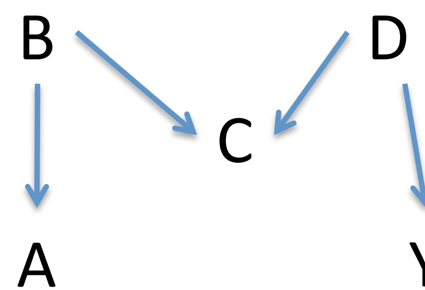


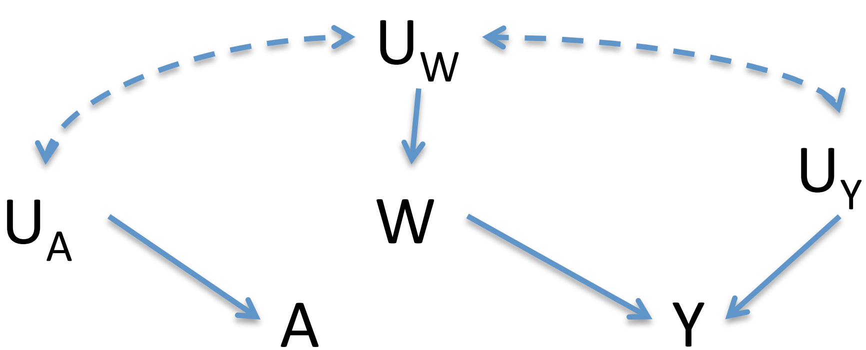
If neither of these sources of dependence are present, A and Y will be independent in every probability distribution P_0 compatible with the SCM

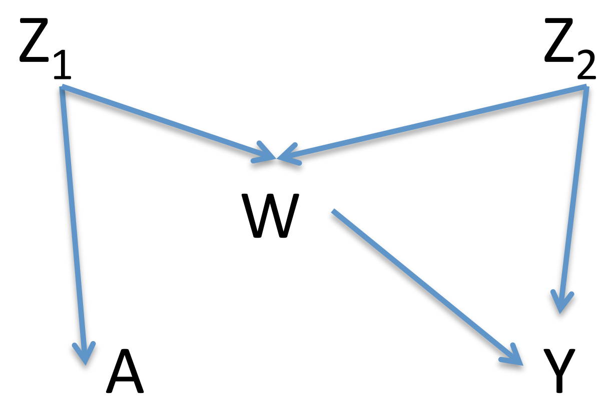
- In other words, any data generating experiment that is compatible with the SCM will give rise to an observed data distribution in which A and Y are independent
  - Regardless of functional form, strength of associations, etc…
  - See also Pearl Causality Chapter 1

Examples: A and Y independent?



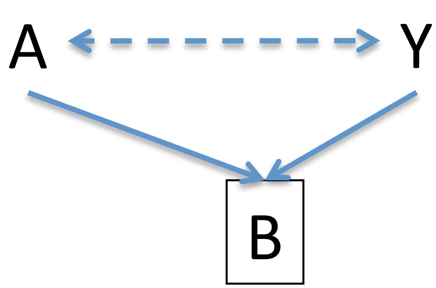


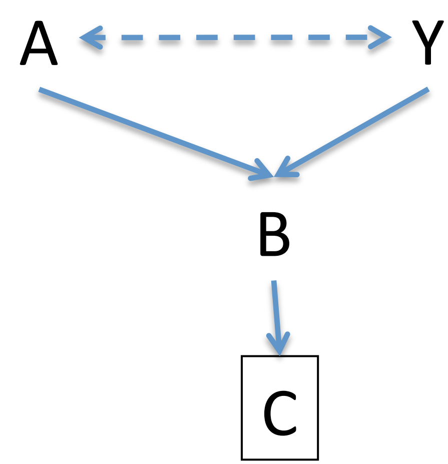




What causal structures can lead to dependence between two observed variables?

3. Conditioning on a Collider
   - Collider= “inverted fork” a->b<-c
   - Conditioning on a common effect (descendent) of A and Y can result in an association between A and Y
     - Berkson’s bias/ selection bias





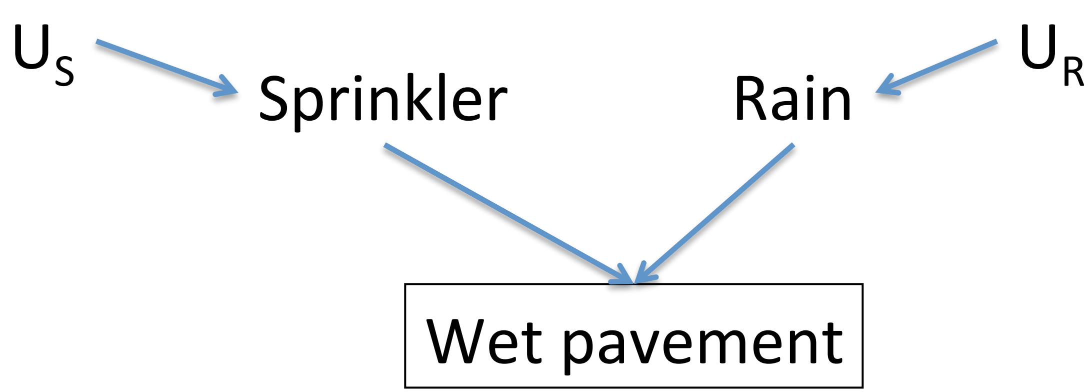
Example: Sprinkler, rain and wet pavement

- Say the sprinkler is on a timer (unknown to us)
- What would an SCM look like for these three variables?
- Does the Graph imply that the sprinkler going off and it raining are independent?
  - If we learn that the sprinkler went off, does it give us any additional information about whether it rained?
  - If we learn that it rained, does it give us any additional information about whether the sprinkler went off?

Pearl, *Causality* 2001

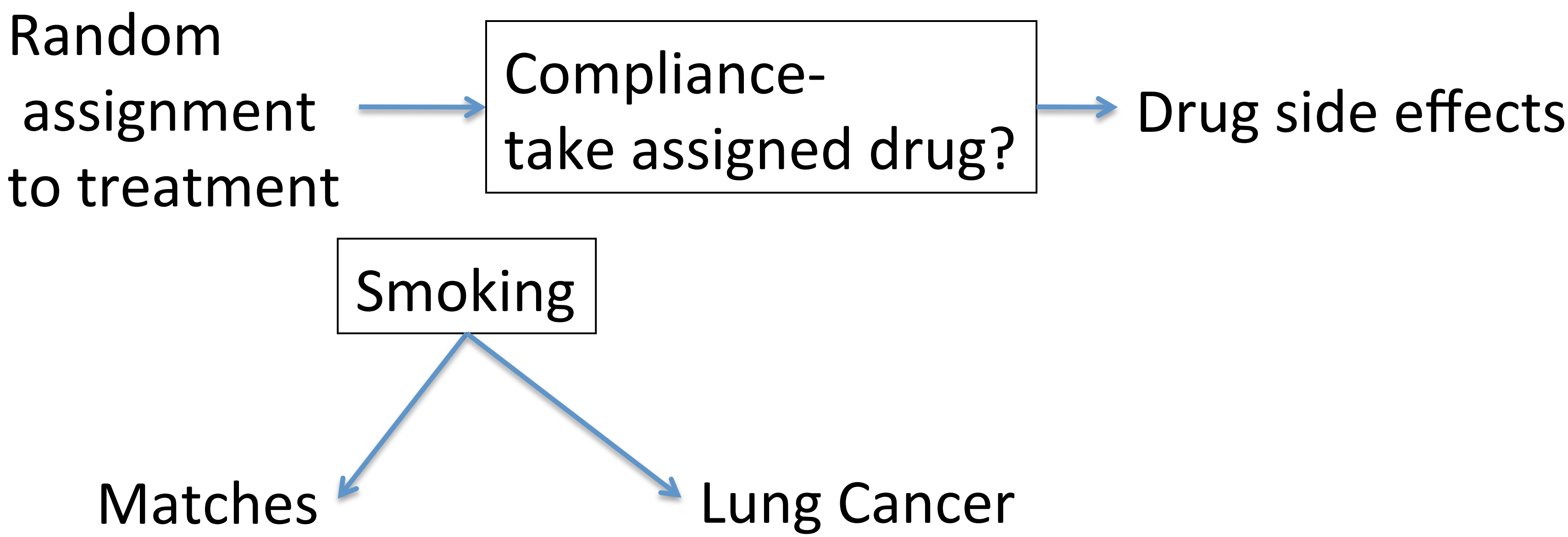
Example: Sprinkler, rain and wet pavement

- Both cause the pavement to be wet
- Are sprinkler and rain independent conditional on whether the pavement is wet?
  - If we see the pavement is wet and learn that the sprinkler went off, does it give us any additional information about whether it rained?



What can remove a source of dependence between variables?

- Conditioning on a causal intermediate or shared common cause between A and Y will remove that source of dependence

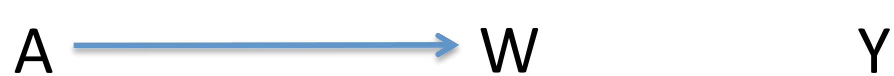


Summing Up: When does a graph/SCM imply that variables/nodes are independent?

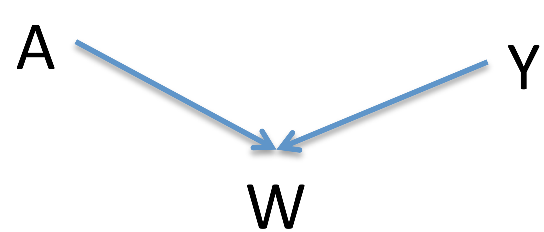
- Path= set of connected edges (any directionality)
- <u>A is (unconditionally) independent of Y</u> if

1. There is <u>no path</u> between A and Y

Or,

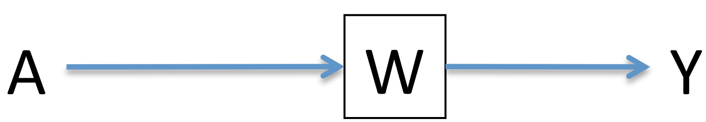


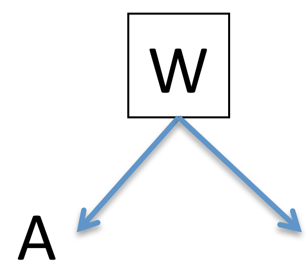
2. All paths between A and Y are "blocked" by a collider



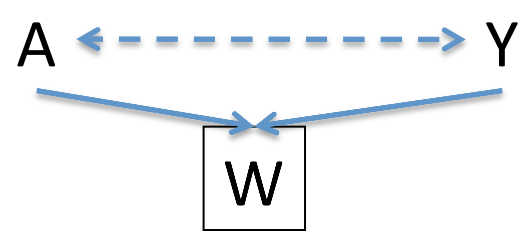
Summing Up: When does a graph imply that variables/nodes are independent?

- A is independent of Y given W if W blocks all unblocked paths and doesn't create any new unblocked paths
  - Conditioning on a non-collider blocks a path





  - Conditioning on a collider (or a descendent of a collider) opens a path



So when is a path blocked?

- If it has a non-collider that has been conditioned on

or

- If it has a collider *and* neither the collider nor a descendent of the collider has been conditioned on

- Absence of unblocked paths implies (conditional) independence

Definition: D-separation

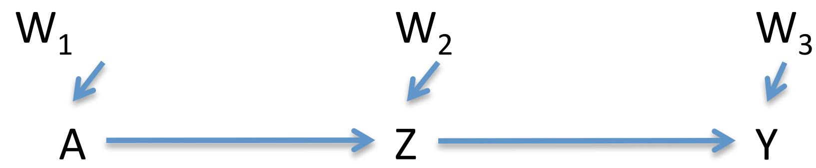
- A set of nodes S blocks a path if either

1) The path has at least one arrow emitting node in S
   - The path is open, but S blocks it
2) The path has at least one collider that is outside of S and has no descendent in S
   - The path is already blocked and S doesn't open it

If S blocks all paths from X to Y then S d-separates X and Y, and X is independent of Y given S

Examples of d-separation

- Path W_1 to Y
  - Blocked by {}?
  - Blocked by W_2?
  - Blocked by Z?
- Path W_1 to W_2
  - Blocked by {}?
  - Blocked by Z?
  - Blocked by Y?



Why is *d-separation* helpful?

- Tells us what restrictions (if any) our causal model puts on the allowed distributions for the observed data
  - d-separation between two variables X and Y conditional on some subset of additional variables S implies X independent of Y given S in every probability distribution compatible with the SCM
- This is testable
- This gives us our statistical model

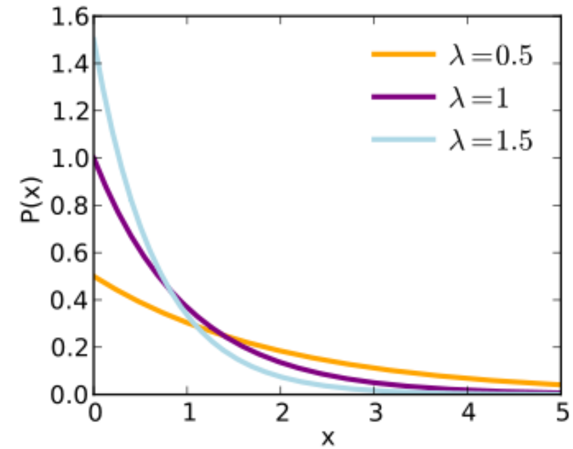
Statistical Models

- The statistical model $\mathcal{M}$ is the set of possible distributions for the observed data: $P_0 \in \mathcal{M}$
- Non-parametric Model: No restrictions on the set of possible distributions P_0
- Semi-parametric Model: Puts some restrictions on the set of possible distributions for P_0
- Parametric Model: Assumes that we know P_0 up to a finite number of unknown parameters

Example: Parametric Model

- O=X~P_0 ; X continuous
- We know that X has exponential distribution

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$



- $\mathcal{M}$ is the set of all exponential distributions
  - Completely described by a single parameter λ
- We know the distribution of X up to a finite number (in this case one) of unknown parameters

Non-Parametric Model

- All we know is that we have n i.i.d. copies of O
  - We may have some additional knowledge about how O was generated (represented in our SCM),
  - This knowledge does not imply any restrictions on the set of possible distributions for P_0
- Example: Non-parametric Model
  - $O=X \sim P_0$; X continuous
  - $\mathcal{M}$ is the set of all possible distributions for X
  - We do not know the distribution of X up to a finite number of unknown parameters

Parametric vs. Non-Parametric Models

- Parametric and Non-parametric models are not mutually exclusive categories
- If O is discrete with finite number of outcomes or possible values then we always know its distribution up to a finite number of unknown parameters
- So a <u>non-parametric model for discrete O</u> (ie a model that does not restrict the set of allowed distributions for O) <u>is</u> <u>also technically parametric</u> (has a finite number of unknown parameters).

Parametric vs. Non-Parametric Models

- Parametric and Non-parametric models are not mutually exclusive categories
- Example: O=(A,W), A and W both binary
  - 4 possible outcomes for O:

    (A=1,W=1); (A=1,W=0); (A=0,W=1); (A=0,W=0)
  - Probabilities of these mutually exclusive events have to sum to 1

  -> Knowing the probability of 3 of these events gives us the whole probability distribution for O

Parametric vs. Non-Parametric Models

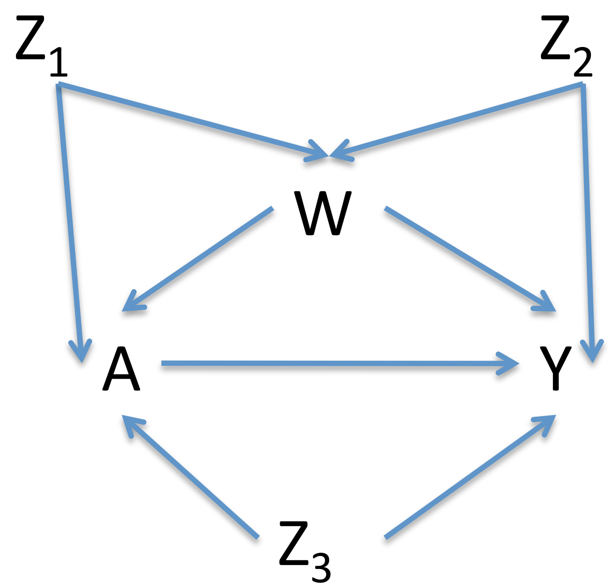
- A <u>non-parametric model for discrete O </u>(ie a model that does not restrict the set of allowed distributions for O) <u>is also parametric </u>(has a finite number of unknown parameters).
  - However, may be many possible parametric models for O that <u>do </u>restrict its set of allowed distributions.
- The point is not that parametric (or semi-parametric) models are bad
- <u>The point is that our model should accurately reflect our knowledge</u>
  - It should contain the true distribution of O

Recap: Why is *d-separation* helpful?

- Tells us what restrictions (if any) our causal model puts on the allowed distributions for the observed data
  - d-separation between two variables X and Y conditional on some subset of additional variables S implies X independent of Y given S in every probability distribution compatible with the SCM
- This is testable
- This gives us our statistical model

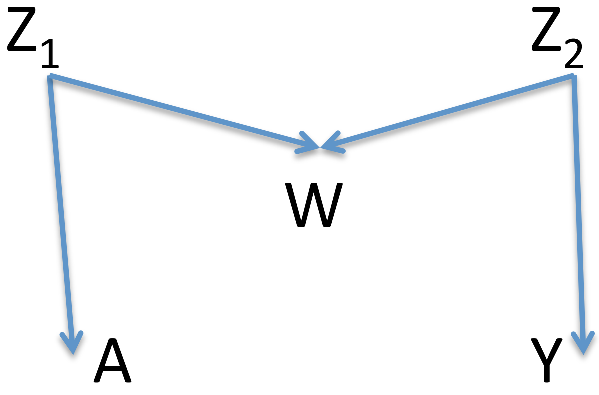
Examples

- O=(W,A,Y)
  - No exclusion restrictions or independence assumptions
- What is the graph?
- What are the testable implications?
- What is the statistical model?



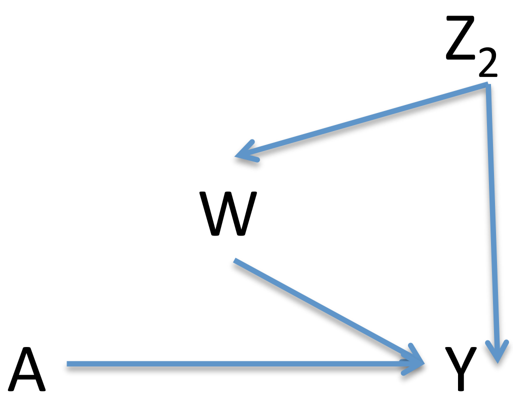
Examples

- O=(W,A,Y)
  - W does not affect A or Y; A does not affect Y
  - U_A and U_Y independent
- What is the graph?
- What are the testable implications?
- What is the statistical model?



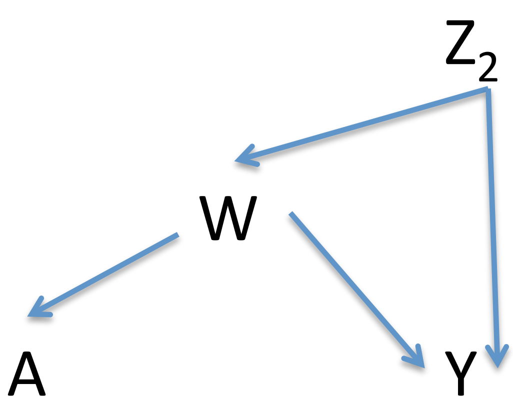
Examples

- O=(W,A,Y)
  - W does not affect A
  - U_A independent of U_Y and U_W
- What is the graph?
- What are the testable implications?
- What is the statistical model?



Examples

- O=(W,A,Y)
  - A does not affect Y
  - U_A independent of U_Y and U_W
- What is the graph?
- What are the testable implications?
- What is the statistical model?



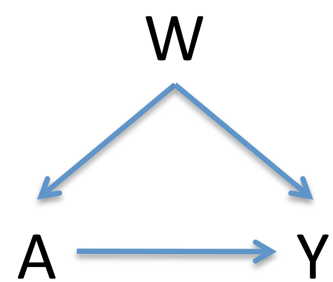
<u>D-separation implies conditional independence:</u>
If A and Y are d-separated by W, then A and Y are independent given W in <u>every</u> distribution compatible with the SCM

But…

<u>Lack of d-separation does not imply conditional dependence …</u>
If A and Y are <u>not</u> d-separated by W, then A and Y dependent given W in <u>at least one</u> distribution compatible with the SCM

Note: Lack of d-separation does not imply dependence…

- Various sources of dependence can cancel each other…
- Example:
  - A causes Y
  - W is a negative confounder of the effect of A on Y (eg W causes A and prevents Y)
  - If these two opposing sources of dependence have the same magnitude, A and Y may still be independent in a distribution compatible with this graph



Implications…Learning graphs from data

- Why not look at independence in observed data distribution and use them to learn about the true causal model?
- In order to make progress, you end up needing a new assumption: "stability" or "faithfulness"
  - Basically- assumes that all the independencies we see in our observed data are structural (rather than reflecting various sources of dependence cancelling each other out)
- This is a large area of active research
  - We will not cover it in this class

What have done so far?

- We specified a structural causal model: $\mathcal{M}^{\mathcal{F}}$
  - Consisted of specifying
    - X variables and U variables
    - Structural equations F (one for each X node)
      - Exclusion restrictions on the set of preceding nodes included in each parent set
    - Set of allowed distributions for P_U
      - Assumptions on the independence of any two or more U variables
  - This gave the allowed distributions for $(U,X) \sim P_{U,X}$

What have we done so far?

- We specified a target causal parameter using counterfactuals
  - Counterfactuals were derived by evaluating the system of structural equations under a specific intervention
  - **The resulting counterfactual parameters are thus parameters of the distribution of (U,X): $\Psi^F(P_{U,X})$**
    - Example

$$\Psi^F(P_{U,X}) = E_{U,X}(Y_1 - Y_0)$$

What have we done so far?

- We linked the observed data to the SCM
  - We assumed that the observed data were drawn from the system we specified using a SCM
  - O is a subset of X
  - $O \sim P_0$
- The model $\mathcal{M}^{\mathcal{F}}$ (which gives us the allowed distributions for $P_{U,X}$) implied a statistical model $\mathcal{M}$ (which gives us the allowed distributions for P_0)

What have we done so far?

- The *d-separation* criteria provided a straightforward way to indentify what restrictions, if any, the SCM places on the allowed distributions for P_0
  - Functional or inequality constraints may also restrict the statistical model
- Often, no restrictions: Non-parametric statistical model
- Sometimes, some restrictions: Semi-parametric statistical model
  - Randomization is a classic example

The Roadmap….

1. Causal Model
Representing background knowledge and uncertainty

Statistical Model
Possible distributions for the Observed data

3. Observed Data
Process that generated the data described by the causal model

2. Question
Translate the scientific question into a formal causal quantity (using counterfactuals)

Key Points

- We assume that the system that generated our observed data is described by our SCM
- Observed data: $O \sim P_0$
- The SCM, which is a model on $P_{U,X}$, implies a model $\mathcal{M}$ on P_0
  - We refer to this as our statistical model
  - d-separation can be used to derive (many of) the testable implications of a SCM
- A non-parametric statistical model places no restrictions on the set of possible distributions for O