Directed acyclic graphs (DAGs) for causal analysis

Ph.D. course in epidemiology

Niels Keiding
February 2014

Simpson (1951): Conditional or marginal effect measures


<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=0</th>
<th>OR = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>A=0</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

C = 1

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=0</th>
<th>OR = 5/6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>A=0</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

C = 0

Supporting text


Simpson: baby playing cards

A = 0 court cards (B, D, K) B = 0 red (heart, diamond)
A = 1 not court cards (A, 2, 3,…, 10) B = 1 black (spade, club)
C = 1 card dirty because baby played with it
C = 0 card clean

Is colour independent of court status?

Yes, marginally OR = 1
No, conditionally on dirtiness OR = 5/6 for C = 1 and for C = 0.

Relevant effect measure: Marginal
Simpson: medical treatment

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>A=0</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

C = 1

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>A=0</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

C = 0

<table>
<thead>
<tr>
<th></th>
<th>B=1</th>
<th>B=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>A=0</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

OR = 1

OR = 5/6

A = 0 not treated  B = 0 not dead  C = 1 male  
A = 1 treated  B = 1 dead  C = 0 female

Does treatment affect death?

No, marginally  OR = 1

Yes for males,  OR of dying = 5/6

Yes for females,  OR of dying = 5/6

Relevant effect measure: Conditional on sex.

Baby playing cards

A     court

B     dirty

C     colour

Even if A and B are independent, they become artificially associated by conditioning on the collider C (selection bias)

Medical treatment

A treatment  B death

Treatment as well as death depends on sex which is a possible confounder that should be controlled for.

Causal explanations of statistical dependence (association) between $X$ and $Y$

1. Random fluctuation
2. $X$ caused $Y$
3. $Y$ caused $X$
4. $X$ and $Y$ share a common cause (includes mixing of two populations where there is independence in each subpopulation)
5. Selection bias (conditioning on common effect of $X$ and $Y$)

Directed Acyclic Graphs: Causal

$X$ direct cause of $Y$

$U$ direct cause of $Y$

$Y$ direct cause of $Z$

$X$ not direct cause of $Z$ but indirect cause of $Z$ via $Y$

$U$ not direct cause of $Z$ but indirect cause of $Z$ via $Y$

There are no common causes not shown in the graph.
Directed Acyclic Graphs: Statistical

Variables are **dependent** if there is an unblocked path between them otherwise **independent**

- $X$ and $Y$ are statistically dependent
- $U$ and $Y$ are statistically dependent
- $Y$ and $Z$ are statistically dependent
- $X$ and $Z$ are statistically dependent
- $U$ and $Z$ are statistically dependent

**Collider**

- $Y$ is a collider between $X$ and $U$: $X \rightarrow Y \leftarrow U$.
- A collider blocks a path
- Conditioning on the collider unblocks the path
- $X$ and $U$ are statistically independent (the only path between them is blocked by the collider $Y$)
- $X$ and $U$ are statistically dependent given $Y$ (conditioning on $Y$ unblocks the path)
- $X$ and $U$ are statistically dependent given $Z$ ($Z$ is descendant of $Y$)
- $U$ and $Z$ are statistically independent given $Y$

Ex: Basketball players are tall and/or fast

**Basic assumptions on DAGs**

*Causal Markov Assumptions:* any variable $X$ is independent of any other variable $Y$ conditional on the direct causes of $X$, unless $Y$ is an effect of $X$.

*Faithfulness:* Two pathways with positive and negative causal effects never perfectly offset one another.

*No measurement error:* Assume large samples

**Example**

- $Z \rightarrow X \rightarrow Y$
- $U \rightarrow X$

*Question:* is there a causal effect of $X$ on $Y$? (that is: should there also be an arrow from $X$ to $Y$?)
*Answer:* $U$ is a confounder, must be ‘controlled’ (we must condition on $U$)

*Question:* is there a causal effect of $Z$ on $Y$?
*Answer:* the relation between $Z$ and $Y$ is unconfounded, no need for conditioning on $U$
Confounding

Treatments and outcomes share a common cause (which may be unmeasured)

Heart disease → Drug → Stroke

Atherosclerosis (unmeasured)

Confounding by indication

Personality & SES (unmeasured)

Smoking

Death

Exercise

Confounding and DAGs

There is a backdoor path between X and Y and this has to be blocked to make the study unconfounded.

Sufficient set Z of covariates to control for confounding:
1. no variable in Z is a descendant of X
2. every path between X and Y that contains an arrow pointing into X is blocked by Z

The M structure

Assume binary exposure: \( a = 0, 1 \)

Each individual \( i \) has two outcomes \( \gamma_i^0 \) if \( a \) is assigned to be 0

\( \gamma_i^1 \) if \( a \) is assigned to be 1

We would like to know \( \gamma_i^1 - \gamma_i^0 \), the causal effect.

Counterfactual model of causality

Not possible

Average causal effect over the population

\[
E(\gamma_i^1) - E(\gamma_i^0)
\]
**Randomised trial**

Half of the population is assigned to \( a = 0 \), half to \( a = 1 \).

\[ E (Y^1) \] is well estimated by average of effect for those assigned to \( a = 1 \)

\[ E (Y^0) \] is well estimated by average of effect for those assigned to \( a = 0 \)

Average causal effect over the population is estimable in a randomized trial.

**Selection bias**


\[
\begin{array}{c}
X \\
\rightarrow \quad C \\
\end{array}
\]

Conditioning on common effect.

*Example:* \( X \) haplotype \( Y \) smoking \( C \) heart disease

*Question:* Does haplotype affect smoking behaviour?

Both \( X \) and \( Y \) cause heart disease but \( X \) and \( Y \) are unassociated. If we only select patients *with* heart disease, those without the dangerous haplotype will be more likely to be smokers, since the heart disease has to come from somewhere.

*Example* (Berkson bias). \( X \) and \( Y \) diseases; \( C \) = hospitalization.


**Selection bias: example from case-control studies**

\[
\begin{array}{c}
X \\
\rightarrow \quad F \\
Y \\
\rightarrow \quad C \\
\end{array}
\]

\( X = \) oestrogen \quad \( F = \) hip fracture \quad \( C = \) selection \quad \( Y = \) AMI

Oestrogen protects against hip fracture, so controls are less likely to be on oestrogen than cases even when there is no association between \( X \) and \( Y \).
Randomised trials and instrumental variables


\[ Z \rightarrow X \rightarrow Y \]

\[ U \]

We want to assess possible effect of \( X \) on \( Y \), but there may be unmeasured confounders \( U \).

Randomized trial

\( Z \) treatment assigned

\( X \) treatment received

Intention to treat test: compare \( Z = 0 \) to \( Z = 1 \).

Estimation: long story. Effect of \( Z \) on \( Y \) is diluted version of effect of \( X \) on \( Y \).

Instrumental variables

\[ Z \rightarrow X \rightarrow Y \]

\[ U \]

\( Z \) is associated with \( X \)

All unblocked paths between \( Z \) and \( Y \) must pass through \( X \).

There is a statistical methodology allowing estimation of the effect of \( X \) on \( Y \) from the effect of \( Z \) on \( X \) and \( Y \).

Time-dependent confounders, or feed-back

Feed-back from outcome \( Y_i \) to covariate \( X_{i+1} \)

\[ Y_i \rightarrow Y_{i+1} \]

\[ X_{i+1} \rightarrow X_i \]

\( Y_i \) intermediate between \( X_{i+1} \) and \( Y_{i+1} \)

so do not condition on \( Y_i \)

\( Y_i \) confounder for effect of \( X_i \) on \( Y_{i+1} \)

So “control” for \( Y_i \) e.g. by conditioning

Robins: g-computation, marginal structural models


Robins (1986) generalized direct standardization to the time-dependent confounding situation and formulated other approaches using inverse probability weighting such as marginal structural models (2000). A very readable survey is by Daniel et al. (2013).