# Causal inference and control

## Week 6:

## Potential outcomes

Søren Wengel Mogensen
Department of Automatic Control
soren.wengel_mogensen@control.lth.se


LUND
UNIVERSITY

On May 8 we will be in a different room.

# Assignment 2

# Assignment 2

In Problem B, we were considering a *linear Gaussian SCM*. For each $j = 1, \ldots, d$,

$$X_j = f_j(\boldsymbol{PA}_j, N_j) = \sum_{X_i} a_{ji} X_i + N_j$$

where the sum is over $\boldsymbol{PA}_j$ and $N_j$ is Gaussian.

# Valid adjustment sets

**Proposition (ECI, Proposition 6.41)**

*Assume that $Y \notin \boldsymbol{PA}_X$. The following sets are valid adjustment sets for $(X, Y)$.*

- $\boldsymbol{Z} = \boldsymbol{PA}_X$.
- $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$ *such that* $\boldsymbol{Z}$ *contains no descendants of* $X$ *and blocks all back-door paths between* $X$ *and* $Y$.
- *no member of* $\boldsymbol{Z}$ *is a descendant of any* $W \in \boldsymbol{V} \setminus \{X\}$ *which lies on a directed path from* $X$ *to* $Y$, *and* $\boldsymbol{Z}$ *blocks all nondirected paths between* $X$ *and* $Y$.

We say that $W$ is a *descendant* of $X$ if there exists a directed path $X \to \ldots \to W$. We say that a path between $X$ and $Y$ is a *back-door path* if $X \leftarrow \ldots Y$. We say that $\boldsymbol{Z} \subseteq \boldsymbol{X} \setminus \{X, Y\}$ *blocks* a path between $X$ and $Y$ if the path is not $d$-connecting given $\boldsymbol{Z}$.

# Adjustment in linear structural causal model

Let $\boldsymbol{Z}$ be a valid adjustment set for $(X, Y)$. In a zero-mean linear Gaussian SCM, we have

$$E(Y|X = x, \boldsymbol{Z} = \boldsymbol{z}) = ax + \boldsymbol{b}^t \boldsymbol{z}$$

We also have (ECI Problem 6.63)

$$\frac{\partial}{\partial x} E^{do(x)}(Y) = a$$

for a constant $a$.

Each valid adjustment set, $\boldsymbol{Z}$, gives us a least-squares estimator, $\hat{\tau}^{z}_{yx}$, of $a$ using the above regression. So what's the optimal adjustment set in terms of (asymptotic) variance of the estimator?

# Optimal adjustment set

As mentioned there is a solution to this problem (they dispense with the assumption of Gaussian error variables) in

- Henckel, Leonard, Emilija Perković, and Marloes H. Maathuis, *Graphical criteria for efficient total effect estimation via adjustment in causal linear models*. Journal of the Royal Statistical Society Series B: Statistical Methodology 84.2 (2022): 579-599.

We will describe the optimal set defined in this paper (their result also holds for more general graphs and for non-singleton $X$ and $Y$).

# Optimal adjustment set

We define first the *causal nodes*, $\mathrm{cn}(X, Y, \mathcal{G})$, relative to $(X, Y)$ in the graph $\mathcal{G}$ as all nodes on a causal (directed) path from $X$ to $Y$ except $X$ itself. We say that $Z$ is a *descendant* of $W$ if there exists a directed path from $W$ to $Z$ or if $Z = W$. We use $\mathrm{de}(\boldsymbol{Z})$ to denote the set of descendants of the set $\boldsymbol{Z}$.

We define the *forbidden nodes* relative to $(X, Y)$ in the graph $\mathcal{G}$ as

$$\mathrm{forb}(X, Y, \mathcal{G}) = \mathrm{de}(\mathrm{cn}(X, Y, \mathcal{G}), \mathcal{G}) \cup \{X\}.$$

We define

$$\boldsymbol{O}(X, Y, \mathcal{G}) = \mathrm{pa}(\mathrm{cn}(X, Y, \mathcal{G}), \mathcal{G}) \setminus \mathrm{forb}(X, Y, \mathcal{G}).$$

# Optimal adjustment set

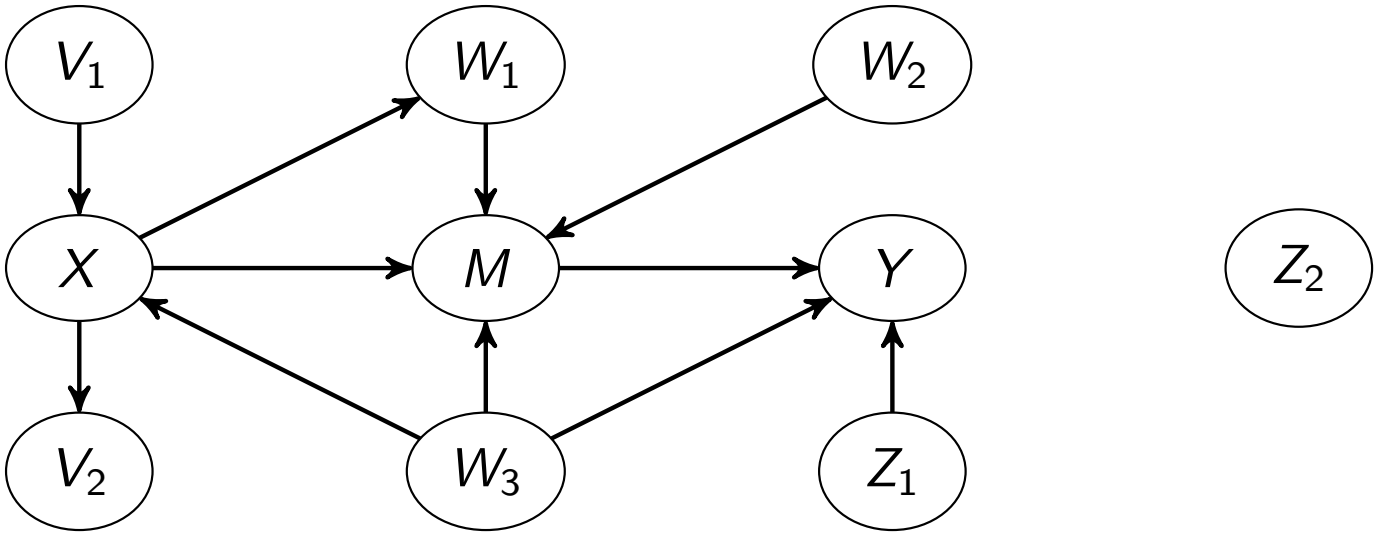
**Theorem (Theorem 3.13, Henckel et al.)**

*Assume $Y$ is a descendant of $X$ in $\mathcal{G}$. The set $\boldsymbol{O}(X, Y, \mathcal{G})$ is a valid adjustment set and it attains the minimal asymptotic variance among all valid adjustment sets, $\boldsymbol{Z}$,*

$$a.var(\hat{\tau}^{o}_{yx}) \leq a.var(\hat{\tau}^{z}_{yx}).$$

*Under faithfulness and for a valid adjustment $\boldsymbol{Z}$, if $a.var(\hat{\tau}^{o}_{yx}) = a.var(\hat{\tau}^{z}_{yx})$, then $\boldsymbol{O} \subseteq \boldsymbol{Z}$.*

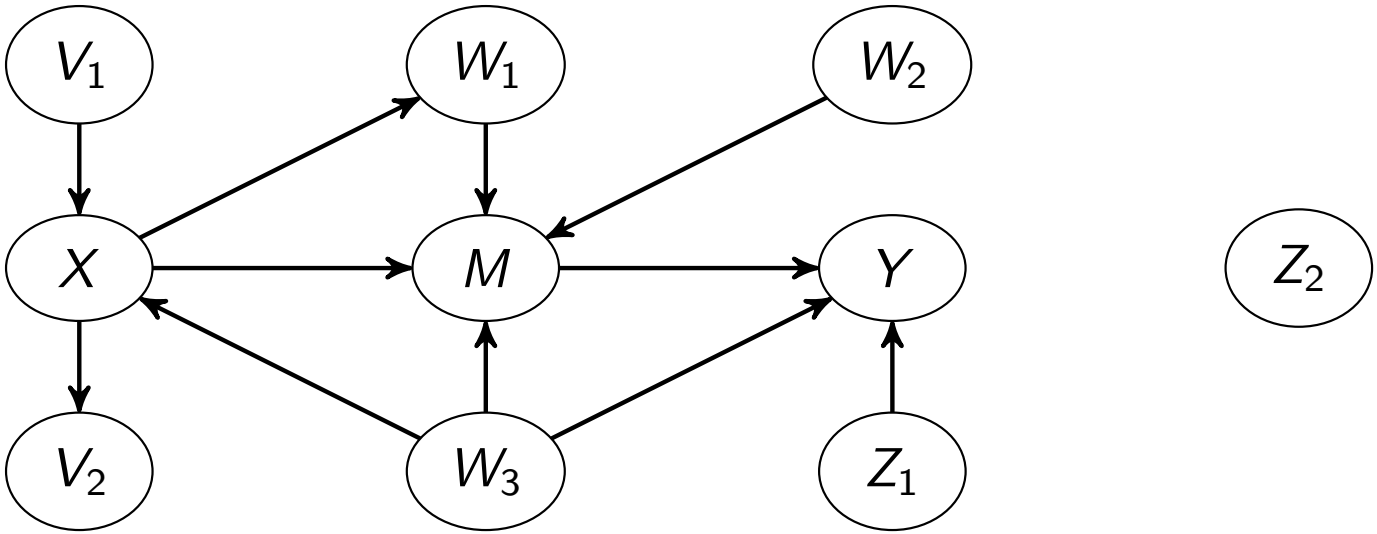
# Optimal adjustment set, example




We first find the causal nodes, $\text{cn}(X, Y, \mathcal{G}) = \{W_1, M, Y\}$. We then find $\text{pa}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) = \{X, W_1, W_2, W_3, M, Z_1\}$, and $\text{forb}(X, Y, \mathcal{G}) = \text{de}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) \cup \{X\} = \{W_1, M, Y, X\}$. Finally,

$$\boldsymbol{O}(X, Y, \mathcal{G}) = \{X, W_1, W_2, W_3, M, Z_1\} \setminus \{W_1, M, Y, X\} = \{W_2, W_3, Z_1\}.$$

# Optimal adjustment set, example




We first find the causal nodes, $\text{cn}(X, Y, \mathcal{G}) = \{W_1, M, Y\}$. We then find $\text{pa}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) = \{X, W_1, W_2, W_3, M, Z_1\}$, and $\text{forb}(X, Y, \mathcal{G}) = \text{de}(\text{cn}(X, Y, \mathcal{G}), \mathcal{G}) \cup \{X\} = \{W_1, M, Y, X\}$. Finally,

$$\boldsymbol{O}(X, Y, \mathcal{G}) = \{X, W_1, W_2, W_3, M, Z_1\} \setminus \{W_1, M, Y, X\} = \{W_2, W_3, Z_1\}.$$

Including parents of $Y$ and of other causal nodes is good (if allowed) and including parents/children of $X$ is bad.

# Graphical marginalization

# Graphical marginalization

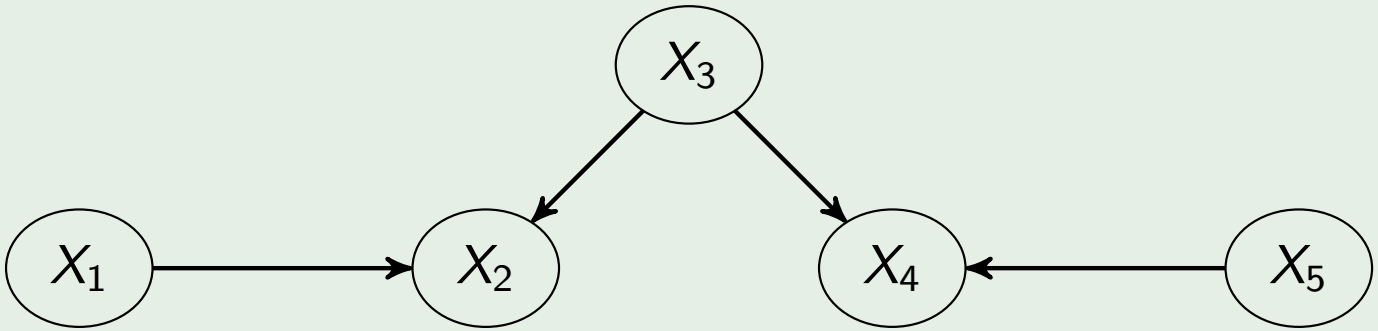
In causal modeling, the idea of *hidden* variables is central. In Week 4, we looked at identification methods that do not require full observation.

In causal discovery, we may be interested in learning a 'marginal' of the causal graph when there are hidden variables, $\boldsymbol{H}$, as well as observed variables, $\boldsymbol{O}$. One natural requirement is that the implied conditional independences are the same when restricting to $\boldsymbol{O}$.

# Graphical marginalization

**Example (DAGs are not closed under marginalization, Richardson and Spirtes [2002])**

Assume $X_3$ is unobserved. There is no DAG on nodes $\{X_1, X_2, X_4, X_5\}$ that encodes the same conditional independences (using $d$-separation).



# Acyclic directed mixed graphs

We say that a graph is a *directed acyclic mixed graph* (ADMG) if every edge is either *directed*, $\rightarrow$, or *bidirected*, $\leftrightarrow$.

The extension of *d*-separation to ADMGs is known as *m-separation*.

# Latent projection

Let $\mathcal{G} = (\boldsymbol{V}, E)$ be an ADMG, $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{H}$. We define the following transformation.

**Definition (Latent projection)**

We define $m(\mathcal{G}, \boldsymbol{O})$ as the graph such that for $X, Y \in \boldsymbol{O}$

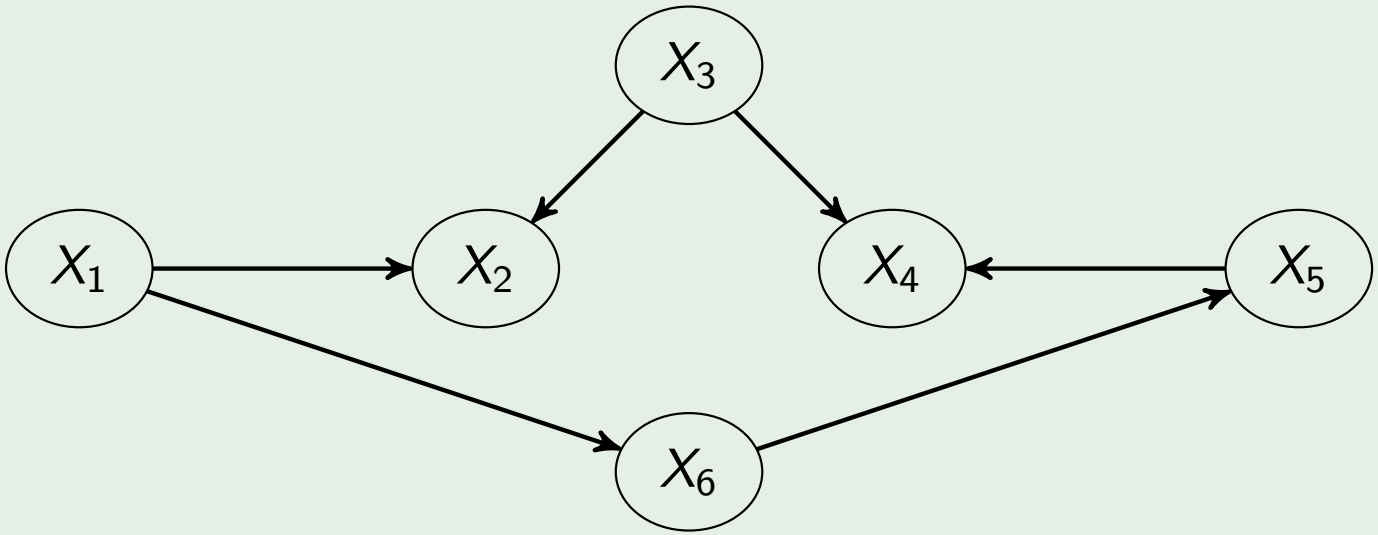
- $X \to Y$ in $m(\mathcal{G}, \boldsymbol{O})$ if there is a directed path $X \to \ldots \to Y$ in $\mathcal{G}$ such that every non-endpoint node is in $\boldsymbol{H}$,
- $X \leftrightarrow Y$ in $m(\mathcal{G}, \boldsymbol{O})$ if there is a path between $X$ and $Y$ such that all non-endpoint nodes are in $\boldsymbol{H}$, all non-endpoint nodes are non-colliders, and there are arrowheads at both $X$ and $Y$.

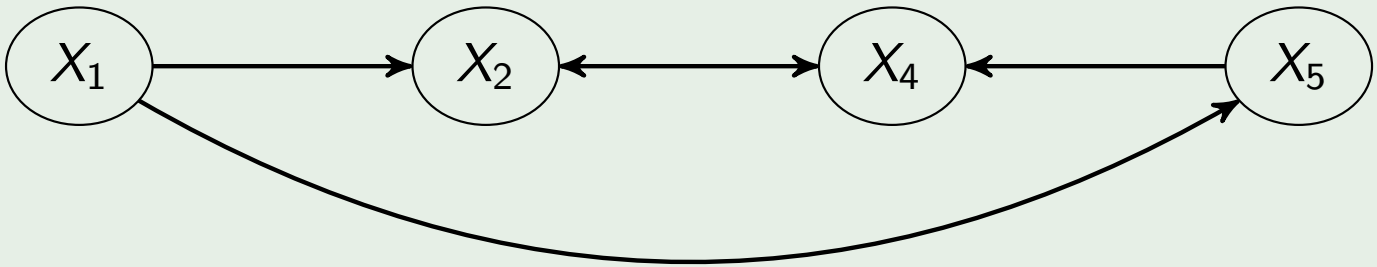
The latent projection is also an ADMG!

# Latent projection

## Example

Let $\boldsymbol{V} = \{X_1, X_2, X_3, X_4, X_5, X_6\}$ and $\boldsymbol{O} = \{X_1, X_2, X_4, X_5\}$



# Latent projection as a marginal

Let $\mathcal{G} = (\boldsymbol{O} \cup \boldsymbol{H}, E)$ be a DAG and let $\mathcal{M} = m(\mathcal{G}, \boldsymbol{O})$.

**Proposition**

*Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \subseteq \boldsymbol{O}$. We have that $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{C}$ in $\mathcal{G}$ if and only if $\boldsymbol{A}$ and $\boldsymbol{B}$ are m-separated by $\boldsymbol{C}$ in $m(\mathcal{G}, \boldsymbol{O})$.*

# Potential outcomes

# SCM, 'procedural' explanation

A central part of the definition of a structural causal model is the *modularity* of assignment mechanisms, $f_j$, i.e., they remain the same unless we intervene upon them.

$$
\begin{aligned}
X_1 &= f_1(PA_1, N_1) = \bar{f}_1(N_1) \\
X_2 &= f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2) \\
X_3 &= f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3) \\
&\dots \\
X_{d-1} &= f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1}) \\
X_d &= f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)
\end{aligned}
$$

# Interventions, 'procedural' explanation

If we have an (atomic) intervention $do(X_j = a)$,

$$
\begin{aligned}
X_1 &= f_1(PA_1, N_1) = \bar{f}_1(N_1) \\
X_2 &= f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2) \\
X_3 &= f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3) \\
&\dots \\
X_j &= a \\
&\dots \\
X_{d-1} &= f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1}) \\
X_d &= f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)
\end{aligned}
$$

# Interventions, 'procedural' explanation

If we have an (atomic) intervention $do(X_j = a)$,

$$
\begin{aligned}
X_1 &= f_1(PA_1, N_1) = \bar{f}_1(N_1) \\
X_2 &= f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2) \\
X_3 &= f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3) \\
&\dots \\
X_j &= a \\
&\dots \\
X_{d-1} &= f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1}) \\
X_d &= f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)
\end{aligned}
$$

These are assignment mechanisms, not ordinary equalities. Interventions are reassignments.

# Potential outcomes

The *potential outcomes* framework (or *Rubin causal model*) takes a different approach to defining causal notions [Neyman, 1923, Fisher, 1925, Rubin, 1974].

It essentially formulates causal inference as a *missing data problem*.

# ECI Example 3.4

We start by looking at Example 3.4. We assume we have a population of patients with a specific eye disease. There is a treatment, $T$, and 99% of patients are cured ($B = 0$) if treated ($T = 1$), and otherwise they go blind ($B = 1$). The 1% go blind if and only if they are treated. The difference in effect of treatment is explained by a rare condition, $N_B$. Treatment is independent of $N_B$ as the value of $N_B$ is unknown to the doctor.

An SCM representing the data-generating process,

$$
\begin{aligned}
T &= N_T \\
B &= T \cdot N_B + (1 - T) \cdot (1 - N_B)
\end{aligned}
$$

where $N_B \sim \text{Ber}(0.01)$.

# Potential outcomes

The potential outcomes framework uses a different formalism to define causal notions. Assume we have a fixed population of individuals (or *units*), $n = 200$. Each of these individuals, $u$, has *two* potential outcomes, one for each level of treatment.

$B_u(t = 1)$ is the value of $B$ unit $u$ would have if they receive treatment.
$B_u(t = 0)$ is the value of $B$ unit $u$ would have if they do not receive treatment.
We assume first that these are deterministic (could also be more general random variables).

When the potential outcomes are deterministic, only the assignment of treatment introduces randomness. That is, all of these potential outcomes are assumed to exist, however, the assignment of treatment will reveal some of them to us.

After assignment of treatment, one of the potential outcomes become *counterfactual* in the sense that is is unobservable and can actually never be observed.

# Potential outcomes

| $u$ | $T$ | $B_u(t=0)$ | $B_u(t=1)$ | $B_u(t=1) - B_u(t=0)$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | -1 |
| 2 | 1 | 1 | 0 | -1 |
| 3 | 0 | 0 | 1 | 1 |
| $\vdots$ | | | | |

Grey is observed information. The quantity

$$B_u(t=1) - B_u(t=0)$$

is the *unit-level causal effect*. The *average causal effect* is

$$\frac{1}{n}\sum_{u=1}^{n} B_u(t=1) - B_u(t=0)$$

# Potential outcomes

Assume we have data from a randomized experiment. If $u \in U_0$ received no treatment and $u \in U_1$ received treatment then

$$\frac{1}{|U_1|} \sum_{u \in U_1} B_u(t = 1) - \frac{1}{|U_0|} \sum_{u \in U_0} B_u(t = 0)$$

is an unbiased estimator of the average causal effect.

# SUTVA

Assume that $\boldsymbol{T} = (T_1, \ldots, T_n)$ is a vector of treatment assignments. The potential outcome of $u$ could depend on all of these treatment assignments, $B_u(\boldsymbol{T})$. In our example, SUTVA (stable-unit treatment value assumption) is the assumption that

$$B_u(\boldsymbol{T}) = B_u(\boldsymbol{T}') \quad \text{if} \quad T_u = T'_u.$$

for all $T$, $T'$ and that for all $\omega \in \Omega$

$$T_u(\omega) = t \Rightarrow B_u(T = t)(\omega) = B_u(\omega)$$

(if unit $u$ receives treatment $t$, then the observed outcome is $B_u(T = t)$). This first is 'no interference' and the second is known as consistency.

# SUTVA

The table implicitly used the SUTVA assumption. Otherwise, we would need a potential outcome for each $u$ for every vector of treatment assignments.

| $u$ | $T$ | $B_u(t = (0, 0, \ldots, 0))$ | $B_u(t = (1, 0, \ldots, 0))$ | $\ldots$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 |
| 3 | 0 | 0 | 1 | 1 |
| $\vdots$ | | | | |

In some settings, interference is very likely to occur in which case one can use weaker stability assumptions.

The consistency assumption links the counterfactuals to observable data. If $T_u = t$, then we observe the potential outcome $B_u(T = t)$.

# Ignorability

In this framework, the treatment assignment is *unconfounded* or *ignorable* if

$$B(t=0), B(t=1) \perp T$$

(we lost the unit subscripts).

There is also a conditional version of this assumption. Under this (and the previous assumptions),

$$E(B|T=t) = E(B(T=t)|T=t) = E(B(T=t))$$

so the average treatment effect is identified,

$$E(B(T=1) - B(T=0)) = E(B|T=1) - E(B|T=0).$$

# References I


Ronald Aylmer Fisher. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, UK, 1925.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9 (translated). *Statistical Science*, 5: 465–480, 1923.

Thomas S. Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66(5):688, 1974.