Statistical methods in public health
Confounding and graphical models

Tommi Härkänen

National Institute for Health and Welfare (THL)
Department of Health (TERO)

September 29, 2015

Contents

Pearl J (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

Graphical model

Graph

A graph consists of a set V of *nodes* and a set E of *edges* connecting the nodes.

Directed graph: Edges have a direction e.g. $X \to Y$.

Acyclic graph: There are no *cycles* i.e. it is not possible to follow the directed edges starting from a node and to end up to the same node: $X \to \cdots \to X$.

Directed acyclic graphs (DAG) are commonly used to describe *causality*.

Relationships of nodes are often described using terms like *parent* $\to$ *child*. Other terms:

- Family: consists of a node and all its parents.
- Root: node with no parents
- Sink: node with no children.
- Tree: all nodes have at most one parent.
- Chain: a tree in which all nodes have at most one child.

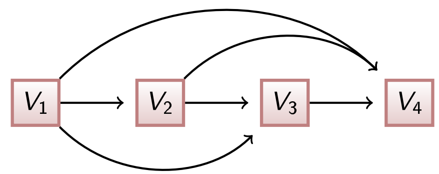
Conditional distribution and a graphical model

The joint probability distribution of m random variables $V := \{V_1, V_2, V_3, \ldots, V_m\}$ can be expressed as a product of conditional distributions. E.g. the *chain rule*

$$\mathbb{P}\{V_1, V_2, V_3, \ldots, V_m\} = \mathbb{P}\{V_m \mid V_1, \ldots, V_{m-1}\} \times \cdots \times \mathbb{P}\{V_2 \mid V_1\} \times \mathbb{P}\{V_1\},$$

where V_1 is the parent of V_2. V_3 is a child of V_1 and V_2, etc.

$m = 4$:



Conditional independence

Conditional independence

Let X, Y and Z be subsets of random variables V. Sets Y and Z are conditionally independent (given X), if $\mathbb{P}\{Y \mid X, Z\} = \mathbb{P}\{Y \mid X\}$.

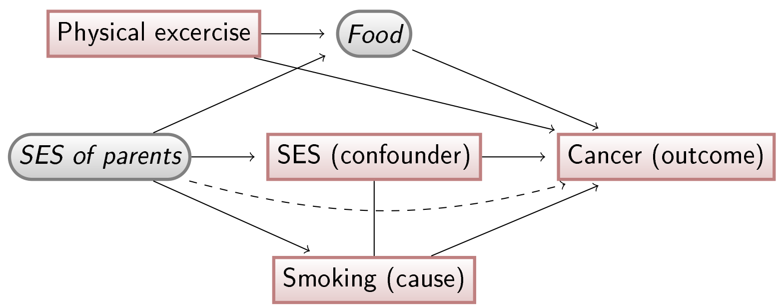
Conditional independence means that if we know X, information about Z does not provide additional information about Y.
In a DAG this can be depicted as $Z \to X \to Y$.

Markovian parents

A set of variables X is called the Markovian parents of node Y, if X is the *minimal* set of variables which render Y independent of all its other predecessors.

Smoking and risk of cancer

Example: Socio-economic status (SES) $\Rightarrow$ Smoking $\Rightarrow$ Cancer



d-separation criterion

A path is a sequence of consecutive edges (any direction).

d-separation

A path p is said to be separated (or blocked) by a set of nodes Z if and only if

1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that m is in Z, or
2. p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that m and its descendants are **not** in Z.

A set Z is said to d-separate X from Y
if Z blocks every path from a node in X to a node in Y.

Condition 1 can be interpreted as *conditional independence*, and condition 2 as *selection bias*.

Back-door criterion

In lecture 4 confounding was defined.
But do we need to adjust for all potential confounders?

Back-door

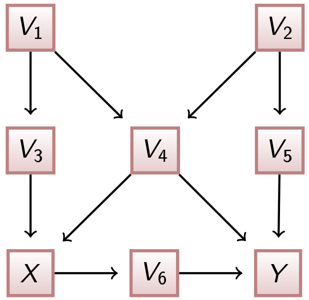
A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables X and Y in a DAG if:

1. no node in Z is a descendant of X and
2. Z blocks every path between X and Y that contains an arrow into X.

Similarly, if X and Y are two disjoint subsets of nodes in a DAG, then Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, Y_j) such that $X_i \in X$ and $Y_j \in Y$.

Back-door criterion

Example



Which sets $Z \subset \{V_1, \ldots, V_6\}$ meet the back-door criterion?

- $Z_1 = \{V_3, V_4\}$ **Yes**
- $Z_2 = \{V_4, V_5\}$ **Yes**
- $Z_3 = \{V_4\}$ **No** (path X, X_3, X_1, X_4, X_2, X_5, Y is not blocked)

Back-door adjustment

Back-door adjustment

If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula

$$\mathbb{P}\{y \mid x\} = \sum_z \mathbb{P}\{y \mid x, z\} \mathbb{P}\{z\}. \tag{1}$$

Note that in (1)

Intervention $\mathbb{P}\{y \mid x\}$ is the predicted probability of $Y = y$ when the value of X is fixed to x. (Pearl uses notation $do(x)$)

(Direct) standardization The r.h.s. is a weighted average of probabilities $\mathbb{P}\{y \mid x, z\}$ estimated from subsets (x, z) and $\mathbb{P}\{z\}$ prevalence of the blocking variables $Z = z$.

Smoking and risk of lung cancer

The back-door criterion

Which factors need to be adjusted for in the analysis?

1. Connect all parents with lines and remove all arrow tips.
2. Adjust for variables which block all paths via black arrows or red lines from smoking (cause) to cancer (outcome).