Statistical and causal inference with sampling selection

Confounding and selection bias are among the greatest challenges to empirical science. These notes focus on alleviation of sampling selection bias in statistical (observation) studies as well as causal (intervention/policy) studies.[1]

The approach is graphical and draws upon Pearl's do-calculus to establish causality. A binary node S is added to the graph (typically a DAG, directed acyclic graph) denoting inclusion in the sample when $S = 1$ and exclusion is denoted $S = 0$. Selective sampling implies that observable variables lead to inclusion. Hence, arcs are included in the graph from these nodes to S. Then, the graph can be used to assess whether selection bias can be alleviated.

Graphical identification is completely general, in other words, nonparametric. In contrast, say, to Hechman's inverse-Mills strategy which is parametric (in particular, normally distributed unobservables in the selection mechanism). The objective of statistical inference then is to identify $p(y \mid x)$ from $p(y \mid x, S = 1)$.

Similarly, the objective of causal inference is to identify $p(y \mid do(x))$ from $p(y \mid do(x), S = 1)$ where $do(x)$ refers to intervention that sets the value of X to x.. Further, causal inference strives to alleviate confounding such that $p(y \mid do(x))$ is identified from observables such as via the back-door adjustment where some components may come from external data.

$$p(y \mid do(x)) = \sum_z p(y \mid x, z)\, p(z)$$

-

Statistical inference

Statistical inference, $p(y \mid x)$, is recoverable without external data if and only if

$$(Y \perp S \mid X)$$

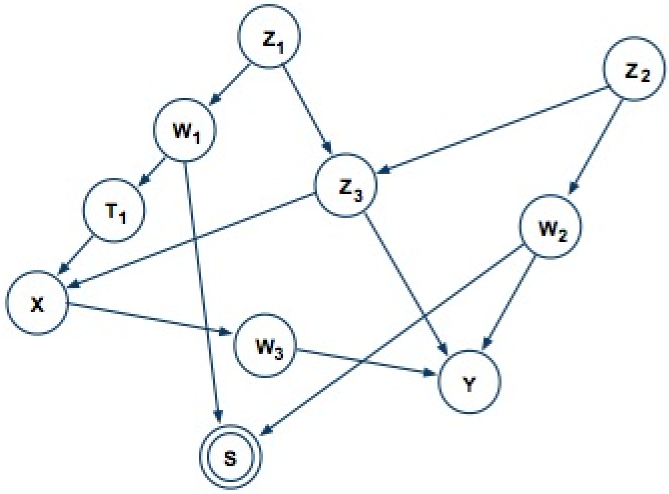
where $\perp$ refers to stochastic independence or d-separation in the graph. The corollary to the above is Y cannot be ancestor to S. Then, $p(y \mid x)$ is identifed by $p(y \mid x, S = 1)$ where $S = 1$ refers to the selective (potentially contaminated) sample.

Statistical inference is recoverable with external data if there exists a set of observables C such that

$$(Y \perp S \mid X, C)$$

Consider the DAG below.

[1] These notes are a synthesis of several papers by Bareinboim, Tian, and Pearl including "Recovering from selection bias in causal and statistical inference," Proceedings of the twenty-eighth AAAI conference on artificial intelligence, 2014.



The set $C \in \{W_1, W_2\}$ satisfies conditional independence of Y and S. Then,

$$p\left(y \mid x, S=1\right) = \sum_{w_1, w_2} p\left(y \mid x, w_1, w_2\right) p\left(w_1, w_2 \mid x, S=1\right)$$

Since C cannot be d-separated from S, $p\left(w_1, w_2 \mid x\right)$ will need to come from external data when inferring the statistical quantity of interest and identification is

$$p\left(y \mid x\right) = \sum_{w_1, w_2} p\left(y \mid x, w_1, w_2, S=1\right) p\left(w_1, w_2 \mid x\right)$$

where the first term can utilize the selective sample.

Causal inference

Causal inference is ultimately settled by the rules of do-calculus where in addition to selection we must address confoundedness.

Rule 1 (insertion/deletion of observations):

$$\Pr\left(y \mid do\left(x\right), z, w\right) = \Pr\left(y \mid do\left(x\right), w\right) \quad if\left(Y \perp Z \mid X, W\right)_{G_{\overline{X}}}$$

Rule 2 (action/observation exchange):

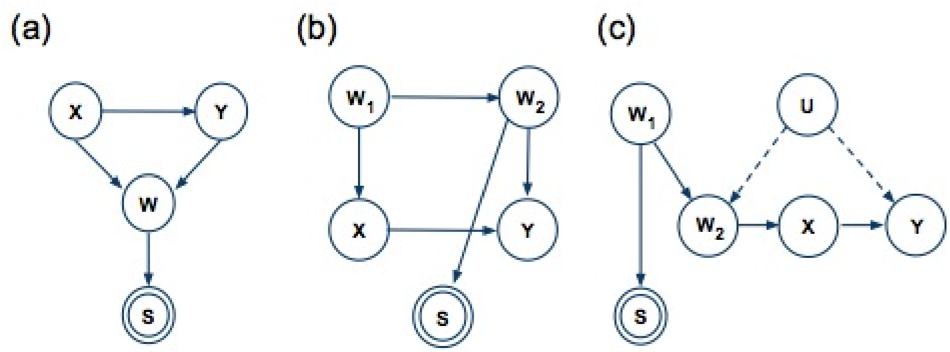
$$\Pr\left(y \mid do\left(x\right), do\left(z\right), w\right) = \Pr\left(y \mid do\left(x\right), z, w\right) \quad if\left(Y \perp Z \mid X, W\right)_{G_{\overline{X}\underline{Z}}}$$

Rule 3 (insertion/deletion of actions/interventions):

$$\Pr\left(y \mid do\left(x\right), do\left(z\right), w\right) = \Pr\left(y \mid do\left(x\right), w\right) \quad if\left(Y \perp Z \mid X, W\right)_{G_{\overline{X}, \overline{Z(W)}}}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-nodes in $G_{\overline{X}}$. Similar to statistical inference, if there are no confounding variables and X makes Y and S conditionally independent then the causal effect can be identified without external sources. Otherwise, causal inference involves employing adjustment variables to address selection and/or confoundedness. Further, adjustment variables cannot be descendants of outcome as conditioning on them results in reverse causality bias.

The three graphs below are illustrative.



The causal effect $p(y \mid do(x))$ is not recoverable from selection bias in DAG (a) even though W block S from Y. Conditioning on W, a collider, confounds the effect of X on Y.

The causal effect in DAG (b) is recoverable. Conditioning on W_2 separates S from Y and blocks the confounding back-door into X connecting Y.

$$p(y \mid do(x), S = 1) = \sum_{w_2} p(y \mid x, w_2) \, p(w_2 \mid S = 1)$$

Rule 2 of do-calculus allows substitution of observation for action to give the first term on the right hand side. Rule 3 of do-calculus allows deletion of $do(x)$ as the arc from W_1 to X is deleted in the subgraph and Y is a collider blocking the other path between W_2 and X. Again, $p(w_2)$ requires external data as W_2 is not separable from selection and identification is

$$p(y \mid do(x)) = \sum_{w_2} p(y \mid x, w_2, S = 1) \, p(w_2)$$

Notice, the conditional causal effect does not require external data.

$$p(y \mid do(x), w_2) = p(y \mid x, w_2) = p(y \mid x, w_2, S = 1)$$

The causal effect in DAG (c) is also recoverable albeit with a bit more creativity. First, W_2 blocks the back-door into X leading to the back-door adjustment within the selective sample.

$$p\left(y \mid do\left(x\right), S=1\right) = \sum_{w_2} p\left(y \mid x, w_2, S=1\right) p\left(w_2 \mid S=1\right)$$

Selection can be dealt with in each term on the right hand side as W_1 blocks all paths connecting S and Y.

$$p\left(y \mid x, w_2, S=1\right) = \sum_{w_1} p\left(y \mid x, w_2, w_1\right) p\left(w_1 \mid w_2, S=1\right)$$

$$p\left(w_2 \mid S=1\right) = \sum_{w_1} p\left(w_2 \mid w_1\right) p\left(w_1 \mid S=1\right)$$

This implies two quantities come from external sources: $p\left(w_1 \mid w_2\right)$ and $p\left(w_1\right)$ and identification is

$$p\left(y \mid do\left(x\right)\right) = \sum_{w_1, w_2} p\left(y \mid x, w_2, w_1, S=1\right) p\left(w_1 \mid w_2\right) \sum_{w_1} p\left(w_2 \mid w_1, S=1\right) p\left(w_1\right)$$

DAG (c) merits a few more observations. If the arc between W_1 and W_2 is reversed, identification of the causal effect is analogous to DAG (b) — W_2 adjusts for both confounding and selection.

Further, statistical inference in DAG (c) requires adjustment by W_1 as is the case for causal inference (however, causal inference also requires adjustment by W_2 for confounding). It might be tempting to adjust for selection by X alone as W_2 is a collider with respect to W_1 and U. However, X is a descendant of the collider W_2 and conditioning on X opens the collider path $S \leftarrow W_1 \rightarrow W_2 \leftarrow U \rightarrow Y$. Therefore, Y is not independent of S conditional on X but is conditional on W_1. Again, $p\left(w_1\right)$ comes from external sources.[2]

[2]Conditional independence tests distinguish between DAG (c) and other DAGs in which W_2 is not a collider with respect to U and W_1. In DAG (c) $X \perp W_1, S \mid W_2$ but when W_2 is not a collider $Y, X \perp W_1, S \mid W_2$.