Ralph's Transportability[1]

Ralph would like to explore causal effects involving private organizations but only has complete data for public organizations. Ralph knows the ability to transport evidence from one setting to another (also known as external validity) is a causal problem as well as a hallmark of scientific progress.

The challenge seems daunting, nonetheless structural causal models (SCM) drawing on graphical algorithms along with do-calculus provide a systematic means to evaluate transportability. Do-calculus is summarized via the following three rules where G refers to a DAG (directed acyclic graph), $G_{\overline{X}}$ means all arrows into X are removed, and $G_{\underline{X}}$ means all arrows out of X are removed.

Rule 1 (insertion/deletion of observations):

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

where $\perp$ refers to stochastic independence or d-separation in the graph and $do\left(x\right)$ refers to intervention that sets the value of X to 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 $Z\left(W\right)$ is the set of Z-nodes that are not ancestors of any W-nodes in $G_{\overline{X}}$. Both nonparametric identification and transportability utilize do-calculus. Transportability does not guarantee nonparametric identification of the causal effect $X \rightarrow Y$ but identification does support transportability.

It is quite likely that differences such as size exist between the source (public) setting and the target (private) setting. Differences in mechanisms by which two populations differ are denoted selection variables and denoted S.[2] Switching between populations involves conditioning on different values of S variables.

The rules of do-calculus above provide the key to transportability in that they allow the analyst to write

$$P^{*}\left(Y \mid do\left(x\right)\right) = P\left(Y \mid do\left(x\right), s\right)$$

in terms that are do-free when they involve the target population so that only the source population involves do operations where P^{*} refers to the target probability distribution and P refers to the source probability distribution.

[1] This example is based on Pearl and Bareinboim, 2014, *Statistical Science*, "External validity: From do-calculus to transportability across populations," 579-595.

[2] This is related by distinct from sampling selection (see Ralph's Technology). Sampling selection points the arrow toward S.

Three typical (but not the only) transportation formulae are as follows.

$$P^* \left(Y \mid do \left(x \right) \right) = P \left(Y \mid do \left(x \right) \right) \tag{1}$$

This is the conventional version of external validity where no adjustments are required to transport evidence from source to target.

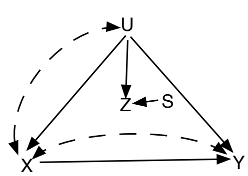
$$P^* \left(Y \mid do \left(x \right) \right) = \sum_z P \left(Y \mid do \left(x \right), z \right) P^* \left(z \right) \tag{2}$$

In this instance, population differences between source and target exist in Z variables requiring adjustment and transportability is supported if $\left(S \perp Y \mid X, Z \right)_{G_{\overline{X}}}$. Notice, the first term derives from source evidence while the second term adjusts the weights based on the target population.

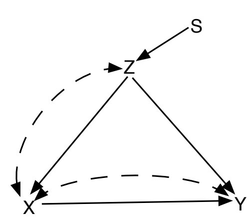
$$P^* \left(Y \mid do \left(x \right) \right) = \sum_z P \left(Y \mid do \left(x \right), z \right) P^* \left(z \mid x \right) \tag{3}$$

This is similar to the case above except adjusting for population differences in Z variables involves conditioning on X. Again, the first term derives from source evidence while the second term adjusts the weights based on the target (conditional) population.

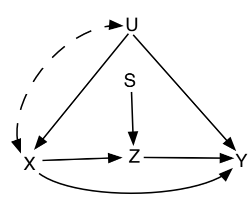
Ralph is contemplating transportability in the graphs below in connection with his public to private causal analysis. U variables are unobservable or unmeasured, dashed arcs (bows) refer to latent variables lying between other variables. Otherwise, X, Y, Z, W variables are observable/measured where $X \to Y$ is the focal causal effect.



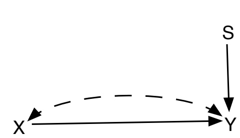
(a)



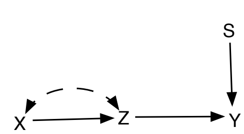
(b)



(c)



(d)



(e)

Suggested:

1. For DAG (a), use do-calculus to show that the transportation formula is (1).

2. For DAG (b), use do-calculus to show that the transportation formula is (2).

3. For DAG (c), use do-calculus to show that the transportation formula is (3).

4. For DAG (d), use do-calculus to show that no transportation formula exists while for DAG (e) the transportation formula is

$$P^{*}\left(Y \mid do\left(x\right)\right) = \sum_{z} P^{*}\left(Y \mid z\right) P\left(z \mid do\left(x\right)\right)$$

-