Causal Analysis in Theory and Practice

The Match-Maker Paradox

Filed under: Discussion, Matching, Selection Bias — moderator @ 6:30 am, 01/08/2012

The Match-Maker Paradox

The following paradox was brought to our attention by Pablo Lardelli from Granada (Spain).

Pablo writes:
1. Imagine that you design a cohort study to assess the causal effect of X on Y, E[Y|do(X=x)]. Prior knowledge informs you that variable M is a possible confounder of the process X ---> Y, which leads you to assume X <--- M ---> Y.

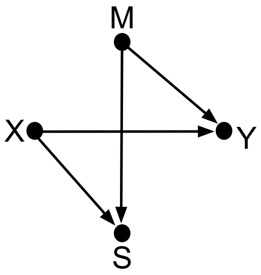
To adjust for the effect of this confounder, you decide to design a matched cohort study, matching on M non exposed to exposed. You know that matching breaks down the association between X and M in the sample.

(In other words, if you make sure that every exposed unit is paired with one unexposed unit with the same value of M, and unpaired units are purged from the sample, simple frequency count tells you that X and M are independent in the sample)

Furthermore, it is well known (for example, see page 173 of "Modern Epidemiology" by Rothman, Greenland and Lash) that, in a matched cohort design, it is not necessary to further condition on M because, since X and M are no longer associated, the relation between X and Y is no longer confounded.

All this is fine with me, so, what is the problem?

2. The problem arises when you draw the DAG representing the associations between variables after matching on M. To my understanding, this DAG would be as follows:



with S indicating selection to the study (S=1: included; S=0 excluded), and it is a descendent of both X and M.

Looking at this DAG, you realize that S is a collider on the path X--->S<---M and, since we are conditioning on S (because the study sample is restricted to S=1) we are in fact opening a non causal path between X and Y (through M) in the sample.

3. Therefore, the final result of the matching is to introduce selection bias in the sample, which must be controlled by conditioning on M. But this stands in contradiction to what we are told by the textbook (see above): "It is NOT necessary to condition on M in a matched cohort design"

Here we have a paradox.

Sincerely yours,
Pablo Lardelli

Paradox Resolved

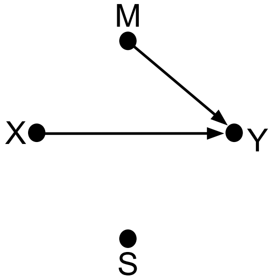
Dear Pablo,

Like all paradoxes, yours too represents a clash between two reliable sources of advice: a textbook and a DAG. Since the former is written by knowledgeable authors and the latter provides mathematically proven results, we know that one of them is improperly applied.

As you can tell from Causality (page 340-41), my natural inclination would be to question textbooks but, in this case, the trouble lies with the way the DAG is applied to the problem.

Since we wish the DAG to represent the post-matching, not the unmatched population there is no further selection done on these cohorts; S=1 for every unit in the matched sample, independently on X or M.

The proper DAG for the matched population would be



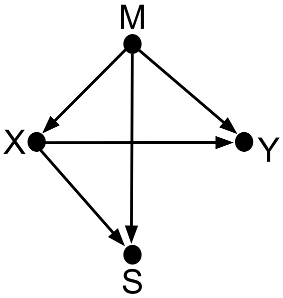
with S not connected to either X or M.

Paradox resolved.

Is it really solved? Unfolding...

But, like most paradoxes, this one too has an auxiliary sub-paradox. Suppose someone says: I do not want to think about the selection process in two separate stages, first matching and then bias removal. DAGs are famous for representing sequential processes all at once. So, let us start with the unmatched sample and use a DAG to represent two processes in tandem; first matching (with S=1) and then the resulting disappearance of bias. Can we do that with one DAG?

Doing it this way would result in the DAG:



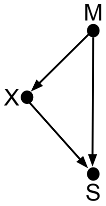
Conditioning on S=1, leaves us with two open non-causal paths:

X <--- M ----> Y and X ----> S <---- M---> Y

So, how come the DAG does not represent what we know to be true: zero bias without conditioning on M???

This sub-paradox has two answers; one quick and one more general.

The quick answer is: **unfaithfulness**. In other words, We never expect dags to represent dependencies. All they can guarantee to us are the **independencies**, and we always need to watch out for the possibility that two or more open paths will exhibit "accidental cancellation."

This is the way Robins and Hernan explain matching in their book-in-progress, http://www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book/ Conditioning on S=1 in the triangle () makes the M--->X association cancel the one created by the collider X ----> S <---- M and the problem is explained away. Case dismissed.

We could quit here, except that I can imagine an inquisitive reader asking a follow-up question:

Wait a minute! This is not an ordinary ***unfaithfulness****, nor an ordinary selection. Ordinary unfaithfulness was labeled "unstable" in Causality (page 48) because we expect it to go away as soon as the system undergoes a slight change in parameters (Causality, page 63). The unfaithfulness we are witnessing with matching is persistent; no matter how we change the joint distribution P(X=x, M=m) in the unmatched sample, we are going to end up with a matched population, namely, X and M will be independent conditional on S=1. This is a non-standard selection process, one that was not considered in Chapter 2 of Causality. If such persistently unfaithful processes occur often in nature, how can we ever hope to learn causal structures from data.*

The answer is: Matching is not a natural process but a man-made one. It violates a key assumption we make when we consider the DAG as a stand-alone representation of natural processes: that units are selected independently of one another (i.i.d.).

We in fact make an even stronger assumption, that, for every variable V in the system, the value V(u) for unit u be independent of the properties of other units in the population. (Reminiscent of SUTVA, or "no-contagion"). In matching, the variable S does not meet this requirement; selection of unit u depends, not merely on the values X(u) and M(u), but also on the values X and M of other units in the sample; after all, this is what "matching" means. There is no selection process P(S=1|X=x, M=m) = f(x,m) in the world that would result in a "persistent unfaithfulness" of the kind we see in matching.

I hope this brings the sub-paradox to a resolution.

======Judea

(Sander Greenland, Miguel Hernan and Pablo Lardelli participated in an email discussion, sprinkled with minor disagreements, leading to this posting).