

CAUSAL INFERENCE: PROPENSITY SCORES, INSTRUMENTAL VARIABLES, AND SENSITIVITY ANALYSES

ŠAUNAK SEN*

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

sen@biostat.ucsf.edu

INTRODUCTION

One of the key features of the scientific method is its ability to assign cause. It is one of the most important methods for accumulating knowledge that informs and drives much of biomedical research. Indeed evidence provided by randomized clinical trials are essential for most drugs to be approved by the FDA (Food and Drug Administration). For many biomedical investigations, conducting an experiment may be impossible, inordinately expensive, or unethical. So, investigators have to find ways to provide evidence in favor of causal effects via observational studies.[†]

THE COUNTERFACTUAL FRAMEWORK

Causal inference has inspired much philosophical debate in the literature, much of which is outside our scope. Some hold that causal inference is impossible without experiments. The moderate viewpoint is that some causal inference is possible even without experiments. To participate in that discussion it is helpful to use the language of randomized experiments, and to examine the counterfactual framework that has come to dominate the theory behind causal inference.

Subject	Control	Treatment	Effect
Cab	220	210	-10
Dab	230	220	-10
Fab	240	220	-20
Gab	250	230	-20
Jab	260	240	-20
Lab	270	250	-20
Nab	280	270	-10
Tab	290	280	-10

In this example, the average treatment effect is -15. However, this is a hypothetical reality, not what actually happens, or what we can observe. In reality, we will see something like this

Subject	Control	Treatment
Cab	220	
Dab	230	
Fab	240	
Gab	250	
Jab		240
Lab		250
Nab		270
Tab		280

*© 2006 Šaunak Sen; Last updated April 16, 2006.

[†]Winship C, Morgan SL (1999) The estimation of causal effects from observational studies. *Annual Reviews of Sociology*, 25:659-706

Thus we get to see only half of the counterfactual data, the rest is missing data. In this example, the estimated treatment effect would be:

$$\frac{(240 + 250 + 270 + 280)}{4} - \frac{(220 + 230 + 240 + 250)}{4} = 260 - 235 = 25.$$

This underestimates the treatment effect because the subjects receiving the treatment were heavier than those in the control group (even in the absence of treatment). Even if everyone in the treatment and control groups are of the same weight without treatment, we can still have bias if those in the treatment group are more likely to benefit, as in the following example:

Subject	Control	Treatment
Cab	220	
Dab	230	
Fab		220
Gab		230
Jab		240
Lab		250
Nab	280	
Tab	290	

The estimated treatment effect would be:

$$\frac{(220 + 230 + 280 + 290)}{4} - \frac{(220 + 230 + 240 + 250)}{4} = 255 - 235 = -20.$$

This answer is also biased.

Thus, there are two sources of bias. The control and treatment groups may have a different average outcomes even in the absence of treatment. Or, the average treatment effect may be different in the treatment and control groups (those who are more likely to benefit from the treatment may get the treatment).

Why does the randomized trial give the right answer? Well, the randomized trial doesn't necessarily give the correct answer. It gives the right answer *on average*, the average being over all possible realizations of the experiment (the randomization). So I did a randomized assignment of treatments and calculated the effect. It came to be -35. Then I did it 10 times. Then 100 times, then 1000 times, and then 10000 times. Results tabulated below.

Number of replications	Average effect	Standard error
1	-35.0	
10	-10.0	4.87
100	-14.7	1.67
1000	-14.4	0.55
10000	-14.8	0.17

Random error and non-random error At the end of a randomized trial, we summarize results using estimates and confidence intervals. The confidence intervals are intended to summarize the *random error* that we have incurred. For a randomized trial, the random error cancels out, on average, and becomes progressively smaller as the size of the trial increases.

For an observational study, the random error becomes progressively small with study size, but the non-random bias (as in the examples above) does not decrease, and remains roughly constant. Thus, the magnitude of bias relative to random error actually *increases* with study size.

Thus it becomes more important to adjust (or compensate for) systematic biases as our study size becomes larger. Here are some common strategies.

PROPENSITY SCORES

The propensity score method can be used if we can assume that we have observed all factors (confounders) that affect both treatment assignment, and the outcome. This assumption is very important, and the estimated treatment effects will be biased if this assumption does not hold.

The first step is to construct a *propensity score*. This is the probability of treatment assignment given the confounders. Typically, one does not know this quantity, and we estimate this using logistic regression.

The second step is the adjustment step, which can be done in several ways. The two most common methods are covariate adjustment, and matching. In the former, we enter the propensity score in a regression model for the outcome just as we would enter any adjusting variable. For the latter method we create strata by grouping individuals by their propensity score, and calculating the treatment effect within those strata. The advantage of the latter method is that it makes fewer assumptions of the nature of the relationship of the confounders to the outcome.

INSTRUMENTAL VARIABLES

A disadvantage of the propensity score method is that we need to assume that we have observed all potential confounders. We can never be sure that we have. One way around that is to find a variable (the instrumental variable) that affects treatment assignment, but not the outcome. The key assumption is that the instrument does not affect the outcome directly, except through treatment assignment. Thus, we may have unobserved confounders that affect both treatment assignment, as well as outcome, and still get a consistent estimate of the treatment effect.

If the instrument is binary, we estimate the treatment effect as a ratio of two quantities. The first is the difference in the mean outcome in the two instrument groups. The second is the difference in the proportion of treatment assignments in the two instrument groups. The ratio is our estimate of treatment effect.

Note that if the instrument is *weak*, i.e. the proportion of treatment assignments in the two instrument groups is about the same (instrument is weakly correlated with treatment assignment), then the above ratio estimate becomes unstable (or undefined), and hence has high variance. Thus, sometimes, a biased estimate from a propensity score analysis may be more desirable than a (potentially) unbiased estimate from using an instrumental variable.

SENSITIVITY ANALYSES

Sensitivity analyses quantify how hidden biases of various magnitudes due to unmeasured confounders might alter the conclusions of a study. The first formal sensitivity analysis was conducted in a paper discussing cigarette smoking and cancer.[‡] The question they sought to answer was, can an unmeasured factor which causes lung cancer, but is more common in smokers than in non-smokers, explain the association between smoking and lung cancer? Their sensitivity analysis concluded that such an unmeasured confounder would have to be 9 times more common in smokers than non-smokers to explain the observed association. Such a huge disparity between smokers and non-smoker, although possible, is unlikely. Thus, their analysis shows that the association between smoking and lung cancer is likely to be causal.

There are many ways of performing sensitivity analyses. One method[§] works as follows. Assuming that there is hidden bias of a specified magnitude (that affects the distribution of subjects to treatment groups), the method provides an upper and lower bound for how the p-value of association might be affected. By varying the assumed magnitude of hidden bias, we can find out how much hidden bias is necessary to render our p-value "not significant." This parallels original approach taken in the discussion of smoking and cancer but has the advantage of accommodating sampling variation. The key to the method is that it estimates a p-value by assuming that the randomization assignment of treatments was biased by an unobserved variable. It then mathematically derives how much that randomization distribution can be slanted to give bounds for the p-value.

[‡]Cornfield J, Haenszel W, Hammond E, Lilienfeld A, Shimkin M, AND Wynder E (1959) Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173-203.

[§]Rosenbaum PR (1995) *Observational studies*. Springer-Verlag.