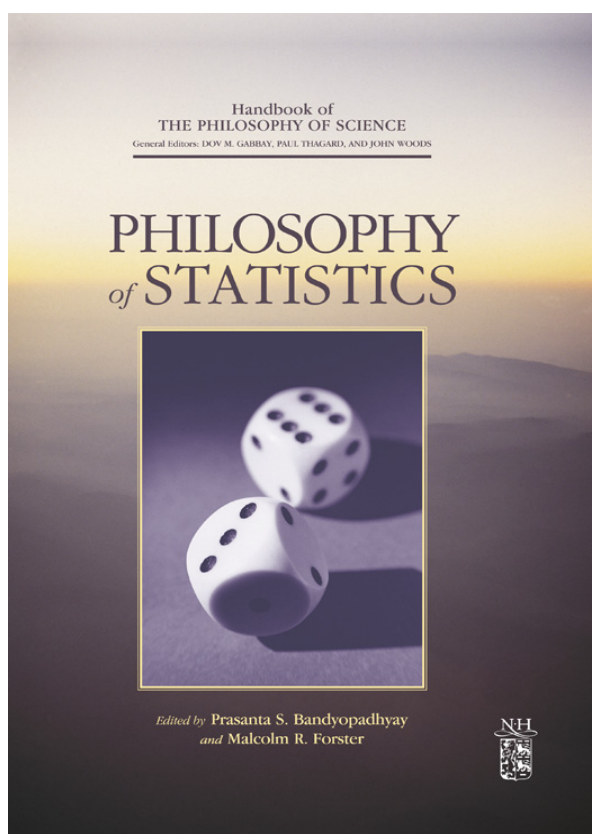


Provided for non-commercial research and educational use only. Not for reproduction, distribution or commercial use.

This chapter was originally published in the book *Handbook of The Philosophy of Science: Philosophy of Statistics*. The copy attached is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research, and educational use. This includes without limitation use in instruction at your institution, distribution to specific colleagues, and providing a copy to your institution's administrator.



All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>

From Greenland Sander, The Logic and Philosophy of Causal Inference: A Statistical Perspective. In: Dov M. Gabbay and John Woods, editors, Handbook of The Philosophy of Science: Philosophy of Statistics. San Diego: North Holland, 2011, pp. 813-830.

ISBN: 978-0-444-51862-0

© Copyright 2011 Elsevier B. V.

North Holland.

THE LOGIC AND PHILOSOPHY OF CAUSAL INFERENCE: A STATISTICAL PERSPECTIVE

Sander Greenland

The topic of causality is a vast one in science and philosophy, so vast that even a limited review would require a book. Yet most theories have not found favor among empirical researchers – by whom I mean those whose primary job is to collect and analyze data, as opposed to philosophers or theoreticians. This chapter will thus concern only the few statistical theories for causation and causal inference that have produced methods now widespread in practice.

In an attempt to avoid the confusion that often accompanies narrative descriptions of causation and causal inference (especially in applied sciences), this chapter uses rather stark and purely logical descriptions, and will assume the reader has at least some familiarity with probability and statistics. Detailed illustrations and applications, along with philosophical discussions, can be found in the references. Special emphasis will be given to issues of causal inference from uncontrolled observations (observational studies), in which the effect under study becomes difficult to separate from other, extraneous phenomena; the latter are often called “bias sources” or “systematic errors”.

Before embarking on these descriptions, I will touch briefly on the relevance (or possible lack thereof) for statistics of philosophies of causation.

DO WE NEED PHILOSOPHY OF CAUSATION FOR A STATISTICAL THEORY OF CAUSAL INFERENCE?

It is possible to distinguish two kinds of inference: Inference to causal models from observations, and inference from causal models to the effects of manipulations. Inference to causal models may be viewed as trying to construct a general set of laws from existing observations that can be tested with and applied to new observations. In statistics this problem is subsumed under the topic of model specification or model building. Inference from causal models may be viewed as deducing tests and making decisions based on proposed or accepted laws, which in statistics is subsumed under topics of testing, estimation, and decision theory.

In applied statistics, the feedback between these two directions of inference is often summarized as a cycle of model proposal → model test → model revision → model test that continues until available tests cease to have practical impact on the model [Box, 1980]. There are familiar controversies about whether cycles of this form lead toward “truth” or simply toward effective tools for prediction and

manipulation (e.g., [Kuhn, 1970a; 1970b]), and whether the philosophical debate surrounding causal inference stems from the fact that the word “causation” evokes some notion of a deeper truth about the world hidden from current view.

Of interest then is that the most successful statistical model of causation, the potential-outcomes model discussed below, has attracted theoretical criticisms precisely because it contains counterfactual elements hidden from randomized-experimental test (e.g., [Dawid, 2000]). These criticisms have been dismissed by applied statisticians (see the discussion following [Dawid, 2000]), who understand that the manipulative account inherent in potential-outcomes models fits well with the more instrumentalist or predictive view of causation than critics admit. Indeed, these models can be and have been used to great success with no worry about whether their hidden elements need to be taken seriously [Greenland, 2004], just as the celestial cogs and wheels once used to display the Ptolemaic model of celestial motions were no obstacle to its considerable predictive success.

Given this instrumentalist view, it might seem that causal inference maybe distinguished from other inferences only due to its emphasis on manipulation rather than prediction. From a statistical viewpoint, the distinction between prediction and causal inference is semantic, not philosophical: Causal inference is merely special case of prediction in which we are concerned with predicting outcomes under alternative manipulations. Because only one of the alternatives can be carried out, only one of the outcomes can be observed, resulting in nonidentification. But the solution to this problem is no different than in problems of pure prediction: We simply assume some limited form of isotropy, in which predictive regularities (whether labeled “predictive” or “causal”) persist over the space and time spans of interest, at least enough to justify generalizations across the spans. Whether a deeper analysis is warranted for practice remains to be seen.

POTENTIAL OUTCOMES AND STRUCTURAL EQUATIONS

Manipulative accounts of causation, including those with counterfactuals, have deep roots in the history of modern science. Informal outlines for causal inference may be traced as far back as the development of experimental science. After all, typical definitions of “experiment” include an element of experimenter “control” of conditions, implying that such control will affect the outcome. Early in these developments, however, Hume [1739; 1748, p. 115] recognized that the definition of “cause” implicit in much usage carried the seeds of intractable underdetermination or, as known in statistics, *nonidentification*; that is, observation alone could not determine or identify whether one condition caused another.

To formalize the notion of causation and delineate the identification problem, consider the following model (introduced by Neyman [1923]) which became established in the experimental literature in the mid-20th century, was later extended and popularized for observational research [Rubin, 1990], and is now standard in much of statistics where it is called the *potential-outcomes* or “counterfactual” model. Suppose we observe a subject i to have a particular outcome after being

given a treatment. (“Subject” is here merely a term for observational unit; it may be a plot of land, a laboratory animal, or a population, as opposed to a person.) Let X be the variable ranging over the treatment possibilities, and let x and x^* be two distinct treatments (that is, values for X with $x^* \neq x$). Let Y be the variable ranging over the outcome (response) possibilities, and let y and y^* be two distinct outcomes (that is, values for Y with $y^* \neq y$). As an example, X might encode a range of treatment options for women with perimenopausal complaints, such as unopposed estrogen therapy, opposed estrogen therapy, placebo treatment, and no treatment, while Y could indicate survival over the decade following treatment initiation ($Y = 1$ if the woman survives, 0 if not), or Y could be the survival time (lifespan) following treatment initiation.

A common notion of cause and effect is then captured as follows: Receiving treatment x (i.e., having $X = x$) *caused* the outcome Y to be y for the subject relative to having instead $X = x^*$, if the actual outcome and treatment were y and x , but would have been y^* had x^* been administered instead. The two outcomes y and y^* are then called the *potential outcomes* corresponding to treatments x and x^* for the subject, and the difference $y - y^*$ is called a measure of the *effect* on the subject of giving $X = x$ instead of (or relative to) $X = x^*$. In the example, $y - y^*$ could be the difference of survival time with unopposed estrogen therapy (x) versus placebo (x^*).

The nonidentification problem reflects that the subject’s response y^* to treatment x^* is not observed if the received treatment x is not equal to x^* , and therefore the effect $y - y^*$ cannot be computed from the observations. For example, we cannot observe how long a woman would survive under placebo therapy if in reality she receives unopposed estrogen, and so we cannot compute the effect that receiving unopposed estrogen rather than placebo had on this woman. This problem in causal statements is often highlighted by noting that the premise of the conditional “If X had been x^* instead of x , Y would have been y^* ” is counterfactual (contrary to fact) [Lewis, 1973ab]. Any statistical inference about the effect must therefore invoke physical assumptions that give precise meaning to the counterfactual conditional. It must also invoke “identification” assumptions that allow construction of estimates and tests of the average or expected effect $E(Y - Y^*)$ from the data actually observed.

For further details and citations on the model see [Greenland *et al.*, 1999; Greenland, 2004], or some of the many other reviews (e.g., [Morgan and Winship, 2007; Pearl, 2009]). The reader is warned however that (notwithstanding the enormous contributions by Rubin to the model), much of the sociologic literature misattributes the model to Rubin, some going so far as to call the model the “Rubin Causal Model” (e.g., [Holland, 1986]), and thus misses large segments of the literature on the model in the experimental, econometric, and health sciences.

Causal Laws and Structural Equations

Basic science often provides an empirical pattern or a more formal physical theory that predicts the subject's outcome Y as a function of the treatment X . This function might be subject-specific, tailored to specifics of the subject's characteristics. For example, suppose the subject is an ordinary ceramic dish, the treatment is dropping it flat on a concrete floor from height x , and the outcome is breakage ($Y = 1$) or not ($Y = 0$). Ordinary experience provides us a rough theory that says dropping the dish two meters will cause breakage whereas dropping it one millimeter will not; that is, $X = 2000\text{mm}$ will cause $Y = 1$ relative to $X = 1\text{mm}$. A different theory would apply to a steel dish. Especially in physics, such experience may eventually give rise to a mathematical "law" or model $f(x)$ relating Y to X for each of a broad class of subjects. Examples include laws governing behavior of charged particles in response to an electric field of a given strength, or more limited laws governing the size of predator populations in response to prey abundance.

Such theoretical mechanisms or laws shift the uncertainty about the counterfactuals (which are unobserved potential outcomes) to uncertainty about the mechanisms or laws connecting the outcome variable Y to the antecedent variable X . Indeed, that shift is often promoted as a major force for progress in experimental science, as follows: Suppose one proposes a general physical theory that says or implies that $Y = f(x)$ whenever $X = x$ for each subject in a given class. Observing these predictions fail — that is, observing enough subjects in the class who have $Y \neq f(x)$ — can then be grounds for discarding or modifying the theory.

A functional relation $Y = f_i(x)$ supplying the outcome of subject i under different possible treatments is often called a *structural equation* for the subject. It is important to note that the equation gives the variation in the outcome Y as X varies *within a single subject*; that is, it shows how Y varies as X is varied while i is held constant. This single-subject feature captures the counterfactual nature of causal laws.

Although it is not the defining feature of a structural equation, the within-subject property distinguishes the equation from ordinary "regression" functions of statistics, which describe *associations*. An association is the variation in Y as one moves *across* subjects with different X ; that is, both X and i are varying in a regression. Nonetheless, because analytic statistical methodology is heavily invested in estimating associations, and because associations are all that are statistically identified in nonexperimental settings, much of statistical theory for causal inference comprises delineation of assumptions that allow deduction of regression equations from structural equations [Berk, 2003; Pearl, 2009].

Causal Null Hypotheses

A large portion of standard statistical theory concerns testing of associational "null hypotheses," which assert absence of association in a population or distribution from which the observations are supposed to have arisen. In causal inference these

nulls become no-effect hypotheses or “causal nulls,” which in their strong or strict form state that the structural equation is constant for each subject; that is, $f_i(x) = c_i$ for each subject i . This hypothesis need not correspond to the hypothesis of no association, which asserts that Y does not change across different subjects with different X values. But much of statistical theory for causal inference comprises delineation of assumptions that allow deduction of associational null hypotheses from causal null hypotheses. When such an associational null is identifiable, its rejection by a statistical test implies that the causal null hypotheses should be rejected as well.

The Causal Identification Problem

As just described, there is one element that makes a theory causal and which demands more than mere observation for inference: The multiple possible values for Y for each subject i , given different values of X . Let x_i be the observed value of X for subject i , and suppose we have a theory that predicts Y as a function $f_i(x_i)$ of X . No matter how many subjects yield the predicted outcome $f_i(x_i)$ upon passive observation, and so are in accord with the causal theory that $Y = f_i(x)$, that theory cannot be deemed more than a good *description* of how Y and X will be associated as one looks *across* subjects. In terms of passive observation, it merely predicts how plots the pairs x_i, y_i will look as i varies across subjects.

The theory’s predictions may be borne out among the treatment assignments (X distribution) we observe, but we cannot be sure the theory would have succeeded under other, counterfactual treatment assignments. In algebraic terms, we might see $Y_i = f_i(x_i)$ for all the subjects, even though Y_i might not have equaled $f_i(x^*)$ for some $x^* \neq x_i$. In sum, a theory may be a very good description of what we observe but a poor predictor of what we *would* observe upon intervening to change treatment assignment X . In other words, no matter how well it predicts associations across subjects, it need not tell us how the outcome Y would change upon changing X *within* subjects.

Confounding and Randomization

Whether by the investigator, nature, or another party, the treatment x_i given subject i might have been determined in a way that is associated with Y across subjects, apart from any causal (within-subject) link from X to Y . The term *confounding* is often used to refer to this condition [Greenland *et al.*, 1999], although it is also known as “nonignorability of the treatment-assignment mechanism” [Rubin, 1991]. Another way to describe this condition is that there is between-subject variation in Y that is not due to within-subject variation of Y with X ; in other words, given a fixed value x for X , it is variation in the potential outcome Y_x with X across subjects. Earlier, informal discussions of causal inference described this condition as “extraneous variation,” or that portion of between-subject association of X and Y that is not due to an effect of X on Y . They recognized that such

covariation of X and Y would remain present even if the causal null hypothesis were correct, and thus would distort causal inferences or tests of that hypothesis [Mill, 1843].

Sources of extraneous variation are sometimes called “confounders,” although the term *confounder* is often defined in more strict terms [Greenland and Pearl, 2007]; see the section on causal diagrams below). Recognizing the impossibility of eliminating or even knowing all confounders in biological work, R.A. Fisher [1932, 1935] developed an elegant theory of *randomized experiments* to allow statistical “control for” confounding. This control is accomplished by enforcing a known distribution for confounding under sufficiently strict causal hypotheses (such as the null hypothesis).

In its basic form, randomization theory replaces vague ignorance about the degree of confounding with a fully specified probability distribution for the observed outcomes under the sharp null hypothesis of no effect $Y_i = f_i(x) = c_i$ (Y constant across X within i). Consider classical permutation inference (e.g., [Cox and Hinkley, 1974, Ch.6]) in an experiment that assigns values of X among N subjects, and let R be the treatment-assignment variable. R ranges over possible values for X ; thus, when subject i is assigned to have $X = x$, the subject’s value r_i for R is equal to x . Note however that the subject may deviate from assigned treatment, resulting in the actual value x_a of X not being equal to the assigned value x (in which case $X \neq R$).

If assignment R has no effect on Y , the observed outcomes y_1, \dots, y_N should be the same regardless of the assignments r_1, \dots, r_N ; in other words, they should be the c_i in the null model $Y_i = f_i(x) = c_i$. Thus, given this causal null hypothesis, we may regard the outcome list (y_1, \dots, y_N) as if it were fixed from the start of the experiment and thus independent of a subsequent random treatment allocation. A known randomization scheme then allows one to compute exact probabilities for any allocation of these fixed outcomes among the treatment levels, including the allocation observed.

From these probabilities, we can compute null distributions of test statistics (such as the sample sum of cross products $\sum_i x_i y_i$ used for trend tests) and thus compute P -values (“observed significance levels”) for testing the causal null hypothesis. If the possible treatment allocations are merely permutations of the actual treatment-assignment list r_1, \dots, r_N , the resulting test is known as a *permutation test* for the effect. Test of means may also be used, noting that the total-sample mean of Y remains fixed under the null hypothesis, regardless of allocation.

As a special case of the foregoing methods, consider binary X and Y with perfect random allocation of N_1 subjects to $X = 1$ and N_0 subjects to $X = 0$, and let $N = N_1 + N_0$. Then $\sum_i x_i y_i$ is the number of subjects in the $X = 1, Y = 1$ cell of the 2 by 2 table of X and Y . Under the causal null hypothesis, (y_1, \dots, y_N) is a fixed vector of potential outcomes, and so the $Y = 1$ marginal total in the experiment is fixed at $y_+ = \sum_i y_i$ (so is the sample mean of Y , y_+/N , which is the proportion of subjects having $Y = 1$). Under the null, randomization becomes nothing more

than randomly allocating the fixed outcomes (y_1, \dots, y_N) to the $X = 1$ and $X = 0$ categories, which induces a hypergeometric distribution for $\sum_i x_i y_i$, as derived by Fisher for his exact test [Cox and Hinkley, 1974, Ch. 5].

Causality and Conditionality

The conditionality problem illustrates how the introduction of a causal component into a statistical model can resolve previous ambiguities in choice of a statistical procedure. This resolution comes from explicitly modeling the otherwise hidden within-subject dimension underlying causal questions, and shows how statistical questions can arise even when no ordinary sample-to-population inference problem exists.

There has been a long-running controversy in statistics concerning whether the use of permutation tests is justifiable when the sample distribution of Y (comprising the observed values y_1, \dots, y_N) is not fixed in advance by the investigator. Some of the arguments for these tests appeal to rather abstruse and somewhat controversial principles of ancillarity or conditionality [Little, 1989] while others are based on favorable repeated-sampling (frequency) properties when y_1, \dots, y_N represent a sample from a distribution for Y .

If one is concerned only with the observed N units, the question of association of X and Y among those sampled is purely empirical rather than inferential, in that it is answered by simply plotting or cross-tabulating the observed pairs (x_i, y_i) . In this regard, it is no different a question than asking the heights of the N tallest mountains on earth: Accept the reported measurements and you have your answer.

The absence of a statistical inference problem in this descriptive question has led many to automatically identify the fixed-margin controversy and even causal inference with the problem of inference to a larger population. Add however the causal dimension and we have a problem of inference from observed properties of the sample (the observed distribution of the pairs x_i, y_i) to unobserved properties of the same sample, namely the unobserved potential outcomes of the sampled subjects. As described above, the fixed Y margin can be deduced directly from the causal null hypothesis for the observed sample of units ($i = 1, \dots, N$), with no reference to sampling from a larger population. In particular, the fixed Y margin is just a physical property of the sample under the causal null hypothesis among those sampled [Greenland, 1991].

There is however a connection to inference about a population from which the observed units are sampled. First, note that rejection of the null for the sample implies rejection for the population: The sample is part of the population, and thus finding an effect in the sample implies that there is an effect in the population (namely, in the part that composes the sample). The converse condition is that failure to reject the null for the sample should correspond to failure to reject for the population. This converse is not logically necessary, but violating it would amount to asserting that causation exists in a population even though we would not assert causation exists in the portion we observed. At the very least such an

assertion would seem paradoxical [Greenland, 1991].

The General Causal Inference Problem as a Missing-Data Problem

To extend statistical reasoning about causation beyond the null hypothesis, it is essential to add a model for the distributions of the potential outcomes. In the general form of this model, the univariate outcome variable Y is replaced by a fixed, baseline covariate vector \mathbf{Y} with components Y_x indexed by values x of the treatment X ; these Y_x are the potential-outcome variables, one for each possible value x of X . (More generally, as described later, X may be a vector \mathbf{X} that indexes the potential-outcome vector \mathbf{Y} .)

The treatment-allocation variable becomes a vector \mathbf{R} of indicators with components R_x , where $R_x = 1$ if and only if component Y_x of \mathbf{Y} is observed. The observed value \mathbf{r}_i of \mathbf{R} for subject i thus displays which component of \mathbf{Y} was observed (if any), and is a vector of zeros except possibly a single 1 at the component corresponding to the actual treatment. As a consequence, $r_{+i} = \sum_x r_{xi}$ is either 0 or 1 (no or one component of \mathbf{Y} observed). This conceptual framework allows one to view causal-inference problems as special types of missing-data problems, in which only one component of \mathbf{Y} can be observed (i.e., no more than one component of \mathbf{R} can be 1), leaving the rest of \mathbf{Y} as “missing data” [Rubin, 1991].

Because only one component of \mathbf{Y} can be observed, the joint distribution $\Pr(\mathbf{Y}=\mathbf{y})$ of the components of \mathbf{Y} is not statistically identified — that is, distinct distributions for \mathbf{Y} can lead to exactly the same distribution for the observations. Assuming that every subject has an observed outcome, those observations are the subject-specific dot products $\mathbf{R}'\mathbf{Y} = \sum_x R_x Y_x$, which equal the observed outcomes (the Y_x for which $R_x = 1$), and \mathbf{R} , which shows the treatment that was received (i.e., the R_x that equals 1). Without further assumptions or experimental control, all we may identify is $\Pr(\mathbf{R}'\mathbf{Y} = \mathbf{r}'\mathbf{y}, \mathbf{R}=\mathbf{r})$ and functions of it, including the conditional distributions $\Pr(Y_x = y | R_x = 1)$. In the nonexperimental settings, it may be prudent to search for the weakest identification assumptions consonant with our practical goals.

Standardization and Inference on Marginal Effects

The goal of most statistical causal inference is to compare the marginal distributions $\Pr(Y_x = y)$ of the Y_x across X . These comparisons are identified under the independence assumption that for all x , $\Pr(Y_x = y) = \Pr(Y_x = y | R_x = 1)$, sometimes called “weak ignorability” (the stronger but inessential condition that \mathbf{R} is independent of \mathbf{Y} is then called “strong ignorability”), and which corresponds to absence of confounding. The paradigmatic example arises when \mathbf{R} is randomized by the investigator, for then \mathbf{R} is independent of everything, including any potential-outcome vector \mathbf{Y} .

In nonrandomized studies, ignorability conditions are usually unacceptable assumptions. One strategy for this situation is to pretend instead that \mathbf{R} and \mathbf{Y} are

independent conditional on a vector of fully observed covariates \mathbf{Z} , called *strong ignorability given \mathbf{Z}* . A weaker condition sufficient for practical applications is the analogous set of X -specific relations,

$$(1) \quad \Pr(Y_x = y | R_x = 1, \mathbf{Z} = \mathbf{z}) = \Pr(Y_x = y | \mathbf{Z} = \mathbf{z})$$

for all x , called *weak ignorability given \mathbf{Z}* .

Now consider the following “standardization” formula, which biostatisticians, demographers, and epidemiologists may recognize as a modern version of the classical formula for “direct standardization” to the covariate distribution $\Pr(\mathbf{Z} = \mathbf{z})$:

$$(2) \quad \Pr(Y_x = y) = \sum_z \Pr(Y_x = y | \mathbf{Z} = \mathbf{z}) \Pr(\mathbf{Z} = \mathbf{z})$$

This equation is just a basic probability relation displaying $\Pr(Y_x = y)$ as a covariate-probability weighted average of the covariate-specific potential-outcome probabilities $\Pr(Y_x = y | \mathbf{Z} = \mathbf{z})$. Applying assumption (1) to equation (2) yields

$$(3) \quad \Pr(Y_x = y) = \sum_z \Pr(Y_x = y | R_x = 1, \mathbf{Z} = \mathbf{z}) \Pr(\mathbf{Z} = \mathbf{z}).$$

This equation provides the desired marginals $\Pr(Y_x = y)$ in terms of the observable distribution $\Pr(Y_x = y | R_x = 1, \mathbf{Z} = \mathbf{z})$, assuming the standardization is adequate to remove confounding. Equation (3) is thus sometimes termed “no confounding of the marginal effects given \mathbf{Z} ”. It is a weaker condition than assumption (1) (weak ignorability given \mathbf{Z}) because deviations from (1) may average to zero over \mathbf{Z} and thus preserve equation (3), or at least leave it an acceptable approximation. In other words, in theory, stratification by \mathbf{Z} need not be sufficient for estimating conditional effects in order to be sufficient for estimating marginal effects.

Using the fact that $\Pr(R_x = 1 | \mathbf{Z} = \mathbf{z}) = \Pr(R_x = 1, \mathbf{Z} = \mathbf{z}) / \Pr(\mathbf{Z} = \mathbf{z})$, equation (3) may be rewritten in an equivalent form

$$(4) \quad \Pr(Y_x = y) = \sum_z \Pr(Y_x = y, R_x = 1, \mathbf{Z} = \mathbf{z}) / \Pr(R_x = 1 | \mathbf{Z} = \mathbf{z}).$$

This equation displays $\Pr(Y_x = y)$ as an *inverse-probability weighted* (IPW) average of the unconditional observation probabilities $\Pr(Y_x = y, R_x = 1, \mathbf{Z} = \mathbf{z})$. The allocation probabilities $\Pr(R_x = 1 | \mathbf{Z} = \mathbf{z})$ that form the inverse weights are sometimes called “propensity scores” and can be generalized to continuous and time-dependent treatment processes [Robins, 1999ab]. Equation (4) shows how these scores, if known or at least identifiable, can be used to estimate a marginal potential-outcome distribution $\Pr(Y_x = y)$ under the \mathbf{Z} -conditional weak ignorability assumption, which licenses the derivation of equation (3) and hence (4) from equation (2).

Summary

Classical permutation arguments are based on enforcing a distribution for the observation-indicator vector \mathbf{R} (e.g., by treatment randomization) and then using

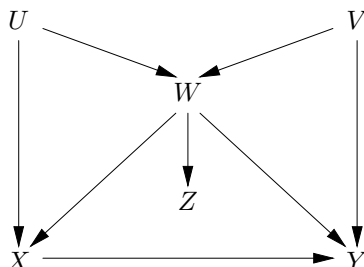


Figure 1. Example of a DAG

this distribution as the source of subsequent probability statements about the data. These arguments are not heavily emphasized in most statistical training, and yet are the ones most directly linked to potential-outcome models of causation. They should be contrasted with regression statistics, which base their probability statement on assumed distributions for the observed outcome $\mathbf{R}'\mathbf{Y}$ given observed covariates \mathbf{Z} .

The advantage of the treatment-based approach to causal inference is clear when indeed the distribution of \mathbf{R} is known or at least identifiable, as in experiments: It seems far better to use an identified distribution than one that is merely assumed. Nonetheless, in nonexperimental (observational) research, all distributions become mere assumptions when not based on known mechanisms. The choice between approaches then comes down to judgments about which assumptions are more plausible or at least more palatable. Current research on statistical methods for causal inference includes development of “multiply robust” procedures, which retain their statistical validity under broader conditions than either treatment-based or outcome-based modeling [Kang and Shafer, 2007].

CAUSAL SYSTEMS AND CAUSAL DIAGRAMS

Suppose now we have a time-sequenced set of structural equations, or *causal system*, in which (for example) the output w of a function $f(u, v)$ may become an input to a later function $g(u, w)$ with output x . (Equivalently, suppose that for a variable W with corresponding potential-outcome vector \mathbf{W} , the observed value $W = \mathbf{R}'_W \mathbf{W}$ may be part of the index vectors for the potential outcomes of a subsequent variable.) We may then illustrate the system using a *directed acyclic graph* (DAG) in which arrows connect input variables to output variables [Pearl, 2009; Glymour and Greenland, 2008; Spirtes *et al.*, 2001]. Such graph is a *causal diagram* if (as here) the arrows are interpreted as links in causal chains.

Figure 1 provides an example, illustrating a system of structural equations

$$\begin{aligned} U &= f_U(\varepsilon_U), & V &= f_V(\varepsilon_V), & W &= f_W(u, v, \varepsilon_W), \\ X &= f_X(u, w, \varepsilon_X), & Y &= f_Y(v, w, x, \varepsilon_Y), & Z &= f_Z(w, \varepsilon_Z), \end{aligned}$$

where the inputs $\varepsilon_U, \varepsilon_V, \varepsilon_W, \varepsilon_X, \varepsilon_Y, \varepsilon_Z$ are “purely random disturbances,” that is, inputs that are independent random variables (but not necessarily identically distributed). Traditionally, such disturbances are left implicit, i.e., understood to be present but not shown.

The entire system may be viewed as a multivariate model for the graphed variables, with the graph encoding various constraints on the joint distribution of these variables [Lauritzen, 1996; Spirtes *et al.*, 2001; Pearl, 2009]. In particular, the distribution of the disturbances induces a joint distribution of the graphed variables which obeys the *Markov decomposition*. That is, the joint distribution of the graphed variables decomposes into factors, one for each graphed variable, that give the probability of each variable given its graphical/functional inputs (“parents”). For Figure 1 the decomposition is (in an abbreviated notation)

$$\Pr(u, v, w, x, y, z) = \Pr(u)\Pr(v)\Pr(w|u, v)\Pr(x|u, w)\Pr(y|v, w, x)\Pr(z|w).$$

Because U and V have no inputs within the system (they are “exogenous”), their factors $\Pr(u)$ and $\Pr(v)$ are unconditional.

The decomposition provides not only the original joint distribution, but also a formula for the effect on that distribution of shifting the functional relations or distributions of any subset of the variables. For example, *randomization* of W usually refers to an intervention that replaces $W = f_W(u, w, \varepsilon_W)$ by $W = f_W(\varepsilon_W^*)$ where the distribution of ε_W^* (and hence W) is determined by the investigator and hence is known. The resulting joint distribution of the variables is then

$$\Pr(u, v, w, x, y, z) = \Pr(u)\Pr(v)\Pr(w)\Pr(x|u, w)\Pr(y|v, w, x)\Pr(z|w),$$

the factor $\Pr(w|u, v)$ being replaced by the new randomization distribution for $W, \Pr(w)$. The corresponding graph lacks arrows into W .

Suppose that instead of randomizing W we intervene to force all values of W to a particular value w_0 , without altering any other functional relation (if this can be done). We then replace the equation $W = f_W(u, w, \varepsilon_W)$ by the equation $W = w_0$. The resulting joint distribution of the variables is then 0 except at $W = w_0$, where it is

$$\Pr(u, v, w_0, x, y, z) = \Pr(u)\Pr(v)\Pr(x|u, w_0)\Pr(y|v, w_0, x)\Pr(z|w_0),$$

the factor $\Pr(w|u, v)$ being replaced by the new, forced distribution $\Pr(w_0) = 1$. The corresponding graph lacks arrows into W , and has $W = w_0$ in place of W . Repeating this exercise for other values of W shows how the system responds to various interventions that fix or set W to particular values, again presuming this is can be done without disturbing other systemic relations [Pearl, 2009].

Some Useful Elements of Graph Theory

To describe further consequences of the Markov decomposition relevant for causal inference we need several concepts from graphical probability theory. Two variables in a graph are *adjacent* if they are connected by an arrow. Consider a

“path” in a graph, a nonrepeating sequence of variables in which successive sequence members are adjacent. Effects of one variable on another are transmitted by causal sequences or *causal pathways*, which in a causal diagram are *causal* or *directed* paths in which each arrow points to the tail of the next arrow in the path. More precisely, given a causal diagram, the existence of a directed path from one variable to another is a necessary but not sufficient condition for an effect to occur. Thus, in Figure 1, $U \rightarrow W \rightarrow Y$ means U *can* affect Y via its effect on W , but *not* that U does affect Y . The graph is *acyclic* if (as assumed here) no variable in the graph affects itself (meaning there is no feedback loop in the graph).

A central problem of causal inference is distinguishing causation from mere probabilistic dependence, or “association” as it is often called. Graph theory provides a quick visual distinction, via the following concepts. A variable is a *collider* on a path if it is at a meeting of two arrowheads along the path; otherwise it is a *noncollider* on the path. In Figure 1, W is a collider on the path UWV but a noncollider on the paths UWY and XWY . A path is said to be *closed* or *blocked* at a collider, and *open* or *unblocked* at a noncollider. The entire path is open if it has no collider, otherwise it is closed. Open paths include but are not limited to causal pathways, a fact which (as discussed below) reflects classic problems in causal inference.

It is often helpful to think of associations as signals flowing through the graph. Given a graph, associations can flow through or be transmitted by open paths. Open paths themselves are merely conduits for the transmission, however. More precisely, given a graph, the existence of an open path between two variables is a necessary but not sufficient condition for an association between them. (It should be noted however that the presence of an open path will in practice almost always correspond to the presence of an association, although more likely a miniscule one if the path is long [Greenland, 2003].)

Conversely, a sufficient (but not necessary) condition for two variables to be unassociated (independent) is that there is no open path between (that is, any path between them is closed). Thus, we can immediately spot the independencies that must hold in the graphed distribution by just seeing whether two variables have no open path between them. If a statistical test of these independencies rejects them (that is, detects associations where none should be, according to the graph), that result may be taken as evidence against the posited causal system that gave rise to the graph.

Biases and Confounding

Now suppose we are interested in an effect of one variable (the target antecedent) on another variable (the target outcome). Any open path between these target variables that is not part of this effect is a *biasing path*, because it provides a pathway for association between the target variables that is not due to the target effect.

To illustrate, suppose our interest is in the net (total) effect of the antecedent W

on the outcome Y in Figure 1. This effect equals the net association transmitted through all the causal paths from W to Y , which are WY and WXY . There are, however, two other open paths from W to Y : $WUXY$ and WVY . Thus, the association we observe will be the net transmission through all four paths, which may be far from the net transmission through the two target paths WY and WXY ; that is, the signal of interest may be seriously corrupted by unwanted signals through $WUXY$ and WVY . These unwanted signals (transmissions along biasing paths) are examples of *biases*, although the concept of “bias” subsumes other phenomena as well (such as distortions due to measurement error).

Unconditionally, a biasing path for a net effect in a DAG must pass through a shared (“common”) cause of the target variables. What is more, it must consist of two segments, one being a causal path from the shared cause to the target antecedent, and the other a causal path from the shared cause to the target outcome that does not include the target antecedent. For example, suppose again that our inferential target is the net effect of W on Y in Figure 1, which has biasing paths $WUXY$ and WVY . $WUXY$ can be decomposed into a causal path UW from U to W and a causal path UXY from U to Y , joined at the shared cause U of W and Y . Similarly, WVY can be decomposed into a causal path VW from V to W and a causal path VY from V to Y , joined at the shared cause V of W and Y .

Any bias that is transmitted via a common cause of the target variables is an example of confounding, in that it contributes to the association of the target antecedent variable with the potential target outcomes (that is, it contributes to nonignorability). More generally, confounding arises from association that is transmitted along biasing paths that terminate with an effect on the target outcome. Thus, as described earlier, confounding is association due to “extraneous” effects on the target outcome. Those effects are said to “confound” the target effect. Correspondingly, a *confounding path* is a biasing path that terminates with an arrow into the outcome of interest [Greenland and Pearl, 2007]. In Figure 1, both of the biasing paths for the effect of W on Y ($WUXY$ and WVY) are confounding paths, because both terminate with an arrow into (effect on) the target outcome Y .

If confounding occurs, variables within the responsible confounding paths are often called *confounders*. Because the entire confounding path is open, any confounder must be linked by open paths to both target variables, and must have associations with both target variables. The converse is not correct, however: A variable associated with both target variables need not be a confounder. For example, when examining the net effect of U on Y in Figure 1, X could be associated with U and Y but would not be a confounder because it does not lie on a confounding path.

Given a DAG with no conditioning, it can be shown that all biasing paths are confounding paths, and vice-versa. Conditioning, however, may open biasing paths, some of which may not be confounding paths. We thus now turn to the concept of conditioning in graphs.

Conditioning and Control

Let G and C be disjoint subsets of variables in the graph, with g and c being sets of values for G and C . Independencies in the conditional distribution $\Pr(g|c)$ implied by the graph may then be seen using just a few more concepts. One key notion is that the open/closed status of a variable along a path is reversed by conditioning (stratifying) on the variable: A collider becomes open and noncollider becomes closed. As a consequence, the status of paths may reverse. For example, if we condition on W in Figure 1, the closed path $XUWVY$ becomes open and now can transmit associations; it thus becomes a confounding path for the X effect on Y . At the same time, the open paths $XUWY$, $XWVY$, and XWY become closed and can no longer transmit associations, and thus are no longer confounding paths.

It should be noted that, in accord with ordinary language, experimental sciences use the term “control” to refer to a physical alteration of a system to remove sources of bias, such as randomization. In observational sciences, however, “control of a variable” is often used more broadly to include conditioning on a variable, whether it removes bias or creates bias. Thus, conditioning on W in Figure 1 will “control” (remove) any confounding of the X effect on Y that was present in the original system, but at the same time may introduce new confounding by opening paths that were previously closed.

Not all biasing paths opened by conditioning are confounding paths, however. For example, suppose our target effect is the net effect of U on V in Figure 1. Because there is no causal pathway from U to V , this effect is zero, or “null.” But conditioning on W will open the path UWV , allowing unwanted association to flow from U to V . In other words, conditioning on W transforms UWV into a biasing path which is not a confounding path, because it does not terminate with an arrow into V . Bias that results from such conditioning on a shared effect of the target variables is often called “Berksonian,” in honor of the discoverer of this type of bias, Joseph Berkson [Glymour and Greenland, 2008].

Conditioning can also produce bias due to closing target paths. For example, suppose our target effect is the net effect of U on Y in Figure 1 (associations transmitted via UXY , UWY , and $UWXY$). This target effect is equal to the unconditional association of U and Y , because there is no unconditional biasing path. Conditioning on W will open the path $UWVY$, allowing unwanted association to flow from U to Y . In other words, conditioning on W transforms $UWVY$ into a biasing path (which is a confounding path, since it terminates in an arrow into Y). But conditioning on W will also close the target paths UWY and $UWXY$, blocking part of the effect (signal) of interest. The association of U and Y conditional on W may thus bear little resemblance to the target effect.

Suppose instead that our interest is only in that part of the effect of U on Y not mediated by W (UXY in Figure 1). A standard strategy in the social-science literature is to then condition on W in order to block the effects mediated by W (UWY and $UWXY$ in Figure 1). Unfortunately, this literature almost always overlooks the fact that the same strategy can introduce bias via the pathways

opened by conditioning on the intermediate W ($UWVY$ in Figure 1)

Conditioning on effects of a variable can also partially reverse its status on paths. For example, conditioning on a variable affected by a collider on a path from X to Y can partially open the path and hence can result in new bias if the rest of the path is open after the conditioning. For example, if our target is the net effect of X on Y in Figure 1, conditioning on Z (affected by W) can partially close the confounding paths passing through W ($XUWY$, XWY , and $XWVY$) yet partially open the path $XUWVY$, which becomes a confounding path.

Collider Bias, Response Bias, and Selection Bias

Any biasing path opened by conditioning (whether full or partial) must pass through at least one collider; hence any bias that results from a newly opened path may be called *collider bias* [Greenland, 2003]. Of particular interest are those instances in which collider bias results from the process that determines how subjects come to be included in the statistical analysis of a target effect or association. The process is usually described in terms of subject response (to requests for participation) or subject selection (whether selection by the researcher or self-selection by the subject). Any bias that results from the process is thus often called “response bias” or “selection bias.” As mentioned above, “Berksonian bias” usually refers to situations in which both the target variables affect inclusion.

A more general and accurate term for all these biases is *inclusion bias*. To describe their shared structural and graphical representation, suppose Z is an indicator variable for inclusion in the analysis, with $Z = 1$ if a subject is included and $Z = 0$ if not. All associations observed must then be conditional on $Z = 1$. To say inclusion is random (“random sampling” for analysis) means that the structural equation for Z is $Z = \varepsilon_Z$, where ε_Z is a random indicator independent of all other random disturbances; Z will then have no causes in the corresponding graph (it will be exogenous in the graph). But if inclusion is affected by more than one causal pathway, Z will appear in the graph as a collider or as a variable affected by a collider, and observed associations may suffer considerable bias from the forced conditioning on $Z = 1$.

To illustrate, consider again Figure 1, with Z the inclusion indicator, and again with the target being the net effect of U on Y . Here, W may influence whether a subject is included or not. As a consequence, the forced conditioning on $Z = 1$ can partially open the path $UWVY$, which becomes a confounding path, and partially close the target paths UWY and $UWXY$. The net bias from these changes may be considerable, if Z is strongly affected by W , or minor if Z is only weakly affected by W .

DISCUSSION

The above models can be extended to deal with another major source of bias in observational research, measurement error. Because the topic is quite involved and brings in many elements not related to causality, it has not been included here. Notably however, measurement error expands the nonidentification problem to noncausal associations, thus further weakening identification of causal effects from observed associations; see [Greenland, 2005a; 2009; 2010] for examples and discussion.

The above models also extend to consideration of longitudinal (time-varying) treatments such as medical regimens, but again many technical elements arise in applications [Robins, 1999ab]. For a discussion of relations of the models to the sufficient-component cause model common in epidemiology (which is equivalent to the INUS model of Mackie [1965]) see [Greenland and Brumback, 2002; VanderWeele and Robins, 2008].

As may be apparent from the above presentation, statistical and structural representations of causation bypass most of the philosophic subtleties associated with the complex topic of cause and effect. This bypass has facilitated applications and may reflect the task-oriented attitude of most scientists. Nonetheless, it should not lead one to overlook some serious practical problems that are usually ignored.

Perhaps the largest problem is the possible ambiguity in what it means to intervene on a variable or to “change” its level. This ambiguity can render ambiguous the concept of a potential-outcome vector \mathbf{Y} for X . After all, if Y_x is a counterfactual component of \mathbf{Y} (as all but one component must be), its value may depend in a dramatic fashion on exactly how X would come to be x if x is counterfactual [Greenland, 2005b; Hernán, 2005]. In causal diagrams the same problem is utterly invisible. These are not fatal objections to the models, for the models have proven useful whenever the meaning of interventions and outcomes is unambiguous (for example when X is measles vaccination and Y is the subsequent occurrence of measles). But they are quite disconcerting when the models are used to make claims about the impact of (for example) “eradication of childhood disease”: The effect of such an ambiguous action depends dramatically on exactly how it is carried out (e.g., by vaccinating, by curing, or by killing children).

A related problem of less practical concern, but nonetheless disconcerting, is that potential-outcome models and their structural-equation generalizations seem to use an informal notion of causation to define actual effects. In particular, “setting” or forcing a treatment variable X to a particular level x (whether that is done in response to a random-number generator or in response to extraneous factors) is a causal command left undefined in the account. Here again the accompanying causal-diagram theory is silent, taking the causal interpretation of its arrows as a primitive.

Regardless of any objections and problems, the statistics and observational science literature employing potential-outcome models has exploded over the past few decades, while causal diagrams have spread rapidly in the wake. It thus seems

important that those interested in issues of causality become familiar with these formal yet practical tools for causal inference.

ACKNOWLEDGEMENTS

I am grateful to Katherine Hoggatt and a referee for helpful comments on the initial draft of this paper.

BIBLIOGRAPHY

- [Berk, 2004] R. A. Berk. *Regression analysis: a constructive critique*. Newbury Park, CA: Sage, 2004.
- [Box, 1980] G. E. P. Box. Sampling and Bayes inference in scientific modeling and robustness. *J R Stat Soc Ser A*; 143:383–430, 1980.
- [Cox and Hinkley, 1974] D. R. Cox and D. V. Hinkley. *Theoretical statistics*. New York: Chapman and Hall, 1974.
- [Dawid, 2000] A. P. Dawid. Causal inference without counterfactuals (with comments). *J Am Stat Assoc* 95: 407–428, 2000.
- [Fisher, 1932] R. A. Fisher. *Statistical methods for research workers*, 4th ed. London: 1932.
- [Fisher, 1935] R. A. Fisher. *The design of experiments*. Oliver and Boyd, Edinburgh, 1935.
- [Glymour and Greenland, 2008] M. M. Glymour and S. Greenland. Causal diagrams. In K. J. Rothman, S. Greenland and T. L. Lash. *Modern Epidemiology*, 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins, 2008.
- [Greenland, 1991] S. Greenland. On the logical justification of conditional tests for two-by-two contingency tables. *Am Statist*, 45:248–251, 1991.
- [Greenland, 2003] S. Greenland. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14:300–306, 2003.
- [Greenland, 2004] S. Greenland. An overview of methods for causal inference from observational studies. In A. Gelman and X. L. Meng, eds. *Applied Bayesian modeling and causal inference from an incomplete-data perspective*. New York: Wiley, 2004.
- [Greenland, 2005a] S. Greenland. Multiple-bias modeling for analysis of observational data (with discussion). *J R Stat Soc series A*, 168:267–308, 2005a.
- [Greenland, 2005b] S. Greenland. Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). *Emerg Themes Epidemiol*, 2:1–4, 2005b. (Originally published as “Causality theory for policy uses of epidemiologic measures”. Chapter 6.2 in: C. J. L. Murray, J. A. Salomon, C. D. Mathers, and A. D. Lopez, eds. *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/WHO, 291–302.)
- [Greenland, 2009] S. Greenland. Relaxation penalties and priors for plausible modeling of non-identified bias sources. *Statistical Science*, 24:195–210, 2009.
- [Greenland, 2010] S. Greenland. Overthrowing the tyranny of null hypotheses hidden in causal diagrams. Ch. 22 in R. Dechter, H. Geffner, and J. Y. Halpern, eds. *Heuristics, Probabilities, and Causality: A Tribute to Judea Pearl*, pp. 365–382. London: College Publications, 2010.
- [Greenland and Brumback, 2002] S. Greenland and B. A. Brumback. An overview of relations among causal modeling methods. *Int J Epidemiol*, 31:1030–1037, 2002.
- [Greenland and Pearl, 2007] S. Greenland and J. Pearl. Causal diagrams. In S. Boslaugh, ed. *Encyclopedia of epidemiology*. Thousand Oaks, CA: Sage Publications, 2007: 149–156.
- [Greenland et al., 1999] S. Greenland, J. M. Robins and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14:29–46, 1999.
- [Hernán, 2005] M. A. Hernán. Hypothetical interventions to define causal effects—afterthought or prerequisite? *Am J Epidemiol*, 162:618–620, 2005.
- [Holland, 1986] P. W. Holland. Statistics and causal inference (with discussion). *J Am Stat Assoc*, 81:945–970, 1986.
- [Hume, 1978] D. Hume. *A treatise of human nature*. Oxford: Oxford University Press, 1888; 2nd ed, 1978. (Original publication, 1739.)

- [Hume, 1988] D. Hume. *An Enquiry Concerning Human Understanding*. Open Court Publishing Company, Chicago, 1988, p. 115. (Original publication 1748.)
- [Kang and Schafer, 2007] J. D. Y. Kang and J. L. Schafer. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:477-580, 2007.
- [Kuhn, 1970a] T. S. Kuhn. *Reflections on my critics*. In: Lakatos I, Musgrave A, eds. *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press, 1970a.
- [Kuhn, 1970b] T. S. Kuhn. *The structure of scientific revolutions*, 2nd ed. Chicago: University of Chicago Press, 1970b, Chapter XIII.
- [Lauritzen, 1996] S. Lauritzen. *Graphical Models*. New York: Oxford, 1996.
- [Lewis, 1973a] D. Lewis. Causation. *J Philos*, 70:556-567, 1973a. (Reprinted with postscript in: Lewis D. Philosophical papers. New York: Oxford University Press, 1986.)
- [Lewis, 1973b] D. Lewis. *Counterfactuals*. Blackwell, Oxford, 1973b.
- [Little, 1989] R. J. A. Little. On testing equality of two independent binomial proportions. *Am Statist*, 43:283-288, 1989.
- [Mackie, 1965] J. L. Mackie. Causes and conditions. *Am Philos Q*, 2:245-255, 1965. Reprinted in Sosa E, Tooley M, eds. *Causation*. New York: Oxford, 1993, 33-55.
- [Mill, 1956] J. S. Mill. *A System of Logic, Ratiocinative and Inductive*. Longmans Green, London 1956, Chapter X. (Original publication 1843)
- [Morgan and Winship, 2007] S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press, 2007.
- [Neyman, 1990] J. Neyman. On the application of probability theory to agricultural experiments: Essay on principles, Section 9. 1923; Partial translation from the original French in *Statistical Science*, 5, 465-480, 1990.
- [Pearl, 2009] J. Pearl. *Causality*, 2nd ed. New York: Cambridge, 2009.
- [Robins, 1999a] J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, M.E. Halloran and D. Berry. eds., IMA Volume 116, pp 95-134. New York: Springer-Verlag, 1999a.
- [Robins, 1999b] J. M. Robins. Association, causation, and marginal structural models. *Synthese*, 121: 151-179, 1999b.
- [Rubin, 1923] D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5:472-480, 1990.
- [Rubin, 1991] D. B. Rubin. Practical implications of modes of statistical inference for causal effects, and the critical role of the assignment mechanism. *Biometrics*, 47:1213-1234, 1991.
- [Spirtes et al., 2001] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Cambridge, MA, MIT Press, 2001.
- [VanderWeele and Robins, 2008] T. J. VanderWeele and J. M. Robins. Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika*, 95:49-61, 2008.