Causality

Causality refers to the relationship between events where one set of events (the effects) is a direct consequence of another set of events (the causes). Causal inference is the process by which one can use data to make claims about causal relationships. Since inferring causal relationships is one of the central tasks of science, it is a topic that has been heavily debated in philosophy, statistics, and the scientific disciplines. In this article, we review the models of causation and tools for causal inference most prominent in the social sciences, including regularity approaches, associated with David Hume, and counterfactual models, associated with Jerzy Neyman, Donald Rubin, and David Lewis, among many others. One of the most notable developments in the study of causation is the increasing unification of disparate methods around a common conceptual and mathematical language that treats causality in counterfactual terms---i.e., the Neyman-Rubin model. We discuss how counterfactual models highlight the deep challenges involved in making the move from correlation to causation, particularly in the social sciences where controlled experiments are relatively rare.

Regularity Models of Causation

Until the advent of counterfactual models, causation was primarily defined in terms of observable phenomena. It was philosopher David Hume in the eighteenth century who began the modern tradition of regularity models of causation by defining causation in terms of repeated "conjunctions" of events. In *An Enquiry into Human Understanding* (1751), Hume argued that the labeling of two particular events as being causally related rested on an untestable metaphysical assumption. Consequently, Hume argued that causality could only be adequately defined in terms of empirical regularities

involving classes of events. How could we know that a flame caused heat, Hume asked? Only by calling "to mind their constant conjunction in all past instances. Without further ceremony, we call the one cause and the other effect, and infer the existence of one from that of the other." Hume argued that three empirical phenomenon were necessary for inferring causality: contiguity ("the cause and effect must be contiguous in time and space"), succession ("the cause must be prior to the effect"), and constant conjunction ("there must be a constant union betwixt the cause and effect"). Under this framework, causation was defined purely in terms of empirical criteria, rather than unobservable assumptions. In other words, Hume's definition of causation and his mode of inference were one and the same.

John Stewart Mill, who shared the regularity view of causation with David Hume, elaborated basic tools for causal inference that were highly influential in the social sciences. For Mill, the goal of science was the discovery of regular empirical laws. To that end, Mill proposed in his 1843 *A System of Logic*, a series of rules or "canons" for inductive inference. These rules entailed a series of research designs that examined whether there existed covariation between a hypothesized cause and its effect, time precedence of the cause, and no plausible alternative explanation of the effect under study. Mill argued that these research designs were only effective when combined with a manipulation in an experiment. Recognizing that manipulation was unrealistic in many areas of the social sciences, Mill expressed skepticism about possibility of causal inference for questions not amenable to experiments.

The mostly widely used of Mill's canons, the "Direct Method of Difference", entailed the comparison of two units identical in all respects except for some manipulable

treatment. The method of difference involves creating a counterfactual control unit for a treated unit under the assumption that the units are exactly alike prior to treatment, an early example of counterfactual reasoning applied to causal inference. Mill stated the method as follows:

> If an instance in which the phenomenon... occurs and an instance in which it does not... have every circumstance save one in common... [then] the circumstance [in] which the two instances differ is the... cause or a necessary part of the cause (III, sec. 8).

The weakness of this research design is that in practice, particularly in the social sciences, it is very difficult to eliminate all heterogeneity in the units under study. Even in the most controlled environments, two units will rarely be the same on all background conditions. Consequently, inferences made under this method require strong assumptions.

Mill's and related methods have been criticized on a variety of grounds. His cannons and related designs assume that the relationship between cause and effect is *unique* and deterministic. These conditions allow neither for more than one cause of an effect nor for interaction among causes. The assumption that causal relationships are deterministic or perfectly regular precludes the possibility of measurement error. If outcomes are measured with error, as they often are in the social sciences, then methods predicated on detecting constant conjunctions will fail. Furthermore, the causal relationships typically studied in the social and biological sciences are rarely, if ever, unique. Causes in these fields are more likely to have highly contingent effects, making regular causal relationships very rare.

Counterfactual Models of Causation

Regularity models of causation have largely been abandoned in favor of counterfactual models. Rather than defining causality purely in reference to observable events, counterfactual models define causation in terms of a comparison of observable and unobservable events. Linguistically, counterfactual statements are most naturally expressed using subjunctive conditional statements such as "if India had not been democratic, periodic famines would have continued". Thus, the counterfactual approach to causality begins with the idea that some of the information required for inferring causal relationships is and will always be unobserved, and therefore some assumptions must be made. In stark contrast to the regularity approach of Hume, the fact of counterfactual causation is fundamentally separate from the tools used to infer it. As a result, philosophers like David Lewis (1973) could write about the meaning of causality with little discussion of how it might be inferred. It was statisticians, beginning with Jerzy Neyman in 1923 and continued most prominently by Donald Rubin, who began to clarify the conditions under which causal inferences were possible if causation was fundamentally a "missing data problem".

Counterfactual Models within Philosophy

Within philosophy, counterfactual models of causation were largely absent until the 1970's due to W.V. Quine's dismissal of the approach in his *Methods of Logic* (1950) when he pointed out that counterfactual statements could be nonsensical. He illustrated this point by his famous comparison of the conditional statements "If Bizet and Verdi had been compatriots, Bizet would have been Italian" and "If Bizet and Verdi had been compatriots, Verdi would have been French." For Quine, the incoherence of the two

statements implied that subjective conditionals lacked clear and objective truth conditions. Quine's suspicion of conditional statements was also rooted in his skepticism of evaluating the plausibility of counterfactual "feigned worlds", as he explained in *Word and Object* (1960):

> The subjunctive conditional depends, like indirect quotation and more so, on a dramatic projection: we feign belief in the antecedent and see how convincing we then find the consequent. What traits of the real world to suppose preserved in the feigned world of the contrary-to-fact antecedent can only be guessed from a sympathetic sense of the fabulist's likely purpose in spinning his fable (pg. 222).

Perhaps because of this view of counterfactuals, Quine had a dim view of the concept of causality. He argued that as science advanced, vague notions of causal relationships would disappear and be replaced by Humean "concomitances"---i.e., regularities.

In philosophy, David Lewis popularized the counterfactual approach to causality fifty years after it first appeared in statistics with Jerzy Neyman's 1923 paper on agricultural experiments. For Lewis, Quine's examples only revealed problems with *vague* counterfactuals, not counterfactuals in general. A cause, according to Lewis in his 1973 article "Causation", was "something that makes a difference, and the difference it makes must be a difference from what would have happened without it". More specifically, he defined causality in terms of "possible" (counterfactual) worlds. As a

primitive, he postulated that one can order possible worlds with respect to their closeness with the actual world. Counterfactual statements can be defined as followed:

> If A were the case, C would be the case" is true in the actual world if and only if (i) there are no possible A-worlds; or (ii) some A-world where C holds is closer to the actual world than is any A-world where C does not hold.

More intuitively, causal inferences arise by comparing the actual world to the closest possible world. If C occurs both in the actual and the closest possible world without A, according to Lewis, then A is not the cause of C. If, on the other hand, C does not occur in the closest possible world without A, then A is a cause of C.

Lewis's theory was concerned with ontology, not epistemology. As a result, one might argue that his work has limited use to empirical research since he provided little practical guidance on how one could conjure closest possible worlds to use as comparison cases. Without additional assumptions, Lewis's model suggests that causal inference is a fruitless endeavor given our inability to observe non-existent counterfactual worlds.

Statistical Models of Causation

Fortunately, statisticians beginning with Jerzy Neyman in 1923, elaborated a model of causation that allowed one to treat causation in counterfactual terms and provided guidance on how empirical researchers could create observable counterfactuals. Say we are interested in inferring the effect of some cause T on a parameter $\bar{Y}$ of the distribution of outcome Y in population A relative to treatment C (control). Population A is composed of a finite number of units and $\bar{Y}_{A,T}$ is simply a summary of the distribution

of that population when exposed to T, such as the mean. If treatment C (control) were to be applied to population A, then we would observe $\bar{Y}_{A,C}$. To use Lewis's terminology, in the actual world, we observe $\bar{Y}_{A,T}$ and in the counterfactual world, we would observe $\bar{Y}_{A,C}$. The causal effect of T relative to C for population A is a measure of the difference between $\bar{Y}_{A,T}$ and $\bar{Y}_{A,C}$, such as $\bar{Y}_{A,T} - \bar{Y}_{A,C}$. Of course, we can only observe the parameter that summarizes the actual world and not the counterfactual world.

The key insight of statistical models of causation is that under special circumstances we can use another population, B, that was exposed to control, to act as the closest possible world of A. If we believe that $\bar{Y}_{A,C} = \bar{Y}_{B,C}$, then we no longer need to rely on a unobserved counterfactual world to make causal inferences, we can simply look at the difference between the observed $\bar{Y}_{A,T}$ and $\bar{Y}_{B,C}$. In most cases $\bar{Y}_{A,C} \neq \bar{Y}_{B,C}$, however, so any inferences made by comparing the two populations will be *confounded*. What are the special circumstances that allow us to construct a suitable counterfactual population and make unconfounded inferences? As discussed below, the most reliable method is through randomization of treatment assignment, but counterfactual inferences with observational data are possible---albeit more hazardous---as well. In either case, causes are defined in reference to some real or imagined *intervention*, which makes the counterfactuals well defined.

The Neyman-Rubin Model

The counterfactual model of causation in statistics originated with Neyman's 1923 model which is non-parametric for a finite number of treatments where each unit has a potential outcome for each possible treatment condition. In the simplest case with two treatment conditions, each unit has two potential outcomes, one if the unit is treated

and the other if untreated. In this case, a causal effect is defined as the difference between the two potential outcomes, but only one of the two potential outcomes is observed. In the 1970s, Donald Rubin developed the model into a general framework for causal inference with implications for observational research. Paul Holland in 1986 wrote an influential review article that highlighted some of the philosophical implications of the framework. Consequently, instead of the "Neyman-Rubin model", the model is often simply called the Rubin causal model or sometimes the Neyman-Rubin-Holland model or the Neyman-Holland-Rubin model.

The Neyman-Rubin model is more than just the math of the original Neyman model. Unlike Neyman's original formulation, it does not rely upon an urn model motivation for the observed potential outcomes, but rather the random assignment of treatment. For observational studies, one relies on the assumption that the assignment of treatment can be treated as-if it were random. In either case, the mechanism by which treatment is assigned is of central importance. The realization that the primacy of the assignment mechanism holds true for observational data no less than for experimental, is due to Donald Rubin. This insight has been turned into a motto: "no causation without manipulation".

Let Y_{iT} denote the potential outcome for unit i if the unit receives treatment, and let Y_{iC} denote the potential outcome for unit i in the control regime. The treatment effect for observation i is defined by $\tau_i = Y_{iT} - Y_{iC}$. Causal inference is a missing data problem because Y_{iT} and Y_{iC} are never both observed. This remains true regardless of the methodology used to make inferential progress—regardless of whether we use quantitative or qualitative methods of inference. The fact that we cannot observe both

potential outcomes at the same time is commonly referred to as the "fundamental problem of causal inference".

Let T_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise. The observed outcome for observation i is then:

$$Y_i = T_i Y_{iT} + (1 - T_i) Y_{iC}$$

The average causal effect τ is the difference between the expected values $E(Y_T)$ and $E(Y_C)$. We only observe the conditional expectations $E(Y_T \mid T = 1)$ and $E(Y_C \mid T = 0)$, not the unconditional expectations required for obtaining τ. Until we assume that $E(Y_T \mid T = 1) = E(Y_T)$ and $E(Y_C \mid T = 0) = E(Y_C)$, we cannot calculate the average treatment effect. Note that the estimand of interest, such as the average treatment effect, is conceptually distinct from the estimators used to infer it from data, such as difference-in-means, linear regression, or other techniques.

Experiments

To estimate the average treatment effect, we require the assumption of *independence*. The singular virtue of experiments is that physical randomization of an intervention ensures independence between treatment status and potential outcomes. R.A. Fisher, in the 1920s and 1930s, first emphasized the importance of random assignment for eliminating bias, calling randomization of treatment the "reasoned basis for inference". From a Lewsian perspective, the control group in an experiment functions as an observable "possible world". With the independence assumption, the average treatment effect can be estimated from observables using the following expression:

$$\tau = E(Y_{iT} \mid T = 1) - E(Y_{iC} \mid T = 0) = E(Y_{iT}) - E(Y_{iC})$$

Under randomization, the assumption that T_i is independent of Y_{iT} and Y_{iC} is plausible, making the treatment and control groups exchangeable in expectation.

One of the assumptions which randomization by itself does not justify is that the response of one unit should be unaffected by the particular assignment of treatments to the other units. This "no interference between units" is often called the Stable Unit Treatment Value Assumption (SUTVA). SUTVA implies that the potential outcomes for a given unit do not vary with the treatments assigned to any other unit, and that there are not different versions of treatment.

Observational Data

In observational data, stronger assumptions are usually required to estimate causal effects. In observational studies, the causal quantity of interest is often the "average treatment effect on the treated" or ATT, which is the average effect conditional on being in the treatment regime. The parameter of interest is:

$$\tau \mid (T = 1) = \mathrm{E}(Y_{iT} \mid T = 1) - \mathrm{E}(Y_{iC} \mid T = 1)$$

Since the counterfactual control units, $\mathrm{E}(Y_{iC} \mid T = 1)$, are not observed, a control group must be constructed. The two assumptions required to construct a valid control group are conditional independence of the potential outcomes and treatment assignment and overlap, or:

1. $Y_{iT}, Y_{iC} \perp T \mid X$
2. $0 < \Pr(T = 1 \mid X) < 1$

When these two conditions hold, we can say that treatment assignment is *strongly ignorable*. Once a control group is constructed that enables us to satisfy these two conditions, the average treatment effect on the treated can be estimated as:

$$\tau \mid (T = 1) = \mathrm{E}[\mathrm{E}(Y_{iT} \mid T = 1) - \mathrm{E}(Y_{iC} \mid T = 0)] \mid T = 1$$

It is important to note that the outer expectation is taken over the distribution of $X \mid (T = 1)$, which is the distribution of covariates among the treated units.

Note that the ATT estimator is changing how individual observations are weighted, and that observations which are outside of common support receive zero weights. That is, if some covariate values are only observed for control observations, those observations will be irrelevant for estimating ATT and are effectively dropped. Therefore, the overlap assumption for ATT only requires that the support of X for treatment observations be a subset of the support of X for control observations. More generally, one would also want to drop treatment observations if they have covariate values which do not overlap with control observations. In such cases, it is unclear exactly what estimand one is estimating because it is no longer ATT as some treatment observations have been dropped along with some control observations.

The key assumption being made here is strong ignorability. Even thinking about this assumption presupposes some rigor in the research design. For example, is it clear what is pre- and what is post-treatment? If not, one is unable even to form the relevant questions, the most useful of which may be the one suggested by H.F. Dorn in 1953 who proposed that the designer of every observational study should ask "[h]ow would the study be conducted if it were possible to do it by controlled experimentation?" This clear question also appears in Cochran's 1965 Royal Statistical Society discussion paper on the planning of observational studies of human populations. Dorn's question has become one which researchers in the tradition of the Neyman-Rubin model ask themselves and their students. The question forces the researcher to focus on a clear manipulation and then on

the selection problem at hand. Only then can one even begin to think clearly about how plausible the strong ignorability assumption may or may not be.

Structural Equation Modeling

Another prominent approach to causal inference using counterfactuals is structural equation modeling, a method most associated with the work of Judea Pearl. Structural equation modeling is an old enterprise that has a rich history, including foundational work on causality in systems of structural equations by geneticist Sewall Wright (1921), economist Trygve Haavelmo (1943) and political scientist Herbert Simon (1953). Modern advocates of structural equation modeling argue that the probability calculus approach to causal modeling used by researchers in the Neyman-Rubin tradition is too narrow in that it does not explicitly take into account knowledge about the mechanisms linking background, independent, and dependent variables. Rather than modeling causality in relation to experiments, structural equation modelers prefer to write out a more elaborate causal model of the relationships under investigation through a system of structural functions. A system of such functions are said to be structural if they are assumed to be invariant to possible changes in the form of the other functions. Under this framework, the effects of treatments are understood as interventions in a pre-specified structural model.

For structural equation modelers, hypothetical interventions should be explicitly and formally related to the causal mechanisms under study. In Pearl's version of structural equation modeling, for example, the mathematical operator "$do(x)$" is used to represent physical interventions in a set of equations that deletes certain functions from the model, replaces them by a constant, and preserves the rest of the model. The

counterfactual conditional of "if X had been x" is interpreted as an instruction to modify the original model so that some causal variable X is set to x by some intervention, experimental or otherwise. This operator is accompanied by a set of rules called "do calculus" that helps a researcher judge whether or not sufficient information exists to identify the effect of the intervention of interest. Rather than identifying one all-encompassing assumption---strong ignorability---as in the Neyman-Rubin approach, Pearl proposes that researchers adopt a series of local assumptions about how an intervention interacts with a pre-specified structural model to identify causal quantities. Despite the rather substantial conceptual differences between these two approaches, however, they are mutually compatible. This compatibility arises from their shared reliance on counterfactual understandings of causality.

Causal Mechanisms

The Neyman-Rubin counterfactual approach is primarily concerned with defining *what* the effect of a cause is, not explaining *how* causes affect outcomes. The apparatus of most statistical models of causation have no formal role for social theory, explanation, or causal mechanisms. Given social scientists' interest in these issues, a common critique of the Neyman-Rubin model and its cousins are that they are too narrow for social sciences. Advocates of the statistical approach have countered that counterfactual models of causation can be augmented to take into account causal mechanisms.

While experiments have the virtue of credibly identifying the causal effect of an intervention, they are sometimes criticized as "black boxes". To understand the pathways by which interventions affect the outcome, social scientists have relied on a method known as "mediation analysis", which models the relationship between a treatment, a

potentially post-treatment variable, and the outcome ultimately of interest. An important distinction in this literature is whether or not the "mediator" is treated as post-treatment or not. If the mediator is not affected by treatment, the effect of interest is how the manipulable mediator affects or moderates the outcome when the main treatment variable is fixed, known as the "controlled direct effect".

This controlled direct effect is not always the effect of interest, however, since mediation analysis is often intended to shed light on the role of mechanisms, which in this framework, can be defined as a process that can transmit, at least partially, the effect of a treatment on an outcome. An important distinction between a manipulation (a "treatment") and a mechanism is that the former involves an external intervention, while the latter does not. The goal of this type of mediation analysis is to estimate what fraction of a causal effect is "indirect", i.e. due to the treatment changing the level of the mediator and consequently the outcome, and what fraction is "direct", i.e. due to the treatment affecting the outcome through other pathways. Expressed in counterfactual language, an "uncontrolled" indirect effect is a comparison between the outcome when the mediator is set at the value realized in the treatment condition and the outcome when the mediator is set to the value that would be observed under the control condition, while holding treatment status constant.

Uncontrolled mediation effects are often of great interest, but unfortunately, even with a randomized intervention, their identification rest on strong assumptions. In mediation analyses, the level of the mediator is generally assumed to be independent of the counterfactual outcomes conditional on treatment assignment, i.e. the mediator is assigned "as if" random. Given that an uncontrolled mediator variable, by definition, is

not randomly assigned, this assumption is strong indeed. While the identification assumptions may be warranted in special circumstances, the main lesson of the statistical literature is that the quantitative study of causal mechanisms is an enterprise fraught with difficulties, even in the context of randomized experiments.

Qualitative Evidence and Theory Falsification

While the quantitative study of causality is well-developed and increasingly unified under counterfactual models, many social scientist supplement statistical methods with qualitative reasoning to aid causal inference. Sometimes called "causal process observations", qualitative evidence can be an important source of leverage for both the design of causal analyses and the interpretation of their findings. Within the social sciences, for example, most evidence for mechanisms is qualitative, not quantitative. Qualitative researchers argue that by direct observation of causal processes, a researcher can discern potentially important mechanisms that may have escaped notice. Insight derived from observations that are poorly suited for rectangular datasets may then lead to more formal investigations using experimental and observational quantitative methods. The health sciences, for example, are replete with examples of qualitative observation paving the way for groundbreaking experiments.

Given that most questions in the social sciences are studied using observational research designs, another role for qualitative insight is the justification of the conditional independence assumption. Although many and perhaps most observational studies pay inadequate attention to justifying the adequacy of their designs, careful observational research must identify important confounders and uncover fortuitous "natural" experiments for making well grounded inferences. Qualitative evidence can be used to

identify appropriate confounders to adjust for, as well as justify any claim that treatment was allocated "as if" random.

For many questions in the social sciences, however, a research design guaranteeing the validity of causal inferences is difficult to obtain. When this is the case, researchers can attempt to defend hypothesized causal relationships by seeking data that subjects their theory to repeated falsification. Karl Popper famously argued that the degree to which we have confidence in a hypothesis is not necessarily a function of the number of tests it has withstood, but rather the severity of the tests to which the hypothesis has been subjected. A test of a hypothesis with a design susceptible to hidden bias is not particularly severe or determinative. If the implication is tested in many contexts, however, with different designs that have distinct sources of bias, and the hypothesis is still not rejected, then one may have more confidence that the causal relationship is genuine. Note that repeatedly testing a hypothesis with research designs suffering from similar types of bias does not constitute a severe test, since each repetition will merely replicate the biases of the original design. In cases where randomized experiments are infeasible or credible natural experiments are unavailable, the inferential difficulties facing researchers are large. In such circumstances, only creative and severe falsification tests can make the move from correlation to causation convincing.

F. Daniel Hidalgo, UC Berkeley

Jasjeet S. Sekhon, UC Berkeley

Cross-references: endogeneity; experiments, field; experiments, quasi (natural); matching; models, statistical; quantitative methods, basic assumptions; statistics

Further Readings

Brady, H. E. (2008). Causation and Explanation in Social Science. In J. Box-Steffensmeier, H. Brady, & D. Collier (Eds.). *The Oxford Handbook of Political Methodology (*217-70). New York: Oxford University Press.

Cochran, W. G., & Chambers, S. P. (1965). The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society. Series A (General)*, *128*(2), 234-266.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh, London: Oliver and Boyd.

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and Collapsibility in Causal Inference. *Statistical Science*, *14*(1), 29-46.

Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, *11*(1), 1-12.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Hume, D. (1902). *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*. (Lewis, Amherst Selby-Bigge, Ed.) (2nd ed.). Oxford: Clarenden Press.

Lewis, D. K. (2001). *Counterfactuals*. Malden, Mass: Wiley-Blackwell.

Mill, J. S. (1884). *A System of Logic, Ratiocinative and Inductive*. London: Longmans, Green, and Co.

Pearl, J. (2000). *Causality*. New York: Cambridge University Press.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Routledge.

Quine, W. V. O. (1974). *Methods of Logic*. London: Taylor & Francis.

Robins, J. M., & Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, *3*(2), 143-155.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*(5), 688-701.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Splawa-Neyman, J. (1990) [1923]. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (D. M. Dabrowska & T. P. Speed, Trans.) *Statistical Science*, *5*(4), 465-472.