CHAPTER ELEVEN

Reflections, Elaborations, and Discussions with Readers

As X-rays are to the surgeon,
graphs are for causation.
The author

In this chapter, I reflect back on the material covered in Chapters 1 to 10, discuss issues that require further elaboration, introduce new results obtained in the past eight years, and answer questions of general interest posed to me by readers of the first edition. These range from clarification of specific passages in the text, to conceptual and philosophical issues concerning the controversial status of causation, how it is taught in classrooms and how it is treated in textbooks and research articles.

The discussions follow roughly the order in which these issues are presented in the book, with section numbers indicating the corresponding chapters.

11.1 CAUSAL, STATISTICAL, AND GRAPHICAL VOCABULARY

11.1.1 Is the Causal–Statistical Dichotomy Necessary?

***Question to Author* (from many readers)**

Chapter 1 (Section 1.5) insists on a sharp distinction between statistical and causal concepts; the former are definable in terms of a joint distribution function (of observed variables), the latter are not. Considering that many concepts which the book classifies as "causal" (e.g., "randomization," "confounding," and "instrumental variables") are commonly discussed in the statistical literature, is this distinction crisp? Is it necessary? Is it useful?

Author Answer

The distinction is crisp,[1] necessary, and useful, and, as I tell audiences in all my lectures: "If you get nothing out of this lecture except the importance of keeping statistical and causal concepts apart, I would consider it a success." Here, I would dare go even further:

[1] The basic distinction has been given a variety of other nomenclatures, e.g., descriptive vs. etiological, associational vs. causal, empirical vs. theoretical, observational vs. experimental, and many others. I am not satisfied with any of these surrogates, partly because they were not as crisply defined, partly because their boundaries got blurred through the years, and partly because the concatenation "nonstatistical" triggers openness to new perspectives.

"If I am remembered for no other contribution except for insisting on the causal–statistical distinction, I would consider my scientific work worthwhile."

The distinction is embarrassingly crisp and simple, because it is based on the fundamental distinction between statics and kinematics. Standard statistical analysis, typified by regression, estimation, and hypothesis-testing techniques, aims to assess parameters of a static distribution from samples drawn of that distribution. With the help of such parameters, one can infer associations among variables, estimate the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer not only the likelihood of events under static conditions, but also the dynamics of events under *changing conditions,* for example, changes induced by treatments or external interventions, or by new policies or new experimental designs.

This distinction implies that causal and statistical concepts do not mix. There is nothing in the joint distribution of symptoms and diseases to tell us that curing the former would or would not cure the latter. More generally, there is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change – say, from observational to experimental setup – because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by extra assumptions that identify what in the distribution remains invariant when the specified modification takes place. The sum total of these extra assumptions is what we call "causal knowledge."

These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: behind every causal conclusion there must lie some causal assumption that is not discernible from the distribution function.

Take the concept of randomization – why is it not statistical? Assume we are given a bivariate density function $f(x,y)$, and we are told that one of the variables is randomized; can we tell which one it is by just examining $f(x, y)$? Of course not; therefore, following our definition, randomization is a causal, not a statistical concept. Indeed, every randomized experiment is based on external *intervention*; that is, subjects are "forced" to take one treatment or another in accordance with the experimental protocol, regardless of their natural inclination. The presence of intervention immediately qualifies the experimental setup, as well as all relationships inferred from that setup, as causal.

Note, however, that the purpose of the causal–statistical demarcation line (as stated in Section 1.4, p. 40) is not to exclude causal concepts from the province of statistical analysis but, rather, to encourage investigators to treat causal concepts distinctly, with the proper set of mathematical and inferential tools. Indeed, statisticians were the first to conceive of randomized experiments, and have used them successfully since the time of Fisher (1926). However, both the assumptions and conclusions in those studies were kept implicit, in the mind of ingenious investigators; they did not make their way into the mathematics. For example, one would be extremely hard pressed to find a statistics textbook, even at the graduate level, containing a mathematical proof that randomization indeed produces unbiased estimates of the quantities we wish estimated – i.e., efficacy of treatments or policies.

As a related example, very few statistics teachers today can write down a formula stating that "randomized experiments prove drug x_1 to be twice as effective as drug x_2."

Of course, they can write: $P(y \mid x_1)/P(y \mid x_2) = 2$ (y being the desirable outcome), but then they must keep in mind that this ratio applies to a specific randomized condition, and should not be confused with likelihood ratios prevailing in observational studies. Scientific progress requires that such distinctions be expressed mathematically.[2]

The most important contribution of causal analysis in the past two decades has been the emergence of mathematical languages in which not merely the data, but the experimental design itself can be given mathematical description. Such description is essential, in fact, if one wishes the results of one experiment to serve as premises in another, or to predict outcomes of one design from data obtained under another, or merely to decide if we are in possession of sufficient knowledge to render such cross-design predictions possible.

Is the Distinction Necessary?

Science thrives on distinctions, especially those that do not readily mix. The distinction between rational and irrational numbers, for example, is extremely important in number theory, for it spares us futile efforts to define the latter through some arithmetic operations on the former. The same can be said about the distinctions between prime, composite, algebraic, and transcendental numbers. Logicians, among them George Boole (1815–1864) and Augustus De Morgan (1806–1871), wasted half a century trying to prove syllogisms of first-order logic (e.g., all men are mortal) using the machinery of propositional logic; the distinction between the two was made crisp only at the end of the nineteenth century.

A similar situation occurred in the history of causality. Philosophers have struggled for half a century trying to reduce causality to probabilities (Section 7.5) and have gotten nowhere, except for traps such as "evidential decision theory" (Section 4.1). Epidemiologists have struggled for half a century to define "confounding" in the language of associations (Chapter 6, pp. 183, 194). Some are still struggling (see Section 11.6.4). This effort could have been avoided by appealing to first principles: If confounding were a statistical concept, we would have been able to identify confounders from features of nonexperimental data, adjust for those confounders, and obtain unbiased estimates of causal effects. This would have violated our golden rule: behind any causal conclusion there must be some causal assumption, untested in observational studies. That epidemiologists did not recognize in advance the futility of such attempts is a puzzle that can have only two explanations: they either did not take seriously the causal–statistical divide, or were afraid to classify "confounding" – a simple, intuitive concept – as "nonstatistical."

Divorcing simple concepts from the province of statistics - the most powerful formal language known to empirical scientists – can be traumatic indeed. Social scientists have been laboring for half a century to evaluate public policies using statistical analysis, anchored in regression techniques, and only recently have confessed, with great disappointment, what should have been recognized as obvious in the 1960's: "Regression analyses typically do nothing more than produce from a data set a collection of conditional means and conditional variances" (Berk 2004, p. 237). Economists have gone through a

[2] The potential-outcome approach of Neyman (1923) and Rubin (1974) does offer a notational distinction, by writing $P(Y_{x_1} = y)/P(Y_{x_2} = y) = 2$ for the former, and $P(y \mid x_1)/P(y \mid x_2) = 2$ for the latter. However, the opaqueness of this notation and the incomplete state of its semantics (see Sections 3.6.3 and 11.3.2) have prevented it from penetrating classrooms, textbooks, and laboratories.

similar trauma with the concept of exogeneity (Section 5.4.3). Even those who recognized that a strand of exogeneity (i.e., superexogeneity) is of a causal variety came back to define it in terms of distributions (Maddala 1992; Hendry 1995) – crossing the demarcation line was irresistible. And we understand why; defining concepts in term of prior and conditional distributions – the ultimate oracles of empirical knowledge – was considered a mark of scientific prudence. We know better now.

Is the Distinction Useful?

I am fairly confident that today, enlightened by failed experiments in philosophy, epidemiology, and economics, no reputable discipline would waste half a century chasing after a distribution-based definition of another causal concept, however tempted by prudence or intuition. Today, the usefulness of the demarcation line lies primarily in helping investigators trace the assumptions that are needed to support various types of scientific claims. Since every claim invoking causal concepts must rely on some judgmental premises that invoke causal vocabulary, and since causal vocabulary can only be formulated in causally distinct notation, the demarcation line provides notational tools for identifying the judgmental assumptions to which every causal claim is vulnerable.

Statistical assumptions, even untested, are testable in principle, given a sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference stands out in Bayesian analysis. Though the priors that Bayesians commonly assign to statistical parameters are untested quantities, the sensitivity to these priors tends to diminish with increasing sample size. In contrast, sensitivity to prior causal assumptions – say, that treatment does not change gender – remains high regardless of sample size.

This makes it doubly important that the notation we use for expressing causal assumptions be meaningful and unambiguous so that scientists can clearly judge the plausibility or inevitability of the assumptions articulated.

How Does One Recognize Causal Expressions in the Statistical Literature?

Those versed in the potential-outcome notation (Neyman 1923; Rubin 1974; Holland 1988) can recognize such expressions through the subscripts that are attached to counterfactual events and variables, e.g., $Y_x(u)$ or Z_{xy}. (Some authors use parenthetical expressions, e.g., $Y(x, u)$ or $Z(x, y)$.) (See Section 3.6.3 for semantics.)

Alternatively, this book also uses expressions of the form $P(Y = y \mid do(X = x))$ or $P(Y_x = y)$ to denote the probability (or frequency) that event $(Y = y)$ would occur if treatment condition $X = x$ were enforced uniformly over the population. (Clearly, $P(Y = y \mid do(X = x))$ is equivalent to $P(Y_x = y)$.) Still a third formal notation is provided by graphical models, where the arrows represent either causal influences, as in Definition 1.3.1, or functional (i.e., counterfactual) relationships, as in Figure 1.6(c).

These notational devices are extremely useful for detecting and tracing the causal premises with which every causal inference study must commence. Any causal premise that is cast in standard probability expressions, void of graphs, counterfactual subscripts, or $do(*)$ operators, can safely be discarded as inadequate. Consequently, any article describing an empirical investigation that does not commence with expressions involving graphs, counterfactual subscripts, or $do(*)$ can safely be proclaimed as inadequately written.

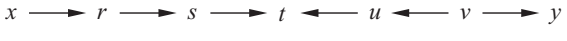

Figure 11.1 A graph containing a collider at t.

While this harsh verdict may condemn valuable articles in the empirical literature to the province of inadequacy, it can save investigators endless hours of confusion and argumentation in deciding whether causal claims from one study are relevant to another. More importantly, the verdict should encourage investigators to visibly explicate causal premises, so that they can be communicated unambiguously to other investigators and invite professional scrutiny, deliberation, and refinement.