a surrogate variable Z that is easier to control than X. For example, if we are interested in assessing the effect of cholesterol levels (X) on heart disease (Y), a reasonable experiment to conduct would be to control subjects' diet (Z), rather than exercising direct control over cholesterol levels in subjects' blood.

Formally, this problem amounts to transforming $P(y \mid \hat{x})$ into expressions in which only members of Z obtain the hat symbol. Using Theorem 3.4.1, it can be shown that the following conditions are sufficient for admitting a surrogate variable Z:

(i) X intercepts all directed paths from Z to Y; and

(ii) $P(y \mid \hat{x})$ is identifiable in $G_{\overline{Z}}$.

Indeed, if condition (i) holds, then we can write $P(y \mid \hat{x}) = P(y \mid \hat{x}, \hat{z})$, because $(Y \perp\!\!\!\perp Z \mid X)_{G_{\overline{X}\overline{Z}}}$. But $P(y \mid \hat{x}, \hat{z})$ stands for the causal effect of X on Y in a model governed by $G_{\overline{z}}$, which – by condition (ii) – is identifiable. Translated to our cholesterol example, these conditions require that there be no direct effect of diet on heart conditions and no confounding of cholesterol levels and heart disease, unless we can neutralize such confounding by additional measurements.

Figures 3.9(e) and 3.9(h) (in Section 3.5.2) illustrate models in which both conditions hold. With Figure 3.9(e), for example, we obtain this estimand

$$P(y \mid \hat{x}) = P(y \mid x, \hat{z}) = \frac{P(y, x \mid \hat{z})}{P(x \mid \hat{z})}. \tag{3.45}$$

This can be established directly by first applying Rule 3 to add $\hat{z}$,

$$P(y \mid \hat{x}) = P(y \mid \hat{x}, \hat{z}) \quad \text{because } (Y \perp\!\!\!\perp Z \mid X)_{G_{\overline{X}\overline{Z}}},$$

and then applying Rule 2 to exchange $\hat{x}$ with x:

$$P(y \mid \hat{x}, \hat{z}) = P(y \mid x, \hat{z}) \quad \text{because } (Y \perp\!\!\!\perp X \mid Z)_{G_{\underline{X}\overline{Z}}}.$$

According to (3.45), only one level of Z suffices for the identification of $P(y \mid \hat{x})$ for any values of y and x. In other words, Z need not be varied at all; it can simply be held constant by external means and, if the assumptions embodied in G are valid, the r.h.s. of (3.45) should attain the same value regardless of the (constant) level at which Z is being held. In practice, however, several levels of Z will be needed to ensure that enough samples are obtained for each desired value of X. For example, if we are interested in the difference $E(Y \mid \hat{x}) - E(Y \mid \hat{x}')$, where x and x' are two treatment levels, then we should choose two values z and z' of Z that maximize the number of samples in x and x' (respectively) and then estimate

$$E(Y \mid \hat{x}) - E(Y \mid \hat{x}') = E(Y \mid x, \hat{z}) - E(Y \mid x', \hat{z}').$$

3.5 GRAPHICAL TESTS OF IDENTIFIABILITY

Figure 3.7 shows simple diagrams in which $P(y \mid \hat{x})$ cannot be identified owing to the presence of a "bow" pattern – a confounding arc (dashed) embracing a causal link between X and Y. A confounding arc represents the existence in the diagram of a back-door

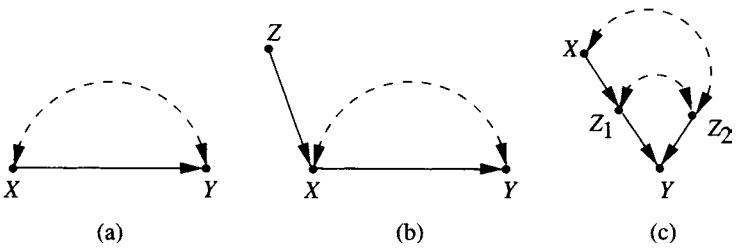

Figure 3.7 (a) A bow pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $P(y \mid \hat{x})$ even in the presence of an instrumental variable Z, as in (b). (c) A bowless graph that still prohibits the identification of $P(y \mid \hat{x})$.

path that contains only unobserved variables and has no converging arrows. For example, the path X, Z_0, B, Z_3 in Figure 3.1 can be represented as a confounding arc between X and Z_3. A bow pattern represents an equation $y = f_Y(x, u, \varepsilon_Y)$, where U is unobserved and dependent on X. Such an equation does not permit the identification of causal effects, since any portion of the observed dependence between X and Y may always be attributed to spurious dependencies mediated by U.

The presence of a bow pattern prevents the identification of $P(y \mid \hat{x})$ even when it is found in the context of a larger graph, as in Figure 3.7(b). This is in contrast to linear models, where the addition of an arc to a bow pattern can render $P(y \mid \hat{x})$ identifiable (see Chapter 5, Figure 5.9). For example, if Y is related to X via a linear relation $y = bx + u$, where U is an unobserved disturbance possibly correlated with X, then $b = \frac{\partial}{\partial x} E(Y \mid \hat{x})$ is not identifiable. However, adding an arc $Z \rightarrow X$ to the structure (i.e., finding a variable Z that is correlated with X but not with U) would facilitate the computation of $E(Y \mid \hat{x})$ via the instrumental variable formula (Bowden and Turkington 1984; see also Chapter 5):

$$b \triangleq \frac{\partial}{\partial x} E(Y \mid \hat{x}) = \frac{E(Y \mid z)}{E(X \mid z)} = \frac{r_{YZ}}{r_{XZ}}. \tag{3.46}$$

In nonparametric models, adding an instrumental variable Z to a bow pattern (Figure 3.7(b)) does not permit the identification of $P(y \mid \hat{x})$. This is a familiar problem in the analysis of clinical trials in which treatment assignment (Z) is randomized (hence, no link enters Z) but compliance is imperfect (see Chapter 8). The confounding arc between X and Y in Figure 3.7(b) represents unmeasurable factors that influence subjects' choice of treatment (X) as well as subjects' response to treatment (Y). In such trials, it is not possible to obtain an unbiased estimate of the treatment effect $P(y \mid \hat{x})$ without making additional assumptions on the nature of the interactions between compliance and response (as is done, for example, in the analyses of Imbens and Angrist 1994 and Angrist et al. 1996). Although the added arc $Z \rightarrow X$ permits us to calculate bounds on $P(y \mid \hat{x})$ (Robins 1989, sec. lg; Manski 1990; Balke and Pearl 1997) and the upper and lower bounds may even coincide for certain types of distributions $P(x, y, z)$ (Section 8.2.4), there is no way of computing $P(y \mid \hat{x})$ for *every* positive distribution $P(x, y, z)$, as required by Definition 3.2.4.

In general, the addition of arcs to a causal diagram can impede, but never assist, the identification of causal effects in nonparametric models. This is because such addition

reduces the set of d-separation conditions carried by the diagram; hence, if a causal effect derivation fails in the original diagram, it is bound to fail in the augmented diagram as well. Conversely, any causal effect derivation that succeeds in the augmented diagram (by a sequence of symbolic transformations, as in Corollary 3.4.2) would succeed in the original diagram.

Our ability to compute $P(y_1 \mid \hat{x})$ and $P(y_2 \mid \hat{x})$ for pairs (Y_1, Y_2) of singleton variables does not ensure our ability to compute joint distributions, such as $P(y_1, y_2 \mid \hat{x})$. Figure 3.7(c), for example, shows a causal diagram where both $P(z_1 \mid \hat{x})$ and $P(z_2 \mid \hat{x})$ are computable yet $P(z_1, z_2 \mid \hat{x})$ is not. Consequently, we cannot compute $P(y \mid \hat{x})$. It is interesting to note that this diagram is the smallest graph that does not contain a bow pattern and still presents an uncomputable causal effect.

Another interesting feature demonstrated by Figure 3.7(c) is that computing the effect of a joint intervention is often easier than computing the effects of its constituent singleton interventions.[6] Here, it is possible to compute $P(y \mid \hat{x}, \hat{z}_2)$ and $P(y \mid \hat{x}, \hat{z}_1)$, yet there is no way of computing $P(y \mid \hat{x})$. For example, the former can be evaluated by invoking Rule 2 in $G_{\overline{X}\underline{Z}_2}$, giving

$$\begin{aligned} P(y \mid \hat{x}, \hat{z}_2) &= \sum_{z_1} P(y \mid z_1, \hat{x}, \hat{z}_2)\, P(z_1 \mid \hat{x}, \hat{z}_2) \\ &= \sum_{z_1} P(y \mid z_1, x, z_2)\, P(z_1 \mid x). \end{aligned} \tag{3.47}$$

However, Rule 2 cannot be used to convert $P(z_1 \mid \hat{x}, z_2)$ into $P(z_1 \mid x, z_2)$ because, when conditioned on Z_2, X and Z_1 are d-connected in $G_{\underline{X}}$ (through the dashed lines). A general approach to computing the effect of joint and sequential interventions was developed by Pearl and Robins (1995) and is described in Chapter 4 (Section 4.4).

3.5.1 Identifying Models

Figure 3.8 shows simple diagrams in which the causal effect of X on Y is identifiable (where X and Y are single variables). Such models are called "identifying" because their structures communicate a sufficient number of assumptions (missing links) to permit the identification of the target quantity $P(y \mid \hat{x})$. Latent variables are not shown explicitly in these diagrams; rather, such variables are implicit in the confounding arcs (dashed). Every causal diagram with latent variables can be converted to an equivalent diagram involving measured variables interconnected by arrows and confounding arcs. This conversion corresponds to substituting out all latent variables from the structural equations of (3.2) and then constructing a new diagram by connecting any two variables X_i and X_j by (i) an arrow from X_j to X_i whenever X_j appears in the equation for X_i, and (ii) a confounding arc whenever the same ε term appears in both f_i and f_j. The result is a diagram in which all unmeasured variables are exogenous and mutually independent.

Several features should be noted from examining the diagrams in Figure 3.8.

[6] This was brought to my attention by James Robins, who has worked out many of these computations in the context of sequential treatment management (Robins 1986, p. 1423).

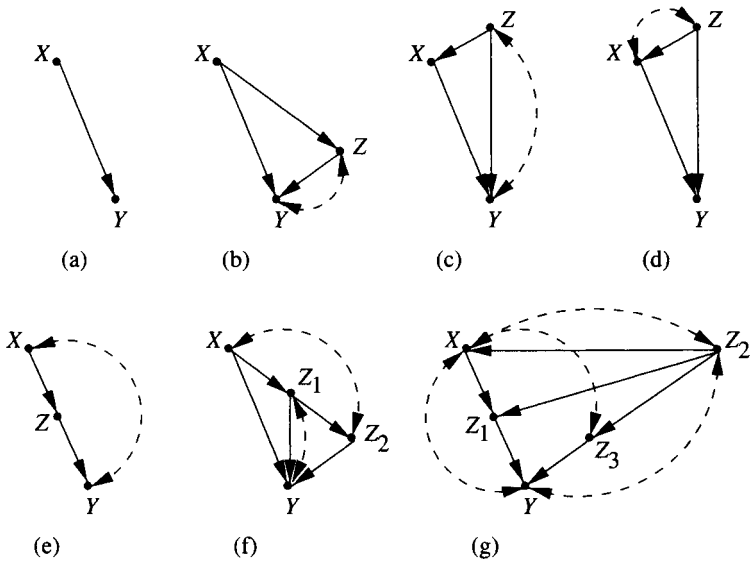

Figure 3.8 Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

1. Since the removal of any arc or arrow from a causal diagram can only assist the identifiability of causal effects, $P(y \mid \hat{x})$ will still be identified in any edge subgraph of the diagrams shown in Figure 3.8. Likewise, the introduction of mediating observed variables onto any edge in a causal graph can assist, but never impede, the identifiability of any causal effect. Therefore, $P(y \mid \hat{x})$ will still be identified from any graph obtained by adding mediating nodes to the diagrams shown in Figure 3.8.
2. The diagrams in Figure 3.8 are maximal in the sense that the introduction of any additional arc or arrow onto an existing pair of nodes would render $P(y \mid \hat{x})$ no longer identifiable. Note the conformity with Theorem 3.6.1, p. 105.
3. Although most of the diagrams in Figure 3.8 contain bow patterns, none of these patterns emanates from X (as is the case in Figures 3.9(a) and (b) to follow). In general, a necessary condition for the identifiability of $P(y \mid \hat{x})$ is the absence of a confounding arc between X and any child of X that is an ancestor of Y.
4. Diagrams (a) and (b) in Figure 3.8 contain no back-door paths between X and Y and thus represent experimental designs in which there is no confounding bias between the treatment (X) and the response (Y); hence, $P(y \mid \hat{x}) = P(y \mid x)$. Likewise, diagrams (c) and (d) in Figure 3.8 represent designs in which observed covariates Z block every back-door path between X and Y (i.e., X is "conditionally ignorable" given Z, in the lexicon of Rosenbaum and Rubin 1983); hence, $P(y \mid \hat{x})$ is obtained by standard adjustment for Z (as in (3.19)):
$$P(y \mid \hat{x}) = \sum_z P(y \mid x, z)\, P(z).$$
5. For each of the diagrams in Figure 3.8, we readily obtain a formula for $P(y \mid \hat{x})$ by using symbolic derivations patterned after those in Section 3.4.3. The derivation

is often guided by the graph topology. For example, diagram (f) in Figure 3.8 dictates the following derivation. Writing

$$P(y \mid \hat{x}) = \sum_{z_1, z_2} P(y \mid z_1, z_2, \hat{x})\, P(z_1, z_2 \mid \hat{x}),$$

we see that the subgraph containing $\{X, Z_1, Z_2\}$ is identical in structure to that of diagram (e), with (Z_1, Z_2) replacing (Z, Y), respectively. Thus, $P(z_1, z_2 \mid \hat{x})$ can be obtained from (3.43). Likewise, the term $P(y \mid z_1, z_2, \hat{x})$ can be reduced to $P(y \mid z_1, z_2, x)$ by Rule 2, since $(Y \perp\!\!\!\perp X \mid Z_1, Z_2)_{G_{\underline{X}}}$. We therefore have

$$P(y \mid \hat{x}) = \sum_{z_1, z_2} P(y \mid z_1, z_2, x)\, P(z_1 \mid x) \sum_{x'} P(z_2 \mid z_1, x')\, P(x'). \tag{3.48}$$

Applying a similar derivation to diagram (g) of Figure 3.8 yields

$$P(y \mid \hat{x}) = \sum_{z_1} \sum_{z_2} \sum_{x'} P(y \mid z_1, z_2, x')\, P(x' \mid z_2) \times P(z_1 \mid z_2, x)\, P(z_2). \tag{3.49}$$

Note that the variable Z_3 does not appear in (3.49), which means that Z_3 need not be measured if all one wants to learn is the causal effect of X on Y.

6. In diagrams (e), (f), and (g) of Figure 3.8, the identifiability of $P(y \mid \hat{x})$ is rendered feasible through observed covariates Z that are affected by the treatment X (since members of Z are descendants of X). This stands contrary to the warning – repeated in most of the literature on statistical experimentation – to refrain from adjusting for concomitant observations that are affected by the treatment (Cox 1958; Rosenbaum 1984; Pratt and Schlaifer 1988; Wainer 1989). It is commonly believed that a concomitant Z that is affected by the treatment must be excluded from the analysis of the total effect of the treatment (Pratt and Schlaifer 1988). The reason given for the exclusion is that the calculation of total effects amounts to integrating out Z, which is functionally equivalent to omitting Z to begin with. Diagrams (e), (f), and (g) show cases where the total effects of X are indeed the target of investigation and, even so, the measurement of concomitants that are affected by X (e.g., Z or Z_1) is still necessary. However, the adjustment needed for such concomitants is nonstandard, involving two or more stages of the standard adjustment of (3.19) (see (3.28), (3.48), and (3.49)).
7. In diagrams (b), (c), and (f) of Figure 3.8, Y has a parent whose effect on Y is not identifiable; even so, the effect of X on Y is identifiable. This demonstrates that local identifiability is not a necessary condition for global identifiability. In other words, to identify the effect of X on Y we need not insist on identifying each and every link along the paths from X to Y.

3.5.2 Nonidentifying Models

Figure 3.9 presents typical diagrams in which the total effect of X on Y, $P(y \mid \hat{x})$, is not identifiable. Noteworthy features of these diagrams are as follows.

1. All graphs in Figure 3.9 contain unblockable back-door paths between X and Y, that is, paths ending with arrows pointing to X that cannot be blocked by observed nondescendants of X. The presence of such a path in a graph is, indeed,

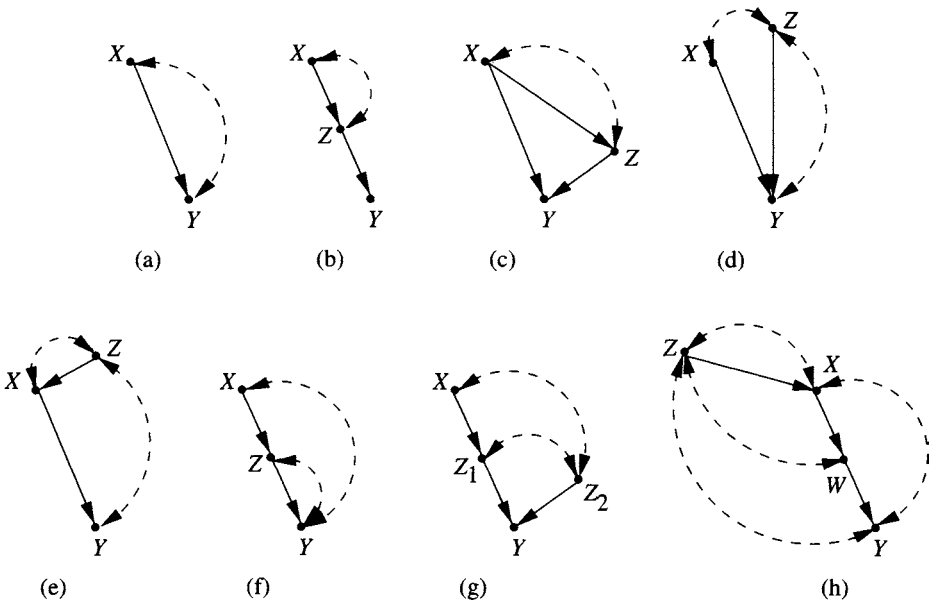

Figure 3.9 Typical models in which $P(y \mid \hat{x})$ is not identifiable.

a necessary test for nonidentifiability (see Theorem 3.3.2). That it is not a sufficient test is demonstrated by Figure 3.8(e), in which the back-door path (dashed) is unblockable and yet $P(y \mid \hat{x})$ is identifiable.

2. A sufficient condition for the nonidentifiability of $P(y \mid \hat{x})$ is the existence of a confounding arc between X and any of its children on a path from X to Y, as shown in Figures 3.9(b) and (c). A stronger sufficient condition is that the graph contain any of the patterns shown in Figure 3.9 as an edge subgraph.
3. Graph (g) in Figure 3.9 (same as Figure 3.7(c)) demonstrates that local identifiability is not sufficient for global identifiability. For example, we can identify $P(z_1 \mid \hat{x})$, $P(z_2 \mid \hat{x})$, $P(y \mid \hat{z}_1)$, and $P(y \mid \hat{z}_2)$ but not $P(y \mid \hat{x})$. This is one of the main differences between nonparametric and linear models; in the latter, all causal effects can be determined from the structural coefficients and each coefficient represents the causal effect of one variable on its immediate successor.

3.6 DISCUSSION

3.6.1 Qualifications and Extensions

The methods developed in this chapter facilitate the drawing of quantitative causal inferences from a combination of qualitative causal assumptions (encoded in the diagram) and nonexperimental observations. The causal assumptions in themselves cannot generally be tested in nonexperimental studies, unless they impose constraints on the observed distributions. The most common type of constraints appears in the form of conditional independencies, as communicated through the d-separation conditions in the diagrams. Another type of constraints takes the form of numerical inequalities. In Chapter 8, for example, we show that the assumptions associated with instrumental variables (Figure 3.7(b)) are

subject to falsification tests in the form of inequalities on conditional probabilities (Pearl 1995b). Still, such constraints permit the testing of merely a small fraction of the causal assumptions embodied in the diagrams; the bulk of those assumptions must be substantiated from domain knowledge as obtained from either theoretical considerations (e.g., that falling barometers do not cause rain) or related experimental studies. For example, the experimental study of Moertel et al. (1985), which refuted the hypothesis that vitamin C is effective against cancer, can be used as a substantive assumption in observational studies involving vitamin C and cancer patients; it would be represented as a missing link (between vitamin C and cancer) in the associated diagram. In summary, the primary use of the methods described in this chapter lies not in testing causal assumptions but in providing an effective language for making those assumptions precise and explicit. Assumptions can thereby be isolated for deliberation or experimentation and then (once validated) be integrated with statistical data to yield quantitative estimates of causal effects.

An important issue that will be considered only briefly in this book (see Section 8.5) is sampling variability. The mathematical derivation of causal effect estimands should be considered a first step toward supplementing these estimands with confidence intervals and significance levels, as in traditional analysis of controlled experiments. We should remark, though, that having obtained nonparametric estimands for causal effects does not imply that one should refrain from using parametric forms in the estimation phase of the study. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in (3.28) can be converted to the product $E(Y \mid \hat{x}) = r_{ZX} r_{YZ \cdot X} x$, where $r_{YZ \cdot X}$ is the standardized regression coefficient (Section 5.3.1); the estimation problem then reduces to that of estimating regression coefficients (e.g., by least squares). More sophisticated estimation techniques can be found in Rosenbaum and Rubin (1983), Robins (1989, sec. 17), and Robins et al. (1992, pp. 331–3). For example, the "propensity score" method of Rosenbaum and Rubin (1983) was found useful when the dimensionality of the adjusted covariates is high (Section 11.3.5). Robins (1999) shows that, rather than estimating individual factors in the adjustment formula of (3.19), it is often more advantageous to use $P(y \mid \hat{x}) = \sum_z \frac{P(x, y, z)}{P(x \mid z)}$, where the preintervention distribution remains unfactorized. One can then separately estimate the denominator $P(x \mid z)$, weigh individual samples by the inverse of this estimate, and treat the weighted samples as if they were drawn at random from the postintervention distribution $P(y \mid \hat{x})$. Postintervention parameters, such as $\frac{\partial}{\partial x} E(Y \mid \hat{x})$, can then be estimated by ordinary least squares. This method is especially advantageous in longitudinal studies with time-varying covariates, as in the problems discussed in Sections 3.2.3 (see (3.18)) and 4.4.3.

Several extensions of the methods proposed in this chapter are noteworthy. First, the identification analysis for atomic interventions can be generalized to complex time-varying policies in which a set X of controlled variables is made to respond in a specified way to some set Z of covariates via functional or stochastic strategies, as in Sections 3.2.3 and 4.4.3. In Chapter 4 (Section 4.4.3) it is shown that identifying the effect of such policies requires a sequence of back-door conditions in the associated diagram.

A second extension concerns the use of the intervention calculus (Theorem 3.4.1) in nonrecursive models, that is, in causal diagrams involving directed cycles or feedback loops. The basic definition of causal effects in terms of "wiping out" equations from the model (Definition 3.2.1) still carries over to nonrecursive systems (Strotz and Wold

1960; Sobel 1990), but then two issues must be addressed. First, the analysis of identification must ensure the stability of the remaining submodels (Fisher 1970). Second, the *d*-separation criterion for DAGs must be extended to cover cyclic graphs as well. The validity of *d*-separation has been established for nonrecursive linear models (Spirtes 1995) as well as for nonlinear systems involving discrete variables (Pearl and Dechter 1996). However, the computation of causal effect estimands will be harder in cyclic nonlinear systems, because symbolic reduction of $P(y \mid \hat{x})$ to hat-free expressions may require the solution of nonlinear equations. In Chapter 7 (Section 7.2.1) we demonstrate the evaluation of policies and counterfactuals in nonrecursive linear systems (see also Balke and Pearl 1995a).

A third extension concerns generalizations of intervention calculus (Theorem 3.4.1) to situations where the data available is not obtained under i.i.d. (independent and identically distributed) sampling. One can imagine, for instance, a physician who prescribes a certain treatment to patients only when the fraction of survivors among previous patients drops below some threshold. In such cases, it is required to estimate the causal effect $P(y \mid \hat{x})$ from nonindependent samples. Vladimir Vovk (1996) gave conditions under which the rules of Theorem 3.4.1 will be applicable when sampling is not i.i.d., and he went on to cast the three inference rules as a logical production system.

3.6.2 Diagrams as a Mathematical Language

The benefit of incorporating substantive background knowledge into probabilistic inference was recognized as far back as Thomas Bayes (1763) and Pierre Laplace (1814), and its crucial role in the analysis and interpretation of complex statistical studies is generally acknowledged by most modern statisticians. However, the mathematical language available for expressing background knowledge has remained in a rather pitiful state of development.

Traditionally, statisticians have approved of only one way of combining substantive knowledge with statistical data: the Bayesian method of assigning subjective priors to distributional parameters. To incorporate causal information within this framework, plain causal statements such as "Y is not affected by X" must be converted into sentences or events capable of receiving probability values (e.g., counterfactuals). For instance, to communicate the innocent assumption that mud does not cause rain, we would have to use a rather unnatural expression and say that the probability of the counterfactual event "rain if it were not muddy" is the same as the probability of "rain if it were muddy." Indeed, this is how the potential-outcome approach of Neyman and Rubin has achieved statistical legitimacy: causal judgments are expressed as constraints on probability functions involving counterfactual variables (see Section 3.6.3).

Causal diagrams offer an alternative language for combining data with causal information. This language simplifies the Bayesian route by accepting plain causal statements as its basic primitives. Such statements, which merely indicate whether a causal connection between two variables of interest exists, are commonly used in ordinary discourse and provide a natural way for scientists to communicate experience and organize knowledge.[7]

[7] Remarkably, many readers of this chapter (including two referees of this book) classified the methods presented here as belonging to the "Bayesian camp" and as depending on a "good prior." This

It can be anticipated, therefore, that the language of causal graphs will find applications in problems requiring substantial domain knowledge.

The language is not new. The use of diagrams and structural equation models to convey causal information has been quite popular in the social sciences and econometrics. Statisticians, however, have generally found these models suspect, perhaps because social scientists and econometricians have failed to provide an unambiguous definition of the empirical content of their models – that is, to specify the experimental conditions, however hypothetical, whose outcomes would be constrained by a given structural equation. (Chapter 5 discusses the history of structural equations in the social sciences and economics.) As a result, even such basic notions as "structural coefficients" or "missing links" become the object of serious controversy (Freedman 1987; Goldberger 1992) and misinterpretations (Whittaker 1990, p. 302; Wermuth 1992; Cox and Wermuth 1993).

To a large extent, this history of controversy and miscommunication stems from the absence of an adequate mathematical notation for defining basic notions of causal modeling. For example, standard probabilistic notation cannot express the empirical content of the coefficient b in the structural equation $y = bx + \varepsilon_Y$, even if one is prepared to assume that ε_Y (an unobserved quantity) is uncorrelated with X.[8] Nor can any probabilistic meaning be attached to the analyst's excluding from the equation variables that do not "directly affect" Y.[9]

The notation developed in this chapter gives these (causal) notions a clear empirical interpretation, because it permits one to specify precisely what is being held constant and what is merely measured in a given experiment. (The need for this distinction was recognized by many researchers, most notably Pratt and Schlaifer 1988 and Cox 1992.) The meaning of b is simply $\frac{\partial}{\partial x}E(Y \mid \hat{x})$, that is, the rate of change (in x) of the expectation of Y in an experiment where X is held at x by external control. This interpretation holds regardless of whether ε_Y and X are correlated (e.g., via another equation $x = ay + \varepsilon_x$). Likewise, the analyst's decision as to which variables should be included in a given equation can be based on a hypothetical controlled experiment: A variable Z is excluded from the equation for Y if (for every level of ε_Y) Z has no influence on Y when all other variables (S_{YZ}) are held constant; this implies $P(y \mid \hat{z}, \hat{s}_{YZ}) = P(y \mid \hat{s}_{YZ})$. Specifically, variables that are excluded from the equation $y = bx + \varepsilon_Y$ are not conditionally independent of Y given measurements of X but instead are *causally* irrelevant to Y given settings of X. The operational meaning of the "disturbance term" ε_Y is likewise demystified: ε_Y is defined as the difference $Y - E(Y \mid \hat{s}_Y)$. Two disturbance terms, ε_X and ε_Y, are correlated if $P(y \mid \hat{x}, \hat{s}_{XY}) \neq P(y \mid x, \hat{s}_{XY})$, and so on (see Chapter 5, Section 5.4, for further elaboration).

The distinctions provided by the hat notation clarify the empirical basis of structural equations and should make causal models more acceptable to empirical researchers.

---

classification is misleading. The method does depend on subjective assumptions (e.g., mud does not cause rain), but such assumptions are causal, not statistical, and cannot be expressed as prior probabilities on parameters of joint distributions.

8 Voluminous literature on the subject of "exogeneity" (e.g., Richard 1980; Engle et al. 1983; Hendry 1995) has emerged from economists' struggle to give statistical interpretation to the causal assertion "X and ε_Y are uncorrelated" (Aldrich 1993; see Section 5.4.3).

9 The bitter controversy between Goldberger (1992) and Wermuth (1992) revolves around Wermuth's insistence on giving a statistical interpretation to the zero coefficients in structural equations (see Section 5.4.1).

Moreover, since most scientific knowledge is organized around the operation of "holding X fixed" rather than "conditioning on X," the notation and calculus developed in this chapter should provide an effective means for scientists to communicate substantive information and to infer its logical consequences.

3.6.3 Translation from Graphs to Potential Outcomes

This chapter uses two representations of causal information: graphs and structural equations, where the former is an abstraction of the latter. Both representations have been controversial for almost a century. On the one hand, economists and social scientists have embraced these modeling tools, but they continue to question and debate the causal content of the parameters they estimate (see Sections 5.1 and 5.4 for details); as a result, the use of structural models in policy-making contexts is often viewed with suspicion. Statisticians, by and large, reject both representations as problematic (Freedman 1987) if not meaningless (Wermuth 1992; Holland 1995), and they sometimes resort to the Neyman–Rubin potential-outcome notation when pressed to communicate causal information (Rubin 1990).[10] A detailed formal analysis of the relationships between the structural and potential-outcome approaches is offered in Chapter 7 (Section 7.4.4) and proves their mathematical equivalence – a theorem in one entails a theorem in the other. In this section we highlight the key methodological differences.

The primitive object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y(x, u)$ or $Y_x(u)$, read: "the value that Y would obtain in unit u, had X been x." This counterfactual expression has a formal interpretation in structural equations models. Consider a structural model M that contains a set of equations

$$x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n, \tag{3.50}$$

as in (3.4). Let U stand for the vector $(U_1, \ldots, U_n)$ of background variables, let X and Y be two disjoint subsets of observed variables, and let M_x be the submodel created by replacing the equations corresponding to variables in X with $X = x$, as in Definition 3.2.1. The structural interpretation of $Y(x, u)$ is given by

$$Y(x, u) \triangleq Y_{Mx}(u). \tag{3.51}$$

That is, $Y(x, u)$ is the (unique) solution of Y under the realization $U = u$ in the submodel M_x of M. Although the term *unit* in the potential-outcome literature normally stands for the identity of a specific individual in a population, a unit may also be thought of as the set of attributes that characterize that individual, the experimental conditions under study, the time of day, and so on – all of which are represented as components of the vector u in structural modeling. In fact, the only requirements on U are (i) that it represent as many background factors as needed to render the relations among endogenous variables deterministic and (ii) that the data consist of independent samples drawn from $P(u)$. The

[10] A parallel framework was developed in the econometrics literature under the rubric "switching regression" (Manski 1995, p. 38), which Heckman (1996) attributed to Roy (1951) and Quandt (1958); but lacking the formal semantics of (3.51), did not progress beyond a "framework."

identity of an individual person in an experiment is often sufficient for this purpose because it represents the anatomical and genetic makings of that individual, which are often sufficient for determining that individual's response to treatments or other programs of interest.

Equation (3.51) defines the key formal connection between the opaque English phrase "the value that Y would obtain in unit u, had X been x" and the physical processes that transfer changes in X into changes in Y. The formation of the submodel M_x explicates precisely how the hypothetical phrase "had X been x" could be realized, as well as what process must give in to make $X = x$ a reality.

Given this interpretation of $Y(x, u)$, it is instructive to contrast the methodologies of causal inference in the potential-outcome versus structural frameworks. If U is treated as a random variable, then the value of the counterfactual $Y(x, u)$ becomes a random variable as well, denoted as $Y(x)$ or Y_x. The potential-outcome analysis proceeds by imagining the observed distribution $P(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function P^* defined over both observed and counterfactual variables. Queries about causal effects (written $P(y \mid \hat{x})$ in our structural analysis) are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $P^*(Y(x) = y)$. The new hypothetical entities $Y(x)$ are treated as ordinary random variables; for example, they are assumed to obey the axioms of probability calculus, the laws of conditioning, and the axioms of conditional independence. Moreover, these hypothetical entities are assumed to be connected to observed variables via consistency constraints (Robins 1986) such as[11]

$$X = x \Longrightarrow Y(x) = Y, \tag{3.52}$$

which states that, for every u, if the actual value of X turns out to be x, then the value that Y would take on if X were x is equal to the actual value of Y. Thus, whereas the structural approach views the intervention $do(x)$ as an operation that changes the model (and the distribution) but keeps all variables the same, the potential-outcome approach views the variable Y under $do(x)$ to be a different variable, $Y(x)$, loosely connected to Y through relations such as (3.52). In Chapter 7 we show, using the structural interpretation of $Y(x, u)$, that it is indeed legitimate to treat counterfactuals as random variables in all respects and, moreover, that consistency constraints like (3.52) follow as theorems from the structural interpretation, and no other constraint need ever be considered.

To communicate substantive causal knowledge, the potential-outcome analyst must express causal assumptions as constraints on P^*, usually in the form of conditional independence assertions involving counterfactual variables. For example, to communicate the understanding that – in a randomized clinical trial with imperfect compliance (see Figure 3.7(b)) – the way subjects react (Y) to treatments (X) is statistically independent of the treatment assignment (Z), the potential-outcome analyst would write $Y(x) \perp\!\!\!\perp Z$. Likewise, to convey the understanding that the assignment is randomized and hence independent of how subjects comply with the assignment, the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp X\ (z)$.

[11] Gibbard and Harper (1976, p. 156) expressed this constraint as $A \supset [(A \mathrel{\square\!\!\rightarrow} S) \equiv S]$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest; in other cases, only bounds on the solution can be obtained. For example, if one can plausibly assume that a set Z of covariates satisfies the conditional independence

$$Y(x) \perp\!\!\!\perp X \,|\, Z \tag{3.53}$$

(an assumption that was termed "conditional ignorability" by Rosenbaum and Rubin 1983), then the causal effect $P^*(Y(x) = y)$ can readily be evaluated, using (3.52), to yield[12]

$$\begin{aligned} P^*(Y(x) = y) &= \sum_z P^*(Y(x) = y \mid z)\, P(z) \\ &= \sum_z P^*(Y(x) = y \mid x, z)\, P(z) \\ &= \sum_z P^*(Y = y \mid x, z)\, P(z) \\ &= \sum_z P(y \mid x, z)\, P(z). \end{aligned} \tag{3.54}$$

The last expression contains no counterfactual quantities (thus permitting us to drop the asterisk from P^*) and coincides precisely with the adjustment formula of (3.19), which obtains from the back-door criterion. However, the assumption of conditional ignorability (equation (3.53)) – the key to the derivation of (3.54) – is not straightforward to comprehend or ascertain. Paraphrased in experimental metaphors, this assumption reads: The way an individual with attributes Z would react to treatment $X = x$ is independent of the treatment actually received by that individual.

Section 3.6.2 explains why this approach may appeal to traditional statisticians, even though the process of eliciting judgments about counterfactual dependencies has been extremely difficult and error-prone; instead of constructing new vocabulary and new logic for causal expressions, all mathematical operations in the potential-outcome framework are conducted within the safe confines of probability calculus. The drawback lies in the requirement of using independencies among counterfactual variables to express plain causal knowledge. When counterfactual variables are not viewed as by-products of a deeper, process-based model, it is hard to ascertain whether *all* relevant counterfactual independence judgments have been articulated,[13] whether the judgments articulated are redundant, or whether those judgments are self-consistent. The elicitation of such counterfactual judgments can be systematized by using the following translation from graphs (see Section 7.1.4 for additional relationships).

Graphs encode substantive information in both the equations and the probability function $P(u)$; the former is encoded as missing arrows, the latter as missing dashed arcs.

[12] Gibbard and Harper (1976, p. 157) used the "ignorability assumption" $Y(x) \perp\!\!\!\perp X$ to derive the equality $P(Y(x) = y) = P(y \mid x)$.

[13] A typical oversight in the example of Figure 3.7(b) has been to write $Z \perp\!\!\!\perp Y(x)$ and $Z \perp\!\!\!\perp X(z)$ instead of $Z \perp\!\!\!\perp \{Y(x), X(z)\}$, as dictated by (3.56).

Each parent–child family (PA_i, X_i) in a causal diagram G corresponds to an equation in the model M of (3.50). Hence, missing arrows encode exclusion assumptions, that is, claims that adding excluded variables to an equation will not change the outcome of the hypothetical experiment described by that equation. Missing dashed arcs encode independencies among disturbance terms in two or more equations. For example, the absence of dashed arcs between a node Y and a set of nodes $\{Z_1, \ldots, Z_k\}$ implies that the corresponding background variables, U_Y and $\{U_{Z_1}, \ldots, U_{Z_k}\}$, are independent in $P(u)$.

These assumptions can be translated into the potential-outcome notation using two simple rules (Pearl 1995a, p. 704); the first interprets the missing arrows in the graph, the second, the missing dashed arcs.

1. *Exclusion restrictions:* For every variable Y having parents PA_Y and for every set of variables S disjoint of PA_Y, we have

$$Y(pa_Y) = Y(pa_Y, s). \tag{3.55}$$

2. *Independence restrictions:* If $Z_1, \ldots, Z_k$ is any set of nodes not connected to Y via dashed arcs, we have[14]

$$Y(pa_Y) \perp\!\!\!\perp \{Z_1(pa_{Z_1}), \ldots, Z_k(pa_{Z_k})\}. \tag{3.56}$$

The independence restriction translates the independence between U_Y and $\{U_{Z_1}, \ldots, U_{Z_k}\}$ into independence between the corresponding potential-outcome variables. This follows from the observation that, once we set their parents, the variables in $\{Y, Z_1, \ldots, Z_k\}$ stand in functional relationships to the U terms in their corresponding equations.

As an example, the model shown in Figure 3.5 displays the following parent sets:

$$PA_X = \{\emptyset\}, \quad PA_Z = \{X\}, \quad PA_Y = \{Z\}. \tag{3.57}$$

Consequently, the exclusion restrictions translate into:

$$Z(x) = Z(y, x), \tag{3.58}$$

$$X(y) = X(x, y) = X(z) = X, \tag{3.59}$$

$$Y(z) = Y(z, x); \tag{3.60}$$

the absence of a dashed arc between Z and $\{Y, X\}$ translates into the independence restriction

$$Z(x) \perp\!\!\!\perp \{Y(z), X\}. \tag{3.61}$$

Given a sufficient number of such restrictions on P^*, the analyst attempts to compute causal effects $P^*(Y(x) = y)$ using standard probability calculus together with the logical constraints (e.g., (3.52)) that couple counterfactual variables with their measurable counterparts. These constraints can be used as axioms, or rules of inference, in attempting to

[14] The restriction is in fact stronger, jointly applying to all instantiations of the PA variables. For example, $X \perp\!\!\!\perp Y\ (pa_Z)$ should be interpreted as $X \perp\!\!\!\perp \{Y(pa'_Z), Y(pa''_Z), Y(pa'''_Z), \ldots\}$, where $(pa'_Z), (pa''_Z), (pa'''_Z), \ldots$ are the values that the set PA_Z may take on.

transform causal effect expressions of the form $P^*(Y(x) = y)$ into expressions involving only measurable variables. When such a transformation is found, the corresponding causal effect is identifiable, since P^* then reduces to P.

The question naturally arises of whether the constraints used by potential-outcome analysts are *complete* – that is, whether they are sufficient for deriving every valid statement about causal processes, interventions, and counterfactuals. To answer this question, the validity of counterfactual statements must be defined relative to more basic mathematical objects, such as possible worlds (Section 1.4.4) or structural equations (equation (3.51)). In the standard potential-outcome framework, however, the question of completeness remains open, because $Y(x, u)$ is taken as a primitive notion and because consistency constraints such as (3.52) – although they appear plausible for the English expression "had X been x" – are not derived from a deeper mathematical object. This question of completeness is settled in Chapter 7, where a necessary and sufficient set of axioms is derived from the structural semantics given to $Y(x, u)$ by (3.51).

In assessing the historical development of structural equations and potential-outcome models, one cannot overemphasize the importance of the conceptual clarity that structural equations offer vis-à-vis the potential-outcome model. The reader may appreciate this importance by attempting to judge whether the condition of (3.61) holds in a given familiar situation. This condition reads: "the value that Z would obtain had X been x is jointly independent of both X and the value that Y would obtain had Z been x." (In the structural representation, the sentence reads: "Z shares no cause with either X or Y, except for X itself, as shown in Figure 3.5.") The thought of having to express, defend, and manage formidable counterfactual relationships of this type may explain why the enterprise of causal inference is currently viewed with such awe and despair among rank-and-file epidemiologists and statisticians – and why most economists and social scientists continue to use structural equations instead of the potential-outcome alternatives advocated in Holland (1988), Angrist et al. (1996), and Sobel (1998). On the other hand, the algebraic machinery offered by the potential-outcome notation, once a problem is properly formalized, can be quite powerful in refining assumptions, deriving probabilities of counterfactuals, and verifying whether conclusions follow from premises – as we demonstrate in Chapter 9. The translation given in (3.51)–(3.56) is the key to unifying the two camps and should help researchers combine the best features of the two approaches.

3.6.4 Relations to Robins's *G*-Estimation

Among the investigations conducted in the potential-outcome framework, the one closest in spirit to the structural analysis described in this chapter is Robins's work on "causally interpreted structured tree graphs" (Robins 1986, 1987). Robins was the first to realize the potential of Neyman's counterfactual notation $Y(x)$ as a general mathematical language for causal inference, and he used it to extend Rubin's (1978) "time-independent treatment" model to studies with direct and indirect effects and time-varying treatments, concomitants, and outcomes.

Robins considered a set $V = \{V_1, \ldots, V_M\}$ of temporally ordered discrete random variables (as in Figure 3.3) and asked under what conditions one can identify the effect of control policy $g : X = x$ on outcomes $Y \subseteq V \setminus X$, where $X = \{X_1, \ldots, X_K\} \subseteq V$ are

the temporally ordered and potentially manipulable treatment variables of interest. The causal effect of $X = x$ on Y was expressed as the probability

$$P(y \mid g = x) \triangleq P\{Y(x) = y\},$$

where the counterfactual variable $Y(x)$ stands for the value that outcome variables Y would take had the treatment variables X been x.

Robins showed that $P(y \mid g = x)$ is identified from the distribution $P(v)$ if each component X_k of X is "assigned at random, given the past," a notion explicated as follows. Let L_k be the variables occurring between X_{k-1} and X_k, with L_1 being the variables preceding X_1, Write $\overline{L}_k = (L_1, \ldots, L_k)$, $L = \overline{L}_K$, and $\overline{X}_k = (X_1, \ldots, X_k)$, and define $\overline{X}_0$, $\overline{L}_0$, $\overline{V}_0$ to be identically zero. The treatment $X_k = x_k$ is said to be *assigned at random, given the past*, if the following relation holds:

$$(Y(x) \perp\!\!\!\perp X_k \mid \overline{L}_k, \overline{X}_{k-1} = \overline{x}_{k-1}). \tag{3.62}$$

Robins further proved that, if (3.62) holds for every k, then the causal effect is given by

$$P(y \mid g = x) = \sum_{\overline{l}_k} P(y \mid \overline{l}_K, \overline{x}_K) \prod_{k=1}^{K} P(l_k \mid \overline{l}_{k-1}, \overline{x}_{k-1}), \tag{3.63}$$

an expression he called the "G-computation algorithm formula." This expression can be derived by applying condition (3.62) iteratively, as in the derivation of (3.54). If X is univariate, then (3.63) reduces to the standard adjustment formula

$$P(y \mid g = x) = \sum_{l_1} P(y \mid x, l_1)\, P(l_1),$$

paralleling (3.54). Likewise, in the special structure of Figure 3.3, (3.63) reduces to (3.18).

To place this result in the context of our analysis in this chapter, we need to focus attention on condition (3.62), which facilitated Robins's derivation of (3.63), and ask whether this formal counterfactual independency can be given a meaningful graphical interpretation. The answer will be given in Chapter 4 (Theorem 4.4.1), where we derive a graphical condition for identifying the effect of a plan, i.e., a sequential set of actions. The condition reads as follows: $P(y \mid g = x)$ is identifiable and is given by (3.63) if every action-avoiding back-door path from X_k to Y is blocked by some subset L_k of nondescendants of X_k. (By "action-avoiding" we mean a path containing no arrow entering an X variable later than X_k.) Chapter 11 (Section 11.4.2) shows by examples that this "sequential back-door" criterion is more general than that given in (3.62).

The structural analysis introduced in this chapter supports and generalizes Robins's result from a new theoretical perspective. First, on the technical front, this analysis offers systematic ways of managing models where Robins's starting assumption (3.62) is inapplicable. Examples are Figures 3.8(d)–(g).

Second, on the conceptual front, the structural framework represents a fundamental shift from the vocabulary of counterfactual independencies, to the vocabulary of

processes and mechanisms, in which human knowledge is encoded. The former requires humans to affirm esoteric relationships such as (3.62), while the latter expresses those same relationships in vivid graphical terms of missing links. Robins's pioneering research has shown that, to properly manage multistage problems with time-varying treatments, the opaque condition of "ignorability" (3.53) should be broken down to its sequential constituents. This has led to the sequential back-door criterion of Theorem 4.4.5.

Personal Remarks and Acknowledgments

The work recounted in this chapter sprang from two simple ideas that totally changed my attitude toward causality. The first idea arose in the summer of 1990, while I was working with Tom Verma on "A Theory of Inferred Causation" (Pearl and Verma 1991; see also Chapter 2). We played around with the possibility of replacing the parents–child relationship $P(x_i \mid pa_i)$ with its functional counterpart $x_i = f_i(pa_i, u_i)$ and, suddenly, everything began to fall into place: we finally had a mathematical object to which we could attribute familiar properties of physical mechanisms instead of those slippery epistemic probabilities $P(x_i \mid pa_i)$ with which we had been working so long in the study of Bayesian networks. Danny Geiger, who was writing his dissertation at that time, asked with astonishment: "Deterministic equations? Truly deterministic?" Although we knew that deterministic structural equations have a long history in econometrics, we viewed this representation as a relic of the past. For us at UCLA in the early 1990s, the idea of putting the semantics of Bayesian networks on a deterministic foundation seemed a heresy of the worst kind.

The second simple idea came from Peter Spirtes's lecture at the International Congress of Philosophy of Science (Uppsala, Sweden, 1991). In one of his slides, Peter illustrated how a causal diagram would change when a variable is manipulated. To me, that slide of Spirtes's – when combined with the deterministic structural equations – was the key to unfolding the manipulative account of causation and led to most of the explorations described in this chapter.

I should really mention another incident that contributed to this chapter. In early 1993 I read the fierce debate between Arthur Goldberger and Nanny Wermuth on the meaning of structural equations (Goldberger 1992; Wermuth 1992). It suddenly hit me that the century-old tension between economists and statisticians stems from simple semantic confusion: Statisticians read structural equations as statements about $E(Y \mid x)$, while economists read them as $E(Y \mid do(x))$. This would explain why statisticians claim that structural equations have no meaning and why economists retort that statistics has no substance. I wrote a technical report, "On the Statistical Interpretation of Structural Equations" (Pearl 1993c), hoping to see the two camps embrace in reconciliation. Nothing of the sort happened. The statisticians in the dispute continued to insist that anything that is not interpreted as $E(Y \mid x)$ simply lacks meaning. The economists, in contrast, are still trying to decide if it was $do(x)$ that they have been meaning to say all along.

Encouraging colleagues receive far too little credit in official channels, considering the immense impact they have on the encouraged. I must take this opportunity to acknowledge four colleagues who saw clarity shining through the $do(x)$ operator before it gained popularity: Steffen Lauritzen, David Freedman, James Robins, and Philip Dawid.

Phil showed special courage in printing my paper in *Biometrika* (Pearl 1995a), the journal founded by causality's worst adversary – Karl Pearson.

~~Postscript for the Second Edition~~

Complete identification results

A key identification condition, which generalizes all the criteria established in this chapter, has been derived by Jin Tian. It reads:

Theorem 3.6.1 (Tian and Pearl, 2002a)
A sufficient condition for identifying the causal effect $P(y \mid do(x))$ is that there exists no bi-directed path (i.e., a path composed entirely of bi-directed arcs) between X and any of its children.[15]

Remarkably, the theorem asserts that, as long as every child of X (on the pathways to Y) is not reachable from X via a bi-directed path, then, regardless of how complicated the graph, the causal effect $P(y \mid do(x))$ is identifiable. All identification criteria discussed in this chapter are special cases of the one defined in this theorem. For example, in Figure 3.5 $P(y \mid do(x))$ can be identified because the two paths from X to Z (the only child of X) are not bi-directed. In Figure 3.7, on the other hand, there is a path from X to Z_1 traversing only bi-directed arcs, thus violating the condition of Theorem 3.6.1, and $P(y \mid do(x))$ is not identifiable.

Note that all graphs in Figure 3.8 and none of those in Figure 3.9 satisfy the condition above. Tian and Pearl (2002a) further showed that the condition is both sufficient and necessary for the identification of $P(\boldsymbol{v} \mid do(x))$, where V includes all variables except X. A necessary and sufficient condition for identifying $P(w \mid do(z))$, with W and Z two arbitrary sets, was established by Shpitser and Pearl (2006b). Subsequently, a complete graphical criterion was established for determining the identifiability of *conditional* interventional distributions, namely, expressions of the type $P(y \mid do(x), z)$ where X, Y, and Z are arbitrary sets of variables (Shpitser and Pearl 2006a).

These results constitute a complete characterization of causal effects in graphical models. They provide us with polynomial time algorithms for determining whether an arbitrary quantity invoking the $do(x)$ operator is identified in a given semi-Markovian model and, if so, what the estimand of that quantity is. Remarkably, one corollary of these results also states that the *do*-calculus is complete, namely, a quantity $Q = P(y \mid do(x), z)$ is identified if and only if it can be reduced to a *do*-free expression using the three rules of Theorem 3.4.1.[16] Tian and Shpitser (2010) provide a comprehensive summary of these results.

Applications and Critics

Gentle introductions to the concepts developed in this chapter are given in (Pearl 2003c) and (Pearl 2009a). Applications of causal graphs in epidemiology are reported in Robins

[15] Before applying this criterion, one may delete from the causal graph all nodes that are not ancestors of Y.
[16] This was independently established by Huang and Valtorta (2006).

(2001), Hernán et al. (2002), Hernán et al. (2004), Greenland and Brumback (2002), Greenland et al. (1999a,b) Kaufman et al. (2005), Petersen et al. (2006), Hernández-Díaz et al. (2006), VanderWeele and Robins (2007) and Glymour and Greenland (2008).

Interesting applications of the front-door criterion (Section 3.3.2) were noted in social science (Morgan and Winship 2007) and economics (Chalak and White 2006).

Some advocates of the "potential outcome" approach have been most resistant to accepting graphs or structural equations as the basis for causal analysis and, lacking these conceptual tools, were unable to address the issue of covariate selection (Rosenbaum 2002, p. 76; Rubin 2007, 2008a) and were led to dismiss important scientific concepts as "ill-defined," "deceptive," "confusing" (Holland 2001; Rubin 2004, 2008b), and worse (Rubin 2009). Lauritzen (2004) and Heckman (2005) have criticized this attitude; Pearl (2009a,b, 2010a) illuminates its damaging consequences.

Equally puzzling are concerns of some philosophers (Cartwright 2007; Woodward 2003) and economists (Heckman 2005) that the *do*-operator is too local to model complex, real-life policy interventions, which sometimes affect several mechanisms at once and often involve conditional decisions, imperfect control, and multiple actions. These concerns emerge from conflating the mathematical definition of a relationship (e.g., causal effect) with the technical feasibility of testing that relationship in the physical world. While the *do*-operator is indeed an ideal mathematical tool (not unlike the *derivative* in differential calculus), it nevertheless permits us to specify and analyze interventional strategies of great complexity. Readers will find examples of such strategies in Chapter 4, and a further discussion of this issue in Chapter 11 (Sections 11.4.3–11.4.6 and Section 11.5.4).

Chapter Road Map to the Main Results

The three key results in this chapter are: 1. The control of confounding, 2. The evaluation of policies, and 3. The evaluation of counterfactuals.

1. The problem of controlling confounding bias is resolved through the back-door condition (Theorem 3.3.2, pp.79–80) – a criterion for selecting a set of covariates that, if adjusted for, would yield an unbiased estimate of causal effects.
2. The policy evaluation problem – to predict the effect of interventions from nonexperimental data – is resolved through the *do*-calculus (Theorem 3.4.1, pp. 85–86) and the graphical criteria that it entails (Theorem 3.3.4, p. 83; Theorem 3.6.1, p. 105). The completeness of *do*-calculus implies that any (nonparametric) policy evaluation problem that is not supported by an identifying graph, or an equivalent set of causal assumptions, can be proven "unsolvable."
3. Finally, equation (3.51) provides a formal semantics for counterfactuals, through which joint probabilities of counterfactuals can be defined and evaluated in the framework of scientific theories (see Chapter 7). This semantics will enable us to develop a number of techniques for counterfactual analyses (Chapters 8–11), including the Mediation Formula (equations (4.17)–(4.18)) – a key tool for assessing causal pathways in nonlinear models.