CHAPTER ONE

Introduction to Probabilities, Graphs, and Causal Models

Chance gives rise to thoughts,
and chance removes them.
Pascal (1670)

1.1 INTRODUCTION TO PROBABILITY THEORY

1.1.1 Why Probabilities?

Causality connotes lawlike necessity, whereas probabilities connote exceptionality, doubt, and lack of regularity. Still, there are two compelling reasons for starting with, and in fact stressing, probabilistic analysis of causality; one is fairly straightforward, the other more subtle.

The simple reason rests on the observation that causal utterances are often used in situations that are plagued with uncertainty. We say, for example, "reckless driving causes accidents" or "you will fail the course because of your laziness" (Suppes 1970), knowing quite well that the antecedents merely tend to make the consequences more likely, not absolutely certain. Any theory of causality that aims at accommodating such utterances must therefore be cast in a language that distinguishes various shades of likelihood – namely, the language of probabilities. Connected with this observation, we note that probability theory is currently the official mathematical language of most disciplines that use causal modeling, including economics, epidemiology, sociology, and psychology. In these disciplines, investigators are concerned not merely with the presence or absence of causal connections but also with the relative strengths of those connections and with ways of inferring those connections from noisy observations. Probability theory, aided by methods of statistical analysis, provides both the principles and the means of coping with – and drawing inferences from – such observations.

The more subtle reason concerns the fact that even the most assertive causal expressions in natural language are subject to exceptions, and those exceptions may cause major difficulties if processed by standard rules of deterministic logic. Consider, for example, the two plausible premises:

1. My neighbor's roof gets wet whenever mine does.
2. If I hose my roof it will get wet.

Taken literally, these two premises imply the implausible conclusion that my neighbor's roof gets wet whenever I hose mine.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

Such paradoxical conclusions are normally attributed to the finite granularity of our language, as manifested in the many exceptions that are implicit in premise 1. Indeed, the paradox disappears once we take the trouble of explicating those exceptions and write, for instance:

1*. My neighbor's roof gets wet whenever mine does, except when it is covered with plastic, or when my roof is hosed, etc.

Probability theory, by virtue of being especially equipped to tolerate unexplicated exceptions, allows us to focus on the main issues of causality without having to cope with paradoxes of this kind.

As we shall see in subsequent chapters, tolerating exceptions solves only some of the problems associated with causality. The remaining problems – including issues of inference, interventions, identification, ramification, confounding, counterfactuals, and explanation – will be the main topic of this book. By portraying those problems in the language of probabilities, we emphasize their universality across languages. Chapter 7 will recast these problems in the language of deterministic logic and will introduce probabilities merely as a way to express uncertainty about unobserved facts.

1.1.2 Basic Concepts in Probability Theory

The bulk of the discussion in this book will focus on systems with a finite number of discrete variables and thus will require only rudimentary notation and elementary concepts in probability theory. Extensions to continuous variables will be outlined but not elaborated in full generality. Readers who want additional mathematical machinery are invited to study the many excellent textbooks on the subject – for example, Feller (1950), Hoel et al. (1971), or the appendix to Suppes (1970). This section provides a brief summary of elementary probability concepts, based largely on Pearl (1988b), with special emphasis on Bayesian inference and its connection to the psychology of human reasoning under uncertainty. Such emphasis is generally missing from standard textbooks.

We will adhere to the Bayesian interpretation of probability, according to which probabilities encode degrees of belief about events in the world and data are used to strengthen, update, or weaken those degrees of belief. In this formalism, degrees of belief are assigned to propositions (sentences that take on true or false values) in some language, and those degrees of belief are combined and manipulated according to the rules of probability calculus. We will make no distinction between sentential propositions and the actual events represented by those propositions. For example, if A stands for the statement "Ted Kennedy will seek the nomination for president in year 2012," then $P(A \mid K)$ stands for a person's subjective belief in the event described by A given a body of knowledge K, which might include that person's assumptions about American politics, specific proclamations made by Kennedy, and an assessment of Kennedy's age and personality. In defining probability expressions, we often simply write $P(A)$, leaving out the symbol K. However, when the background information undergoes changes, we need to identify specifically the assumptions that account for our beliefs and explicitly articulate K (or some of its elements).

In the Bayesian formalism, belief measures obey the three basic axioms of probability calculus:

$$0 \leq P(A) \leq 1, \tag{1.1}$$

$$P(\text{sure proposition}) = 1, \tag{1.2}$$

$$P(A \text{ or } B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive.} \tag{1.3}$$

The third axiom states that the belief assigned to any set of events is the sum of the beliefs assigned to its nonintersecting components. Because any event A can be written as the union of the joint events $(A \wedge B)$ and $(A \wedge \neg B)$, their associated probabilities are given by[1]

$$P(A) = P(A, B) + P(A, \neg B), \tag{1.4}$$

where $P(A, B)$ is short for $P(A \wedge B)$. More generally, if B_i, $i = 1, 2, \ldots, n$, is a set of exhaustive and mutually exclusive propositions (called a *partition* or a *variable*), then $P(A)$ can be computed from $P(A, B_i)$, $i = 1, 2, \ldots, n$, by using the sum

$$P(A) = \sum_i P(A, B_i), \tag{1.5}$$

which has come to be known as the "law of *total* probability." The operation of summing up probabilities over all B_i is also called "marginalizing over B"; and the resulting probability, $P(A)$, is called the *marginal* probability of A. For example, the probability of A, "The outcomes of two dice are equal," can be computed by summing over the joint events $(A \wedge B_i)$, $i = 1, 2, \ldots, 6$, where B_i stands for the proposition "The outcome of the first die is i." This yields

$$P(A) = \sum_i P(A, B_i) = 6 \times \frac{1}{36} = \frac{1}{6}. \tag{1.6}$$

A direct consequence of (1.2) and (1.4) is that a proposition and its negation must be assigned a total belief of unity,

$$P(A) + P(\neg A) = 1, \tag{1.7}$$

because one of the two statements is certain to be true.

The basic expressions in the Bayesian formalism are statements about *conditional probabilities* – for example, $P(A \mid B)$ – which specify the belief in A under the assumption that B is known with absolute certainty. If $P(A \mid B) = P(A)$, we say that A and B are *independent,* since our belief in A remains unchanged upon learning the truth of B. If $P(A \mid B, C) = P(A \mid C)$, we say that A and B are *conditionally independent* given C; that is, once we know C, learning B would not change our belief in A.

Contrary to the traditional practice of defining conditional probabilities in terms of joint events,

$$P(A \mid B) = \frac{P(A, B)}{P(B)}, \tag{1.8}$$

[1] The symbols $\wedge$, $\vee$, $\neg$, $\Rightarrow$ denote the logical connectives *and, or, not,* and *implies,* respectively.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

Bayesian philosophers see the conditional relationship as more basic than that of joint events – that is, more compatible with the organization of human knowledge. In this view, B serves as a pointer to a context or frame of knowledge, and $A \mid B$ stands for an event A in the context specified by B (e.g., a symptom A in the context of a disease B). Consequently, empirical knowledge invariably will be encoded in conditional probability statements, whereas belief in joint events (if it is ever needed) will be computed from those statements via the product

$$P(A, B) = P(A \mid B)\, P(B), \tag{1.9}$$

which is equivalent to (1.8). For example, it was somewhat unnatural to assess

$$P(A, B_i) = \frac{1}{36}$$

directly in (1.6). The mental process underlying such assessment presumes that the two outcomes are independent, so to make this assumption explicit the probability of the joint event (equality, B_i) should be assessed from the conditional event (equality $\mid B_i$) via the product

$$\begin{aligned} P(\text{equality} \mid B_i)\, P(B_i) &= P(\text{outcome of second die is } i \mid B_i) P(B_i) \\ &= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}. \end{aligned}$$

As in (1.5), the probability of any event A can be computed by conditioning it on any set of exhaustive and mutually exclusive events B_i, $i = 1, 2, \ldots, n$, and then summing:

$$P(A) = \sum_i P(A \mid B_i) P(B_i). \tag{1.10}$$

This decomposition provides the basis for hypothetical or "assumption-based" reasoning. It states that the belief in any event A is a weighted sum over the beliefs in all the distinct ways that A might be realized. For example, if we wish to calculate the probability that the outcome X of the first die will be greater than the outcome Y of the second, we can condition the event $A : X > Y$ on all possible values of X and obtain

$$\begin{aligned} P(A) &= \sum_{i=1}^{6} P(Y < X \mid X = i) P(X = i) \\ &= \sum_{i=1}^{6} P(Y < i) \frac{1}{6} = \sum_{i=1}^{6} \sum_{j=1}^{i-1} P(Y = j) \frac{1}{6} \\ &= \frac{1}{6} \sum_{i=2}^{6} \frac{i-1}{6} = \frac{5}{12}. \end{aligned}$$

It is worth reemphasizing that formulas like (1.10) are always understood to apply in some larger context K, which defines the assumptions taken as common knowledge (e.g., the fairness of dice rolling). Equation (1.10) is really a shorthand notation for the statement

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

$$P(A \mid K) = \sum_i P(A \mid B_i, K)P(B_i \mid K). \tag{1.11}$$

This equation follows from the fact that every conditional probability $P(A \mid K)$ is itself a genuine probability function; hence it satisfies (1.10).

Another useful generalization of the product rule (equation (1.9)) is the *chain rule* formula. It states that if we have a set of n events, $E_1, E_2,\ldots, E_n$, then the probability of the joint event $(E_1, E_2,\ldots, E_n)$ can be written as a product of n conditional probabilities:

$$P(E_1, E_2, \ldots, E_n) = P(E_n \mid E_{n-1}, \ldots, E_2, E_1) \ \ldots \ P(E_2 \mid E_1)\, P(E_1). \tag{1.12}$$

This product can be derived by repeated application of (1.9) in any convenient order.

The heart of Bayesian inference lies in the celebrated inversion formula,

$$P(H \mid e) = \frac{P(e \mid H)P(H)}{P(e)}, \tag{1.13}$$

which states that the belief we accord a hypothesis H upon obtaining evidence e can be computed by multiplying our previous belief $P(H)$ by the likelihood $P(e \mid H)$ that e will materialize if H is true. This $P(H \mid e)$ is sometimes called the posterior probability (or simply *posterior*), and $P(H)$ is called the prior probability (or *prior*). The denominator $P(e)$ of (1.13) hardly enters into consideration because it is merely a normalizing constant $P(e) = P(e \mid H)P(H) + P(e \mid \neg H)P(\neg H)$, which can be computed by requiring that $P(H \mid e)$ and $P(\neg H \mid e)$ sum to unity.

Whereas formally (1.13) might be dismissed as a tautology stemming from the definition of conditional probabilities,

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \quad \text{and} \quad P(B \mid A) = \frac{P(A, B)}{P(A)}, \tag{1.14}$$

the Bayesian subjectivist regards (1.13) as a normative rule for updating beliefs in response to evidence. In other words, although conditional probabilities can be viewed as purely mathematical constructs (as in (1.14)), the Bayes adherent views them as primitives of the language and as faithful translations of the English expression "…, given that I know A." Accordingly, (1.14) is not a definition but rather an empirically verifiable relationship between English expressions. It asserts, among other things, that the belief a person attributes to B after discovering A is never lower than that attributed to $A \wedge B$ before discovering A. Also, the ratio between these two beliefs will increase proportionally with the degree of surprise $[P(A)]^{-1}$ one associates with the discovery of A.

The importance of (1.13) is that it expresses a quantity $P(H \mid e)$ – which people often find hard to assess – in terms of quantities that often can be drawn directly from our experiential knowledge. For example, if a person at the next gambling table declares the outcome "twelve," and we wish to know whether he was rolling a pair of dice or spinning a roulette wheel, our models of the gambling devices readily yield the quantities $P(\text{twelve} \mid \text{dice})$ and $P(\text{twelve} \mid \text{roulette})$: 1/36 for the former and 1/38 for the latter. Similarly, we can judge the prior probabilities $P(\text{dice})$ and $P(\text{roulette})$ by estimating the number of roulette wheels and dice tables at the casino. Issuing a direct judgment of

P(dice | twelve) would have been much more difficult; only a specialist in such judgments, trained at the very same casino, could do it reliably.

In order to complete this brief introduction, we must discuss the notion of *probabilistic model* (also called *probability space*). A probabilistic model is an encoding of information that permits us to compute the probability of every well-formed sentence S in accordance with the axioms of (1.1)–(1.3). Starting with a set of atomic propositions A, B, C,..., the set of well-formed sentences consists of all Boolean formulas involving these propositions, for example, $S = (A \wedge B) \vee \neg C$. The traditional method of specifying probabilistic models employs a *joint distribution function*, which is a function that assigns nonnegative weights to every *elementary event* in the language (an elementary event being a conjunction in which every atomic proposition or its negation appears once) such that the sum of the weights adds up to 1. For example, if we have three atomic propositions, A, B, and C, then a joint distribution function should assign nonnegative weights to all eight combinations – $(A \wedge B \wedge C), (A \wedge B \wedge \neg C), \ldots, (\neg A \wedge \neg B \wedge \neg C)$ – such that the eight weights sum to 1.

The reader may recognize the set of elementary events as the *sample space* in probability textbooks. For example, if A, B, and C correspond to the propositions that coins 1, 2, and 3 will come up heads, then the sample space will consist of the set {HHH, HHT, HTH,..., TTT}. Indeed, it is sometimes convenient to view the conjunctive formulas corresponding to elementary events as *points* (or *worlds* or *configurations*), and to regard other formulas as *sets* made up of these points. Since every Boolean formula can be expressed as a disjunction of elementary events, and since the elementary events are mutually exclusive, we can always compute $P(S)$ using the additivity axiom (equation (1.3)). Conditional probabilities can be computed the same way, using (1.14). Thus, any joint probability function represents a complete probabilistic model.

Joint distribution functions are mathematical constructs of great importance. They allow us to determine quickly whether we have sufficient information to specify a complete probabilistic model, whether the information we have is consistent, and at what point additional information is needed. The criteria are simply to check (i) whether the information available is sufficient for uniquely determining the probability of every elementary event in the domain and (ii) whether the probabilities add up to 1.

In practice, however, joint distribution functions are rarely specified explicitly. In the analysis of continuous random variables, the distribution functions are given by algebraic expressions such as those describing normal or exponential distributions; for discrete variables, indirect representation methods have been developed where the overall distribution is inferred from local relationships among small groups of variables. Graphical models, the most popular of these representations, provide the basis of discussion throughout this book. Their use and formal characterization will be discussed in the next few sections.

1.1.3 Combining Predictive and Diagnostic Supports

The essence of Bayes's rule (equation 1.13)) is conveniently portrayed using the *odds* and *likelihood ratio* parameters. Dividing (1.13) by the complementary form for $P(\neg H \mid e)$, we obtain

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

$$\frac{P(H \mid e)}{P(\neg H \mid e)} = \frac{P(e \mid H)}{P(e \mid \neg H)} \frac{P(H)}{P(\neg H)}. \tag{1.15}$$

Defining the *prior odds* on H as

$$O(H) = \frac{P(H)}{P(\neg H)} = \frac{P(H)}{1 - P(H)} \tag{1.16}$$

and the *likelihood ratio* as

$$L(e \mid H) = \frac{P(e \mid H)}{P(e \mid \neg H)}, \tag{1.17}$$

the *posterior odds*

$$O(H \mid e) = \frac{P(H \mid e)}{P(\neg H \mid e)} \tag{1.18}$$

are given by the product

$$O(H \mid e) = L(e \mid H) O(H). \tag{1.19}$$

Thus, Bayes's rule dictates that the overall strength of belief in a hypothesis H, based on both our previous knowledge K and the observed evidence e, should be the product of two factors: the prior odds $O(H)$ and the likelihood ratio $L(e \mid H)$. The first factor measures the *predictive* or *prospective* support accorded to H by the background knowledge alone, while the second represents the *diagnostic* or *retrospective* support given to H by the evidence actually observed.[2]

Strictly speaking, the likelihood ratio $L(e \mid H)$ might depend on the content of the tacit knowledge base K. However, the power of Bayesian techniques comes primarily from the fact that, in causal reasoning, the relationship $P(e \mid H)$ is fairly local: given that H is true, the probability of e can be estimated naturally since it is usually not dependent on many other propositions in the knowledge base. For example, once we establish that a patient suffers from a given disease H, it is natural to estimate the probability that she will develop a certain symptom e. The organization of medical knowledge rests on the paradigm that a symptom is a stable characteristic of the disease and should therefore be fairly independent of other factors, such as epidemic conditions, previous diseases, and faulty diagnostic equipment. For this reason the conditional probabilities $P(e \mid H)$, as opposed to $P(H \mid e)$, are the atomic relationships in Bayesian analysis. The former possess modularity features similar to logical rules. They convey a degree of confidence in rules such as "If H then e," a confidence that persists regardless of what other rules or facts reside in the knowledge base.

Example 1.1.1 Imagine being awakened one night by the shrill sound of your burglar alarm. What is your degree of belief that a burglary attempt has taken place? For

[2] In epidemiology, if H stands for exposure and e stands for disease, then the likelihood ratio L is called the "risk ratio" (Rothman and Greenland 1998, p. 50). Equation (1.18) would then give the odds that a person with disease e had been exposed to H.

illustrative purposes we make the following judgments: (a) There is a 95% chance that an attempted burglary will trigger the alarm system – $P(\text{alarm} \mid \text{burglary}) = 0.95$; (b) based on previous false alarms, there is a slight (1%) chance that the alarm will be triggered by a mechanism other than an attempted burglary – $P(\text{alarm} \mid \text{no burglary}) = 0.01$; (c) previous crime patterns indicate that there is a one in ten thousand chance that a given house will be burglarized on a given night – $P(\text{burglary}) = 10^{-4}$.

Putting these assumptions together using (1.19), we obtain

$$O(\text{burglary} \mid \text{alarm}) = L(\text{alarm} \mid \text{burglary})\, O(\text{burglary})$$

$$= \frac{0.95}{0.01} \frac{10^{-4}}{1 - 10^{-4}} = 0.0095.$$

So, from

$$P(A) = \frac{O(A)}{1 + O(A)} \tag{1.20}$$

we have

$$P(\text{burglary} \mid \text{alarm}) = \frac{0.0095}{1 + 0.0095} = 0.00941.$$

Thus, the retrospective support imparted to the burglary hypothesis by the alarm evidence has increased its degree of belief almost a hundredfold, from one in ten thousand to 94.1 in ten thousand. The fact that the belief in burglary is still below 1% should not be surprising, given that the system produces a false alarm almost once every three months. Notice that it was not necessary to estimate the absolute values of the probabilities $P(\text{alarm} \mid \text{burglary})$ and $P(\text{alarm} \mid \text{no burglary})$. Only their ratio enters the calculation, so a direct estimate of this ratio could have been used instead.

1.1.4 Random Variables and Expectations

By a *variable* we will mean an attribute, measurement or inquiry that may take on one of several possible outcomes, or *values*, from a specified domain. If we have beliefs (i.e., probabilities) attached to the possible values that a variable may attain, we will call that variable a *random variable*.[3] For example, the color of the shoes that I will wear tomorrow is a random variable named "color," and the values it may take come from the domain {yellow, green, red,…}.

Most of our analysis will concern a finite set V of random variables (also called *partitions*) where each variable $X \in V$ may take on values from a finite domain D_X. We will use capital letters (e.g., X, Y, Z) for variable names and lowercase letters (x, y, z)

[3] This is a minor generalization of the textbook definition, according to which a random variable is a mapping from the sample space (e.g., the set of elementary events) to the real line. In our definition, the mapping is from the sample space to any set of objects called "values," which may or may not be ordered.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

as generic symbols for specific values taken by the corresponding variables. For example, if X stands for the color of an object, then x will designate any possible choice of an element from the set {yellow, green, red,…}. Clearly, the proposition X = yellow describes an *event,* namely, a subset of possible states of affair that satisfy the proposition "the color of the object is yellow." Likewise, each variable X can be viewed as a partition of the states of the world, since the statement $X = x$ defines a set of exhaustive and mutually exclusive sets of states, one for each value of x.

In most of our discussions, we will not make notational distinction between variables and sets of variables, because a set of variables essentially defines a compound variable whose domain is the Cartesian product of the domains of the individual constituents in the set. Thus, if Z stands for the set $\{X, Y\}$, then z stands for pairs (x, y) such that $x \in D_X$ and $y \in D_Y$. When the distinction between variables and sets of variables requires special emphasis, indexed letters (say, $X_1, X_2,\ldots, X_n$ or $V_1, V_2,\ldots, V_n$) will be used to represent individual variables.

We shall consistently use the abbreviation $P(x)$ for the probabilities $P(X = x)$, $x \in D_x$. Likewise, if Z stands for the set $\{X, Y\}$, then $P(z)$ will be defined as

$$P(z) \triangleq P(Z = z) = P(X = x, Y = y), \quad x \in D_X, \quad y \in D_Y.$$

When the values of a random variable X are real numbers, X is called a *real* random variable; one can then define the *mean* or *expected value* of X as

$$E(X) \triangleq \sum_x xP(x) \tag{1.21}$$

and the *conditional mean* of X, given event $Y = y$, as

$$E(X \mid y) \triangleq \sum_x xP(x \mid y). \tag{1.22}$$

The expectation of any function g of X is defined as

$$E[g(X)] \triangleq \sum_x g(x)P(x). \tag{1.23}$$

In particular, the function $g(X) = (X - E(X))^2$ has received much attention; its expectation is called the *variance* of X, denoted σ_X^2;

$$\sigma_X^2 \triangleq E[(X - E(X))^2].$$

The conditional mean $E(X \mid Y = y)$ is the *best estimate* of X, given the observation $Y = y$, in the sense of minimizing the expected square error $\sum_x (x - x')^2 P(x \mid y)$ over all possible x'.

The expectation of a function $g(X, Y)$ of two variables, X and Y, requires the joint probability $P(x, y)$ and is defined as

$$E[g(X, Y)] \triangleq \sum_{x, y} g(x, y)P(x, y)$$

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

(cf. equation (1.23)). Of special importance is the expectation of the product $g(X, Y) = (X - E(X))(Y - E(Y))$, which is known as the *covariance* of X and Y,

$$\sigma_{XY} \triangleq E\,[(X - E(X))(Y - E(Y))],$$

and which is often normalized to yield the *correlation coefficient*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and the *regression coefficient* (of X on Y)

$$r_{XY} \triangleq \rho_{XY} \frac{\sigma_X}{\sigma_Y} = \frac{\sigma_{XY}}{\sigma_Y^2}.$$

The *conditional* variance, covariance, and correlation coefficient, given $Z = z$, are defined in a similar manner, using the conditional distribution $P(x, y \mid z)$ in taking expectations. In particular, the *conditional correlation coefficient*, given $Z = z$, is defined as

$$\rho_{XY|z} = \frac{\sigma_{XY|z}}{\sigma_{X|z}\,\sigma_{Y|z}}. \tag{1.24}$$

Additional properties, specific to normal distributions, will be reviewed in Chapter 5 (Section 5.2.1).

The foregoing definitions apply to discrete random variables – that is, variables that take on finite or denumerable sets of values on the real line. The treatment of expectation and correlation is more often applied to continuous random variables, which are characterized by a *density function* $f(x)$ defined as follows:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

for any two real numbers a and b with $a < b$. If X is discrete, then $f(x)$ coincides with the probability function $P(x)$, once we interpret the integral through the translation

$$\int_{-\infty}^{\infty} f(x)dx \Longleftrightarrow \sum_x P(x). \tag{1.25}$$

Readers accustomed to continuous analysis should bear this translation in mind whenever summation is used in this book. For example, the expected value of a continuous random variable X can be obtained from (1.21), to read

$$E(X) = \int_{-\infty}^{\infty} xf\,(x)\,dx,$$

with analogous translations for the variance, correlation, and so forth.

We now turn to define *conditional independence* relationships among variables, a central notion in causal modelling.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

1.1.5 Conditional Independence and Graphoids

Definition 1.1.2 (Conditional Independence)
Let $V = \{V_1, V_2, \ldots\}$ be a finite set of variables. Let $P(\cdot)$ be a joint probability function over the variables in V, and let X, Y, Z stand for any three subsets of variables in V. The sets X and Y are said to be conditionally independent given *Z if*

$$P(x \mid y, z) = P(x \mid z) \quad \textit{whenever} \quad P(y, z) > 0. \tag{1.26}$$

In words, learning the value of Y does not provide additional information about X, once we know Z. (Metaphorically, Z "screens off" X from Y.)

Equation (1.26) is a terse way of saying the following: For any configuration x of the variables in the set X and for any configurations y and z of the variables in Y and Z satisfying $P(Y = y, Z = z) > 0$, we have

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z). \tag{1.27}$$

We will use Dawid's (1979) notation $(X \perp\!\!\!\perp Y \mid Z)_P$ or simply $(X \perp\!\!\!\perp Y \mid Z)$ to denote the conditional independence of X and Y given Z; thus,

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{iff} \quad P(x \mid y, z) = P(x \mid z) \tag{1.28}$$

for all values x, y, z such that $P(y, z) > 0$. Unconditional independence (also called *marginal independence*) will be denoted by $(X \perp\!\!\!\perp Y \mid \emptyset)$; that is,

$$(X \perp\!\!\!\perp Y \mid \emptyset) \text{ iff } P(x \mid y) = P(x) \qquad \text{whenever} \quad P(y) > 0 \tag{1.29}$$

("iff" is shorthand for "if and only if"). Note that $(X \perp\!\!\!\perp Y \mid Z)$ implies the conditional independence of all pairs of variables $V_i \in X$ and $V_j \in Y$, but the converse is not necessarily true.

The following is a (partial) list of properties satisfied by the conditional independence relation $(X \perp\!\!\!\perp Y \mid Z)$. (We use YW to abbreviate $Y \cup W$.)

Symmetry: $(X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (Y \perp\!\!\!\perp X \mid Z)$.

Decomposition: $(X \perp\!\!\!\perp YW \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid Z)$.

Weak union: $(X \perp\!\!\!\perp YW \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid ZW)$.

Contraction: $(X \perp\!\!\!\perp Y \mid Z)\ \&\ (X \perp\!\!\!\perp W \mid ZY) \Longrightarrow (X \perp\!\!\!\perp YW \mid Z)$.

Intersection: $(X \perp\!\!\!\perp W \mid ZY)\ \&\ (X \perp\!\!\!\perp Y \mid ZW) \Longrightarrow (X \perp\!\!\!\perp YW \mid Z)$.

(Intersection is valid in strictly positive probability distributions.)

The proof of these properties can be derived by elementary means from (1.28) and the basic axioms of probability theory.[4] These properties were called *graphoid axioms* by

[4] These properties were first introduced by Dawid (1979) and Spohn (1980) in a slightly different form, and were independently proposed by Pearl and Paz (1987) to characterize the relationships between graphs and informational relevance. Geiger and Pearl (1993) present an in-depth analysis.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

Pearl and Paz (1987) and Geiger et al. (1990) and have been shown to govern the concept of informational relevance in a wide variety of interpretations (Pearl 1988b). In graphs, for example, these properties are satisfied if we interpret $(X \perp\!\!\!\perp Y \mid Z)$ to mean "all paths from a subset X of nodes to a subset Y of nodes are intercepted by a subset Z of nodes."

The intuitive interpretation of the graphoid axioms is as follows (Pearl 1988b, p. 85). The *symmetry* axiom states that, in any state of knowledge Z, if Y tells us nothing new about X, then X tells us nothing new about Y. The *decomposition* axiom asserts that if two combined items of information are judged irrelevant to X, then each separate item is irrelevant as well. The *weak union* axiom states that learning irrelevant information W cannot help the irrelevant information Y become relevant to X. The *contraction* axiom states that if we judge W irrelevant to X after learning some irrelevant information Y, then W must have been irrelevant before we learned Y. Together, the weak union and contraction properties mean that irrelevant information should not alter the relevance status of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant. The *intersection* axiom states that if Y is irrelevant to X when we know W and if W is irrelevant to X when we know Y, then neither W nor Y (nor their combination) is relevant to X.

1.2 GRAPHS AND PROBABILITIES

1.2.1 Graphical Notation and Terminology

A graph consists of a set V of *vertices* (or *nodes*) and a set E of *edges* (or *links*) that connect some pairs of vertices. The vertices in our graphs will correspond to variables (whence the common symbol V), and the edges will denote a certain relationship that holds in pairs of variables, the interpretation of which will vary with the application. Two variables connected by an edge are called *adjacent*.

Each edge in a graph can be either directed (marked by a single arrowhead on the edge), or undirected (unmarked links). In some applications we will also use "bidirected" edges to denote the existence of unobserved common causes (sometimes called *confounders*). These edges will be marked as dotted curved arcs with two arrowheads (see Figure 1.1(a)). If all edges are directed (see Figure 1.1(b)), we then have a *directed* graph. If we strip away all arrowheads from the edges in a graph G, the resultant undirected graph is called the *skeleton* of G. A *path* in a graph is a sequence of edges (e.g., $((W, Z), (Z, Y), (Y, X), (X, Z))$ in Figure 1.1(a)) such that each edge starts with the vertex ending the preceding edge. In other words, a path is any unbroken, nonintersecting route traced out along the edges in a graph, which may go either along or against the arrows. If every edge in a path is an arrow that points from the first to the second vertex of the pair, we have a *directed path*. In Figure 1.1(a), for example, the path $((W, Z), (Z, Y))$ is directed, but the paths $((W, Z), (Z, Y), (Y, X))$ and $((W, Z), (Z, X))$ are not. If there exists a path between two vertices in a graph, then the two vertices are said to be *connected*; else they are *disconnected*.

Directed graphs may include directed cycles (e.g., $X \rightarrow Y, Y \rightarrow X$), representing mutual causation or feedback processes, but not self-loops (e.g., $X \rightarrow X$). A graph (like the two in Figure 1.1) that contains no directed cycles is called *acyclic*. A graph that is

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

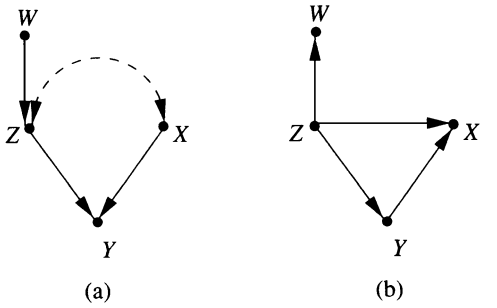

Figure 1.1 (a) A graph containing both directed and bidirected edges. (b) A directed acyclic graph (DAG) with the same skeleton as (a).

both directed and acyclic (Figure 1.1(b)) is called a *directed acyclic graph* (DAG), and such graphs will occupy much of our discussion of causality. We make free use of the terminology of kinship (e.g., *parents, children, descendants, ancestors, spouses*) to denote various relationships in a graph. These kinship relations are defined along the full arrows in the graph, including arrows that form directed cycles but ignoring bidirected and undirected edges. In Figure 1.1(a), for example, Y has two parents (X and Z), three ancestors (X, Z, and W), and no children, while X has no parents (hence, no ancestors), one spouse (Z), and one child (Y). A *family* in a graph is a set of nodes containing a node and all its parents. For example, $\{W\}$, $\{Z, W\}$, $\{X\}$, and $\{Y, Z, X\}$ are the families in the graph of Figure 1.1(a).

A node in a directed graph is called a *root* if it has no parents and a *sink* if it has no children. Every DAG has at least one root and at least one sink. A connected DAG in which every node has at most one parent is called a *tree*, and a tree in which every node has at most one child is called a *chain*. A graph in which every pair of nodes is connected by an edge is called *complete*. The graph in Figure 1.1(a), for instance, is connected but not complete, because the pairs (W, X) and (W, Y) are not adjacent.

1.2.2 Bayesian Networks

The role of graphs in probabilistic and statistical modeling is threefold:

1. to provide convenient means of expressing substantive assumptions;
2. to facilitate economical representation of joint probability functions; and
3. to facilitate efficient inferences from observations.

We will begin our discussion with item 2.

Consider the task of specifying an arbitrary joint distribution, $P(x_1,\ldots, x_n)$, for n dichotomous variables. To store $P(x_1,\ldots, x_n)$ explicitly would require a table with 2^n entries, an unthinkably large number by any standard. Substantial economy can be achieved when each variable depends on just a small subset of other variables. Such dependence information permits us to decompose large distribution functions into several small distributions – each involving a small subset of variables – and then to piece them together coherently to answer questions of a global nature. Graphs play an essential role in such decomposition, for they provide a vivid representation of the sets of variables that are relevant to each other in any given state of knowledge.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

Both directed and undirected graphs have been used by researchers to facilitate such decomposition. Undirected graphs, sometimes called *Markov networks* (Pearl 1988b), are used primarily to represent symmetrical spatial relationships (Isham 1981; Cox and Wermuth 1996; Lauritzen 1996). Directed graphs, especially DAGs, have been used to represent causal or temporal relationships (Lauritzen 1982; Wermuth and Lauritzen 1983; Kiiveri et al. 1984) and came to be known as *Bayesian networks*, a term coined in Pearl (1985) to emphasize three aspects: (1) the subjective nature of the input information; (2) the reliance on Bayes's conditioning as the basis for updating information; and (3) the distinction between causal and evidential modes of reasoning, a distinction that underscores Thomas Bayes's paper of 1763. Hybrid graphs (involving both directed and undirected edges) have also been proposed for statistical modeling (Wermuth and Lauritzen 1990), but in this book our main interest will focus on directed acyclic graphs, with occasional use of directed cyclic graphs to represent feedback cycles.

The basic decomposition scheme offered by directed acyclic graphs can be illustrated as follows. Suppose we have a distribution P defined on n discrete variables, which we may order arbitrarily as $X_1, X_2, \ldots, X_n$. The chain rule of probability calculus (equation (1.12)) always permits us to decompose P as a product of n conditional distributions:

$$P(x_1, \ldots, x_n) = \prod_j P(x_j \mid x_1, \ldots, x_{j-1}). \tag{1.30}$$

Now suppose that the conditional probability of some variable X_j is not sensitive to all the predecessors of X_j but only to a small subset of those predecessors. In other words, suppose that X_j is independent of all other predecessors, once we know the value of a select group of predecessors called PA_j. We can then write

$$P(x_j \mid x_1, \ldots, x_{j-1}) = P(x_j \mid pa_j) \tag{1.31}$$

in the product of (1.30), which will considerably simplify the input information required. Instead of specifying the probability of X_j conditional on all possible realizations of its predecessors $X_1, \ldots, X_{j-1}$, we need only concern ourselves with the possible realizations of the set PA_j. The set PA_j is called the *Markovian parents* of X_j, or *parents* for short. The reason for the name becomes clear when we build graphs around this concept.

Definition 1.2.1 (Markovian Parents)
Let $V = \{X_1, \ldots, X_n\}$ be an ordered set of variables, and let $P(v)$ be the joint probability distribution on these variables. A set of variables PA_j is said to be Markovian parents *of X_j if PA_j is a minimal set of predecessors of X_j that renders X_j independent of all its other predecessors. In other words, PA_j is any subset of $\{X_1, \ldots, X_{j-1}\}$ satisfying*

$$P(x_j \mid pa_j) = P(x_j \mid x_1, \ldots, x_{j-1}) \tag{1.32}$$

and such that no proper subset of PA_j satisfies (1.32).[5]

[5] Lowercase symbols (e.g., x_j, pa_j) denote particular realizations of the corresponding variables (e.g., X_j, PA_j).

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

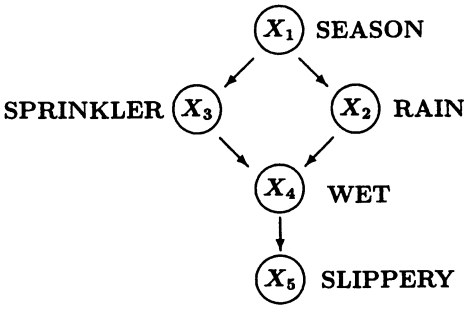

Figure 1.2 A Bayesian network representing dependencies among five variables.

Definition 1.2.1 assigns to each variable X_j a select set PA_j of preceding variables that are sufficient for determining the probability of X_j; knowing the values of other preceding variables is redundant once we know the values pa_j of the parent set PA_j. This assignment can be represented in the form of a DAG in which variables are represented by nodes and arrows are drawn from each node of the parent set PA_j toward the child node X_j. Definition 1.2.1 also suggests a simple recursive method for constructing such a DAG: Starting with the pair (X_1, X_2), we draw an arrow from X_1 to X_2 if and only if the two variables are dependent. Continuing to X_3, we draw no arrow in case X_3 is independent of $\{X_1, X_2\}$; otherwise, we examine whether X_2 screens off X_3 from X_1 or X_1 screens off X_3 from X_2. In the first case, we draw an arrow from X_2 to X_3; in the second, we draw an arrow from X_1 to X_3. If no screening condition is found, we draw arrows to X_3 from both X_1 and X_2. In general: at the jth stage of the construction, we select any minimal set of X_j's predecessors that screens off X_j from its other predecessors (as in equation (1.32)), call this set PA_j and draw an arrow from each member in PA_j to X_j. The result is a directed acyclic graph, called a Bayesian network, in which an arrow from X_i to X_j assigns X_i, as a Markovian parent of X_j, consistent with Definition 1.2.1.

It can be shown (Pearl 1988b) that the set PA_j is unique whenever the distribution $P(v)$ is strictly positive (i.e., involving no logical or definitional constraints), so that every configuration v of variables, no matter how unlikely, has some finite probability of occurring. Under such conditions, the Bayesian network associated with $P(v)$ is unique, given the ordering of the variables.

Figure 1.2 illustrates a simple yet typical Bayesian network. It describes relationships among the season of the year (X_1), whether rain falls (X_2), whether the sprinkler is on (X_3), whether the pavement would get wet (X_4), and whether the pavement would be slippery (X_5). All variables in this figure are binary (taking a value of either true or false) except for the root variable X_1, which can take one of four values: spring, summer, fall, or winter. The network was constructed in accordance with Definition 1.2.1, using causal intuition as a guide. The absence of a direct link between X_1 and X_5, for example, captures our understanding that the influence of seasonal variations on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement). This intuition coincides with the independence condition of (1.32), since knowing X_4 renders X_5 independent of $\{X_1, X_2, X_3\}$.

The construction implied by Definition 1.2.1 defines a Bayesian network as a carrier of conditional independence relationships along the order of construction. Clearly, every distribution satisfying (1.32) must decompose (using the chain rule of (1.30)) into the product

$$P(x_1, \ldots, x_n) = \prod_i P(x_i \mid pa_i). \tag{1.33}$$

For example, the DAG in Figure 1.2 induces the decomposition

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2, x_3)P(x_5 \mid x_4). \tag{1.34}$$

The product decomposition in (1.33) is no longer order-specific since, given P and G, we can test whether P decomposes into the product given by (1.33) without making any reference to variable ordering. We therefore conclude that a necessary condition for a DAG G to be a Bayesian network of probability distribution P is for P to admit the product decomposition dictated by G, as given in (1.33).

Definition 1.2.2 (Markov Compatibility)
If a probability function P admits the factorization of (1.33) *relative to DAG G, we say that G* represents *P, that G and P are* compatible, *or that P is* Markov relative *to G.*[6]

Ascertaining compatibility between DAGs and probabilities is important in statistical modeling primarily because compatibility is a necessary and sufficient condition for a DAG *G to explain* a body of empirical data represented by P, that is, to describe a stochastic process capable of *generating P* (e.g., Pearl 1988b, pp. 210–23). If the value of each variable X_i is chosen at random with some probability $P_i(x_i \mid pa_i)$, based solely on the values pa_i previously chosen for PA_i, then the overall distribution P of the generated instances $x_1, x_2, \ldots, x_n$ will be Markov relative to G. Conversely, if P is Markov relative to G, then there exists a set of probabilities $P_i(x_i \mid pa_i)$ according to which we can choose the value of each variable X_i such that the distribution of the generated instances $x_1, x_2, \ldots, x_n$ will be equal to P. (In fact, the correct choice of $P_i(x_i \mid pa_i)$ would be simply $P(x_i \mid pa_i)$.)

A convenient way of characterizing the set of distributions compatible with a DAG G is to list the set of (conditional) independencies that each such distribution must satisfy. These independencies can be read off the DAG by using a graphical criterion called *d-separation* (Pearl 1988b; the *d* denotes *directional*), which will play a major role in many discussions in this book.

1.2.3 The *d*-Separation Criterion

Consider three disjoint sets of variables, X, Y, and Z, which are represented as nodes in a directed acyclic graph G. To test whether X is independent of Y given Z in any distribution compatible with G, we need to test whether the nodes corresponding to variables Z "block" all paths from nodes in X to nodes in Y. By *path* we mean a sequence of consecutive edges (of any directionality) in the graph, and blocking is to be interpreted as stopping the flow of information (or of dependency) between the variables that are connected by such paths, as defined next.

Definition 1.2.3 (*d*-Separation)
A path p is said to be d-separated (*or* blocked) *by a set of nodes Z if and only if*

[6] The latter expression seems to gain strength in recent literature (e.g., Spirtes et al. 1993; Lauritzen 1996). Pearl (1988b, p. 116) used "G is an *I-map* of *P*."

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

1. *p contains a chain* $i \to m \to j$ *or a fork* $i \leftarrow m \to j$ *such that the middle node m is in Z, or*
2. *p contains an inverted fork* (*or* collider) $i \to m \leftarrow j$ *such that the middle node m is not in Z and such that no descendant of m is in Z.*

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

The intuition behind *d*-separation is simple and can best be recognized if we attribute causal meaning to the arrows in the graph. In causal chains $i \to m \to j$ and causal forks $i \leftarrow m \to j$, the two extreme variables are marginally dependent but become independent of each other (i.e., blocked) once we condition on (i.e., know the value of) the middle variable. Figuratively, conditioning on m appears to "block" the flow of information along the path, since learning about i has no effect on the probability of j, given m. Inverted forks $i \to m \leftarrow j$, representing two causes having a common effect, act the opposite way; if the two extreme variables are (marginally) independent, they will become dependent (i.e., connected through unblocked path) once we condition on the middle variable (i.e., the common effect) or any of its descendants. This can be confirmed in the context of Figure 1.2. Once we know the season, X_3 and X_2 are independent (assuming that sprinklers are set in advance, according to the season); whereas finding that the pavement is wet or slippery renders X_2 and X_3 dependent, because refuting one of these explanations increases the probability of the other.

In Figure 1.2, $X = \{X_2\}$ and $Y = \{X_3\}$ are *d*-separated by $Z = \{X_1\}$, because both paths connecting X_2 and X_3 are blocked by Z. The path $X_2 \leftarrow X_1 \to X_3$ is blocked because it is a fork in which the middle node X_1 is in Z, while the path $X_2 \to X_4 \leftarrow X_3$ is blocked because it is an inverted fork in which the middle node X_4 and all its descendants are outside Z. However, X and Y are not *d*-separated by the set $Z' = \{X_1, X_5\}$: the path $X_2 \to X_4 \leftarrow X_3$ (an inverted fork) is not blocked by Z', since X_5, a descendant of the middle node X_4, is in Z'. Metaphorically, learning the value of the consequence X_5 renders its causes X_2 and X_3 dependent, as if a pathway were opened along the arrows converging at X_4.

At first glance, readers might find it a bit odd that conditioning on a node not lying on a blocked path may unblock the path. However, this corresponds to a general pattern of causal relationships: observations on a common consequence of two independent causes tend to render those causes dependent, because information about one of the causes tends to make the other more or less likely, given that the consequence has occurred. This pattern is known as *selection bias* or *Berkson's paradox* in the statistical literature (Berkson 1946) and as the *explaining away effect* in artificial intelligence (Kim and Pearl 1983). For example, if the admission criteria to a certain graduate school call for either high grades as an undergraduate or special musical talents, then these two attributes will be found to be correlated (negatively) in the student population of that school, even if these attributes are uncorrelated in the population at large. Indeed, students with low grades are likely to be exceptionally gifted in music, which explains their admission to the graduate school.

Figure 1.3 illustrates more elaborate examples of *d*-separation: example (a) contains a bidirected arc $Z_1 \dashleftarrow\!\!\dashrightarrow Z_3$, and (b) involves a directed cycle $X \to Z_2 \to Z_1 \to X$. In

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

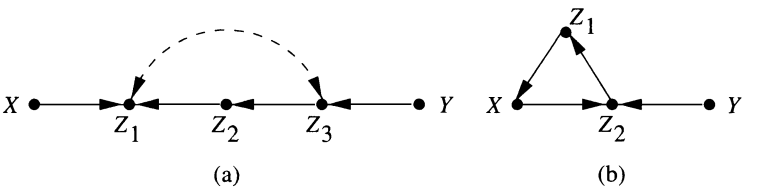

Figure 1.3 Graphs illustrating *d*-separation. In (a), X and Y are *d*-separated given Z_2 and *d*-connected given Z_1. In (b), X and Y cannot be *d*-separated by any set of nodes.

Figure 1.3(a), the two paths between X and Y are blocked when none of $\{Z_1, Z_2, Z_3\}$ is measured. However, the path $X \rightarrow Z_1 \blacktriangleleft\!\!-\!\!-\!\!\blacktriangleright Z_3 \leftarrow Y$ becomes unblocked when Z_1 is measured. This is so because Z_1 unblocks the "colliders" at both Z_1 and Z_3; the first because Z_1 is the collision node of the collider, the second because Z_1 is a descendant of the collision node Z_3 through the path $Z_1 \leftarrow Z_2 \leftarrow Z_3$. In Figure 1.3(b), X and Y cannot be *d*-separated by any set of nodes, including the empty set. If we condition on Z_2, we block the path $X \leftarrow Z_1 \leftarrow Z_2 \leftarrow Y$ yet unblock the path $X \rightarrow Z_2 \leftarrow Y$. If we condition on Z_1, we again block the path $X \leftarrow Z_1 \leftarrow Z_2 \leftarrow Y$ and unblock the path $X \rightarrow Z_2 \leftarrow Y$, because Z_1 is a descendant of the collision node Z_2.

The connection between *d*-separation and conditional independence is established through the following theorem due to Verma and Pearl (1988; see also Geiger et al. 1990).

Theorem 1.2.4 (Probabilistic Implications of *d*-Separation)
If sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every distribution compatible with G. Conversely, if X and Y are not d-separated by Z in a DAG G, then X and Y are dependent conditional on Z in at least one distribution compatible with G.

The converse part of Theorem 1.2.4 is in fact much stronger – the absence of *d*-separation implies dependence in *almost all* distributions compatible with G. The reason is that a precise tuning of parameters is required to generate independency along an unblocked path in the diagram, and such tuning is unlikely to occur in practice (see Spirtes et al. 1993 and Sections 2.4 and 2.9.1).

In order to distinguish between the probabilistic notion of conditional independence $(X \perp\!\!\!\perp Y \mid Z)_P$ and the graphical notion of *d*-separation, for the latter we will use the notation $(X \perp\!\!\!\perp Y \mid Z)_G$. We can thereby express Theorem 1.2.4 more succinctly as follows.

Theorem 1.2.5
For any three disjoint subsets of nodes (X, Y, Z) in a DAG G and for all probability functions P, we have:

(i) $(X \perp\!\!\!\perp Y \mid Z)_G \Longrightarrow (X \perp\!\!\!\perp Y \mid Z)_P$ *whenever G and P are compatible; and*

(ii) *if $(X \perp\!\!\!\perp Y \mid Z)_P$ holds in all distributions compatible with G, it follows that $(X \perp\!\!\!\perp Y \mid Z)_G$.*

An alternative test for *d*-separation has been devised by Lauritzen et al. (1990), based on the notion of ancestral graphs. To test for $(X \perp\!\!\!\perp Y \mid Z)_G$, delete from G all nodes except those in $\{X, Y, Z\}$ and their ancestors, connect by an edge every pair of nodes that share

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

a common child, and remove all arrows from the arcs. Then $(X \perp\!\!\!\perp Y \mid Z)_G$ holds if and only if Z intercepts all paths between X and Y in the resulting undirected graph.

Note that the ordering with which the graph was constructed does not enter into the d-separation criterion; it is only the topology of the resulting graph that determines the set of independencies that the probability P must satisfy. Indeed, the following theorem can be proven (Pearl 1988b, p. 120).

Theorem 1.2.6 (Ordered Markov Condition)
A necessary and sufficient condition for a probability distribution P to be Markov relative to a DAG G is that, conditional on its parents in G, each variable be independent of all its predecessors in some ordering of the variables that agrees with the arrows of G.

A consequence of this theorem is an order-independent criterion for determining whether a given probability P is Markov relative to a given DAG G.

Theorem 1.2.7 (Parental Markov Condition)
A necessary and sufficient condition for a probability distribution P to be Markov relative to a DAG G is that every variable be independent of all its nondescendants (in G), conditional on its parents. (We exclude X_i when speaking of its "nondescendants.")

This condition, which Kiiveri et al. (1984) and Lauritzen (1996) called the "local" Markov condition, is sometimes taken as the definition of Bayesian networks (Howard and Matheson 1981). In practice, however, the ordered Markov condition is easier to use.

Another important property that follows from d-separation is a criterion for determining whether two given DAGs are observationally equivalent – that is, whether every probability distribution that is compatible with one of the DAGs is also compatible with the other.

Theorem 1.2.8 (Observational Equivalence)
Two DAGs are observationally equivalent if and only if they have the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow (Verma and Pearl 1990).[7]

Observational equivalence places a limit on our ability to infer directionality from probabilities alone. Two networks that are observationally equivalent cannot be distinguished without resorting to manipulative experimentation or temporal information. For example, reversing the direction of the arrow between X_1 and X_2 in Figure 1.2 would neither introduce nor destroy a v-structure. Therefore, this reversal yields an observationally equivalent network, and the directionality of the link $X_1 \rightarrow X_2$ cannot be determined from probabilistic information. The arrows $X_2 \rightarrow X_4$ and $X_4 \rightarrow X_5$, however, are of different nature; there is no way of reversing their directionality without creating a new v-structure. Thus, we see that some probability functions P (such as the one responsible for the construction of the Bayesian network in Figure 1.2), when unaccompanied

[7] An identical criterion was independently derived by Frydenberg (1990) in the context of chain graphs, where strict positivity is assumed.

by temporal information, can constrain the directionality of some arrows in the graph. The precise meaning of such directionality constraints – and the possibility of using these constraints for inferring causal relationships from data – will be formalized in Chapter 2.

1.2.4 Inference with Bayesian Networks

Bayesian networks were developed in the early 1980s to facilitate the tasks of prediction and "abduction" in artificial intelligence (AI) systems. In these tasks, it is necessary to find a coherent interpretation of incoming observations that is consistent with both the observations and the prior information at hand. Mathematically, the task boils down to the computation of $P(y \mid x)$, where X is a set of observations and Y is a set of variables that are deemed important for prediction or diagnosis.

Given a joint distribution P, the computation of $P(y \mid x)$ is conceptually trivial and invokes straightforward application of Bayes's rule to yield

$$P(y \mid x) = \frac{\sum_s P(y, x, s)}{\sum_{y,s} P(y, x, s)}, \tag{1.35}$$

where S stands for the set of all variables *excluding* X and Y. Because every Bayesian network defines a joint probability P (given by the product in (1.33)), it is clear that $P(y \mid x)$ can be computed from a DAG G and the conditional probabilities $P(x_i \mid pa_i)$ defined on the families of G.

The challenge, however, lies in performing these computations efficiently and within the representation level provided by the network topology. The latter is important in systems that generate explanations for their reasoning processes. Although such inference techniques are not essential to our discussion of causality, we will nevertheless survey them briefly, for they demonstrate (i) the effectiveness of organizing probabilistic knowledge in the form of graphs and (ii) the feasibility of performing coherent probabilistic calculations (and approximations thereof) on such organization. Details can be found in the references cited.

The first algorithms proposed for probabilistic calculations in Bayesian networks used message-passing architecture and were limited to trees (Pearl 1982; Kim and Pearl 1983). With this technique, each variable is assigned a simple processor and permitted to pass messages asynchronously to its neighbors until equilibrium is achieved (in a finite number of steps). Methods have since been developed that extend this tree propagation (and some of its synchronous variants) to general networks. Among the most popular are Lauritzen and Spiegelhalter's (1988) method of join-tree propagation and the method of cut-set conditioning (Pearl 1988b, pp. 204–10; Jensen 1996). In the join-tree method, we decompose the network into clusters (e.g., cliques) that form tree structures and then treat the set variables in each cluster as a compound variable that is capable of passing messages to its neighbors (which are also compound variables). For example, the network of Figure 1.2 can be structured as a Markov-compatible chain of three clusters:

$$\{X_1, X_2, X_3\} \rightarrow \{X_2, X_3, X_4\} \rightarrow \{X_4, X_5\}.$$

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

In the cut-set conditioning method, a set of variables is instantiated (given specific values) such that the remaining network forms a tree. The propagation is then performed on that tree, and a new instantiation chosen, until all instantiations have been exhausted; the results are then averaged. In Figure 1.2, for example, if we instantiate X_1 to any specific value (say, X_1 = summer), then we break the pathway between X_2 and X_3 and the remaining network becomes tree-structured. The main advantage of the cut-set conditioning method is that its storage-space requirement is minimal (linear in the size of the network), whereas that of the join-tree method might be exponential. Hybrid combinations of these two basic algorithms have also been proposed (Shachter et al. 1994; Dechter 1996) to allow flexible trade-off of storage versus time (Darwiche 2009).

Whereas inference in general networks is "NP-hard" (Cooper 1990), the computational complexity for each of the methods cited here can be estimated prior to actual processing. When the estimates exceed reasonable bounds, an approximation method such as stochastic simulation (Pearl 1988b, pp. 210–23) can be used instead. This method exploits the topology of the network to perform Gibbs sampling on local subsets of variables, sequentially as well as concurrently.

Additional properties of DAGs and their applications to evidential reasoning in expert systems are discussed in Pearl (1988b), Lauritzen and Spiegelhalter (1988), Pearl (1993a), Spiegelhalter et al. (1993), Heckerman et al. (1995), and Darwiche (2009).

1.3 CAUSAL BAYESIAN NETWORKS

The interpretation of direct acyclic graphs as carriers of independence assumptions does not necessarily imply causation; in fact, it will be valid for any set of recursive independencies along any ordering of the variables, not necessarily causal or chronological. However, the ubiquity of DAG models in statistical and AI applications stems (often unwittingly) primarily from their causal interpretation – that is, as a system of processes, one per family, that could account for the generation of the observed data. It is this causal interpretation that explains why DAG models are rarely used in any variable ordering other than those which respect the direction of time and causation.

The advantages of building DAG models around causal rather than associational information are several. First, the judgments required in the construction of the model are more meaningful, more accessible and hence more reliable. The reader may appreciate this point by attempting to construct a DAG representation for the associations in Figure 1.2 along the ordering (X_5, X_1, X_3, X_2, X_4). Such exercises illustrate not only that some independencies are more vividly accessible to the mind than others but also that conditional independence judgments are accessible (hence reliable) only when they are anchored onto more fundamental building blocks of our knowledge, such as causal relationships. In the example of Figure 1.2, our willingness to assert that X_5 is independent of X_2 and X_3 once we know X_4 (i.e., whether the pavement is wet) is defensible because we can easily translate the assertion into one involving causal relationships: that the *influence* of rain and sprinkler on slipperiness is *mediated* by the wetness of the pavement. Dependencies that are not supported by causal links are considered odd or spurious and are even branded "paradoxical" (see the discussion of Berkson's paradox, Section 1.2.3).

We will have several opportunities throughout this book to demonstrate the primacy of causal over associational knowledge. In extreme cases, we will see that people tend to ignore probabilistic information altogether and attend to causal information instead (see Section 6.1.4).[8] This puts into question the ruling paradigm of graphical models in statistics (Wermuth and Lauritzen 1990; Cox and Wermuth 1996), according to which conditional independence assumptions are the primary vehicle for expressing substantive knowledge.[9] It seems that if conditional independence judgments are by-products of stored causal relationships, then tapping and representing those relationships directly would be a more natural and more reliable way of expressing what we know or believe about the world. This is indeed the philosophy behind causal Bayesian networks.

The second advantage of building Bayesian networks on causal relationships – one that is basic to the understanding of causal organizations – is the ability to represent and respond to external or spontaneous *changes*. Any local reconfiguration of the mechanisms in the environment can be translated, with only minor modification, into an isomorphic reconfiguration of the network topology. For example, to represent a disabled sprinkler in the story of Figure 1.2, we simply delete from the network all links incident to the node Sprinkler. To represent the policy of turning the sprinkler off if it rains, we simply add a link between Rain and Sprinkler and revise $P(x_3 \mid x_1, x_2)$. Such changes would require much greater remodeling efforts if the network were not constructed along the causal direction but instead along (say) the order $(X_5, X_1, X_3, X_2, X_4)$. This remodeling flexibility may well be cited as the ingredient that marks the division between deliberative and reactive agents and that enables the former to manage novel situations instantaneously, without requiring training or adaptation.

1.3.1 Causal Networks as Oracles for Interventions

The source of this flexibility rests on the assumption that each parent–child relationship in the network represents a stable and autonomous physical mechanism – in other words, that it is conceivable to change one such relationship *without* changing the others. Organizing one's knowledge in such *modular* configurations permits one to predict the effect of external interventions with a minimum of extra information. Indeed, causal models (assuming they are valid) are much more informative than probability models. A joint distribution tells us how probable events are and how probabilities would change with subsequent observations, but a causal model also tells us how these probabilities would change as a result of external interventions – such as those encountered in policy analysis, treatment management, or planning everyday activity. Such changes cannot be deduced from a joint distribution, even if fully specified.

The connection between modularity and interventions is as follows. Instead of specifying a new probability function for each of the many possible interventions, we specify

[8] The Tversky and Kahneman (1980) experiments with causal biases in probability judgment constitute another body of evidence supporting this observation. For example, most people believe that it is more likely for a girl to have blue eyes, given that her mother has blue eyes, than the other way around; the two probabilities are in fact equal.

[9] The author was as guilty of advocating the centrality of conditional independence as were his colleagues in statistics; see Pearl (1988b, p. 79).

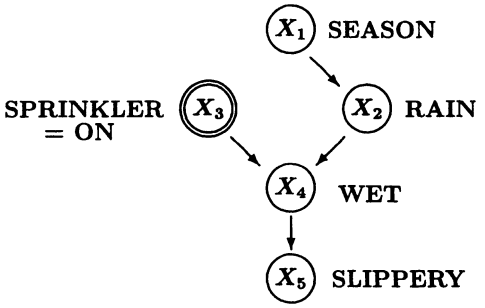

Figure 1.4 Network representation of the action "turning the sprinkler On."

merely the immediate change implied by the intervention and, by virtue of autonomy, we assume that the change is local, and does not spread over to mechanisms other than those specified. Once we know the identity of the mechanism altered by an intervention and the nature of the alteration, the overall effect of an intervention can be predicted by modifying the corresponding factors in (1.33) and using the modified product to compute a new probability function. For example, to represent the action "turning the sprinkler On" in the network of Figure 1.2, we delete the link $X_1 \rightarrow X_3$ and assign X_3 the value On. The graph resulting from this operation is shown in Figure 1.4, and the resulting joint distribution on the remaining variables will be

$$P_{X_3=\text{On}}(x_1, x_2, x_4, x_5) = P(x_1)\,P(x_2 \mid x_1)\,P(x_4 \mid x_2, X_3 = \text{On})\,P(x_5 \mid x_4), \tag{1.36}$$

in which all the factors on the right-hand side (r.h.s.), by virtue of autonomy, are the same as in (1.34).

The deletion of the factor $P(x_3 \mid x_1)$ represents the understanding that, whatever relationship existed between seasons and sprinklers prior to the action, that relationship is no longer in effect while we perform the action. Once we physically turn the sprinkler on and keep it on, a new mechanism (in which the season has no say) determines the state of the sprinkler.

Note the difference between the action $do(X_3 = \text{On})$ and the observation $X_3 = \text{On}$. The effect of the latter is obtained by ordinary Bayesian conditioning, that is, $P(x_1, x_2, x_4, x_5 \mid X_3 = \text{On})$, while that of the former by conditioning a mutilated graph, with the link $X_1 \rightarrow X_3$ removed. This indeed mirrors the difference between seeing and doing: after *observing* that the sprinkler is on, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of a contemplated *action* "turning the sprinkler On."

The ability of causal networks to predict the effects of actions of course requires a stronger set of assumptions in the construction of those networks, assumptions that rest on causal (not merely associational) knowledge and that ensure the system would respond to interventions in accordance with the principle of autonomy. These assumptions are encapsulated in the following definition of causal Bayesian networks.

Definition 1.3.1 (Causal Bayesian Network)
Let $P(v)$ be a probability distribution on a set V of variables, and let $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ that sets a subset X of variables

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

to constants x.[10] *Denote by* $\boldsymbol{P}_*$ *the set of all interventional distributions* $P_x(v)$, $X \subseteq V$, *including* $P(v)$, *which represents no intervention (i.e.,* $X = \emptyset$*). A DAG G is said to be a* causal Bayesian network *compatible with* $\boldsymbol{P}_*$ *if and only if the following three conditions hold for every* $P_x \in \boldsymbol{P}_*$:

(i) $P_x(v)$ *is Markov relative to G*;

(ii) $P_x(v_i) = 1$ *for all* $V_i \in X$ *whenever* v_i *is consistent with* $X = x$;

(iii) $P_x(v_i \mid pa_i) = P(v_i \mid pa_i)$ *for all* $V_i \notin X$ *whenever* pa_i *is consistent with* $X = x$, i.e., each $P(v_i \mid pa_i)$ *remains invariant to interventions not involving* V_i.

Definition 1.3.1 imposes constraints on the interventional space $\boldsymbol{P}_*$ that permit us to encode this vast space economically, in the form of a single Bayesian network *G*. These constraints enable us to compute the distribution $P_x(v)$ resulting from any intervention $do(X = x)$ as a *truncated factorization*

$$P_x(v) = \prod_{\{i \mid V_i \notin X\}} P(v_i \mid pa_i) \qquad \text{for all } v \text{ consistent with } x, \tag{1.37}$$

which follows from (and implies) conditions (i)–(iii), thus justifying the family deletion procedure on *G*, as in (1.36). It is not hard to show that, whenever *G* is a causal Bayes network with respect to $\boldsymbol{P}_*$, the following two properties must hold.

Property 1
For all i,

$$P(v_i \mid pa_i) = P_{pa_i}(v_i). \tag{1.38}$$

Property 2
For all i and for every subset S of variables disjoint of $\{V_i, PA_i\}$, *we have*

$$P_{pa_i, s}(v_i) = P_{pa_i}(v_i). \tag{1.39}$$

Property 1 renders every parent set PA_i *exogenous* relative to its child V_i, ensuring that the conditional probability $P(v_i \mid pa_i)$ coincides with the effect (on V_i) of setting PA_i to pa_i by external control. Property 2 expresses the notion of invariance; once we control its direct causes PA_i, no other interventions will affect the probability of V_i.

1.3.2 Causal Relationships and Their Stability

This mechanism-based conception of interventions provides a semantical basis for notions such as "causal effects" or "causal influence," to be defined formally and analyzed in Chapters 3 and 4. For example, to test whether a variable X_i has a causal influence on another variable X_j, we compute (using the truncated factorization formula of (1.37)) the (marginal) distribution of X_j under the actions $do(X_i = x_i)$ – namely, $P_{x_i}(x_j)$ for all

[10] The notation $P_x(v)$ will be replaced in subsequent chapters with $P(v \mid do(x))$ and $P(v \mid \hat{x})$ to facilitate algebraic manipulations.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

values x_i of X_i – and test whether that distribution is sensitive to x_i. It is easy to see from our previous examples that only variables that are descendants of X_i in the causal network can be influenced by X_i; deleting the factor $P(x_i|pa_i)$ from the joint distribution turns X_i into a root node in the mutilated graph, and root variables (as the *d*-separation criterion dictates) are independent of all other variables except their descendants.

This understanding of causal influence permits us to see precisely why, and in what way, causal relationships are more "stable" than probabilistic relationships. We expect such difference in stability because causal relationships are *ontological*, describing objective physical constraints in our world, whereas probabilistic relationships are *epistemic,* reflecting what we know or believe about the world. Therefore, causal relationships should remain unaltered as long as no change has taken place in the environment, even when our knowledge about the environment undergoes changes. To demonstrate, consider the causal relationship S_1, "Turning the sprinkler on would not affect the rain," and compare it to its probabilistic counterpart S_2, "The state of the sprinkler is independent of (or unassociated with) the state of the rain." Figure 1.2 illustrates two obvious ways in which S_2 will change while S_1 remains intact. First, S_2 changes from false to true when we learn what season it is (X_1). Second, given that we know the season, S_2 changes from true to false once we observe that the pavement is wet (X_4 = true). On the other hand, S_1 remains true regardless of what we learn or know about the season or about the pavement.

The example reveals a stronger sense in which causal relationships are more stable than the corresponding probabilistic relationships, a sense that goes beyond their basic ontological–epistemological difference. The relationship S_1 will remain invariant to changes in the mechanism that regulates how seasons affect sprinklers. In fact, it remains invariant to changes in *all* mechanisms shown in this causal graph. We thus see that causal relationships exhibit greater robustness to ontological changes as well; they are sensitive to a smaller set of mechanisms. More specifically, and in marked contrast to probabilistic relationships, causal relationships remain invariant to changes in the mechanism that governs the causal variables (X_3 in our example).

In view of this stability, it is no wonder that people prefer to encode knowledge in causal rather than probabilistic structures. Probabilistic relationships, such as marginal and conditional independencies, may be helpful in hypothesizing initial causal structures from uncontrolled observations. However, once knowledge is cast in causal structure, those probabilistic relationships tend to be forgotten; whatever judgments people express about conditional independencies in a given domain are derived from the causal structure acquired. This explains why people feel confident asserting certain conditional independencies (e.g., that the price of beans in China is independent of the traffic in Los Angeles) having no idea whatsoever about the numerical probabilities involved (e.g., whether the price of beans will exceed $10 per bushel).

The element of stability (of mechanisms) is also at the heart of the so-called explanatory accounts of causality, according to which causal models need not encode behavior under intervention but instead aim primarily to provide an "explanation" or "understanding" of how data are generated.[11] Regardless of what use is eventually made

[11] Elements of this explanatory account can be found in the writings of Dempster (1990), Cox (1992), and Shafer (1996); see also King et al. (1994, p. 75).

of our "understanding" of things, we surely would prefer an understanding in terms of durable relationships, transportable across situations, over those based on transitory relationships. The sense of "comprehensibility" that accompanies an adequate explanation is a natural by-product of the transportability of (and hence of our familiarity with) the causal relationships used in the explanation. It is for reasons of stability that we regard the falling barometer as predicting but not explaining the rain; those predictions are not transportable to situations where the pressure surrounding the barometer is controlled by artificial means. True understanding enables predictions in such novel situations, where some mechanisms change and others are added. It thus seems reasonable to suggest that, in the final analysis, the explanatory account of causation is merely a variant of the manipulative account, albeit one where interventions are dormant. Accordingly, we may as well view our unsatiated quest for understanding "how data is generated" or "how things work" as a quest for acquiring the ability to make predictions under a wider range of circumstances, including circumstances in which things are taken apart, reconfigured, or undergo spontaneous change.

1.4 FUNCTIONAL CAUSAL MODELS

The way we have introduced the causal interpretation of Bayesian networks represents a fundamental departure from the way causal models (and causal graphs) were first introduced into genetics (Wright 1921), econometrics (Haavelmo 1943), and the social sciences (Duncan 1975), as well as from the way causal models are used routinely in physics and engineering. In those models, causal relationships are expressed in the form of deterministic, *functional* equations, and probabilities are introduced through the assumption that certain variables in the equations are unobserved. This reflects Laplace's (1814) conception of natural phenomena, according to which nature's laws are deterministic and randomness surfaces owing merely to our ignorance of the underlying boundary conditions. In contrast, all relationships in the definition of causal Bayesian networks were assumed to be inherently stochastic and thus appeal to the modern (i.e., quantum mechanical) conception of physics, according to which all nature's laws are inherently probabilistic and determinism is but a convenient approximation.

In this book, we shall express preference toward Laplace's quasi-deterministic conception of causality and will use it, often contrasted with the stochastic conception, to define and analyze most of the causal entities that we study. This preference is based on three considerations. First, the Laplacian conception is more general. Every stochastic model can be emulated by many functional relationships (with stochastic inputs), but not the other way around; functional relationships can only be approximated, as a limiting case, using stochastic models. Second, the Laplacian conception is more in tune with human intuition. The few esoteric quantum mechanical experiments that conflict with the predictions of the Laplacian conception evoke surprise and disbelief, and they demand that physicists give up deeply entrenched intuitions about locality and causality (Maudlin 1994). Our objective is to preserve, explicate, and satisfy – not destroy – those intuitions.[12]

[12] The often heard argument that human intuitions belong in psychology and not in science or philosophy is inapplicable when it comes to causal intuition – the original authors of causal thoughts

Finally, certain concepts that are ubiquitous in human discourse can be defined only in the Laplacian framework. We shall see, for example, that such simple concepts as "the probability that event B occured *because* of event A" and "the probability that event B would have been *different* if it were not for event A" cannot be defined in terms of purely stochastic models. These so-called *counterfactual* concepts will require a synthesis of the deterministic and probabilistic components embodied in the Laplacian model.

1.4.1 Structural Equations

In its general form, a functional causal model consists of a set of equations of the form

$$x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n, \tag{1.40}$$

where pa_i (connoting *parents*) stands for the set of variables that directly determine the value of X_i and where the U_i represent errors (or "disturbances") due to omitted factors. Equation (1.40) is a nonlinear, nonparametric generalization of the linear structural equation models (SEMs)

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, \ldots, n, \tag{1.41}$$

which have become a standard tool in economics and social science (see Chapter 5 for a detailed exposition of this enterprise). In linear models, pa_i corresponds to those variables on the r.h.s. of (1.41) that have nonzero coefficients.

The interpretation of the functional relationship in (1.40) is the standard interpretation that functions carry in physics and the natural sciences; it is a recipe, a strategy, or a *law* specifying what value nature would assign to X_i in response to every possible value combination that (PA_i, U_i) might take on. A set of equations in the form of (1.40) and in which each equation represents an autonomous mechanism is called a *structural model*; if each variable has a distinct equation in which it appears on the left-hand side (called the *dependent* variable), then the model is called a *structural causal model* or a *causal model* for short.[13] Mathematically, the distinction between structural and algebraic equations is that the former change meaning under solution-preserving algebraic operations (e.g., moving terms from one side of an equation to the other.)

To illustrate, Figure 1.5 depicts a canonical econometric model relating price and demand through the equations

$$q = b_1 p + d_1 i + u_1, \tag{1.42}$$

$$p = b_2 q + d_2 w + u_2, \tag{1.43}$$

where Q is the quantity of household demand for a product A, P is the unit price of product A, I is household income, W is the wage rate for producing product A, and U_1 and

cannot be ignored when the meaning of the concept is in question. Indeed, compliance with human intuition has been the ultimate criterion of adequacy in every philosophical study of causation, and the proper incorporation of background information into statistical studies likewise relies on accurate interpretation of causal judgment.

[13] Formal treatment of causal models, structural equations, and error terms are given in Chapter 5 (Section 5.4.1) and Chapter 7 (Sections 7.1 and 7.2.5).

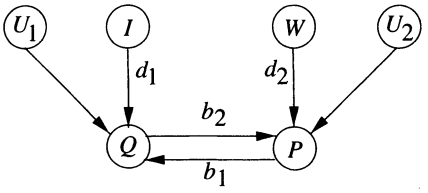

Figure 1.5 Causal diagram illustrating the relationship between price (P), demand (Q), income (I), and wages (W).

U_2 represent error terms – unmodeled factors that affect quantity and price, respectively (Goldberger 1992). The graph associated with this model is cyclic, and the vertices associated with the variables U_1, U_2, I, and W are root nodes, conveying the assumption of mutual independence. The idea of *autonomy* (Aldrich 1989), in this context, means that the two equations represent two loosely coupled segments of the economy, consumers and producers. Equation (1.42) describes how consumers decide what quantity Q to buy, and (1.43) describes how manufacturers decide what price P to charge. Like all feedback systems, this too represents implicit dynamics; today's prices are determined on the basis of yesterday's demand, and these prices will determine the demand in the next period of transactions. The solution to such equations represents a long-term equilibrium under the assumption that the background quantities, U_1 and U_2, remain constant.

The two equations are considered to be "autonomous" relative to the dynamics of changes in the sense that external changes affecting one equation do not imply changes to the others. For example, if government decides on price control and sets the price P at p_0, then (1.43) will be modified to read $p = p_0$ but the relationships in (1.42) will remain intact, yielding $q = b_1 p_0 + d_1 i + u_1$. We thus see that b_1, the "demand elasticity," should be interpreted as the rate of change of Q per unit *controlled* change in P. This is different, of course, from the rate of change of Q per unit *observed* change in P (under uncontrolled conditions), which, besides b_1, is also affected by the parameters of (1.43) (see Section 7.2.1, equation (7.14)). The difference between controlled and observed changes is essential for the correct interpretation of structural equation models in social science and economics, and it will be discussed at length in Chapter 5. If we have reasons to believe that consumer behavior will also change under a price control policy, then this modified behavior would need to be modeled explicitly – for example, by treating the coefficients b_1 and d_1 as dependent variables in auxiliary equations involving P.[14] Section 7.2.1 will present an analysis of policy-related problems using this model.

To illustrate the workings of nonlinear functional models, consider again the causal relationships depicted in Figure 1.2. The causal model associated with these relationships will consist of five functions, each representing an autonomous mechanism governing one variable:

$$x_1 = u_1,$$

$$x_2 = f_2(x_1, u_2),$$

[14] Indeed, consumers normally react to price fixing by hoarding goods in anticipation of shortages (Lucas 1976). Such phenomena are not foreign to structural models, though; they simply call for more elaborate equations to capture consumers' expectations.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

$$\begin{aligned} x_3 &= f_3(x_1, u_3), \\ x_4 &= f_4(x_3, x_2, u_4), \\ x_5 &= f_5(x_4, u_5). \end{aligned} \tag{1.44}$$

The error variables $U_1,\ldots, U_5$ are not shown explicitly in the graph; by convention, this implies that they are assumed to be mutually independent. When some disturbances are judged to be dependent, it is customary to encode such dependencies by augmenting the graph with double-headed arrows, as shown in Figure 1.1(a).

A typical specification of the functions $\{f_1,\ldots, f_5\}$ and the disturbance terms is given by the following Boolean model:

$$\begin{aligned} x_2 &= [(X_1 = \text{winter}) \vee (X_1 = \text{fall}) \vee u_2] \wedge \neg u'_2, \\ x_3 &= [(X_1 = \text{summer}) \vee (X_1 = \text{spring}) \vee u_3] \wedge \neg u'_3, \\ x_4 &= (x_2 \vee x_3 \vee u_4) \wedge \neg u'_4, \\ x_5 &= (x_4 \vee u_5) \wedge \neg u'_5, \end{aligned} \tag{1.45}$$

where x_i stands for X_i = true and where u_i and u'_i stand for triggering and inhibiting abnormalities, respectively. For example, u_4 stands for (unspecified) events that might cause the pavement to get wet (X_4) when the sprinkler is off ($\neg x_3$) and it does not rain ($\neg x_2$) (e.g., a broken water pipe), while u'_4 stands for (unspecified) events that would keep the pavement dry in spite of the rain (x_2), the sprinkler (x_3), and u_4 (e.g., pavement covered with a plastic sheet).

It is important to emphasize that, in the two models just described, the variables placed on the left-hand side of the equality sign (the dependent or output variables) act distinctly from the other variables in each equation. The role of this distinction becomes clear when we discuss interventions, since it is only through this distinction that we can identify which equation ought to be modified under local interventions of the type "fix the price at p_0" ($do(P = p_0)$) or "turn the sprinkler On" ($do(X_3 = \text{true})$).[15]

We now compare the features of functional models as defined in (1.40) with those of causal Bayesian networks defined in Section 1.3. Toward this end, we will consider the processing of three types of queries:

predictions (e.g., would the pavement be slippery if we *find* the sprinkler off?);

interventions (e.g., would the pavement be slippery if we *make sure* that the sprinkler is off?); and

counterfactuals (e.g., would the pavement be slippery *had* the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?).

We shall see that these three types of queries represent a hierarchy of three fundamentally different types of problems, demanding knowledge with increasing levels of detail.

[15] Economists who write the supply–demand equations as $\{q = ap + u_1, q = bp + u_2\}$, with q appearing on the l.h.s. of both equations, are giving up the option of analyzing price control policies unless additional symbolic machinery is used to identify which equation will be modified by the $do(P = p_0)$ operator.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

1.4.2 Probabilistic Predictions in Causal Models

Given a causal model (equation (1.40)), if we draw an arrow from each member of PA_i toward X_i, then the resulting graph G will be called a *causal diagram*. If the causal diagram is acyclic, then the corresponding model is called *semi-Markovian* and the values of the X variables will be uniquely determined by those of the U variables. Under such conditions, the joint distribution $P(x_1,\ldots, x_n)$ is determined uniquely by the distribution $P(u)$ of the error variables. If, in addition to acyclicity, the error terms are jointly independent, the model is called *Markovian*.

A fundamental theorem about Markovian models establishes a connection between causation and probabilities via the parental Markov condition of Theorem 1.2.7.

Theorem 1.4.1 (Causal Markov Condition)
Every Markovian causal model M induces a distribution $P(x_1,\ldots, x_n)$ *that satisfies the parental Markov condition relative to the causal diagram G associated with M; that is, each variable* X_i *is independent of all its nondescendants, given its parents* PA_i *in G* (Pearl and Verma 1991).[16]

The proof is immediate. Considering that the set $\{PA_i, U_i\}$ determines one unique value of X_i, the distribution $P(x_1,\ldots, x_n, u_1,\ldots, u_n)$ is certainly Markov relative to the augmented DAG $G(X, U)$, in which the U variables are represented explicitly. The required Markov condition of the marginal distribution $P(x_1,\ldots, x_n)$ follows by d-separation in $G(X, U)$.

Theorem 1.4.1 shows that the parental Markov condition of Theorem 1.2.7 follows from two causal assumptions: (1) our commitment to include in the model (not in the background) every variable that is a cause of two or more other variables; and (2) Reichenbach's (1956) common-cause assumption, also known as "no correlation without causation," stating that, if any two variables are dependent, then one is a cause of the other *or* there is a third variable causing both. These two assumptions imply that the background factors in U are mutually independent and hence that the causal model is Markovian. Theorem 1.4.1 explains both why Markovian models are so frequently assumed in causal analysis and why the parental Markov condition (Theorem 1.2.7) is so often regarded as an inherent feature of causal models (see, e.g., Kiiveri et al. 1984; Spirtes et al. 1993).[17]

The causal Markov condition implies that characterizing each child–parent relationship as a deterministic function, instead of the usual conditional probability $P(x_i \mid pa_i)$, imposes equivalent independence constraints on the resulting distribution and leads to the same recursive decomposition that characterizes Bayesian networks (see equation (1.33)). More significantly, this holds regardless of the choice of functions $\{f_i\}$ and regardless

[16] Considering its generality and transparency, I would not be surprised if some version of this theorem has appeared earlier in the literature, but I am not aware of any nonparametric version.

[17] Kiiveri et al.'s (1984) paper, entitled "Recursive Causal Models," provides the first proof (for strictly positive distributions) that the parental Markov condition of Theorem 1.2.7 follows from the factorization of (1.33). This implication, however, is purely probabilistic and invokes no aspect of causation. In order to establish a connection between causation and probability we must first devise a model for causation, either in terms of manipulations (as in Definition 1.3.1) or in terms of functional relationships in structural equations (as in Theorem 1.4.1).

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

of the error distributions $P(u_i)$. Thus, we need not specify in advance the functional form of $\{f_i\}$ or the distributions $P(u_i)$; once we measure (or estimate) $P(x_i \mid pa_i)$, all probabilistic properties of a Markovian causal model are determined, regardless of the mechanism that actually generates those conditional probabilities. Druzdzel and Simon (1993) showed that, for every Bayesian network G characterized by a distribution P (as in (1.33)), there exists a functional model (as in (1.40)) that generates a distribution identical to P.[18] It follows that in all probabilistic applications of Bayesian networks – including statistical estimation, prediction, and diagnosis – we can use an equivalent functional model as specified in (1.40), and we can regard functional models as just another way of encoding joint distribution functions.

Nonetheless, the causal–functional specification has several advantages over the probabilistic specification, even in purely predictive (i.e., nonmanipulative) tasks. First and foremost, all the conditional independencies that are displayed by the causal diagram G are guaranteed to be *stable* – that is, invariant to parametric changes in the mechanisms represented by the functions f_i and the distributions $P(u_i)$. This means that agents who choose to organize knowledge using Markovian causal models can make reliable assertions about conditional independence relations without assessing numerical probabilities – a common ability among humanoids[19] and a useful feature for inference. Second, the functional specification is often more meaningful and natural, and it yields a small number of parameters. Typical examples are the linear structural equations used in social science and economics (see Chapter 5) and the "noisy OR gate" that has become quite popular in modeling the effect of multiple dichotomous causes (Pearl 1988b, p. 184). Third (and perhaps hardest for an empiricist to accept), judgmental assumptions of conditional independence among observable quantities are simplified and made more reliable in functional models, because such assumptions are cast directly as judgments about the presence or absence of *unobserved* common causes (e.g., why is the price of beans in China judged to be independent of the traffic in Los Angeles?). In the construction of Bayesian networks, for example, instead of judging whether each variable is independent of all its nondescendants (given its parents), we need to judge whether the parent set contains *all* relevant immediate causes – in particular, whether no factor omitted from the parent set is a cause of another observed variable. Such judgments are more natural because they are discernible directly from a qualitative causal structure, the very structure that our mind has selected for storing stable aspects of experience.

Finally, there is an additional advantage to basing prediction models on causal mechanisms that stems from considerations of stability (Section 1.3.2). When some conditions in the environment undergo change, it is usually only a few causal mechanisms that are affected by the change; the rest remain unaltered. It is simpler then to reassess (judgmentally) or reestimate (statistically) the model parameters knowing that

[18] In Chapter 9 we will show that, except in some pathological cases, there actually exist an infinite number of functional models with this property.

[19] Statisticians who are reluctant to discuss causality yet have no hesitation expressing background information in the form of conditional independence statements would probably be shocked to realize that such statements acquire their validity from none other than the *causal* Markov condition (Theorem 1.4.1). See note 9.

the corresponding symbolic change is also local, involving just a few parameters, than to reestimate the entire model from scratch.[20]

1.4.3 Interventions and Causal Effects in Functional Models

The functional characterization $x_i = f_i(pa_i, u_i)$, like its stochastic counterpart, provides a convenient language for specifying how the resulting distribution would change in response to external interventions. This is accomplished by encoding each intervention as an alteration on a select set of functions instead of a select set of conditional probabilities. The overall effect of the intervention can then be predicted by modifying the corresponding equations in the model and using the modified model to compute a new probability function. Thus, all features of causal Bayesian networks (Section 1.3) can be emulated in Markovian functional models.

For example, to represent the action "turning the sprinkler On" in the model of (1.44), we delete the equation $x_3 = f_3(x_1, u_3)$ and replace it with $x_3 = \text{On}$. The modified model will contain all the information needed for computing the effect of the action on other variables. For example, the probability function induced by the modified model will be equal to that given by (1.36), and the modified diagram will coincide with that of Figure 1.4.

More generally, when an intervention forces a subset X of variables to attain fixed values x, then a subset of equations is to be pruned from the model in (1.40), one for each member of X, thus defining a new distribution over the remaining variables that characterizes the effect of the intervention and coincides with the truncated factorization obtained by pruning families from a causal Bayesian network (equation (1.37)).[21]

The functional model's representation of interventions offers greater flexibility and generality than that of a stochastic model. First, the analysis of interventions can be extended to cyclic models, like the one in Figure 1.5, so as to answer policy-related questions[22] (e.g.: What would the demand quantity be if we control the price at p_0?). Second, interventions involving the modification of equational parameters (like b_1 and d_1 in (1.42)) are more readily comprehended than those described as modifiers of conditional probabilities, perhaps because stable physical mechanisms are normally associated with equations and not with conditional probabilities. Conditional probabilities are perceived to be derivable from, not generators of, joint distributions. Third, the analysis of causal effects in non-Markovian models will be greatly simplified using functional models. The reason is: there are infinitely many conditional probabilities $P(x_i \mid pa_i)$ but only a finite number of functions $x_i = f_i(pa_i, u_i)$ among discrete variables X_i and PA_i. This fact will enable us in Chapter 8 (Section 8.2.2) to use linear-programming techniques to obtain sharp bounds on causal effects in studies involving noncompliance.

---

[20] To the best of my knowledge, this aspect of causal models has not been studied formally; it is suggested here as a research topic for students of adaptive systems.

[21] An explicit translation of interventions to "wiping out" equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970) and Sobel (1990). More elaborate types of interventions, involving conditional actions and stochastic strategies, will be formulated in Chapter 4.

[22] Such questions, especially those involving the control of endogenous variables, are conspicuously absent from econometric textbooks (see Chapter 5).

Finally, functional models permit the analysis of context-specific actions and policies. The notion of causal effect as defined so far is of only minor use in practical policy making. The reason is that causal effects tell us the general tendency of an action to bring about a response (as with the tendency of a drug to enhance recovery in the overall population) but are not specific to actions in a given situation characterized by a set of particular observations that may themselves be affected by the action. A physician is usually concerned with the effect of a treatment on a patient who has already been examined and found to have certain symptoms. Some of those symptoms will themselves be affected by the treatment. Likewise, an economist is concerned with the effect of taxation in a given economic context characterized by various economical indicators, which (again) will be affected by taxation if applied. Such context-specific causal effects cannot be computed by simulating an intervention in a static Bayesian network, because the context itself varies with the intervention and so the conditional probabilities $P(x_i \mid pa_i)$ are altered in the process. However, the functional relationships $x_i = f_i(pa_i, u_i)$ remain invariant, which enables us to compute context-specific causal effects as outlined in the next section (see Sections 7.2.1, 8.3, and 9.3.4 for full details).

1.4.4 Counterfactuals in Functional Models

We now turn to the most distinctive characteristic of functional models – the analysis of *counterfactuals*. Certain counterfactual sentences, as we remarked before, cannot be defined in the framework of stochastic causal networks. To see the difficulties, let us consider the simplest possible causal Bayesian network consisting of a pair of independent (hence unconnected) binary variables X and Y. Such a network ensues, for example, in a controlled (i.e., randomized) clinical trial when we find that a treatment X has no effect on the distribution of subjects' response Y, which may stand for either recovery ($Y = 0$) or death ($Y = 1$). Assume that a given subject, Joe, has taken the treatment and died; we ask whether Joe's death occurred *because of* the treatment, *despite* the treatment, or *regardless of* the treatment. In other words, we ask for the probability Q that Joe would have died had he not been treated.

To highlight the difficulty in answering such counterfactual questions, let us take an extreme case where 50% of the patients recover and 50% die in both the treatment and the control groups; assume further that the sample size approaches infinity, thus yielding

$$P(y \mid x) = 1/2 \quad \text{for all } x \text{ and } y. \tag{1.46}$$

Readers versed in statistical testing will recognize immediately the impossibility of answering the counterfactual question from the available data, noting that Joe, who took the treatment and died, was never tested under the no-treatment condition. Moreover, the difficulty does not stem from addressing the question to a particular individual, Joe, for whom we have only one data point. Rephrasing the question in terms of population frequencies – asking what percentage Q of subjects who died under treatment would have recovered had they not taken the treatment – will encounter the same difficulties because none of those subjects was tested under the no-treatment condition. Such difficulties have prompted some statisticians to dismiss counterfactual questions as metaphysical and to

advocate the restriction of statistical analysis to only those questions that can be answered by direct tests (Dawid 2000).

However, that our scientific, legal, and ordinary languages are loaded with counterfactual utterances indicates clearly that counterfactuals are far from being metaphysical; they must have definite testable implications and must carry valuable substantive information. The analysis of counterfactuals therefore represents an opportunity to anyone who shares the aims of this book: integrating substantive knowledge with statistical data so as to refine the former and interpret the latter. Within this framework, the counterfactual issue demands answers to tough, yet manageable technical questions: What is the empirical content of counterfactual queries? What knowledge is required to answer those queries? How can this knowledge be represented mathematically? Given such representation, what mathematical machinery is needed for deriving the answers?

Chapter 7 (Section 7.2.2) presents an empirical explication of counterfactuals as claims about the temporal persistence of certain mechanisms. In our example, the response to treatment of each (surviving) patient is assumed to be persistent. If the outcome Y were a reversible condition, rather than death, then the counterfactual claim would translate directly into predictions about response to future treatments. But even in the case of death, the counterfactual quantity Q implies not merely a speculation about the hypothetical behavior of subjects who died but also a testable claim about surviving untreated subjects under subsequent treatment. We leave it as an exercise for the reader to prove that, based on (1.46) and barring sampling variations, the percentage Q of deceased subjects from the treatment group who would have recovered had they not taken the treatment precisely equals the percentage Q' of surviving subjects in the nontreatment group who will die if given treatment.[23] Whereas Q is hypothetical, Q' is unquestionably testable.

Having sketched the empirical interpretation of counterfactuals, our next step in this introductory chapter is the question of representation: What knowledge is required to answer questions about counterfactuals? And how should this knowledge be formulated so that counterfactual queries can be answered quickly and reliably? That such representation exists is evident by the swiftness and consistency with which people distinguish plausible from implausible counterfactual statements. Most people would agree that President Clinton's place in history would be different had he not met Monica Lewinsky, but only a few would assert that his place in history would change had he not eaten breakfast yesterday. In the cognitive sciences, such consistency of opinion is as close as one can get to a proof that an effective machinery for representing and manipulating counterfactuals resides someplace in the human mind. What then are the building blocks of that machinery?

A straightforward representational scheme would (i) store counterfactual knowledge in the form of counterfactual premises and (ii) derive answers to counterfactual queries using some logical rules of inference capable of taking us from premises to conclusions. This approach has indeed been taken by the philosophers Robert Stalnaker (1968) and David Lewis (1973a,b), who constructed logics of counterfactuals using closest-world

[23] For example, if Q equals 100% (i.e., all those who took the treatment and died would have recovered had they not taken the treatment), then all surviving subjects from the nontreatment group will die if given treatment (again, barring sampling variations). Such exercises will become routine when we develop the mathematical machinery for analyzing probabilities of causes (see Chapter 9, Theorem 9.2.12, equations (9.11)–(9.12)).

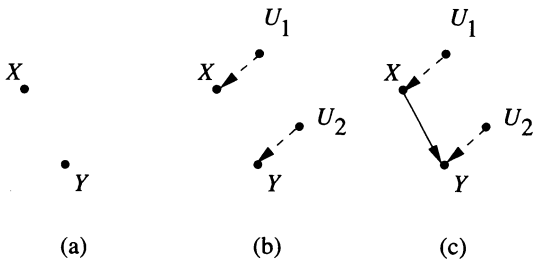

Figure 1.6 (a) A causal Bayesian network that represents the distribution of (1.47). (b) A causal diagram representing the process generating the distribution in (a), according to model 1. (c) Same, according to model 2. (Both U_1 and U_2 are unobserved.)

semantics (i.e., "B would be true if it were A" just in case B is true in the closest possible world (to ours) in which A is true). However, the closest-world semantics still leaves two questions unanswered. (1) What choice of distance measure would make counterfactual reasoning compatible with ordinary conceptions of cause and effect? (2) What mental representation of interworld distances would render the computation of counterfactuals manageable and practical (for both humans and machines)? These two questions are answered by the structural model approach expanded in Chapter 7.

An approach similar to Lewis's (though somewhat less formal) has been pursued by statisticians in the potential-outcome framework (Rubin 1974; Robins 1986; Holland 1988). Here, substantive knowledge is expressed in terms of probabilistic relationships (e.g., independence) among counterfactual variables and then used in the estimation of causal effects. The question of representation shifts from the closest-world to the potential-outcome approach: How are probabilistic relationships among counterfactuals stored or inferred in the investigator's mind? In Chapter 7 (see also Section 3.6.3) we provide an analysis of the closest-world and potential-outcome approaches and compare them to the structural model approach, to be outlined next, in which counterfactuals are *derived* from (and in fact defined by) a functional causal model (equation (1.40)).

In order to see the connection between counterfactuals and structural equations, we should first examine why the information encoded in a Bayesian network, even in its causal interpretation, is insufficient to answer counterfactual queries. Consider again our example of the controlled randomized experiment (equation (1.46)), which corresponds to an edgeless Bayesian network (Figure 1.6(a)) with two independent binary variables and a joint probability:

$$P(y, x) = 0.25 \quad \text{for all } x \text{ and } y. \tag{1.47}$$

We now present two functional models, each generating the joint probability of (1.47) yet each giving a different value to the quantity of interest, Q = the probability that a subject who died under treatment ($x = 1$, $y = 1$) would have recovered ($y = 0$) had he or she not been treated ($x = 0$).

Model 1 (Figure 1.6(b))
Let

$$x = u_1,$$

$$y = u_2,$$

where U_1 and U_2 are two independent binary variables with $P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2}$ (e.g., random coins).

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

Model 1	$u_2 = 0$		$u_2 = 1$		Marginal	
	$x = 1$	$x = 0$	$x = 1$	$x = 0$	$x = 1$	$x = 0$
$y = 1$ (death)	0	0	0.25	0.25	0.25	0.25
$y = 0$ (recovery)	0.25	0.25	0	0	0.25	0.25

Model 2	$u_2 = 0$		$u_2 = 1$		Marginal	
	$x = 1$	$x = 0$	$x = 1$	$x = 0$	$x = 1$	$x = 0$
$y = 1$ (death)	0	0.25	0.25	0	0.25	0.25
$y = 0$ (recovery)	0.25	0	0	0.25	0.25	0.25

Figure 1.7 Contingency tables showing the distributions $P(x, y, u_2)$ and $P(x, y)$ for the two models discussed in the text.

Model 2 (Figure 1.6(c))
Let

$$x = u_1,$$

$$y = xu_2 + (1 - x)(1 - u_2), \tag{1.48}$$

where, as before, U_1 and U_2 are two independent binary variables.

Model 1 corresponds to treatment (X) that has no effect on any of the subjects; in model 2, every subject is affected by treatment. The reason that the two models yield the same distribution is that model 2 describes a mixture of two subpopulations. In one ($u_2 = 1$), each subject dies ($y = 1$) if and only if treated; in the other ($u_2 = 0$), each subject recovers ($y = 0$) if and only if treated. The distributions $P(x, y, u_2)$ and $P(x, y)$ corresponding to these two models are shown in the tables of Figure 1.7.

The value of Q differs in these two models. In model 1, Q evaluates to zero, because subjects who died correspond to $u_2 = 1$ and, since the treatment has no effect on y, changing X from 1 to 0 would still yield $y = 1$. In model 2, however, Q evaluates to unity, because subjects who died under treatment must correspond to $u_2 = 1$ (i.e., those who die if treated), meaning they would recover if and only if not treated.

The first lesson of this example is that stochastic causal models are insufficient for computing probabilities of counterfactuals; knowledge of the actual process behind $P(y \mid x)$ is needed for the computation.[24] A second lesson is that a functional causal model constitutes a mathematical object sufficient for the computation (and definition) of such probabilities. Consider, for example, model 2 of (1.48). The way we concluded that a deceased treated subject ($y = 1, x = 1$) would have recovered if not treated involved three mental steps. First, we applied the evidence at hand, $e : \{y = 1, x = 1\}$, to the model and concluded that e is compatible with only one realization of U_1 and U_2 – namely, $\{u_1 = 1,$

[24] In the potential-outcome framework (Sections 3.6.3 and 7.4.4), such knowledge obtains stochastic appearance by defining distributions over *counterfactual variables* Y_1 and Y_0, which stand for the potential response of an individual to treatment and no treatment, respectively. These hypothetical variables play a role similar to the functions $f_i(pa_i, u_i)$ in our model; they represent the deterministic assumption that every individual possesses a definite response to treatment, regardless of whether that treatment was realized.

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

$u_2 = 1\}$. Second, to simulate the hypothetical condition "had he or she not been treated," we substituted $x = 0$ into (1.48) while ignoring the first equation $x = u_1$. Finally, we solved (1.48) for y (assuming $x = 0$ and $u_2 = 1$) and obtained $y = 0$, from which we concluded that the probability of recovery ($y = 0$) is unity under the hypothetical condition considered.

These three steps can be generalized to any causal model M as follows. Given evidence e, to compute the probability of $Y = y$ under the hypothetical condition $X = x$ (where X is a subset of variables), apply the following three steps to M.

***Step 1* (*abduction*):** Update the probability $P(u)$ to obtain $P(u \mid e)$.

***Step 2* (*action*):** Replace the equations corresponding to variables in set X by the equations $X = x$.

***Step 3* (*prediction*):** Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, this three-step procedure can be interpreted as follows. Step 1 explains the past (U) in light of the current evidence e; step 2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$; finally, step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

Recalling that for each value u of U there is a unique solution for Y, it is clear that step 3 always gives a unique solution for the needed probability; we simply sum up the probabilities $P(u \mid e)$ assigned to all those u that yield $Y = y$ as a solution. Chapter 7 develops effective procedures for computing probabilities of counterfactuals, procedures that are based on probability propagation in "twin" networks (Balke and Pearl 1995a): one network represents the actual world; the other, the counterfactual world.

Note that the hypothetical condition $X = x$ always stands in contradiction to the prevailing values u of U in the model considered (else $X = x$ would actually be realized and thus would not be considered hypothetical). It is for this reason that we invoke (in step 2) an external intervention (alternatively, a "theory change" or a "miracle"; Lewis 1973b), which modifies the model and thus explains the contradiction away. In Chapter 7 we extend this structural–interventional model to give a full semantical and axiomatic account for both counterfactuals and the probability of counterfactuals. In contrast with Lewis's theory, this account is not based on an abstract notion of similarity among hypothetical worlds; rather, it rests on the actual mechanisms involved in the production of the hypothetical worlds considered. Likewise, in contrast with the potential-outcome framework, counterfactuals in the structural account are not treated as undefined primitives but rather as quantities to be derived from the more fundamental concepts of causal mechanisms and their structure.

The three-step model of counterfactual reasoning also uncovers the real reason why stochastic causal models are insufficient for computing probabilities of counterfactuals. Because the U variables do not appear explicitly in stochastic models, we cannot apply step 1 so as to update $P(u)$ with the evidence e at hand. This implies that several ubiquitous notions based on counterfactuals – including probabilities of causes (given the effects), probabilities of explanations, and context-dependent causal effect – cannot be defined in such models. For these, we must make some assumptions about the form of the functions f_i and the probabilities of the error terms. For example, the assumptions of

linearity, normality, and error independence are sufficient for computing all counterfactual queries in the model of Figure 1.5 (see Section 7.2.1). In Chapter 9, we will present conditions under which counterfactual queries concerning probability of causation can be inferred from data when f_i and $P(u)$ are unknown, and only general features (e.g., monotonicity) of these entities are assumed. Likewise, Chapter 8 (Section 8.3) will present methods of *bounding* probabilities of counterfactuals when only stochastic models are available.

The preceding considerations further imply that the three tasks listed in the beginning of this section – prediction, intervention, and counterfactuals – form a natural hierarchy of causal reasoning tasks, with increasing levels of refinement and increasing demands on the knowledge required for accomplishing these tasks. Prediction is the simplest of the three, requiring only a specification of a joint distribution function. The analysis of interventions requires a causal structure in addition to a joint distribution. Finally, processing counterfactuals is the hardest task because it requires some information about the functional relationships and/or the distribution of the omitted factors.

This hierarchy also defines a natural partitioning of the chapters in this book. Chapter 2 will deal primarily with the probabilistic aspects of causal Bayesian networks (though the underlying causal structure will serve as a conceptual guide). Chapters 3–6 will deal exclusively with the interventional aspects of causal models, including the identification of causal effects, the clarification of structural equation models, and the relationships between confounding and collapsibility. Chapters 7–10 will deal with counterfactual analysis, including axiomatic foundation, applications to policy analysis, the bounding of counterfactual queries, the identification of probabilities of causes, and the explication of single-event causation.

I wish the reader a smooth and rewarding journey through these chapters. But first, an important stop for terminological distinctions.

1.5 CAUSAL VERSUS STATISTICAL TERMINOLOGY

This section defines fundamental terms and concepts that will be used throughout this book. These definitions may not agree with those given in standard sources, so it is important to refer to this section in case of doubts regarding the interpretation of these terms.

A ***probabilistic parameter*** is any quantity that is defined in terms[25] of a joint probability function. Examples are the quantities defined in Sections 1.1 and 1.2.

A ***statistical parameter*** is any quantity that is defined in terms of a joint probability distribution of observed variables, making no assumption whatsoever regarding the existence or nonexistence of unobserved variables.

Examples: the conditional expectation $E(Y \mid x)$,
the regression coefficient r_{YX},
the value of the density function at $y = 0, x = 1$.

A ***causal parameter*** is any quantity that is defined in terms of a causal model (as in (1.40)) and is not a statistical parameter.

[25] A quantity Q is said to be *defined in terms of* an object of class C if Q can be computed uniquely from the description of any object in class C (i.e., if Q is defined by a functional mapping from C to the domain of Q).

https://doi.org/10.1017/CBO9780511803161.003 Published online by Cambridge University Press

Examples: the coefficients a_{ik} in (1.41),
whether X_9 has influence on X_3 for some u,
the expected value of Y under the intervention $do(X = 0)$,
the number of parents of variable X_7.

Remark: The exclusion of unmeasured variables from the definition of statistical parameters is devised to prevent one from hiding causal assumptions under the guise of latent variables. Such constructions, if permitted, would qualify any quantity as statistical and would thus obscure the important distinction between quantities that can be estimated from statistical data alone, and those that require additional assumptions beyond the data.

A ***statistical assumption*** is any constraint on a joint distribution of an observed variable; for example, that f is multivariate normal or that P is Markov relative to a given DAG D.

A ***causal assumption*** is any constraint on a causal model that cannot be realized by imposing statistical assumptions; for example, that f_i is linear, that U_i and U_j (unobserved) are uncorrelated, or that x_3 does not appear in $f_4(pa_4, u_4)$. Causal assumptions may or may not have statistical implications. In the former case we say that the assumption is "testable" or "falsifiable." Often, though not always, causal assumptions can be falsified from experimental studies, in which case we say that they are "experimentally testable." For example, the assumption that X has no effect on $E(Y)$ in model 2 of Figure 1.6 is empirically testable, but the assumption that X may cure a given subject in the population is not.

Remark: The distinction between causal and statistical parameters is crisp and fundamental – the two do not mix. Causal parameters cannot be discerned from statistical parameters unless causal assumptions are invoked. The formulation and simplification of these assumptions will occupy a major part of this book.

Remark: Temporal precedence among variables may furnish some information about (the absence of) causal relationships – a later event cannot be the cause of an earlier event. Temporally indexed distributions such as $P(y_t \mid y_{t-1}, x_t)$, $t = 1,\ldots$, which are used routinely in economic analysis, may therefore be regarded as borderline cases between statistical and causal models. We shall nevertheless classify those models as statistical because the great majority of policy-related questions *cannot* be discerned from such distributions, given our commitment to making no assumption regarding the presence or absence of unmeasured variables. Consequently, econometric concepts such as "Granger causality" (Granger 1969) and "strong exogeneity" (Engle et al. 1983) will be classified as statistical rather than causal.[26]

Remark: The terms "theoretical" and "structural" are often used interchangeably with "causal"; we will use the latter two, keeping in mind that some structural models may not be causal (see Section 7.2.5).

[26] Caution must also be exercised in labeling as a "data-generating model" the probabilistic sequence $P(y_t|y_{t-1}\, x_t)$, $t = 1,\ldots$ (e.g., Davidson and MacKinnon 1993, p. 53; Hendry 1995). Such sequences are statistical in nature and, unless causal assumptions of the type developed in Chapter 2 (see Definitions 2.4.1 and 2.7.4) are invoked, they cannot be applied to policy-evaluation tasks.

Causal versus Statistical Concepts

The demarcation line between causal and statistical parameters extends as well to general concepts and will be supported by terminological distinction. Examples of *statistical* concepts are: correlation, regression, conditional independence, association, likelihood, collapsibility, risk ratio, odds ratio, propensity score, Granger's causality, and so on. Examples of *causal* concepts are: randomization, influence, effect, confounding, exogeneity, ignorability, disturbance (e.g., (1.40)), spurious correlation, path coefficients, instrumental variables, intervention, explanation, and so on. The purpose of this demarcation line is not to exclude causal concepts from the province of statistical analysis but, rather, to encourage investigators to treat nonstatistical concepts with the proper set of tools.

Some readers may be surprised by the idea that textbook concepts such as randomization, confounding, spurious correlation, and effects are nonstatistical. Others may be shocked at the idea that controversial concepts such as exogeneity, confounding, and counterfactuals *can* be defined in terms of causal models. This book is written with these readers in mind, and the coming pages will demonstrate that the distinctions just made between causal and statistical concepts are essential for clarifying both.

Two Mental Barriers to Causal Analysis

The sharp distinction between statistical and causal concepts can be translated into a useful principle: behind every causal claim there must lie some causal assumption that is not discernable from the joint distribution and, hence, not testable in observational studies. Such assumptions are usually provided by humans, resting on expert *judgment*. Thus, the way humans organize and communicate experiential knowledge becomes an integral part of the study, for it determines the veracity of the judgments experts are requested to articulate.

Another ramification of this causal–statistical distinction is that any mathematical approach to causal analysis must acquire *new notation*. The vocabulary of probability calculus, with its powerful operators of expectation, conditionalization, and marginalization, is defined strictly in terms of distribution functions and is therefore insufficient for expressing causal assumptions or causal claims. To illustrate, the syntax of probability calculus does not permit us to express the simple fact that "symptoms do not cause diseases," let alone draw mathematical conclusions from such facts. All we can say is that two events are dependent – meaning that if we find one, we can expect to encounter the other, but we cannot distinguish statistical dependence, quantified by the conditional probability $P(disease \mid symptom)$, from causal dependence, for which we have no expression in standard probability calculus.

The preceding two requirements: (1) to commence causal analysis with untested, judgmental assumptions, and (2) to extend the syntax of probability calculus, constitute the two main obstacles to the acceptance of causal analysis among professionals with traditional training in statistics (Pearl 2003c, also sections 11.1.1 and 11.6.4). This book helps overcome the two barriers through an effective and friendly notational system based on symbiosis of graphical and algebraic approaches.