

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.

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 $P(A | K)$ 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:

1.1 Introduction to Probability Theory

3

$$0 \leq P(A) \leq 1, \quad (1.1)$$

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

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

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 by¹

$$P(A) = P(A, B) + P(A, \neg B), \quad (1.4)$$

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

$$P(A) = \sum_i P(A, B_i), \quad (1.5)$$

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 $(A \wedge B_i), i = 1, 2, \dots, 6$, where B_i stands for the proposition “The outcome of the first die is i .” This yields

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

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

$$P(A) + P(\neg A) = 1, \quad (1.7)$$

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, $P(A | B)$ – which specify the belief in A under the assumption that B is known with absolute certainty. If $P(A | B) = P(A)$, we say that A and B are *independent*, since our belief in A remains unchanged upon learning the truth of B . If $P(A | B, C) = P(A | C)$, 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,

$$P(A | B) = \frac{P(A, B)}{P(B)}, \quad (1.8)$$

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

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 $A | B$ 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

$$P(A, B) = P(A | B) P(B), \quad (1.9)$$

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 $| B_i$) via the product

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

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 , $i = 1, 2, \dots, n$, and then summing:

$$P(A) = \sum_i P(A | B_i) P(B_i). \quad (1.10)$$

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

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

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

1.1 Introduction to Probability Theory

5

$$P(A | K) = \sum_i P(A | B_i, K)P(B_i | K). \quad (1.11)$$

This equation follows from the fact that every conditional probability $P(A | K)$ 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, \dots, E_n , then the probability of the joint event (E_1, E_2, \dots, E_n) can be written as a product of n conditional probabilities:

$$P(E_1, E_2, \dots, E_n) = P(E_n | E_{n-1}, \dots, E_2, E_1) \dots P(E_2 | E_1) P(E_1). \quad (1.12)$$

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,

$$P(H | e) = \frac{P(e | H)P(H)}{P(e)}, \quad (1.13)$$

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 $P(e | H)$ that e will materialize if H is true. This $P(H | e)$ 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 $P(e) = P(e | H)P(H) + P(e | \neg H)P(\neg H)$, which can be computed by requiring that $P(H | e)$ and $P(\neg H | e)$ sum to unity.

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

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

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 $P(H | e)$ – 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 $P(\text{twelve} | \text{dice})$ and $P(\text{twelve} | \text{roulette})$: $1/36$ 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

$P(\text{dice} \mid \text{twelve})$ 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, \dots , 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 – $(A \wedge B \wedge C), (A \wedge B \neg C), \dots, (\neg A \wedge \neg B \wedge \neg C)$ – 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 $\{\text{HHH}, \text{HHT}, \text{HTH}, \dots, \text{TTT}\}$. 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

1.1 Introduction to Probability Theory

7

$$\frac{P(H | e)}{P(\neg H | e)} = \frac{P(e | H)}{P(e | \neg H)} \frac{P(H)}{P(\neg H)}. \quad (1.15)$$

Defining the *prior odds* on H as

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

and the *likelihood ratio* as

$$L(e | H) = \frac{P(e | H)}{P(e | \neg H)}, \quad (1.17)$$

the *posterior odds*

$$O(H | e) = \frac{P(H | e)}{P(\neg H | e)} \quad (1.18)$$

are given by the product

$$O(H | e) = L(e | H)O(H). \quad (1.19)$$

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 $L(e | H)$. 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 $L(e | H)$ 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 $P(e | H)$ 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 $P(e | H)$, as opposed to $P(H | e)$, 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

$$\begin{aligned} O(\text{burglary} \mid \text{alarm}) &= L(\text{alarm} \mid \text{burglary})O(\text{burglary}) \\ &= \frac{0.95}{0.01} \frac{10^{-4}}{1 - 10^{-4}} = 0.0095. \end{aligned}$$

So, from

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

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.

1.1 Introduction to Probability Theory

9

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 $X = \text{yellow}$ 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, \dots, X_n or V_1, V_2, \dots, 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

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

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 $E(X | Y = y)$ is the *best estimate* of X , given the observation $Y = y$, in the sense of minimizing the expected square error $\sum_x (x - x')^2 P(x | y)$ 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

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

(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*

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

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

$$\rho_{XY|z} = \frac{\sigma_{XY|z}}{\sigma_{X|z} \sigma_{Y|z}}. \quad (1.24)$$

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

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

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.