Chapter I
Introduction

1. Stochastic Modeling

A quantitative description of a natural phenomenon is called a mathematical model of that phenomenon. Examples abound, from the simple equation $S = \frac{1}{2}gt^2$ describing the distance S traveled in time t by a falling object starting at rest to a complex computer program that simulates a biological population or a large industrial system.

In the final analysis, a model is judged using a single, quite pragmatic, factor, the model's *usefulness*. Some models are useful as detailed quantitative prescriptions of behavior, as for example, an inventory model that is used to determine the optimal number of units to stock. Another model in a different context may provide only general qualitative information about the relationships among and relative importance of several factors influencing an event. Such a model is useful in an equally important but quite different way. Examples of diverse types of stochastic models are spread throughout this book.

Such often mentioned attributes as realism, elegance, validity, and reproducibility are important in evaluating a model only insofar as they bear on that model's ultimate usefulness. For instance, it is both unrealistic and quite inelegant to view the sprawling city of Los Angeles as a geometrical point, a mathematical object of no size or dimension. Yet it is quite useful to do exactly that when using spherical geometry to derive a minimum-distance great circle air route from New York City, another "point."

There is no such thing as the best model for a given phenomenon. The pragmatic criterion of usefulness often allows the existence of two or more models for the same event, but serving distinct purposes. Consider light. The wave form model, in which light is viewed as a continuous flow, is entirely adequate for designing eyeglass and telescope lenses. In contrast, for understanding the impact of light on the retina of the eye, the photon model, which views light as tiny discrete bundles of energy, is preferred. Neither model supersedes the other; both are relevant and useful.

The word "stochastic" derives from the Greed ($\sigma\tau o\chi\acute{\alpha}\zeta\epsilon\sigma\theta\alpha\iota$ to aim, to guess) and means "random" or "chance." The antonym is "sure," "deterministic," or "certain." A deterministic model predicts a single outcome from a given set of circumstances. A stochastic model predicts a set of possible outcomes weighted by their likelihoods, or probabilities. A coin flipped into the air will surely return to earth somewhere. Whether it lands heads or tails is random. For a "fair" coin we consider these alternatives equally likely and assign to each the probability $\frac{1}{2}$.

However, phenomena are not in and of themselves inherently stochastic or deterministic. Rather, to model a phenomenon as stochastic or deterministic is the choice of the observer. The choice depends on the observer's purpose; the criterion for judging the choice is usefulness. Most often the proper choice is quite clear, but controversial situations do arise. If the coin once fallen is quickly covered by a book so that the outcome "heads" or "tails" remains unknown, two participants may still usefully employ probability concepts to evaluate what is a fair bet between them; that is, they may usefully view the coin as random, even though most people would consider the outcome now to be fixed or deterministic. As a less mundane example of the converse situation, changes in the level of a large population are often usefully modeled deterministically, in spite of the general agreement among observers that many chance events contribute to their fluctuations.

Scientific modeling has three components: (i) a natural phenomenon under study, (ii) a logical system for deducing implications about the phenomenon, and (iii) a connection linking the elements of the natural system under study to the logical system used to model it. If we think of these three components in terms of the great-circle air route problem, the natural system is the earth with airports at Los Angeles and New York; the logical system is the mathematical subject of spherical geometry; and the

two are connected by viewing the airports in the physical system as points in the logical system.

The modern approach to stochastic modeling is in a similar spirit. Nature does not dictate a unique definition of "probability," in the same way that there is no nature-imposed definition of "point" in geometry. "Probability" and "point" are terms in pure mathematics, defined only through the properties invested in them by their respective sets of axioms. (See Section 2.8 for a review of axiomatic probability theory.) There are, however, three general principles that are often useful in relating or connecting the abstract elements of mathematical probability theory to a real or natural phenomenon that is to be modeled. These are (i) the principle of equally likely outcomes, (ii) the principle of long run relative frequency, and (iii) the principle of odds making or subjective probabilities. Historically, these three concepts arose out of largely unsuccessful attempts to define probability in terms of physical experiences. Today, they are relevant as guidelines for the assignment of probability values in a model, and for the interpretation of the conclusions of a model in terms of the phenomenon under study.

We illustrate the distinctions between these principles with a long experiment. We will pretend that we are part of a group of people who decide to toss a coin and observe the event that the coin will fall heads up. This event is denoted by H, and the event of tails, by T.

Initially, everyone in the group agrees that $\Pr\{H\} = \frac{1}{2}$. When asked why, people give two reasons: Upon checking the coin construction, they believe that the two possible outcomes, heads and tails, are equally likely; and extrapolating from past experience, they also believe that if the coin is tossed many times, the fraction of times that heads is observed will be close to one-half.

The equally likely interpretation of probability surfaced in the works of Laplace in 1812, where the attempt was made to define the probability of an event A as the ratio of the total number of ways that A could occur to the total number of possible outcomes of the experiment. The equally likely approach is often used today to assign probabilities that reflect some notion of a total lack of knowledge about the outcome of a chance phenomenon. The principle requires judicious application if it is to be useful, however. In our coin tossing experiment, for instance, merely introducing the *possibility* that the coin could land on its edge (E) instantly results in $\Pr\{H\} = \Pr\{T\} = \Pr\{E\} = \frac{1}{3}$.

The next principle, the long run relative frequency interpretation of probability, is a basic building block in modern stochastic modeling, made precise and justified within the axiomatic structure by the *law of large numbers*. This law asserts that the relative fraction of times in which an event occurs in a sequence of independent similar experiments approaches, in the limit, the probability of the occurrence of the event on any single trial.

The principle is not relevant in all situations, however. When the surgeon tells a patient that he has an 80–20 chance of survival, the surgeon means, most likely, that 80 percent of similar patients facing similar surgery will survive it. The patient at hand is not concerned with the long run, but in vivid contrast, is vitally concerned only in the outcome of his, the next, trial.

Returning to the group experiment, we will suppose next that the coin is flipped into the air and, upon landing, is quickly covered so that no one can see the outcome. What is $\Pr\{H\}$ now? Several in the group argue that the outcome of the coin is no longer random, that $\Pr\{H\}$ is either 0 or 1, and that although we don't know which it is, probability theory does not apply.

Others articulate a different view, that the distinction between "random" and "lack of knowledge" is fuzzy, at best, and that a person with a sufficiently large computer and sufficient information about such factors as the energy, velocity, and direction used in tossing the coin could have predicted the outcome, heads or tails, with certainty before the toss. Therefore, even before the coin was flipped, the problem was a lack of knowledge and not some inherent randomness in the experiment.

In a related approach, several people in the group are willing to bet with each other, at even odds, on the outcome of the toss. That is, they are willing to *use* the calculus of probability to determine what is a fair bet, without considering whether the event under study is random or not. The usefulness criterion for judging a model has appeared.

While the rest of the mob were debating "random" versus "lack of knowledge," one member, Karen, looked at the coin. Her probability for heads is now different from that of everyone else. Keeping the coin covered, she announces the outcome "Tails," whereupon everyone mentally assigns the value $\Pr\{H\} = 0$. But then her companion, Mary, speaks up and says that Karen has a history of prevarication.

The last scenario explains why there are horse races; different people assign different probabilities to the same event. For this reason, probabil-

ities used in odds making are often called *subjective* probabilities. Then, odds making forms the third principle for assigning probability values in models and for interpreting them in the real world.

The modern approach to stochastic modeling is to divorce the definition of probability from any particular type of application. Probability theory is an axiomatic structure (see Section 2.8), a part of pure mathematics. Its use in modeling stochastic phenomena is part of the broader realm of science and parallels the use of other branches of mathematics in modeling deterministic phenomena.

To be useful, a stochastic model must reflect all those aspects of the phenomenon under study that are relevant to the question at hand. In addition, the model must be amenable to calculation and must allow the deduction of important predictions or implications about the phenomenon.

1.1. Stochastic Processes

A *stochastic process* is a family of random variables X_t, where t is a parameter running over a suitable index set T. (Where convenient, we will write $X(t)$ instead of X_t.) In a common situation, the index t corresponds to discrete units of time, and the index set is $T = \{0, 1, 2, \ldots\}$. In this case, X_t might represent the outcomes at successive tosses of a coin, repeated responses of a subject in a learning experiment, or successive observations of some characteristics of a certain population. Stochastic processes for which $T = [0, \infty)$ are particularly important in applications. Here t often represents time, but different situations also frequently arise. For example, t may represent distance from an arbitrary origin, and X_t may count the number of defects in the interval $(0, t]$ along a thread, or the number of cars in the interval $(0, t]$ along a highway.

Stochastic processes are distinguished by their *state space*, or the range of possible values for the random variables X_t, by their index set T, and by the dependence relations among the random variables X_t. The most widely used classes of stochastic processes are systematically and thoroughly presented for study in the following chapters, along with the mathematical techniques for calculation and analysis that are most useful with these processes. The use of these processes as models is taught by example. Sample applications from many and diverse areas of interest are an integral part of the exposition.

2. Probability Review*

This section summarizes the necessary background material and establishes the book's terminology and notation. It also illustrates the level of the exposition in the following chapters. Readers who find the major part of this section's material to be familiar and easily understood should have no difficulty with what follows. Others might wish to review their probability background before continuing.

In this section statements frequently are made without proof. The reader desiring justification should consult any elementary probability text as the need arises.

2.1. Events and Probabilities

The reader is assumed to be familiar with the intuitive concept of an *event*. (Events are defined rigorously in Section 2.8, which reviews the axiomatic structure of probability theory.)

Let A and B be events. The event that at least one of A or B occurs is called the *union* of A and B and is written $A \cup B$; the event that both occur is called the *intersection* of A and B and is written $A \cap B$, or simply AB. This notation extends to finite and countable sequences of events. Given events $A_1, A_2, \ldots$, the event that at least one occurs is written $A_1 \cup A_2 \cup \cdots = \bigcup_{i=1}^{\infty} A_i$, the event that all occur is written $A_1 \cap A_2 \cap \cdots = \bigcap_{i=1}^{\infty} A_i$.

The probability of an event A is written $\Pr\{A\}$. The *certain* event, denoted by Ω, always occurs, and $\Pr\{\Omega\} = 1$. The *impossible* event, denoted by $\varnothing$, never occurs, and $\Pr\{\varnothing\} = 0$. It is always the case that $0 \leq \Pr\{A\} \leq 1$ for any event A.

Events A, B are said to be *disjoint* if $A \cap B = \varnothing$; that is, if A and B cannot both occur. For disjoint events A, B we have the *addition law* $\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\}$. A stronger form of the addition law is as follows: Let $A_1, A_2, \ldots$ be events with A_i and A_j disjoint whenever $i \neq j$. Then $\Pr\{\bigcup_{i=1}^{\infty} A_i\} = \sum_{i=1}^{\infty} \Pr\{A_i\}$. The addition law leads directly to the *law*

* Many readers will prefer to omit this review and move directly to Chapter III, on Markov chains. They can then refer to the background material that is summarized in the remainder of this chapter and in Chapter II only as needed.

of total probability: Let $A_1, A_2, \ldots$ be disjoint events for which $\Omega = A_1 \cup A_2 \cup \ldots$. Equivalently, exactly one of the events $A_1, A_2 \ldots$ will occur. The law of total probability asserts that $\Pr\{B\} = \sum_{i=1}^{\infty} \Pr\{B \cap A_i\}$ for any event B. The law enables the calculation of the probability of an event B from the sometimes more easily determined probabilities $\Pr\{B \cap A_i\}$, where $i = 1, 2, \ldots$. Judicious choice of the events A_i is prerequisite to the profitable application of the law.

Events A and B are said to be *independent* if $\Pr\{A \cap B\} = \Pr\{A\} \times \Pr\{B\}$. Events $A_1, A_2, \ldots$ are *independent* if

$$\Pr\{A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}\} = \Pr\{A_{i_1}\} \Pr\{A_{i_2}\} \cdots \Pr\{A_{i_n}\}$$

for every finite set of distinct indices $i_1, i_2, \ldots, i_n$.

2.2 Random Variables

An old-fashioned but very useful and highly intuitive definition describes a *random variable* as a variable that takes on its values by chance. In Section 2.8, we sketch the modern axiomatic structure for probability theory and random variables. The older definition just given serves quite adequately, however, in virtually all instances of stochastic modeling. Indeed, this older definition was the only approach available for well over a century of meaningful progress in probability theory and stochastic processes.

Most of the time we adhere to the convention of using capital letters such as X, Y, Z to denote random variables, and lowercase letters such as x, y, z for real numbers. The expression $\{X \leq x\}$ is the event that the random variable X assumes a value that is less than or equal to the real number x. This event may or may not occur, depending on the outcome of the experiment or phenomenon that determines the value for the random variable X. The probability that the event occurs is written $\Pr\{X \leq x\}$. Allowing x to vary, this probability defines a function

$$F(x) = \Pr\{X \leq x\}, \qquad -\infty < x < +\infty,$$

called the *distribution function* of the random variable X. Where several random variables appear in the same context, we may choose to distinguish their distribution functions with subscripts, writing, for example, $F_X(\xi) = \Pr\{X \leq \xi\}$ and $F_Y(\xi) = \Pr\{Y \leq \xi\}$, defining the distribution

functions of the random variables X and Y, respectively, as functions of the real variable ξ.

The distribution function contains all the information available about a random variable before its value is determined by experiment. We have, for instance, $\Pr\{X > a\} = 1 - F(a)$, $\Pr\{a < X \leq b\} = F(b) - F(a)$, and $\Pr\{X = x\} = F(x) - \lim_{\epsilon \downarrow 0} F(x - \epsilon) = F(x) - F(x-)$.

A random variable X is called *discrete* if there is a finite or denumerable set of distinct values $x_1, x_2, \ldots$ such that $a_i = \Pr\{X = x_i\} > 0$ for $i = 1, 2, \ldots$ and $\Sigma_i a_i = 1$. The function

$$p(x_i) = p_X(x_i) = a_i \qquad \text{for } i = 1, 2, \ldots \tag{2.1}$$

is called the *probability mass function* for the random variable X and is related to the distribution function via

$$p(x_i) = F(x_i) - F(x_i -) \quad \text{and} \quad F(x) = \sum_{x_i \leq x} p(x_i).$$

The distribution function for a discrete random variable is a step function, which increases only in jumps, the size of the jump at x_i being $p(x_i)$.

If $\Pr\{X = x\} = 0$ for every value of x, then the random variable X is called *continuous* and its distribution function $F(x)$ is a continuous function of x. If there is a nonnegative function $f(x) = f_X(x)$ defined for $-\infty < x < \infty$ such that

$$\Pr\{a < X \leq b\} = \int_a^b f(x)\,dx \qquad \text{for } -\infty < a < b < \infty, \tag{2.2}$$

then $f(x)$ is called the *probability density function* for the random variable X. If X has a probability density function $f(x)$, then X is continuous and

$$F(x) = \int_{-\infty}^{x} f(\xi)\,d\xi, \qquad -\infty < x < \infty.$$

If $F(x)$ is differentiable in x, then X has a probability density function given by

$$f(x) = \frac{d}{dx} F(x) = F'(x), \qquad -\infty < x < \infty. \tag{2.3}$$

In differential form, (2.3) leads to the informal statement

$$\Pr\{x < X \leq x + dx\} = F(x + dx) - F(x) = dF(x) = f(x)\,dx. \tag{2.4}$$

We consider (2.4) to be a shorthand version of the more precise statement

$$\Pr\{x < X \leq x + \Delta x\} = f(x)\,\Delta x + o(\Delta x), \qquad \Delta x \downarrow 0, \tag{2.5}$$

where $o(\Delta x)$ is a generic remainder term of order less than Δx as $\Delta x \downarrow 0$. That is, $o(\Delta x)$ represents any term for which $\lim_{\Delta x \downarrow 0} o(\Delta x)/\Delta x = 0$. By the fundamental theorem of calculus, equation (2.5) is valid whenever the probability density function is continuous at x.

While examples are known of continuous random variables that do not possess probability density functions, they do not arise in stochastic models of common natural phenomena.

2.3. Moments and Expected Values

If X is a discrete random variable, then its mth *moment* is given by

$$E[X^m] = \sum_i x_i^m \Pr\{X = x_i\}, \tag{2.6}$$

[where the x_i are specified in (2.1)] provided that the infinite sum converges absolutely. Where the infinite sum diverges, the moment is said not to exist. If X is a continuous random variable with probability density function $f(x)$, then its mth moment is given by

$$E[X^m] = \int_{-\infty}^{+\infty} x^m f(x)\,dx, \tag{2.7}$$

provided that this integral converges absolutely.

The *first moment*, corresponding to $m = 1$, is commonly called the *mean* or *expected value* of X and written m_X or μ_X. The mth *central moment* of X is defined as the mth moment of the random variable $X - \mu_X$, provided that μ_X exists. The first central moment is zero. The second central moment is called the *variance* of X and written σ_X^2 or $\mathrm{Var}[X]$. We have the equivalent formulas $\mathrm{Var}[X] = E[(X - \mu)^2] = E[X^2] - \mu^2$.

The *median* of a random variable X is any value ν with the property that

$$\Pr\{X \geq \nu\} \geq \tfrac{1}{2} \quad \text{and} \quad \Pr\{X \leq \nu\} \geq \tfrac{1}{2}.$$

If X is a random variable and g is a function, then $Y = g(X)$ is also a

random variable. If X is a discrete random variable with possible values $x_1, x_2, \ldots$, then the expectation of $g(X)$ is given by

$$E[g(X)] = \sum_{i=1}^{\infty} g(x_i) \Pr\{X = x_i\}, \tag{2.8}$$

provided that the sum converges absolutely. If X is continuous and has the probability density function f_X, then the expected value of $g(X)$ is evaluated from

$$E[g(X)] = \int g(x) f_X(x)\, dx. \tag{2.9}$$

The general formula, covering both the discrete and continuous cases, is

$$E[g(X)] = \int g(x)\, dF_X(x), \tag{2.10}$$

where F_X is the distribution function of the random variable X. Technically speaking, the integral in (2.10) is a Lebesgue–Stieltjes integral. We do not require knowledge of such integrals in this text, but interpret (2.10) to signify (2.8) when X is a discrete random variable, and to represent (2.9) when X possesses a probability density f_X.

Let $F_Y(y) = \Pr\{Y \leq y\}$ denote the distribution function for $Y = g(X)$. When X is a discrete random variable, then

$$\begin{aligned} E[Y] &= \sum_j y_j \Pr\{Y = y_j\} \\ &= \sum_i g(x_i) \Pr\{X = x_i\} \end{aligned}$$

if $y_i = g(x_i)$ and provided that the second sum converges absolutely. In general,

$$\begin{aligned} E[Y] &= \int y\, dF_Y(y) \\ &= \int g(x)\, dF_X(x). \end{aligned} \tag{2.11}$$

If X is a discrete random variable, then so is $Y = g(X)$. It may be, however, that X is a continuous random variable while Y is discrete (the reader should provide an example). Even so, one may compute $E[Y]$ from either form in (2.11) with the same result.

2.4. Joint Distribution Functions

Given a pair (X, Y) of random variables, their *joint distribution function* is the function F_{XY} of two real variables given by

$$F_{XY}(x, y) = F(x, y) = \Pr\{X \leq x \text{ and } Y \leq y\}.$$

Usually, the subscripts X, Y will be omitted, unless ambiguity is possible. A joint distribution function F_{XY} is said to possess a (joint) probability density if there exists a function f_{XY} of two real variables for which

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(\xi, \eta)\, d\eta\, d\xi \qquad \text{for all } x, y.$$

The function $F_X(x) = \lim_{y\to\infty} F(x, y)$ is a distribution function, called the *marginal distribution function* of X. Similarly, $F_Y(y) = \lim_{y\to\infty} F(x, y)$ is the marginal distribution function of Y. If the distribution function F possesses the joint density function f, then the marginal density functions for X and Y are given, respectively, by

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)\, dx.$$

If X and Y are jointly distributed, then $E[X + Y] = E[X] + E[Y]$, provided only that all these moments exist.

Independence

If it happens that $F(x, y) = F_X(x) \times F_Y(y)$ for every choice of x, y, then the random variables X and Y are said to be *independent*. If X and Y are independent and possess a joint density function $f(x, y)$, then necessarily $f(x, y) = f_X(x)f_Y(y)$ for all x, y.

Given jointly distributed random variables X and Y having means μ_X and μ_Y and finite variances, the *covariance* of X and Y, written σ_{XY} or $\mathrm{Cov}[X, Y]$, is the product moment $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y$, and X and Y are said to be *uncorrelated* if their covariance is zero, that is, $\sigma_{XY} = 0$. Independent random variables having finite variances are uncorrelated, but the converse is not true; there are uncorrelated random variables that are not independent.

Dividing the covariance σ_{XY} by the standard deviations σ_X and σ_Y defines the *correlation coefficient* $\rho = \sigma_{XY}/\sigma_X\sigma_Y$ for which $-1 \leq \rho \leq +1$.

The joint distribution function of any finite collection $X_1, \ldots, X_n$ of random variables is defined as the function

$$F(x_1, \ldots, x_n) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$
$$= \Pr\{X_1 \leq x_1, \ldots, X_n \leq x_n\}.$$

If $F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$ for all values of $x_1, \ldots, x_n$, then the random variables $X_1, \ldots, X_n$ are said to be independent.

A joint distribution function $F(x_1, \ldots, x_n)$ is said to have a probability density function $f(\xi_1, \ldots, \xi_n)$ if

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(\xi_1, \ldots, \xi_n)\, d\xi_n \ldots d\xi_1,$$

for all values of $x_1, \ldots, x_n$.

Expectation

For jointly distributed random variables $X_1, \ldots, X_n$ and arbitrary functions $h_1, \ldots, h_m$ of n variables each,

$$E\left[\sum_{j=1}^{m} h_j(X_1, \ldots, X_n)\right] = \sum_{j=1}^{m} E[h_j(X_1, \ldots, X_n)],$$

provided only that all these moments exist.

2.5. Sums and Convolutions

If X and Y are independent random variables having distribution functions F_X and F_Y, respectively, then the distribution function of their sum $Z = X + Y$ is the *convolution* of F_X and F_Y:

$$F_Z(z) = \int_{-\infty}^{+\infty} F_X(z - \xi)\, dF_Y(\xi) = \int_{-\infty}^{+\infty} F_Y(z - \eta)\, dF_X(\eta). \quad (2.12)$$

If we specialize to the situation where X and Y have the probability densi-

ties f_X and f_Y, respectively, then the density function f_Z of the sum $Z = X + Y$ is the convolution of the densities f_X and f_Y:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - \eta) f_Y(\eta)\, d\eta = \int_{-\infty}^{+\infty} f_Y(z - \xi) f_X(\xi)\, d\xi. \quad (2.13)$$

Where X and Y are nonnegative random variables, the range of integration is correspondingly reduced to

$$f_Z(z) = \int_0^z f_X(z - \eta) f_Y(\eta)\, d\eta = \int_0^z f_Y(z - \xi) f_X(\xi)\, d\xi \qquad \text{for } z \geq 0. \quad (2.14)$$

If X and Y are independent and have respective variances σ_X^2 and σ_Y^2, then the variance of the sum $Z = X + Y$ is the sum of the variances: $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$. More generally, if $X_1, \ldots, X_n$ are independent random variables having variances $\sigma_1^2, \ldots, \sigma_n^2$, respectively, then the variance of the sum $Z = X_1 + \ldots + X_n$ is $\sigma_Z^2 = \sigma_1^2 + \ldots + \sigma_n^2$.

2.6. Change of Variable

Suppose that X is a random variable with probability density function f_X and that g is a strictly increasing differentiable function. Then $Y = g(X)$ defines a random variable, and the event $\{Y \leq y\}$ is the same as the event $\{X \leq g^{-1}(y)\}$, where g^{-1} is the inverse function to g; i.e., $y = g(x)$ if and only if $x = g^{-1}(y)$. Thus we obtain the correspondence $F_Y(y) = \Pr\{Y \leq y\} = \Pr\{X \leq g^{-1}(y)\} = F_X(g^{-1}(y))$ between the distribution function of Y and that of X. Recall the differential calculus formula

$$\frac{dg^{-1}}{dy} = \frac{1}{g'(x)} = \frac{1}{dg/dx}, \qquad \text{where } y = g(x),$$

and use this in the chain rule of differentiation to obtain

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(g^{-1}(y))}{dy} = f_X(x) \frac{1}{g'(x)}, \qquad \text{where } y = g(x).$$

The formula

$$f_Y(y) = \frac{1}{g'(x)} f_X(x), \qquad \text{where } y = g(x). \quad (2.15)$$

expresses the density function for Y in terms of the density for X when g is strictly increasing and differentiable.

2.7. Conditional Probability

For any events A and B, the *conditional probability* of A given B is written $\Pr\{A|B\}$ and defined by

$$\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \qquad \text{if } \Pr\{B\} > 0, \tag{2.16}$$

and is left undefined if $\Pr\{B\} = 0$. [When $\Pr\{B\} = 0$, the right side of (2.16) is the indeterminate quantity $\frac{0}{0}$.]

In stochastic modeling, conditional probabilities are rarely procured via (2.16) but instead are dictated as primary data by the circumstances of the application, and then (2.16) is applied in its equivalent multiplicative form

$$\Pr\{A \cap B\} = \Pr\{A|B\} \Pr\{B\} \tag{2.17}$$

to compute other probabilities. (An example follows shortly.) Central in this role is the *law of total probability*, which results from substituting $\Pr\{A \cap B_i\} = \Pr\{A|B_i\} \Pr\{B_i\}$ into $\Pr\{A\} = \sum_{i=1}^{\infty} \Pr\{A \cap B_i\}$, where $\Omega = B_1 \cup B_2 \cup \cdots$ and $B_i \cap B_j = \varnothing$ if $i \neq j$ (cf. Section 2.1), to yield

$$\Pr\{A\} = \sum_{i=1}^{\infty} \Pr\{A|B_i\} \Pr\{B_i\}. \tag{2.18}$$

Example Gold and silver coins are allocated among three urns labeled I, II, III according to the following table:

Urn	Number of Gold Coins	Number of Silver Coins
I	4	8
II	3	9
III	6	6

An urn is selected at random, all urns being equally likely, and then a coin is selected at random from that urn. Using the notation I, II, III for the events of selecting urns I, II, and III, respectively, and G for the event of selecting a gold coin, then the problem description provides the following probabilities and conditional probabilities as data:

$$\Pr\{\mathrm{I}\} = \tfrac{1}{3}, \qquad \Pr\{G|\mathrm{I}\} = \tfrac{4}{12},$$
$$\Pr\{\mathrm{II}\} = \tfrac{1}{3}, \qquad \Pr\{G|\mathrm{II}\} = \tfrac{3}{12},$$
$$\Pr\{\mathrm{III}\} = \tfrac{1}{3}, \qquad \Pr\{G|\mathrm{III}\} = \tfrac{6}{12},$$

and we *calculate* the probability of selecting a gold coin according to (2.18), viz.

$$\Pr\{G\} = \Pr\{G|\mathrm{I}\}\Pr\{\mathrm{I}\} + \Pr\{G|\mathrm{II}\}\Pr\{\mathrm{II}\} + \Pr\{G|\mathrm{III}\}\Pr\{\mathrm{III}\}$$
$$= \tfrac{4}{12}(\tfrac{1}{3}) + \tfrac{3}{12}(\tfrac{1}{3}) + \tfrac{6}{12}(\tfrac{1}{3}) = \tfrac{13}{36}.$$

As seen here, more often than not conditional probabilities are given as data and are not the end result of calculation.

Discussion of conditional distributions and conditional expectation merits an entire chapter (Chapter II).

2.8. Review of Axiomatic Probability Theory*

For the most part, this book studies random variables only through their distributions. In this spirit, we defined a random variable as a variable that takes on its values by chance. For some purposes, however, a little more precision and structure are needed.

Recall that the basic elements of probability theory are

1. the *sample space*, a set Ω whose elements ω correspond to the possible outcomes of an experiment;
2. the family of events, a collection $\mathcal{F}$ of subsets A of Ω: we say that the event A *occurs* if the outcome ω of the experiment is an element of A; and
3. the *probability measure*, a function P defined on $\mathcal{F}$ and satisfying

(a) $\quad 0 = P[\varnothing] \le P[A] \le P[\Omega] = 1 \qquad$ for $A \in \mathcal{F}$

($\varnothing$ = the empty set)

* The material included in this review of axiomatic probability theory is not used in the remainder of the book. It is included in this review chapter only for the sake of completeness.

Chapter II
Conditional Probability and Conditional Expectation

1. The Discrete Case

The conditional probability $\Pr\{A|B\}$ of the event A given the event B is defined by

$$\Pr\{A|B\} = \frac{\Pr\{A \text{ and } B\}}{\Pr\{B\}} \qquad \text{if } \Pr\{B\} > 0, \tag{1.1}$$

and is not defined, or is assigned an arbitrary value, when $\Pr\{B\} = 0$. Let X and Y be random variables that can attain only countably many different values, say 0, 1, 2, ... The *conditional probability mass function* $p_{X|Y}(x|y)$ of X given $Y = y$ is defined by

$$p_{X|Y}(x|y) = \frac{\Pr\{X = x \text{ and } Y = y\}}{\Pr\{Y = y\}} \qquad \text{if } \Pr\{Y = y\} > 0,$$

and is not defined, or is assigned an arbitrary value, whenever $\Pr\{Y = y\} = 0$. In terms of the joint and marginal probability mass functions $p_{XY}(x, y)$ and $p_Y(y) = \Sigma_x p_{XY}(x, y)$, respectively, the definition is

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)} \qquad \text{if } p_Y(y) > 0; \quad x, y = 0, 1, \ldots. \tag{1.2}$$

Observe that $p_{X|Y}(x|y)$ is a probability mass function in x for each fixed y, i.e., $p_{X|Y}(x|y) \geq 0$ and $\Sigma_\xi p_{X|Y}(\xi|y) = 1$, for all x, y.

The law of total probability takes the form

$$\Pr\{X = x\} = \sum_{y=0}^{\infty} p_{X|Y}(x|y)p_Y(y). \tag{1.3}$$

Notice in (1.3) that the points y where $p_{X|Y}(x|y)$ is not defined are exactly those values for which $p_Y(y) = 0$, and hence, do not affect the computation. The lack of a complete prescription for the conditional probability mass function, a nuisance in some instances, is always consistent with subsequent calculations.

Example Let X have a binomial distribution with parameters p and N, where N has a binomial distribution with parameters q and M. What is the marginal distribution of X?

We are given the conditional probability mass function

$$p_{X|N}(k|n) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

and the marginal distribution

$$p_N(n) = \binom{M}{n} q^n (1-q)^{M-n}, \qquad n = 0, 1, \ldots, M.$$

We apply the law of total probability in the form of (1.3) to obtain

$$\begin{aligned}
\Pr\{X = k\} &= \sum_{n=0}^{M} p_{X|N}(k|n)p_N(n) \\
&= \sum_{n=k}^{M} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \frac{M!}{n!(M-n)!} q^n (1-q)^{M-n} \\
&= \frac{M!}{k!} p^k (1-q)^M \left(\frac{q}{1-q}\right)^k \sum_{n=k}^{M} \frac{1}{(n-k)!(M-n)!} (1-p)^{n-k} \\
&\quad \times \left(\frac{q}{1-q}\right)^{n-k} \\
&= \frac{M!}{k!(M-k)!} (pq)^k (1-q)^{M-k} \left[1 + \frac{q(1-p)}{1-q}\right]^{M-k}
\end{aligned}$$

$$= \frac{M!}{k!(M-k)!}(pq)^k(1-pq)^{M-k}, \qquad k = 0, 1, \ldots, M.$$

In words, X has a binomial distribution with parameters M and pq.

Example Suppose X has a binomial distribution with parameters p and N where N has a Poisson distribution with mean λ. What is the marginal distribution for X?

Proceeding as in the previous example but now using

$$p_N(n) = \frac{\lambda^n e^{-\lambda}}{n!}, \qquad n = 0, 1, \ldots,$$

we obtain

$$\begin{aligned}
\Pr\{X = k\} &= \sum_{n=0}^{\infty} p_{X|N}(k|n)p_N(n) \\
&= \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}\frac{\lambda^n e^{-\lambda}}{n!} \\
&= \frac{\lambda^k e^{-\lambda}p^k}{k!}\sum_{n=k}^{\infty}\frac{[\lambda(1-p)]^{n-k}}{(n-k)!} \\
&= \frac{(\lambda p)^k e^{-\lambda}}{k!}e^{\lambda(1-p)} \\
&= \frac{(\lambda p)^k e^{-\lambda p}}{k!} \qquad \text{for } k = 0, 1, \ldots.
\end{aligned}$$

In words, X has a Poisson distribution with mean λp.

Example Suppose X has a negative binomial distribution with parameters p and N, where N has the geometric distribution

$$p_N(n) = (1-\beta)\beta^{n-1} \qquad \text{for } n = 1, 2, \ldots.$$

What is the marginal distribution for X?

We are given the conditional probability mass function

$$p_{X|N}(k|n) = \binom{n+k-1}{k}p^n(1-p)^k, \; k = 0, 1, \ldots.$$

Using the law of total probability, we obtain

$$\begin{aligned}\Pr\{X = k\} &= \sum_{n=0}^{\infty} p_{X|N}(k|n)p_N(n) \\ &= \sum_{n=1}^{\infty} \frac{(n+k-1)!}{k!(n-1)!} p^n(1-p)^k(1-\beta)\beta^{n-1} \\ &= (1-\beta)(1-p)^k p \sum_{n=1}^{\infty} \binom{n+k-1}{k} (\beta p)^{n-1} \\ &= (1-\beta)(1-p)^k p(1-\beta p)^{-k-1} \\ &= \left(\frac{p-\beta p}{1-\beta p}\right)\left(\frac{1-p}{1-\beta p}\right)^k \qquad \text{for } k = 0, 1, \ldots .\end{aligned}$$

We recognize the marginal distribution of X as being of geometric form.

Let g be a function for which the expectation of $g(X)$ is finite. We define the *conditional* expected value of $g(X)$ given $Y = y$ by the formula

$$E[g(X)|Y = y] = \sum_x g(x)p_{X|Y}(x|y) \qquad \text{if } p_Y(y) > 0, \tag{1.4}$$

and the conditional mean is not defined at values y for which $p_Y(y) = 0$. The law of total probability for conditional expectation reads

$$E[g(X)] = \sum_y E[g(X)|Y = y]p_Y(y). \tag{1.5}$$

The conditional expected value $E[g(X)|Y = y]$ is a function of the real variable y. If we evaluate this function at the random variable Y, we obtain a random variable that we denote by $E[g(X)|Y]$. The law of total probability in (1.5) now may be written in the form

$$E[g(X)] = E\{E[g(X)|Y]\}. \tag{1.6}$$

Since the conditional expectation of $g(X)$ given $Y = y$ is the expectation with respect to the conditional probability mass function $p_{X|Y}(x|y)$, conditional expectations behave in many ways like ordinary expectations. The following list summarizes some properties of conditional expectations. In

this list, with or without affixes, X and Y are jointly distributed random variables; c is a real number; g is a function for which $E[|g(X)|] < \infty$; h is a bounded function; and v is a function of two variables for which $E[|v(X, Y)|] < \infty$. The properties are

$$(1)\quad E[c_1 g_1(X_1) + c_2 g_2(X_2)|Y = y] = c_1 E[g_1(X_1)|Y = y] + c_2 E[g_2(X_2)|Y = y]. \tag{1.7}$$

$$(2)\quad \text{if } g \geq 0, \quad \text{then } E[g(X)|Y = y] \geq 0. \tag{1.8}$$

$$(3)\quad E[v(X, Y)|Y = y] = E[v(X, y)|Y = y]. \tag{1.9}$$

$$(4)\quad E[g(X)|Y = y] = E[g(X)] \quad \text{if } X \text{ and } Y \text{ are independent.} \tag{1.10}$$

$$(5)\quad E[g(X)h(Y)|Y = y] = h(y)E[g(X)|Y = y]. \tag{1.11}$$

$$(6)\quad E[g(X)h(Y)] = \sum_y h(y)E[g(X)|Y = y]p_Y(y) = E\{h(Y)E[g(X)|Y]\}. \tag{1.12}$$

As a consequence of (1.7), (1.11), and (1.12), with either $g \equiv 1$ or $h \equiv 1$, we obtain

$$E[c|Y = y] = c, \tag{1.13}$$

$$E[h(Y)|Y = y] = h(y), \tag{1.14}$$

$$E[g(X)] = \sum_y E[g(X)|Y = y]p_Y(y) = E\{E[g(X)|Y]\}. \tag{1.15}$$

Exercises

1.1. I roll a six-sided die and observe the number N on the uppermost face. I then toss a fair coin N times and observe X, the total number of heads to appear. What is the probability that $N = 3$ and $X = 2$? What is the probability that $X = 5$? What is $E[X]$, the expected number of heads to appear?

1.2. Four nickels and six dimes are tossed, and the total number N of heads is observed. If $N = 4$, what is the conditional probability that exactly two of the nickels were heads?

1.10. *Do men have more sisters than women have?* In a certain society, all married couples use the following strategy to determine the number of children that they will have: If the first child is a girl, they have no more children. If the first child is a boy, they have a second child. If the second child is a girl, they have no more children. If the second child is a boy, they have exactly one additional child. (We ignore twins, assume sexes are equally likely, and the sex of distinct children are independent random variables, etc.) (a) What is the probability distribution for the number of children in a family? (b) What is the probability distribution for the number of girl children in a family? (c) A male child is chosen at random from all of the male children in the population. What is the probability distribution for the number of sisters of this child? What is the probability distribution for the number of his brothers?

2. The Dice Game Craps

An analysis of the dice game known as craps provides an educational example of the use of conditional probability in stochastic modeling. In craps, two dice are rolled and the sum of their uppermost faces is observed. If the sum has value 2, 3, or 12, the player loses immediately. If the sum is 7 or 11, the player wins. If the sum is 4, 5, 6, 8, 9, or 10, then further rolls are required to resolve the game. In the case where the sum is 4, for example, the dice are rolled repeatedly until either a sum of 4 reappears or a sum of 7 is observed. If the 4 appears first, the roller wins; if the seven appears first, he or she loses.

Consider repeated rolls of the pair of dice and let Z_n for $n = 0, 1, \ldots$ be the sum observed on the nth roll. Then $Z_0, Z_1, \ldots$ are independent identically distributed random variables. If the dice are fair, the probability mass function is

$$\begin{aligned} p_Z(2) &= \tfrac{1}{36}, & p_Z(8) &= \tfrac{5}{36}, \\ p_Z(3) &= \tfrac{2}{36}, & p_Z(9) &= \tfrac{4}{36}, \\ p_Z(4) &= \tfrac{3}{36}, & p_Z(10) &= \tfrac{3}{36}, \\ p_Z(5) &= \tfrac{4}{36}, & p_Z(11) &= \tfrac{2}{36}, \\ p_Z(6) &= \tfrac{5}{36}, & p_Z(12) &= \tfrac{1}{36}. \\ p_Z(7) &= \tfrac{6}{36}, \end{aligned} \tag{2.1}$$

Let A denote the event that the player wins the game. By the law of total probability,

$$\Pr\{A\} = \sum_{k=2}^{12} \Pr\{A|Z_0 = k\} p_Z(k). \tag{2.2}$$

Because $Z_0 = 2$, 3, or 12 calls for an immediate loss, then $\Pr\{A|Z_0 = k\} = 0$ for $k = 2$, 3, or 12. Similarly, $Z_0 = 7$ or 11 results in an immediate win, and thus $\Pr\{A|Z_0 = 7\} = \Pr\{A|Z_0 = 11\} = 1$. It remains to consider the values $Z_0 = 4$, 5, 6, 8, 9, and 10, which call for additional rolls. Since the logic remains the same in each of these cases, we will argue only the case in which $Z_0 = 4$. Abbreviate with $\alpha = \Pr\{A|Z_0 = 4\}$. Then α is the probability that in successive rolls $Z_1, Z_2, \ldots$ of a pair of dice, a sum of 4 appears before a sum of 7. Denote this event by B, and again bring in the law of total probability. Then

$$\alpha = \Pr\{B\} = \sum_{k=2}^{12} \Pr\{B|Z_1 = k\} p_Z(k). \tag{2.3}$$

Now $\Pr\{B|Z_1 = 4\} = 1$, while $\Pr\{B|Z_1 = 7\} = 0$. If the first roll results in anything other than a 4 or a 7, the problem is repeated in a statistically identical setting. That is, $\Pr\{B|Z_1 = k\} = \alpha$ for $k \neq 4$ or 7. Substitution into (2.3) results in

$$\begin{aligned} \alpha &= p_Z(4) \times 1 + p_Z(7) \times 0 + \sum_{k \neq 4,7} p_Z(k) \times \alpha \\ &= p_Z(4) + [1 - p_Z(4) - p_Z(7)]\alpha, \end{aligned}$$

or

$$\alpha = \frac{p_Z(4)}{p_Z(4) + p_Z(7)}. \tag{2.4}$$

The same result may be secured by means of a longer, more computational, method. One may partition the event B into disjoint elemental events by writing

$$\begin{aligned} B = &\{Z_1 = 4\} \cup \{Z_1 \neq 4 \text{ or } 7, Z_2 = 4\} \\ &\cup \{Z_1 \neq 4 \text{ or } 7, Z_2 \neq 4 \text{ or } 7, Z_3 = 4\} \cup \cdots, \end{aligned}$$

and then

$$\Pr\{B\} = \Pr\{Z_1 = 4\} + \Pr\{Z_1 \neq 4 \text{ or } 7, Z_2 = 4\} + \Pr\{Z_1 \neq 4 \text{ or } 7, Z_2 \neq 4 \text{ or } 7, Z_3 = 4\} + \cdots.$$

Now use the independence of $Z_1, Z_2, \ldots$ and sum a geometric series to secure

$$\begin{aligned}\Pr\{B\} &= p_Z(4) + [1 - p_Z(4) - p_Z(7)]p_Z(4) \\ &\quad + [1 - p_Z(4) - p_Z(7)]^2 p_Z(4) + \cdots \\ &= \frac{p_Z(4)}{p_Z(4) + p_Z(7)}\end{aligned}$$

in agreement with (2.4).

Extending the result just obtained to the other cases having more than one roll, we have

$$\Pr\{A|Z_0 = k\} = \frac{p_Z(k)}{p_Z(k) + p_Z(7)} \qquad \text{for } k = 4, 5, 6, 8, 9, 10.$$

Finally, substitution into (2.2) yields the total win probability

$$\Pr\{A\} = p_Z(7) + p_Z(11) + \sum_{k=4,5,6,8,9,10} \frac{p_Z(k)^2}{p_Z(k) + p_Z(7)}. \tag{2.5}$$

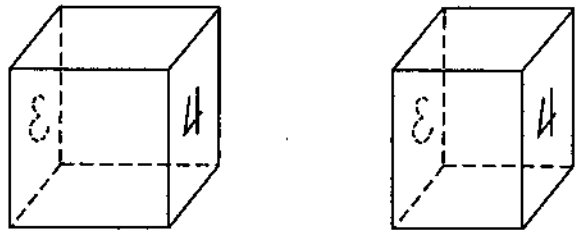
The numerical values for $p_Z(k)$ given in (2.1), together with (2.5), determine the win probability

$$\Pr\{A\} = 0.49292929\cdots.$$

Having explained the computations, let us go on to a more interesting question. Suppose that the dice are not perfect cubes but are shaved so as to be slightly thinner in one dimension than in the other two. The numbers that appear on opposite faces on a single die always sum to 7. That is, 1 is opposite 6, 2 is opposite 5, and 3 is opposite 4. Suppose it is the 3-4 dimension that is smaller than the other two. See Figure 2.1. This will cause 3 and 4 to appear more frequently than the other faces, 1, 2, 5, and 6. To see this, think of the extreme case in which the 3–4 dimension is very thin, leading to a 3 or 4 on almost all tosses. Letting Y denote the result of tossing a single shaved die, we postulate that the probability mass function is given by

$$p_Y(3) = p_Y(4) = \tfrac{1}{6} + 2\varepsilon \equiv p_+,$$
$$p_Y(1) = p_Y(2) = p_Y(5) = p_Y(6) = \tfrac{1}{6} - \varepsilon \equiv p_-,$$

where $\varepsilon > 0$ is a small quantity depending on the amount by which the die has been biased.



A Cubic Die A Shaved Die

Figure 2.1 A cubic die versus a die that has been shaved down in one dimension.

If both dice are shaved in the same manner, the mass function for their sum can be determined in a straightforward manner from the following joint table:

		Die #1				
Die #2	1 p_-	2 p_-	3 p_+	4 p_+	5 p_-	6 p_-
1 p_-	p_-^2	p_-^2	p_+p_-	p_+p_-	p_-^2	p_-^2
2 p_-	p_-^2	p_-^2	p_+p_-	p_+p_-	p_-^2	p_-^2
3 p_+	p_+p_-	p_+p_-	p_+^2	p_+^2	p_+p_-	p_+p_-
4 p_+	p_+p_-	p_+p_-	p_+^2	p_+^2	p_+p_-	p_+p_-
5 p_-	p_-^2	p_-^2	p_+p_-	p_+p_-	p_-^2	p_-^2
6 p_-	p_-^2	p_-^2	p_+p_-	p_+p_-	p_-^2	p_-^2

It is easily seen that the probability mass function for the sum of the dice is

$$
\begin{aligned}
p(2) &= p_-^2 = p(12),\\
p(3) &= 2p_-^2 = p(11),\\
p(4) &= p_-(p_- + 2p_+) = p(10),\\
p(5) &= 4p_+p_- = p(9),\\
p(6) &= p_-^2 + (p_+ + p_-)^2 = p(8),\\
p(7) &= 4p_-^2 + 2p_+^2.
\end{aligned}
$$

To obtain a numerical value to compare to the win probability $0.492929\cdots$ associated with fair dice, let us arbitrarily set $\varepsilon = 0.02$, so that $p_- = 0.146666\cdots$ and $p_+ = 0.206666\cdots$. Then routine substitutions according to the table lead to

$$
\begin{aligned}
p(2) = p(12) &= 0.02151111, \quad p(5) = p(9) = 0.12124445,\\
p(3) = p(11) &= 0.04302222, \quad p(6) = p(8) = 0.14635556,\\
p(4) = p(10) &= 0.08213333, \quad p(7) = 0.17146667,
\end{aligned}
\tag{2.6}
$$

and the win probability becomes $\Pr\{A\} = 0.5029237$.

The win probability of 0.4929293 with fair dice is unfavorable, that is, is less than $\frac{1}{2}$. With shaved dice, the win probability is favorable, now being 0.5029237. What appears to be a slight change becomes, in fact, quite significant when a large number of games are played. See III, Section 5.

Exercises

2.1. A red die is rolled a single time. A green die is rolled repeatedly. The game stops the first time that the sum of the two dice is either 4 or 7. What is the probability that the game stops with a sum of 4?

2.2. Verify the win probability of 0.5029237 by substituting from (2.6) into (2.5).

2.3. Determine the win probability when the dice are shaved on the 1–6 faces and $p_+ = 0.206666\cdots$ and $p_- = 0.146666\cdots$.

5. Martingales*

Stochastic processes are characterized by the dependence relationships that exist among their variables. The martingale property is one such relationship that captures a notion of a game being fair. The martingale property is a restriction solely on the conditional means of some of the variables, given values of others, and does not otherwise depend on the actual distribution of the random variables in the stochastic process. Despite the apparent weakness of the martingale assumption, the consequences are striking, as we hope to suggest.

5.1. The Definition

We begin the presentation with the simplest definition.

Definition A stochastic process $\{X_n; n = 0, 1, \ldots\}$ is a martingale if for $n = 0, 1, \ldots,$

(a) $E[|X_n|] < \infty,$

and

(b) $E[X_{n+1}|X_0, \ldots, X_n] = X_n.$

Taking expectations on both sides of (b),

$$E\{E[X_{n+1}|X_0, \ldots, X_n]\} = E\{X_n\},$$

and using the law of total probability in the form

$$E\{E[X_{n+1}|X_1, \ldots, X_n]\} = E[X_{n+1}]$$

shows that

$$E[X_{n+1}] = E[X_n],$$

and consequently, a martingale has constant mean:

$$E[X_0] = E[X_k] = E[X_n], \qquad 0 \le k \le n. \tag{5.1}$$

* Some problems scattered throughout the text call for the student to identify certain stochastic processes as martingales. Otherwise, the material of this section is not used in the sequel.

A similar conditioning (see Problem 5.1) verifies that the martingale equality (b) extends to future times in the form

$$E[X_m|X_0, \ldots, X_n] = X_n \qquad \text{for } m \geq n. \tag{5.2}$$

To relate the martingale property to concepts of fairness in gambling, consider X_n to be a certain player's fortune after the nth play of a game. The game is "fair" if on average, the player's fortune neither increases nor decreases at each play. The martingale property (b) requires the player's fortune after the next play to equal, on average, his current fortune and not be otherwise affected by previous history. Some early work in martingale theory was motivated in part by problems in gambling. For example, *martingale systems theorems* consider whether an astute choice of betting strategy can turn a fair game into a favorable one, and the name "martingale" derives from a French term for the particular strategy of doubling one's bets until a win is secured. While it remains popular to illustrate martingale concepts with gambling examples, today, martingale theory has such broad scope and diverse applications that to think of it purely in terms of gambling would be unduly restrictive and misleading.

Example *Stock Prices in a Perfect Market* Let X_n be the closing price at the end of day n of a certain publicly traded security such as a share of stock. While daily prices may fluctuate, many scholars believe that, in a perfect market, these price sequences should be martingales. In a perfect market freely open to all, they argue, it should not be possible to predict with any degree of accuracy whether a future price X_{n+1} will be higher or lower than the current price X_n. For example, if a future price could be expected to be higher, then a number of buyers would enter the market, and their demand would raise the current price X_n. Similarly, if a future price could be predicted as lower, a number of sellers would appear and tend to depress the current price. Equilibrium obtains where the future price cannot be predicted, on average, as higher or lower, that is, where price sequences are martingales.

5.2. The Markov Inequality

What does the mean of a random variable tell us about its distribution? For a nonnegative random variable X, Markov's inequality is $\lambda \Pr\{X \geq \lambda\} \leq E[X]$, for any positive constant λ. For example, if $E[X] = 1$, then

$\Pr\{X \geq 4\} \leq \frac{1}{4}$, no matter what the actual distribution of X is. The proof uses two properties: (i) $X \geq 0$ (X is a nonnegative random variable); and (ii) $E[X\mathbf{1}\{X \geq \lambda\}] \geq \lambda \Pr\{X \geq \lambda\}$. (Recall that $\mathbf{1}(A)$ is the *indicator* of an event A and is one if A occurs and zero otherwise. See I, Section 3.1.) Then by the law of total probability,

$$\begin{aligned} E[X] &= E[X\mathbf{1}\{X \geq \lambda\}] + E[X\mathbf{1}\{X < \lambda\}] \\ &\geq E[X\mathbf{1}\{X \geq \lambda\}] \qquad \text{(by (i))} \\ &\geq \lambda \Pr\{X \geq \lambda\}, \qquad \text{(by (ii))} \end{aligned}$$

and Markov's inequality results.

5.3. The Maximal Inequality for Nonnegative Martingales

Because a martingale has constant mean, Markov's inequality applied to a nonnegative martingale immediately yields

$$\Pr\{X_n \geq \lambda\} \leq \frac{E[X_0]}{\lambda}, \lambda > 0.$$

We will extend the reasoning behind Markov's inequality to achieve an inequality of far greater power:

$$\Pr\{\max_{0 \leq n \leq m} X_n \geq \lambda\} \leq \frac{E[X_0]}{\lambda}. \tag{5.3}$$

Instead of limiting the probability of a large value for a single observation X_n, the *maximal inequality* (5.3) limits the probability of observing a large value anywhere in the time interval $0, \ldots, m$, and since the right side of (5.3) does not depend on the length of the interval, the maximal inequality limits the probability of observing a large value at any time in the infinite future of the martingale!

In order to prove the maximal inequality for nonnegative martingales, we need but a single additional fact: If X and Y are jointly distributed random variables and B is an arbitrary set, then

$$E[X\mathbf{1}\{Y \text{ in } B\}] = E[E\{X|Y\}\mathbf{1}\{Y \text{ in } B\}]. \tag{5.4}$$

But (5.4) follows from the conditional expectation property (1.12), $E[g(X)h(Y)] = E\{h(Y)E[g(X)|Y]\}$, with $g(x) = x$ and $h(y) = \mathbf{1}(y \text{ in } B\}$.

We will have need of (5.4) with $X = X_m$ and $Y = (X_0, \ldots, X_n)$, whereupon (5.4) followed by (5.2) then justifies

$$\begin{aligned} E[X_m \mathbf{1}\{X_0 < \lambda, \ldots, X_{n-1} < \lambda, X_n \geq \lambda\}] \\ &= E[E\{X_m | X_0, \ldots, X_n\}\, \mathbf{1}\{X_0 < \lambda, \ldots, X_{n-1} < \lambda, X_n \geq \lambda\}] \quad (5.5) \\ &= E[X_n \mathbf{1}\{X_0 < \lambda, \ldots, X_{n-1} < \lambda, X_n \geq \lambda\}]. \end{aligned}$$

Theorem 5.1 The maximal inequality for nonnegative martingales Let $X_0, X_1, \ldots$ *be a martingale with nonnegative values;* i.e., $\Pr\{X_n \geq 0\} = 1$ for $n = 0, 1, \ldots$. For any $\lambda > 0$,

$$\Pr\{\max_{0 \leq n \leq m} X_n \geq \lambda\} \leq \frac{E[X_0]}{\lambda}, \qquad \text{for } 0 \leq n \leq m \qquad (5.6)$$

and

$$\Pr\{\max_{n \geq 0} X_n > \lambda\} \leq \frac{E[X_0]}{\lambda}, \qquad \text{for all } n. \qquad (5.7)$$

Proof Inequality (5.7) follows from (5.6) because the right side of (5.6) does not depend on m. We begin with the law of total probability, as in I, Section 2.1. Either the $\{X_0, \ldots, X_m\}$ sequence rises above λ for the first time at some index n, or else it remains always below λ. As these possibilities are mutually exclusive and exhaustive, we apply the law of total probability to obtain

$$\begin{aligned} E[X_m] &= \sum_{n=1}^{m} E[X_m \mathbf{1}\{X_0 < \lambda, \ldots, X_{n-1} < \lambda, X_n \geq \lambda\} \\ &\quad + E[X_m \mathbf{1}\{X_0 < \lambda, \ldots, X_m < \lambda\}] \\ &\geq \sum_{n=1}^{m} E[X_m \mathbf{1}\{X_0 < \lambda, \ldots, X_{n-1} < \lambda, X_n \geq \lambda\}] \qquad (X_m \geq 0) \\ &= \sum_{n=1}^{m} E[X_n \mathbf{1}\{X_0 < \lambda, \ldots, X_{n-1} < \lambda, X_n \geq \lambda\}] \qquad (\text{using } 5.5)) \\ &\geq \lambda \sum_{n=0}^{m} \Pr\{X_0 < \lambda, \ldots, X_{n-1} < \lambda, X_n \geq \lambda\} \\ &= \lambda \Pr\{\max_{0 \leq n \leq m} X_n \geq \lambda\}. \end{aligned}$$

Example A gambler begins with a unit amount of money and faces a series of independent fair games. Beginning with $X_0 = 1$, the gambler bets

the amount p, $0 < p < 1$. If the first game is a win, which occurs with probability $\frac{1}{2}$, the gambler's fortune is $X_1 = 1 + pX_0 = 1 + p$. If the first game is a loss, then $X_1 = 1 - pX_0 = 1 - p$. After the nth play and with a current fortune of X_n, the gambler wagers pX_n, and

$$X_{n+1} = \begin{cases} (1+p)X_n & \text{with probability } \frac{1}{2}, \\ (1-p)X_n & \text{with probability } \frac{1}{2}. \end{cases}$$

Then $\{X_n\}$ is a nonnegative martingale, and the maximal inequality (5.6) with $\lambda = 2$, for example, asserts that *the probability that the gambler ever doubles his money is less than or equal to $\frac{1}{2}$, and this holds no matter what the game is, as long as it is fair, and no matter what fraction p of his fortune is wagered at each play.* Indeed, the fraction wagered may vary from play to play, as long as it is chosen without knowledge of the next outcome.

As amply demonstrated by this example, the maximal inequality is a very strong statement. Indeed, more elaborate arguments based on the maximal and other related martingale inequalities are used to show that a nonnegative martingale converges: If $\{X_n\}$ is a nonnegative martingale, then there exists a random variable, let us call it X_∞, for which $\lim_{n\to\infty} X_n = X_\infty$. We cannot guarantee the equality of the expectations in the limit, but the inequality $E[X_0] \geq E[X_\infty] \geq 0$ can be established.

Example In III, Section 8, we will introduce the branching process model for population growth. In this model, X_n is the number of individuals in the population in the nth generation, and $\mu > 0$ is the mean family size, or expected number of offspring of any single individual. The mean population size in the nth generation is $X_0\mu^n$. In this branching process model, X_n/μ^n is a nonnegative martingale (see III, Problem 8.4), and the maximal inequality implies that the probability of the actual population ever exceeding ten times the mean size is less than or equal to 1/10. The nonnegative martingale convergence theorem asserts that the evolution of such a population after many generations may be described by a single random variable X_∞ in the form

$$X_n \approx X_\infty \mu^n, \qquad \text{for large } n.$$

Example *How NOT to generate a uniformly distributed random variable* An urn initially contains one red and one green ball. A ball is drawn at random and it is returned to the urn, together with another ball of the

same color. This process is repeated indefinitely. After the nth play there will be a total of $n + 2$ balls in the urn. Let R_n be the number of these balls that are red, and $X_n = R_n/(n + 2)$ the fraction of red balls. We claim that $\{X_n\}$ is a martingale. First, observe that

$$R_{n+1} = \begin{cases} R_n + 1 & \text{with probability } X_n, \\ R_n & \text{with probability } 1 - X_n, \end{cases}$$

so that

$$E[R_{n+1}|X_n] = R_n + X_n = X_n(2 + n + 1),$$

and finally,

$$E[X_{n+1}|X_n] = \frac{1}{n+3} E[R_{n+1}|X_n] = \frac{2 + n + 1}{n + 3} X_n = X_n.$$

This verifies the martingale property, and because such a fraction is always nonnegative, indeed, between 0 and 1, there must be a random variable X_∞ to which the martingale converges. We will derive the probability distribution of the random limit. It is immediate that R_1 is equally likely to be 1 or 2, since the first ball chosen is equally likely to be red or green. Continuing,

$$\begin{aligned} \Pr\{R_2 = 3\} &= \Pr\{R_2 = 3|R_1 = 2\}\Pr\{R_1 = 2\} \\ &= (\tfrac{2}{3})(\tfrac{1}{2}) = \tfrac{1}{3}; \end{aligned}$$

$$\begin{aligned} \Pr\{R_2 = 2\} &= \Pr\{R_2 = 2|R_1 = 1\}\Pr\{R_1 = 1\} \\ &\quad + \Pr\{R_2 = 2|R_1 = 2\}\Pr\{R_1 = 2\} \\ &= (\tfrac{1}{3})(\tfrac{1}{2}) + (\tfrac{1}{3})(\tfrac{1}{2}) = \tfrac{1}{3}; \end{aligned}$$

and since the probabilities must sum to one,

$$\Pr\{R_2 = 1\} = \tfrac{1}{3}.$$

By repeating these simple calculations, it is easy to see that

$$\Pr\{R_n = k\} = \frac{1}{n+1} \qquad \text{for } k = 1, 2, \ldots, n + 1,$$

and that therefore X_n is uniformly distributed over the values $1/(n + 2)$,

Exercises

2.1. Let A and B be arbitrary, not necessarily disjoint, events. Use the law of total probability to verify the formula

$$\Pr\{A\} = \Pr\{AB\} + \Pr\{AB^c\},$$

where B^c is the complementary event to B. (That is, B^c occurs if and only if B does not occur.)

2.3.

(a) Plot the distribution function

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ x^3 & \text{for } 0 < x < 1, \\ 1 & \text{for } x \geq 1. \end{cases}$$

(b) Determine the corresponding density function $f(x)$ in the three regions (i) $x \leq 0$, (ii) $0 < x < 1$, and (iii) $1 \leq x$.

(c) What is the mean of the distribution?

(d) If X is a random variable following the distribution specified in (a), evaluate $\Pr\{\frac{1}{4} \leq X \leq \frac{3}{4}\}$.

2.5. Let A, B, and C be arbitrary events. Establish the addition law

$$\Pr\{A \cup B \cup C\} = \Pr\{A\} + \Pr\{B\} + \Pr\{C\} - \Pr\{AB\} - \Pr\{AC\} - \Pr\{BC\} + \Pr\{ABC\}.$$

2.10. Let $\mathbf{1}\{A\}$ be the indicator random variable associated with an event A, defined to be one if A occurs, and zero otherwise. Define A^c, the complement of event A, to be the event that occurs when A does not occur. Show

(a) $\mathbf{1}\{A^c\} = 1 - \mathbf{1}\{A\}$.

(b) $\mathbf{1}\{A \cap B\} = \mathbf{1}\{A\}\mathbf{1}\{B\} = \min\{\mathbf{1}\{A\}, \mathbf{1}\{B\}\}$.

(c) $\mathbf{1}\{A \cup B\} = \max\{\mathbf{1}\{A\}, \mathbf{1}\{B\}\}$.

Problems

2.12. Let U, V, and W be independent random variables with equal variances σ^2. Define $X = U + W$ and $Y = V - W$. Find the covariance between X and Y.

Exercises

1.2. Four nickels and six dimes are tossed, and the total number N of heads is observed. If $N = 4$, what is the conditional probability that exactly two of the nickels were heads?

1.6. Suppose U and V are independent and follow the geometric distribution

$$p(k) = \rho(1 - \rho)^k \qquad \text{for } k = 0, 1, \ldots.$$

Define the random variable $Z = U + V$.

(a) Determine the joint probability mass function $p_{U,Z}(u, z) = \Pr\{U = u, Z = z\}$.
(b) Determine the conditional probability mass function for U given that $Z = n$.

Problems

1.6. A dime is tossed repeatedly until a head appears. Let N be the trial number on which this first head occurs. Then a nickel is tossed N times. Let X count the number of times that the nickel comes up tails. Determine $\Pr\{X = 0\}$, $\Pr\{X = 1\}$, and $E[X]$.

1.10. *Do men have more sisters than women have?* In a certain society, all married couples use the following strategy to determine the number of children that they will have: If the first child is a girl, they have no more children. If the first child is a boy, they have a second child. If the second child is a girl, they have no more children. If the second child is a boy, they have exactly one additional child. (We ignore twins, assume sexes are equally likely, and the sex of distinct children are independent random variables, etc.) (a) What is the probability distribution for the number of children in a family? (b) What is the probability distribution for the number of girl children in a family? (c) A male child is chosen at random from all of the male children in the population. What is the probability distribution for the number of sisters of this child? What is the probability distribution for the number of his brothers?