1 Introduction

1.1 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, e.g., 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, such 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 a Greek word ($\sigma\tau o\chi\acute{\alpha}\zeta\epsilon\sigma\theta\alpha\iota$: 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

An Introduction to Stochastic Modeling. DOI: 10.1016/B978-0-12-381416-6.00001-0

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; i.e., 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: (1) a natural phenomenon under study, (2) a logical system for deducing implications about the phenomenon, and (3) 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 1.2.8 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 (1) the principle of equally likely outcomes, (2) the principle of long run relative frequency, and (3) 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% of similar patients facing similar surgery will survive it. The patient at hand is not concerned with the long run, but in vivid contrast, he 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 do not 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, probabilities 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 1.2.8), 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.

1.1.1 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 indicate 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 by 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.

1.2 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.

1.2.1 Events and Probabilities

The reader is assumed to be familiar with the intuitive concept of an *event*. (Events are defined rigorously in Section 1.2.8, 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

* Many readers will prefer to omit this review and move directly to Chapter 3, on Markov chains. They can then refer to the background material that is summarized in the remainder of this chapter and in Chapter 2 only as needed.

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 $\emptyset$, never occurs, and $\Pr\{\emptyset\} = 0$. It is always the case that $0 \leq \Pr\{A\} \leq 1$ for any event A.

Events A and B are said to be *disjoint* if $A \cap B = \emptyset$, i.e., if A and B both cannot occur. For disjoint events A and 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, $\Pr\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \Sigma_{i=1}^{\infty} \Pr\{A_i\}$. The addition law leads directly to the *law of total probability*: Let $A_1, A_2, \ldots$ be disjoint events for which $\Omega = A_1 \cup A_2 \cup \cdots$. Equivalently, exactly one of the events $A_1, A_2, \ldots$ will occur. The law of total probability asserts that $\Pr\{B\} = \Sigma_{i=1}^{\infty} \Pr\{B \cap A_i\}$ 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

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

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

1.2.2 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 1.2.8, 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, e.g., $F_X(\xi) = \Pr\{X \leq \xi\}$ 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

$$p(x_i) = p_X(x_i) = a_i \quad \text{for } i = 1, 2, \ldots \tag{1.1}$$

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

$$\Pr\{a < X \leq b\} = \int_a^b f(x)\mathrm{d}x \quad \text{for} -\infty < a < b < \infty, \tag{1.2}$$

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

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

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

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x) = F'(x), \quad -\infty < x < \infty. \tag{1.3}$$

In differential form, (1.3) leads to the informal statement

$$\Pr\{x < X \leq x + \mathrm{d}x\} = F(x + \mathrm{d}x) - F(x) = \mathrm{d}F(x) = f(x)\mathrm{d}x. \tag{1.4}$$

We consider (1.4) to be a shorthand version of the more precise statement

$$\Pr\{x < X \leq x + \Delta x\} = f(x)\Delta x + o(\Delta x), \quad \Delta x \downarrow 0, \tag{1.5}$$

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 (1.5) 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.

1.2.3 Moments and Expected Values

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

$$E[X^m] = \sum_i x_i^m \Pr\{X = x_i\} \tag{1.6}$$

[where the x_i are specified in (1.1)], 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

$$E[X^m] = \int_{-\infty}^{+\infty} x^m f(x)\mathrm{d}x, \tag{1.7}$$

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 Var[X]. We have the equivalent formulas $\mathrm{Var}[X] = E\left[(X-\mu)^2\right] = E\left[X^2\right] - \mu^2$.

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

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

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

$$E[g(X)] = \sum_{i=1}^{\infty} g(x_i) \Pr\{X = x_i\}, \tag{1.8}$$

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

$$E[g(X)] = \int g(x) f_X(x)\mathrm{d}x. \tag{1.9}$$

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

$$E[g(X)] = \int g(x)\mathrm{d}F_X(x), \tag{1.10}$$

where F_X is the distribution function of the random variable X. Technically speaking, the integral in (1.10) is a Lebesgue–Stieltjes integral. We do not require knowledge of such integrals in this text, but interpret (1.10) to signify (1.8) when X is a discrete random variable, and to represent (1.9) when X possesses a probability density f_X.

Let $F_Y(y) = \Pr\{Y \le y\}$ 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,

$$\begin{aligned} E[Y] &= \int y \, \mathrm{d}F_Y(y) \\ &= \int g(x) \mathrm{d}F_X(x). \end{aligned} \tag{1.11}$$

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 (1.11) with the same result.

1.2.4 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

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

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

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

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

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

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, i.e., $\sigma_{XY} = 0$. 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 $-1 \le \rho \le +1$.

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

$$\begin{aligned} F(x_1, \ldots, x_n) &= F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \\ &= \Pr\{X_1 \le x_1, \ldots, X_n \le x_n\}. \end{aligned}$$

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

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

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.

1.2.5 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:

$$F_Z(z) = \int_{-\infty}^{+\infty} F_X(z-\xi)\mathrm{d}F_Y(\xi) = \int_{-\infty}^{+\infty} F_Y(z-\eta)\mathrm{d}F_X(\eta). \tag{1.12}$$

If we specialize to the situation where X and Y have the probability densities 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:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-\eta)f_Y(\eta)\mathrm{d}\eta = \int_{-\infty}^{+\infty} f_Y(z-\xi)f_X(\xi)\mathrm{d}\xi. \tag{1.13}$$

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

$$f_Z(z) = \int_{0}^{z} f_X(z-\eta)f_Y(\eta)\mathrm{d}\eta = \int_{0}^{z} f_Y(z-\xi)f_X(\xi)\mathrm{d}\xi \quad \text{for } z \geq 0. \tag{1.14}$$

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 $Z = X_1 + \cdots + X_n$ is $\sigma_Z^2 = \sigma_1^2 + \cdots + \sigma_n^2$.

1.2.6 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

$$\frac{\mathrm{d}g^{-1}}{\mathrm{d}y} = \frac{1}{g'(x)} = \frac{1}{\mathrm{d}g/\mathrm{d}x}, \quad \text{where } y = g(x),$$

and use this in the chain rule of differentiation to obtain

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{\mathrm{d}F_X(g^{-1}(y))}{\mathrm{d}y} = f_X(x)\frac{1}{g'(x)}, \quad \text{where } y = g(x).$$

The formula

$$f_Y(y) = \frac{1}{g'(x)} f_X(x), \quad \text{where } y = g(x), \tag{1.15}$$

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

1.2.7 Conditional Probability

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

$$\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \quad \text{if } \Pr\{B\} > 0, \tag{1.16}$$

and is left undefined if $\Pr\{B\} = 0$. [When $\Pr\{B\} = 0$, the right side of (1.16) is the indeterminate quantity $\frac{0}{0}$.]

In stochastic modeling, conditional probabilities are rarely procured via (1.16) but instead are dictated as primary data by the circumstances of the application, and then (1.16) is applied in its equivalent multiplicative form

$$\Pr\{A \cap B\} = \Pr\{A|B\} \Pr\{B\} \tag{1.17}$$

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 $\Pr\{A\} = \Sigma_{i=1}^{\infty} \Pr\{A \cap B_i\}$, where $\Omega = B_1 \cup B_2 \cup \cdots$ and $B_i \cap B_j = \emptyset$ if $i \neq j$ (see Section 1.2.1), to yield

$$\Pr\{A\} = \sum_{i=1}^{\infty} \Pr\{A|B_i\} \Pr\{B_i\}. \tag{1.18}$$

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:

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

and we *calculate* the probability of selecting a gold coin according to (1.18), via

$$\begin{aligned}\Pr\{G\} &= \Pr\{G|\mathrm{I}\}\Pr\{\mathrm{I}\} + \Pr\{G|\mathrm{II}\}\Pr\{\mathrm{II}\} + \Pr\{G|\mathrm{III}\}\Pr\{\mathrm{III}\} \\ &= \frac{4}{12}\left(\frac{1}{3}\right) + \frac{3}{12}\left(\frac{1}{3}\right) + \frac{6}{12}\left(\frac{1}{3}\right) = \frac{13}{36}.\end{aligned}$$

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 2).

1.2.8 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)

$$0 = P[\emptyset] \le P[A] \le P[\Omega] = 1 \quad \text{for } A \in \mathcal{F}$$
$$(\emptyset = \text{the empty set})$$

and

(b)

$$P\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} P[A_n], \tag{1.19}$$

* 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.

if the events $A_1, A_2, \ldots$ are disjoint, i.e., if $A_i \cap A_j = \emptyset$ when $i \neq j$. The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Example When there are only a denumerable number of possible outcomes, say $\Omega = \{\omega_1, \omega_2, \ldots\}$, we may take $\mathcal{F}$ to be the collection of all subsets of Ω. If $p_1, p_2, \ldots$ are nonnegative numbers with $\Sigma_n p_n = 1$, the assignment

$$P[A] = \sum_{\omega_i \in A} p_i$$

determines a probability measure defined on $\mathcal{F}$.

It is not always desirable, consistent, or feasible to take the family of events as the collection of *all* subsets of Ω. Indeed, when Ω is nondenumerably infinite, it may not be possible to define a probability measure on the collection of all subsets maintaining the properties of (1.19). In whatever way we prescribe $\mathcal{F}$ such that (1.19) holds, the family of events $\mathcal{F}$ should satisfy

(a) $\emptyset$ is in $\mathcal{F}$ and Ω is in $\mathcal{F}$;
(b) A^c is in $\mathcal{F}$ whenever A is in $\mathcal{F}$, where $A^c = \{\omega \in \Omega; \omega \notin A\}$ is the complement of A; and
(c) $\bigcup_{n=1}^{\infty} A_n$ is in $\mathcal{F}$ whenever A_n is in $\mathcal{F}$ for $n = 1, 2, \ldots$. (1.20)

A collection $\mathcal{F}$ of subsets of a set Ω satisfying (1.20) is called a σ-*algebra*. If $\mathcal{F}$ is a σ-algebra, then

$$\bigcap_{n=1}^{\infty} A_n = \left(\bigcup_{n=1}^{\infty} A_n^c \right)^c$$

is in $\mathcal{F}$ whenever A_n is in $\mathcal{F}$ for $n = 1, 2, \ldots$. Manifestly, as a consequence, we find that finite unions and finite intersections of members of $\mathcal{F}$ are maintained in $\mathcal{F}$.

In this framework, a real random variable X is a real-valued function defined on Ω fulfilling certain "measurability" conditions given here. The distribution function of the random variable X is formally given by

$$\Pr\{a < X \leq b\} = \mathrm{P}[\{\omega; a < X(\omega) \leq b\}]. \tag{1.21}$$

In words, the probability that the random variable X takes a value in $(a, b]$ is calculated as the probability of the set of outcomes ω for which $a < X(\omega) \leq b$. If relation (1.21) is to have meaning, X cannot be an arbitrary function on Ω, but must satisfy the condition that

$\{\omega; a < X(\omega) \leq b\}$ is in $\mathcal{F}$ for all real $a < b$,

since $\mathcal{F}$ embodies the only sets A for which $P[A]$ is defined. In fact, by exploiting the properties (1.20) of the σ-algebra $\mathcal{F}$, we find that it is enough to require

$$\{\omega; X(\omega) \leq x\} \text{ is in } \mathcal{F} \text{ for all real } x.$$

Let $\mathcal{A}$ be any σ-algebra of subsets of Ω. We say that X is *measurable with respect to* $\mathcal{A}$, or more briefly *$\mathcal{A}$-measurable*, if

$$\{\omega; X(\omega) \leq x\} \text{ is in } \mathcal{A} \text{ for all real } x.$$

Thus, every real random variable is by definition $\mathcal{F}$-measurable. There may, in general, be smaller σ-algebras with respect to which X is also measurable.

The σ-algebra *generated* by a random variable X is defined to be the smallest σ-algebra with respect to which X is measurable. It is denoted by $\mathcal{F}(X)$ and consists exactly of those sets $\mathcal{A}$ that are in every σ-algebra $\mathcal{A}$ for which X is $\mathcal{A}$-measurable. For example, if X has only denumerably many possible values $x_1, x_2, \ldots$, the sets

$$A_i = \{\omega; X(\omega) = x_i\}, \quad i = 1, 2, \ldots,$$

form a countable *partition* of Ω, i.e.,

$$\Omega = \bigcup_{i=1}^{\infty} A_i,$$

and

$$A_i \cap A_j = \emptyset \quad \text{if } i \neq j,$$

and then $\mathcal{F}(X)$ includes precisely $\emptyset$, Ω, and every set that is the union of some of the A_i's.

Example For the reader completely unfamiliar with this framework, the following simple example will help illustrate the concepts. The experiment consists in tossing a nickel and a dime and observing "heads" or "tails." We take Ω to be

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\},$$

where, e.g., (H, T) stands for the outcome "nickel = heads, and dime = tails." We will take the collection of all subsets of Ω as the family of events. Assuming each outcome in Ω to be equally likely, we arrive at the probability measure:

$A \in \mathcal{F}$	$P[A]$	$A \in \mathcal{F}$	$P[A]$
$\emptyset$	0	Ω	1
$\{(H,H)\}$	$\frac{1}{4}$	$\{(H,T),(T,H),(T,T)\}$	$\frac{3}{4}$
$\{(H,T)\}$	$\frac{1}{4}$	$\{(H,H),(T,H),(T,T)\}$	$\frac{3}{4}$
$\{(T,H)\}$	$\frac{1}{4}$	$\{(H,H),(H,T),(T,T)\}$	$\frac{3}{4}$
$\{(T,T)\}$	$\frac{1}{4}$	$\{(H,H),(H,T),(T,H)\}$	$\frac{3}{4}$
$\{(H,H),(H,T)\}$	$\frac{1}{2}$	$\{(T,H),(T,T)\}$	$\frac{1}{2}$
$\{(H,H),(T,H)\}$	$\frac{1}{2}$	$\{(H,T),(T,T)\}$	$\frac{1}{2}$
$\{(H,H),(T,T)\}$	$\frac{1}{2}$	$\{(H,T),(T,H)\}$	$\frac{1}{2}$

The event "nickel is heads" is $\{(H,H),(H,T)\}$ and has, according to the table, probability $\frac{1}{2}$, as it should.

Let X_n be 1 if the nickel is heads, and 0 otherwise; let X_d be the corresponding random variable for the dime; and let $Z = X_n + X_d$ be the total number of heads. As functions on Ω, we have

$\omega \in \Omega$	$X_n(\omega)$	$X_d(\omega)$	$Z(\omega)$
(H,H)	1	1	2
(H,T)	1	0	1
(T,H)	0	1	1
(T,T)	0	0	0

Finally, the σ-algebras generated by X_n and Z are

$$\mathcal{F}(X_n) = \emptyset, \Omega, \{(H,H),(H,T)\}, \{(T,H),(T,T)\},$$

and

$$\begin{aligned}\mathcal{F}(Z) = \emptyset, \Omega, &\{(H,H)\}, \{(H,T),(T,H)\}, \{(T,T)\},\\ &\{(H,T),(T,H),(T,T)\}, \{(H,H),(T,T)\},\\ &\{(H,H),(H,T),(T,H)\}.\end{aligned}$$

$\mathcal{F}(X_n)$ contains four sets and $\mathcal{F}(Z)$ contains eight. Is X_n measurable with respect to $\mathcal{F}(Z)$, or vice versa?

Every pair X, Y of random variables determines a σ-algebra called the σ-algebra generated by X, Y. It is the smallest σ-algebra with respect to which both X and Y are measurable. This σ-algebra comprises exactly those sets A that are in every σ-algebra $\mathcal{A}$ for which X and Y are both $\mathcal{A}$-measurable. If both X and Y assume only

denumerably many possible values, say $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$, respectively, then the sets

$$A_{ij} = \left\{\omega; X(\omega) = x_j, Y(\omega) = y_j\right\}, \quad i, j = 1, 2, \ldots,$$

present a countable partition of Ω, and $\mathcal{F}(X, Y)$ consists precisely of $\emptyset$, Ω, and every set that is the union of some of the A_{ij}'s. Observe that X is measurable with respect to $\mathcal{F}(X, Y)$, and thus $\mathcal{F}(X) \subset \mathcal{F}(X, Y)$.

More generally, let $\{X(t); t \in T\}$ be any family of random variables. Then, the σ-algebra generated by $\{X(t); t \in T\}$ is the smallest σ-algebra with respect to which every random variable $X(t), t \in T$, is measurable. It is denoted by $\mathcal{F}\{X(t); t \in T\}$.

A special role is played by a distinguished σ-algebra of sets of real numbers. The σ-algebra of *Borel sets* is the σ-algebra generated by the identity function $f(x) = x$, for $x \in (-\infty, \infty)$. Alternatively, the σ-algebra of Borel sets is the smallest σ-algebra containing every interval of the form $(a, b]$, $-\infty \le a \le b < +\infty$. A real-valued function of a real variable is said to be *Borel measurable* if it is measurable with respect to the σ-algebra of Borel sets.

Exercises

1.2.1 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 (i.e., B^c occurs if and only if B does not occur).

1.2.2 Let A and B be arbitrary, not necessarily disjoint, events. Establish the general addition law

$$\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{AB\}.$$

Hint: Apply the result of Exercise 1.2.1 to evaluate $\Pr\{AB^c\} = \Pr\{A\} - \Pr\{AB\}$. Then, apply the addition law to the disjoint events AB and AB^c, noting that $A = (AB) \cup (AB^c)$.

1.2.3 **(a)** Plot the distribution function

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

(b) Determine the corresponding density function $f(x)$ in the three regions (1) $x \le 0$, (2) $0 < x < 1$, and (3) $1 \le x$.

(c) What is the mean of the distribution?

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

1.2.4 Let Z be a discrete random variable having possible values $0, 1, 2$, and 3 and probability mass function

$$p(0) = \frac{1}{4}, \quad p(2) = \frac{1}{8},$$
$$p(1) = \frac{1}{2}, \quad p(3) = \frac{1}{8}.$$

(a) Plot the corresponding distribution function.
(b) Determine the mean $E[Z]$.
(c) Evaluate the variance $\mathrm{Var}[Z]$.

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

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

1.2.6 Let X and Y be independent random variables having distribution functions F_X and F_Y, respectively.
(a) Define $Z = \max\{X, Y\}$ to be the larger of the two. Show that $F_Z(z) = F_X(z)F_Y(z)$ for all z.
(b) Define $W = \min\{X, Y\}$ to be the smaller of the two. Show that $F_W(w) = 1 - [1 - F_X(w)][1 - F_Y(w)]$ for all w.

1.2.7 Suppose X is a random variable having the probability density function

$$f(x) = \begin{cases} Rx^{R-1} & \text{for } 0 \le x \le 1, \\ 0 & \text{elsewhere,} \end{cases}$$

where $R > 0$ is a fixed parameter.
(a) Determine the distribution function $F_X(x)$.
(b) Determine the mean $E[X]$.
(c) Determine the variance $\mathrm{Var}[X]$.

1.2.8 A random variable V has the distribution function

$$F(v) = \begin{cases} 0 & \text{for } v < 0, \\ 1 - (1 - v)^A & \text{for } 0 \le v \le 1, \\ 1 & \text{for } v > 1, \end{cases}$$

where $A > 0$ is a parameter. Determine the density function, mean, and variance.

1.2.9 Determine the distribution function, mean, and variance corresponding to the triangular density.

$$f(x) = \begin{cases} x & \text{for } 0 \le x \le 1, \\ 2 - x & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere.} \end{cases}$$

1.2.10 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

1.2.1 Thirteen cards numbered $1, \ldots, 13$ are shuffled and dealt one at a time. Say a *match* occurs on deal k if the kth card revealed is card number k. Let N be the total number of matches that occur in the thirteen cards. Determine $E[N]$.

Hint: Write $N = \mathbf{1}\{A_1\} + \cdots + \mathbf{1}\{A_{13}\}$ where A_k is the event that a match occurs on deal k.

1.2.2 Let N cards carry the distinct numbers $x_1, \ldots, x_n$. If two cards are drawn at random without replacement, show that the correlation coefficient ρ between the numbers appearing on the two cards is $-1/(N-1)$.

1.2.3 A population having N distinct elements is sampled with replacement. Because of repetitions, a random sample of size r may contain fewer than r distinct elements. Let S_r be the sample size necessary to get r distinct elements. Show that

$$E[S_r] = N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{N-r+1}\right).$$

1.2.4 A fair coin is tossed until the first time that the same side appears twice in succession. Let N be the number of tosses required.
(a) Determine the probability mass function for N.
(b) Let A be the event that N is even and B be the event that $N \leq 6$. Evaluate $\Pr\{A\}$, $\Pr\{B\}$, and $\Pr\{AB\}$.

1.2.5 Two players, A and B, take turns on a gambling machine until one of them scores a success, the first to do so being the winner. Their probabilities for success on a single play are p for A and q for B, and successive plays are independent.
(a) Determine the probability that A wins the contest given that A plays first.
(b) Determine the mean number of plays required, given that A wins.

1.2.6 A pair of dice is tossed. If the two outcomes are equal, the dice are tossed again, and the process repeated. If the dice are unequal, their sum is recorded. Determine the probability mass function for the sum.

1.2.7 Let U and W be jointly distributed random variables. Show that U and W are independent if

$$\Pr\{U > u \text{ and } W > w\} = \Pr\{U > u\}\Pr\{W > w\} \quad \text{for all } u, w.$$

1.2.8 Suppose X is a random variable with finite mean μ and variance σ^2, and $Y = a + bX$ for certain constants $a, b \neq 0$. Determine the mean and variance for Y.

1.2.9 Determine the mean and variance for the probability mass function

$$p(k) = \frac{2(n-k)}{n(n-1)} \quad \text{for } k = 1, 2, \ldots, n.$$

1.2.10 Random variables X and Y are independent and have the probability mass functions

$$p_X(0) = \frac{1}{2}, \qquad p_Y(1) = \frac{1}{6},$$
$$p_X(3) = \frac{1}{2}, \qquad p_Y(2) = \frac{1}{3},$$
$$p_Y(3) = \frac{1}{2}.$$

Determine the probability mass function of the sum $Z = X + Y$.

1.2.11 Random variables U and V are independent and have the probability mass functions

$$p_U(0) = \frac{1}{3}, \qquad p_V(1) = \frac{1}{2},$$
$$p_U(1) = \frac{1}{3}, \qquad p_V(2) = \frac{1}{2}.$$
$$p_U(2) = \frac{1}{3},$$

Determine the probability mass function of the sum $W = U + V$.

1.2.12 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.

1.2.13 Let X and Y be independent random variables each with the uniform probability density function

$$f(x) = \begin{cases} 1 & \text{for } 0 < x < 1, \\ 0 & \text{elsewhere.} \end{cases}$$

Find the joint probability density function of U and V, where $U = \max\{X, Y\}$ and $V = \min\{X, Y\}$.

1.3 The Major Discrete Distributions

The most important discrete probability distributions and their relevant properties are summarized in this section. The exposition is brief, since most readers will be familiar with this material from an earlier course in probability.

1.3.1 Bernoulli Distribution

A random variable X following the Bernoulli distribution with parameter p has only two possible values, 0 and 1, and the probability mass function is $p(1) = p$ and $p(0) = 1 - p$, where $0 < p < 1$, and the mean and variance are $E[X] = p$ and $\text{Var}[X] = p(1 - p)$, respectively.

Bernoulli random variables occur frequently as indicators of events. The *indicator* of an event A is the random variable

$$\mathbf{1}(A) = \mathbf{1}_A = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{if } A \text{ does not occur.} \end{cases} \tag{1.22}$$

Then, $\mathbf{1}_A$ is a Bernoulli random variable with parameter $p = E[\mathbf{1}_A] = \Pr\{A\}$.

The simple expedient of using indicators often reduces formidable calculations into trivial ones. For example, let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be arbitrary real numbers and $A_1, A_2, \ldots, A_n$ be events, and consider the problem of showing that

$$\sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_j \alpha_j \Pr\{A_i \cap A_j\} \geq 0. \tag{1.23}$$

Attacked directly, the problem is difficult. But bringing in the indicators $\mathbf{1}(A_i)$ and observing that

$$\begin{aligned} 0 \leq \left\{ \sum_{i=1}^{n} \alpha_i \mathbf{1}(A_i) \right\}^2 &= \left\{ \sum_{i=1}^{n} \alpha_i \mathbf{1}(A_i) \right\} \left\{ \sum_{j=1}^{n} \alpha_j \mathbf{1}(A_j) \right\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \mathbf{1}(A_i) \mathbf{1}(A_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \mathbf{1}(A_i \cap A_j) \end{aligned}$$

gives, after taking expectations,

$$\begin{aligned} 0 \leq E\left[\left\{ \sum_{i=1}^{n} \alpha_i \mathbf{1}(A_i) \right\}^2 \right] &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j E[\mathbf{1}(A_i \cap A_j)] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \Pr\{A_i \cap A_j\}, \end{aligned}$$

and the demonstration of (1.23) is complete.

1.3.2 Binomial Distribution

Consider independent events $A_1, A_2, \ldots, A_n$, all having the same probability $p = \Pr\{A_i\}$ of occurrence. Let Y count the total number of events among $A_1, \ldots, A_n$ that occur.

Then, Y has a binomial distribution with parameters n and p. The probability mass function is

$$\begin{aligned} p_Y(k) &= \Pr\{Y = k\} \\ &= \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n. \end{aligned} \tag{1.24}$$

Writing Y as a sum of indicators in the form $Y = \mathbf{1}(A_1) + \cdots + \mathbf{1}(A_n)$ makes it easy to determine the moments

$$E[Y] = E[\mathbf{1}(A_1)] + \cdots + E[\mathbf{1}(A_n)] = np,$$

and using independence, we can also determine that

$$\mathrm{Var}[Y] = \mathrm{Var}[\mathbf{1}(A_1)] + \cdots + \mathrm{Var}[\mathbf{1}(A_n)] = np(1-p).$$

Briefly, we think of a binomial random variable as counting the number of "successes" in n independent trials where there is a constant probability p of success on any single trial.

1.3.3 Geometric and Negative Binominal Distributions

Let $A_1, A_2, \ldots$ be independent events having a common probability $p = \Pr\{A_i\}$ of occurrence. Say that trial k is a success (S) or failure (F), depending on whether A_k occurs or not, and let Z count the number of *failures* prior to the first success. To be precise, $Z = k$ if and only if $\mathbf{1}(A_1) = 0, \ldots, \mathbf{1}(A_k) = 0$, and $\mathbf{1}(A_{k+1}) = 1$. Then, Z has a geometric distribution with parameter p. The probability mass function is

$$p_Z(k) = p(1-p)^k \quad \text{for } k = 0, 1, \ldots, \tag{1.25}$$

and the first two moments are

$$E[Z] = \frac{1-p}{p}; \quad \mathrm{Var}[Z] = \frac{1-p}{p^2}.$$

Sometimes the term "geometric distribution" is used in referring to the probability mass function

$$p_{Z'}(k) = p(1-p)^{k-1} \quad \text{for } k = 1, 2, \ldots. \tag{1.26}$$

This is merely the distribution of the random variable $Z' = 1 + Z$, the number of *trials* until the first success. Hence $E[Z'] = 1 + E[Z] = 1/p$, and $\mathrm{Var}[Z'] = \mathrm{Var}[Z] = (1-p)/p^2$.

Now fix an integer $r \geq 1$ and let W_r count the number of failures observed before the rth success in $A_1, A_2, \ldots$. Then, W_r has a *negative binominal* distribution with parameters r and p. The event $W_r = k$ calls for (A) exactly $r-1$ successes in the first

$k+r-1$ trials, followed by (B) a success on trial $k+r$. The probability for (A) is obtained from a binomial distribution, and the probability for (B) is simply p, which leads to the following probability mass function for W_r:

$$p(k) = \Pr\{W_r = k\} = \frac{(k+r-1)!}{(r-1)!k!} p^r (1-p)^k, \quad k = 0, 1, \ldots . \tag{1.27}$$

Another way of writing W_r is as the sum $W_r = Z_1 + \cdots + Z_r$, where $Z_1, \ldots, Z_r$ are independent random variables each having the geometric distribution of (1.25). This formulation readily yields the moments

$$E[W_r] = \frac{r(1-p)}{p}; \quad \mathrm{Var}[W_r] = \frac{r(1-p)}{p^2}. \tag{1.28}$$

1.3.4 The Poisson Distribution

If distributions were graded on a scale of one to ten, the Poisson clearly merits a 10. It plays a role in the class of discrete distributions that parallels in some sense that of the normal distribution in the continuous class. The Poisson distribution occurs often in natural phenomena, for powerful and convincing reasons (the law of rare events, see later in this section). At the same time, the Poisson distribution has many elegant and surprising mathematical properties that make analysis a pleasure.

The Poisson distribution with parameter $\lambda > 0$ has the probability mass function

$$p(k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} \quad \text{for } k = 0, 1, \ldots . \tag{1.29}$$

Using this series expansion

$$\mathrm{e}^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots \tag{1.30}$$

we see that $\Sigma_{k\geq 0} p(k) = 1$. The same series helps calculate the mean via

$$\sum_{k=0}^{\infty} k p(k) = \sum_{k=1}^{\infty} k \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} = \lambda \mathrm{e}^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

The same trick works on the variance, beginning with

$$\sum_{k=0}^{\infty} k(k-1) p(k) = \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} = \lambda^2 \mathrm{e}^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.$$

Written in terms of a random variable X having the Poisson distribution with parameter λ, we have just calculated $E[X] = \lambda$ and $E[X(X-1)] = \lambda^2$, whence $E\left[X^2\right] = E[X(X-1)] + E[X] = \lambda^2 + \lambda$ and $\mathrm{Var}[X] = E\left[X^2\right] - \{E[X]\}^2 = \lambda$. That is, the mean and variance are both the same and equal to the parameter λ of the Poisson distribution.

The simplest form of the law of rare events asserts that the binomial distribution with parameters n and p converges to the Poisson with parameter λ if $n \to \infty$ and $p \to 0$ in such a way that $\lambda = np$ remains constant. In words, given an indefinitely large number of independent trials, where success on each trial occurs with the same arbitrarily small probability, then the total number of successes will follow, approximately, a Poisson distribution.

The proof is a relatively simple manipulation of limits. We begin by writing the binomial distribution in the form

$$\begin{aligned}\Pr\{X = k\} &= \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} \\ &= n(n-1)\cdots(n-k+1)\frac{p^k(1-p)^n}{k!\,(1-p)^k}\end{aligned}$$

and then substitute $p = \lambda/n$ to get

$$\begin{aligned}\Pr\{X = k\} &= n(n-1)\cdots(n-k+1)\frac{\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^n}{k!\left(1-\frac{\lambda}{n}\right)^k} \\ &= 1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{\lambda^k\left(1-\frac{\lambda}{n}\right)^n}{k!\left(1-\frac{\lambda}{n}\right)^k}.\end{aligned}$$

Now let $n \to \infty$ and observe that

$$\begin{aligned}1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) &\to 1 \quad \text{as } n \to \infty; \\ \left(1-\frac{\lambda}{n}\right)^n &\to \mathrm{e}^{-\lambda} \quad \text{as } n \to \infty;\end{aligned}$$

and

$$\left(1-\frac{\lambda}{n}\right)^k \to 1 \quad \text{as } n \to \infty;$$

to obtain the Poisson distribution

$$\Pr\{X = k\} = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} \quad \text{for } k = 0, 1, \ldots$$

in the limit. Extended forms of the law of rare events are presented in Chapter 5.

Example *You Be the Judge* In a purse-snatching incident, a woman described her assailant as being seven feet tall and wearing an orange hat, red shirt, green trousers, and yellow shoes. A short while later and a few blocks away a person fitting that description was seen and charged with the crime.

In court, the prosecution argued that the characteristics of the assailant were so rare as to make the evidence overwhelming that the defendant was the criminal.

The defense argued that the description of the assailant was rare, and that, therefore, the number of people fitting the description should follow a Poisson distribution. Since one person fitting the description was found, the best estimate for the parameter is $\lambda = 1$. Finally, they argued that the relevant computation is the conditional probability that there is at least one other person at large fitting the description given that one was observed. The defense calculated

$$\begin{aligned}\Pr\{X \geq 2 | X \geq 1\} &= \frac{1 - \Pr\{X = 0\} - \Pr\{X = 1\}}{1 - \Pr\{X = 0\}} \\ &= \frac{1 - \mathrm{e}^{-1} - \mathrm{e}^{-1}}{1 - \mathrm{e}^{-1}} = 0.4180,\end{aligned}$$

and since this figure is rather large, they argued that the circumstantial evidence arising out of the unusual description was too weak to satisfy the "beyond a reasonable doubt" criterion for guilt in criminal cases.

1.3.5 The Multinomial Distribution

This is a joint distribution of r variables in which only nonnegative integer values $0, \ldots, n$ are possible. The joint probability mass function is

$$\begin{aligned}&\Pr\{X_1 = k_1, \ldots, X_r = k_r\} \\ &\quad = \begin{cases} \dfrac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r} & \text{if } k_1 + \cdots + k_r = n, \\ 0 & \text{otherwise,} \end{cases}\end{aligned} \tag{1.31}$$

where $p_i > 0$ for $i = 1, \ldots, r$ and $p_1 + \cdots + p_r = 1$.

Some moments are $E[X_i] = np_i$, $\mathrm{Var}[X_i] = np_i(1 - p_i)$, and $\mathrm{Cov}[X_i X_j] = -np_i p_j$.

The multinomial distribution generalizes the binomial. Consider an experiment having a total of r possible outcomes, and let the corresponding probabilities be $p_1, \ldots, p_r$, respectively. Now perform n independent replications of the experiment and let X_i record the total number of times that the ith type outcome is observed in the n trials. Then, $X_1, \ldots, X_r$ has the multinomial distribution given in (1.31).

Exercises

1.3.1 Consider tossing a fair coin five times and counting the total number of heads that appear. What is the probability that this total is three?

1.3.2 A fraction $p = 0.05$ of the items coming off a production process are defective. If a random sample of 10 items is taken from the output of the process, what is the probability that the sample contains exactly one defective item? What is the probability that the sample contains one or fewer defective items?

1.3.3 A fraction $p = 0.05$ of the items coming off of a production process are defective. The output of the process is sampled, one by one, in a random manner. What is the probability that the first defective item found is the tenth item sampled?

1.3.4 A Poisson distributed random variable X has a mean of $\lambda = 2$. What is the probability that X equals 2? What is the probability that X is less than or equal to 2?

1.3.5 The number of bacteria in a prescribed area of a slide containing a sample of well water has a Poisson distribution with parameter 5. What is the probability that the slide shows 8 or more bacteria?

1.3.6 The discrete uniform distribution on $\{1, \ldots, n\}$ corresponds to the probability mass function

$$p(k) = \begin{cases} \dfrac{1}{n} & \text{for } k = 1, \ldots, n, \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Determine the mean and variance.

(b) Suppose X and Y are independent random variables, each having the discrete uniform distribution on $\{0, \ldots, n\}$. Determine the probability mass function for the sum $Z = X + Y$.

(c) Under the assumptions of (b), determine the probability mass function for the minimum $U = \min\{X, Y\}$.

Problems

1.3.1 Suppose that X has a discrete uniform distribution on the integers $0, 1, \ldots, 9$, and Y is independent and has the probability distribution $\Pr\{Y = k\} = a_k$ for $k = 0, 1, \ldots$. What is the distribution of $Z = X + Y \pmod{10}$, their sum modulo 10?

1.3.2 The *mode* of a probability mass function $p(k)$ is any value k^* for which $p(k^*) \geq p(k)$ for all k. Determine the mode(s) for

(a) The Poisson distribution with parameter $\lambda > 0$.

(b) The binomial distribution with parameters n and p.

1.3.3 Let X be a Poisson random variable with parameter λ. Determine the probability that X is odd.

1.3.4 Let U be a Poisson random variable with mean μ. Determine the expected value of the random variable $V = 1/(1 + U)$.

1.3.5 Let $Y = N - X$ where X has a binomial distribution with parameters N and p. Evaluate the product moment $E[XY]$ and the covariance $\mathrm{Cov}[X, Y]$.

1.3.6 Suppose (X_1, X_2, X_3) has a multinomial distribution with parameters M and $\pi_i > 0$ for $i = 1, 2, 3$, with $\pi_1 + \pi_2 + \pi_3 = 1$.

(a) Determine the marginal distribution for X_1.

(b) Find the distribution for $N = X_1 + X_2$.

(c) What is the conditional probability $\Pr\{X_1 = k | N = n\}$ for $0 \leq k \leq n$?

1.3.7 Let X and Y be independent Poisson distributed random variables having means μ and ν, respectively. Evaluate the convolution of their mass functions to determine the probability distribution of their sum $Z = X + Y$.

1.3.8 Let X and Y be independent binomial random variables having parameters (N, p) and (M, p), respectively. Let $Z = X + Y$.

(a) Argue that Z has a binomial distribution with parameters $(N+M, p)$ by writing X and Y as appropriate sums of Bernoulli random variables.
(b) Validate the result in (a) by evaluating the necessary convolution.

1.3.9 Suppose that X and Y are independent random variables with the geometric distribution

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

Perform the appropriate convolution to identify the distribution of $Z = X + Y$ as a negative binomial.

1.3.10 Determine numerical values to three decimal places for $\Pr\{X = k\}, k = 0, 1, 2,$ when
(a) X has a binomial distribution with parameters $n = 10$ and $p = 0.1$.
(b) X has a binomial distribution with parameters $n = 100$ and $p = 0.01$.
(c) X has a Poisson distribution with parameter $\lambda = 1$.

1.3.11 Let X and Y be independent random variables sharing the geometric distribution whose mass function is

$$p(k) = (1-\pi)\pi^k \quad \text{for } k = 0, 1, \ldots,$$

where $0 < \pi < 1$. Let $U = \min\{X, Y\}$, $V = \max\{X, Y\}$, and $W = V - U$. Determine the joint probability mass function for U and W and show that U and W are independent.

1.3.12 Suppose that the telephone calls coming into a certain switchboard during a one-minute time interval follow a Poisson distribution with mean $\lambda = 4$. If the switchboard can handle at most 6 calls per minute, what is the probability that the switchboard will receive more calls than it can handle during a specified one-minute interval?

1.3.13 Suppose that a sample of 10 is taken from a day's output of a machine that produces parts of which 5% are normally defective. If 100% of a day's production is inspected whenever the sample of 10 gives 2 or more defective parts, then what is the probability that 100% of a day's production will be inspected? What assumptions did you make?

1.3.14 Suppose that a random variable Z has the geometric distribution

$$p_Z(k) = p(1-p)^k \quad \text{for } k = 0, 1, \ldots,$$

where $p = 0.10$.
(a) Evaluate the mean and variance of Z.
(b) What is the probability that Z strictly exceeds 10?

1.3.15 Suppose that X is a Poisson distributed random variable with mean $\lambda = 2$. Determine $\Pr\{X \leq \lambda\}$.

1.3.16 Consider the generalized geometric distribution defined by

$$p_k = b(1-p)^k \quad \text{for } k = 1, 2, \ldots,$$

and

$$p_0 = 1 - \sum_{k=1}^{\infty} p_k,$$

where $0 < p < 1$ and $p \le b \le p/(1-p)$.

(a) Evaluate p_0 in terms of b and p.

(b) What does the generalized geometric distribution reduce to when $b = p$? When $b = p/(1-p)$?

(c) Show that $N = X + Z$ has the generalized geometric distribution when X is a Bernoulli random variable for which $\Pr\{X = 1\} = \alpha$, $0 < \alpha < 1$, and Z independently has the usual geometric distribution given in (1.25).

1.4 Important Continuous Distributions

For future reference, this section catalogs several continuous distributions and some of their properties.

1.4.1 The Normal Distribution

The *normal distribution* with parameters μ and $\sigma^2 > 0$ is given by the familiar bell-shaped probability density function

$$\phi\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty. \tag{1.32}$$

The density function is symmetric about the point μ, and the parameter σ^2 is the variance of the distribution. The case $\mu = 0$ and $\sigma^2 = 1$ is referred to as the *standard normal distribution*. If X is normally distributed with mean μ and variance σ^2, then $Z = (X - \mu)/\sigma$ has a standard normal distribution. By this means, probability statements about arbitrary normal random variables can be reduced to equivalent statements about standard normal random variables. The standard normal density and distribution functions are given respectively by

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\xi^2/2,} \quad -\infty < \xi < \infty, \tag{1.33}$$

and

$$\Phi(x) = \int_{-\infty}^{x} \phi(\xi) \mathrm{d}\xi, \quad -\infty < x < \infty. \tag{1.34}$$

The *central limit theorem* explains in part the wide prevalence of the normal distribution in nature. A simple form of this aptly named result concerns the partial sums

$S_n = \xi_1 + \cdots + \xi_n$ of independent and identically distributed summands $\xi_1, \xi_2, \ldots$ having finite means $\mu = E[\xi_k]$ and finite variances $\sigma^2 = \mathrm{Var}[\xi_k]$. In this case, the central limit theorem asserts that

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right\} = \Phi(x) \quad \text{for all } x. \tag{1.35}$$

The precise statement of the theorem's conclusion is given by equation (1.35). Intuition is sometimes enhanced by the looser statement that, for large n, the sum S_n is approximately normally distributed with mean $n\mu$ and variance $n\sigma^2$.

In practical terms we expect the normal distribution to arise whenever the numerical outcome of an experiment results from numerous small additive effects, all operating independently, and where no single or small group of effects is dominant.

The Lognormal Distribution

If the natural logarithm of a nonnegative random variable V is normally distributed, then V is said to have a lognormal distribution. Conversely, if X is normally distributed with mean μ and variance σ^2, then $V = e^X$ defines a lognormally distributed random variable. The change-of-variable formula (1.15) applies to give the density function for V to be

$$f_V(v) = \frac{1}{\sqrt{2\pi}\sigma v} \exp\left\{-\frac{1}{2}\left(\frac{\ln v - \mu}{\sigma}\right)^2\right\}, \quad v \ge 0. \tag{1.36}$$

The mean and variance are, respectively,

$$\begin{aligned} E[V] &= \exp\left\{\mu + \frac{1}{2}\sigma^2\right\}, \\ \mathrm{Var}[V] &= \exp\left\{2\left(\mu + \frac{1}{2}\sigma^2\right)\right\}\left[\exp\left\{\sigma^2\right\} - 1\right]. \end{aligned} \tag{1.37}$$

1.4.2 The Exponential Distribution

A nonnegative random variable T is said to have an exponential distribution with parameter $\lambda > 0$ if the probability density function is

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t \ge 0, \\ 0 & \text{for } t < 0. \end{cases} \tag{1.38}$$

The corresponding distribution function is

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t} & \text{for } t \ge 0, \\ 0 & \text{for } t < 0, \end{cases} \tag{1.39}$$

and the mean and variance are given, respectively, by

$$E[T] = \frac{1}{\lambda} \quad \text{and} \quad \mathrm{Var}[T] = \frac{1}{\lambda^2}.$$

Note that the parameter is the reciprocal of the mean and *not* the mean itself.

The exponential distribution is fundamental in the theory of continuous-time Markov chains (see Chapter 5), due in major part to its *memoryless property*, as now explained. Think of T as a lifetime and, given that the unit has survived up to time t, ask for the conditional distribution of the remaining life $T-t$. Equivalently, for $x>0$ determine the conditional probability $\Pr\{T-t>x|T>t\}$. Directly applying the definition of conditional probability (see Section 1.2.7), we obtain

$$\begin{aligned}\Pr\{T-t>x|T>t\} &= \frac{\Pr\{T>t+x, T>t\}}{\Pr\{T>t\}} \\ &= \frac{\Pr\{T>t+x\}}{\Pr\{T>t\}} \quad \text{(because } x>0\text{)} \\ &= \frac{\mathrm{e}^{-\lambda(t+x)}}{\mathrm{e}^{-\lambda t}} \quad \text{[from (1.39)]} \\ &= \mathrm{e}^{-\lambda x}.\end{aligned} \tag{1.40}$$

There is no memory in the sense that $\Pr\{T-t>x|T>t\} = \mathrm{e}^{-\lambda x} = \Pr\{T>x\}$, and an item that has survived for t units of time has a remaining lifetime that is statistically the same as that for a new item.

To view the memoryless property somewhat differently, we introduce the *hazard rate* or *failure rate* $r(s)$ associated with a nonnegative random variable S having continuous density $g(s)$ and distribution function $G(s)<1$. The failure rate is defined by

$$r(s) = \frac{g(s)}{1-G(s)} \quad \text{for } s>0. \tag{1.41}$$

We obtain the interpretation by calculating (see Section 1.2.2)

$$\begin{aligned}\Pr\{s<S\le s+\Delta s|s<S\} &= \frac{\Pr\{s<S\le s+\Delta s\}}{\Pr\{s<S\}} \\ &= \frac{g(s)\Delta s}{1-G(s)} + o(\Delta s) \quad \text{[from (1.5)]} \\ &= r(s)\Delta s + o(\Delta s).\end{aligned}$$

An item that has survived to time s will then fail in the interval $(s, s+\Delta s]$ with conditional probability $r(s)\Delta s + o(\Delta s)$, thus motivating the name "failure rate."

We can invert (4.10) by integrating

$$-r(s) = \frac{-g(s)}{1-G(s)} = \frac{\mathrm{d}[1-G(s)]/\mathrm{d}s}{1-G(s)} = \frac{\mathrm{d}\{\ln[1-G(s)]\}}{\mathrm{d}s}$$

to obtain

$$-\int_0^t r(s)\mathrm{d}s = \ln[1-G(t)],$$

or

$$G(t) = 1 - \exp\left\{-\int_0^t r(s)\mathrm{d}s\right\}, \quad t \geq 0,$$

which gives the distribution function explicitly in terms of the hazard rate.

The exponential distribution is uniquely the continuous distribution with the *constant* failure rate $r(t) \equiv \lambda$. (See Exercise 1.4.8 for the discrete analog.) The failure rate does not vary in time, another reflection of the memoryless property.

Section 1.5 contains several exercises concerning the exponential distribution. In addition to providing practice in relevant algebraic and calculus manipulations, these exercises are designed to enhance the reader's intuition concerning the exponential law.

1.4.3 The Uniform Distribution

A random variable U is uniformly distributed over the interval $[a, b]$, where $a < b$, if it has the probability density function

$$f_U(u) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq u \leq b, \\ 0 & \text{elsewhere.} \end{cases} \tag{1.42}$$

The uniform distribution extends the notion of "equally likely" to the continuous case. The distribution function is

$$F_U(x) = \begin{cases} 0 & \text{for } u \leq a, \\ \dfrac{x-a}{b-a} & \text{for } a < x \leq b, \\ 1 & \text{for } x > b, \end{cases} \tag{1.43}$$

and the mean and variance are, respectively,

$$E[U] = \frac{1}{2}(a+b) \quad \text{and} \quad \mathrm{Var}[U] = \frac{(b-a)^2}{12}.$$

The uniform distribution on the unit interval $[0, 1]$, for which $a = 0$ and $b = 1$, is most prevalent.

1.4.4 The Gamma Distribution

The gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$ has probability density function

$$f(x) = \frac{\lambda}{\Gamma(\alpha)}(\lambda x)^{\alpha-1}\mathrm{e}^{-\lambda x} \quad \text{for } x > 0. \tag{1.44}$$

Given an integer number α of independent exponentially distributed random variables $Y_1, \ldots, Y_\alpha$ having common parameter λ, then their sum $X_\alpha = Y_1 + \cdots + Y_\alpha$ has the

gamma density of (1.44), from which we obtain the moments

$$E[X_\alpha] = \frac{\alpha}{\lambda} \quad \text{and} \quad \text{Var}[X_\alpha] = \frac{\alpha}{\lambda^2},$$

with these moment formulas holding for noninteger α as well.

1.4.5 The Beta Distribution

The beta density with parameters $\alpha > 0$ and $\beta > 0$ is given by

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ 0 & \text{elsewhere.} \end{cases} \tag{1.45}$$

The mean and variance are, respectively,

$$E[X] = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \text{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

(The gamma and beta functions are defined and briefly discussed in Section 1.6.)

1.4.6 The Joint Normal Distribution

Let $\sigma_X, \sigma_Y, \mu_X, \mu_Y$, and ρ be real constants subject to $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 < \rho < 1$. For real variables x and y, define

$$Q(x,y) = \frac{1}{1-\rho^2}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}. \tag{1.46}$$

The joint normal (or bivariate normal) distribution for random variables X, Y is defined by the density function

$$\begin{aligned} \phi_{X,Y}(x,y) = &\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\ &\times \exp\left\{-\frac{1}{2}Q(x,y)\right\}, \quad -\infty < x, y < \infty. \end{aligned} \tag{1.47}$$

The moments are

$$E[X] = \mu_X, \qquad E[Y] = \mu_Y,$$
$$\text{Var}[X] = \sigma_X^2, \quad \text{Var}[Y] = \sigma_Y^2,$$

and

$$\text{Cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = \rho\sigma_X\sigma_Y.$$

The dimensionless parameter ρ is called the *correlation coefficient*. When ρ is positive, then positive values of X are (stochastically) associated with positive values of Y. When ρ is negative, then positive values of X are associated with negative values of Y. If $\rho = 0$, then X and Y are independent random variables.

Linear Combinations of Normally Distributed Random Variables

Suppose X and Y have the bivariate normal density (1.47), and let $Z = aX + bY$ for arbitrary constants a, b. Then Z is normally distributed with mean

$$E[Z] = a\mu_X + b\mu_Y$$

and variance

$$\text{Var}[X] = a^2\sigma_X^2 + 2ab\rho\sigma_X\sigma_Y + b^2\sigma_Y^2.$$

A random vector $X_1, \ldots, X_n$, is said to have a *multivariate normal distribution*, or a *joint normal distribution*, if every linear combination $\alpha_1 X_1 + \cdots + \alpha_n X_n, \alpha_i$ real has a univariate normal distribution. Obviously, if $X_1, \ldots, X_n$ has a joint normal distribution, then so does the random vector $Y_1, \ldots, Y_m$, defined by the linear transformation in which

$$Y_j = \alpha_{j1}X_1 + \cdots + \alpha_{jn}X_n, \quad \text{for } j = 1, \ldots, m,$$

for arbitrary constants α_{ji}.

Exercises

1.4.1 The lifetime, in years, of a certain class of light bulbs has an exponential distribution with parameter $\lambda = 2$. What is the probability that a bulb selected at random from this class will last more than 1.5 years? What is the probability that a bulb selected at random will last exactly 1.5 years?

1.4.2 The median of a random variable X is any value a for which $\Pr\{X \leq a\} \geq \frac{1}{2}$ and $\Pr\{X \geq a\} \geq \frac{1}{2}$. Determine the median of an exponentially distributed random variable with parameter λ. Compare the median to the mean.

1.4.3 The lengths, in inches, of cotton fibers used in a certain mill are exponentially distributed random variables with parameter λ. It is decided to convert all measurements in this mill to the metric system. Describe the probability distribution of the length, in centimeters, of cotton fibers in this mill.

1.4.4 Twelve independent random variables, each uniformly distributed over the interval $(0, 1]$, are added, and 6 is subtracted from the total. Determine the mean and variance of the resulting random variable.

1.4.5 Let X and Y have the joint normal distribution described in equation (1.47). What value of α minimizes the variance of $Z = \alpha X + (1 - \alpha)Y$? Simplify your result when X and Y are independent.

1.4.6 Suppose that U has a uniform distribution on the interval $[0, 1]$. Derive the density function for the random variables

(a) $Y = -\ln(1-U)$.

(b) $W_n = U^n$ for $n \geq 1$.

Hint: Refer to Section 1.2.6.

1.4.7 Given independent exponentially distributed random variables S and T with common parameter λ, determine the probability density function of the sum $R = S + T$ and identify its type by name.

1.4.8 Let Z be a random variable with the geometric probability mass function

$$p(k) = (1-\pi)\pi^k, \quad k = 0, 1, \ldots,$$

where $0 < \pi < 1$.

(a) Show that Z has a constant failure rate in the sense that $\Pr\{Z = k | Z \geq k\} = 1 - \pi$ for $k = 0, 1, \ldots$.

(b) Suppose Z' is a discrete random variable whose possible values are $0, 1, \ldots$, and for which $\Pr\{Z' = k | Z' \geq k\} = 1 - \pi$ for $k = 0, 1, \ldots$. Show that the probability mass function for Z' is $p(k)$.

Problems

1.4.1 Evaluate the moment $E\left[e^{\lambda Z}\right]$, where λ is an arbitrary real number and Z is a random variable following a standard normal distribution, by integrating

$$E[e^{\lambda Z}] \int_{-\infty}^{+\infty} e^{\lambda z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Hint: Complete the square $-\frac{1}{2}z^2 + \lambda z = -\frac{1}{2}\left[(z-\lambda)^2 - \lambda^2\right]$ and use the fact that

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-\lambda)^2/2} dz = 1.$$

1.4.2 Let W be an exponentially distributed random variable with parameter θ and mean $\mu = 1/\theta$.

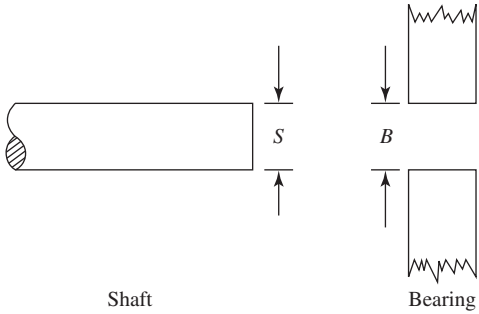
(a) Determine $\Pr\{W > \mu\}$.

(b) What is the mode of the distribution?

1.4.3 Let X and Y be independent random variables uniformly distributed over the interval $\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$ for some fixed θ. Show that $W = X - Y$ has a distribution that is independent of θ with density function

$$f_w(w) = \begin{cases} 1 + w & \text{for } -1 \leq w < 0, \\ 1 - w & \text{for } 0 \leq w \leq 1, \\ 0 & \text{for } |w| > 1. \end{cases}$$

1.4.4 Suppose that the diameters of bearings are independent normally distributed random variables with mean $\mu_B = 1.005$ inch and variance $\sigma_B^2 = (0.003)^2$ inch^2. The diameters of shafts are independent normally distributed random variables having mean $\mu_S = 0.995$ inch and variance $\sigma_S^2 = (0.004)^2$ inch^2.



Let S be the diameter of a shaft taken at random and let B be the diameter of a bearing.

(a) What is the probability $\Pr\{S > B\}$ of interference?

(b) What is the probability of one or fewer interferences in 20 random shaft-bearing pairs?

Hint: The clearance, defined by $C = B - S$, is normally distributed (why?), and interference occurs only if $C < 0$.

1.4.5 If X follows an exponential distribution with parameter $\alpha = 2$, and independently, Y follows an exponential distribution with parameter $\beta = 3$, what is the probability that $X < Y$?

1.5 Some Elementary Exercises

We have collected in this section a number of exercises that go beyond what is usually covered in a first course in probability.

1.5.1 Tail Probabilities

In mathematics, what is a "trick" upon first encounter becomes a basic tool when familiarity through use is established. In dealing with nonnegative random variables, we can often simplify the analysis by the trick of approaching the problem through the upper tail probabilities of the form $\Pr\{X > x\}$. Consider the following example.

A jar has n chips numbered $1, 2, \ldots, n$. A person draws a chip, returns it, draws another, returns it, and so on, until a chip is drawn that has been drawn before. Let X be the number of drawings. Find the probability distribution for X.

It is easier to compute $\Pr\{X > k\}$ first. Then, $\Pr\{X > 1\} = 1$, since at least two draws are always required. The event $\{X > 2\}$ occurs when distinct numbers appear on the first two draws, whence $\Pr\{X > 2\} = (n/n)[(n-1)/n]$. Continuing in this manner, we

obtain

$$\Pr\{X > k\} = 1\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right), \quad \text{for } k = 1, \ldots, n-1. \tag{1.48}$$

Finally,

$$\begin{aligned}
\Pr\{X = k\} &= \Pr\{X > k-1\} - \Pr\{X > k\} \\
&= \left[\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{k-2}{n}\right)\right] \\
&\quad - \left[\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{k-2}{n}\right)\left(1 - \frac{k-1}{n}\right)\right] \\
&= \left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{k-2}{n}\right)\left[1 - \left(1 - \frac{k-1}{n}\right)\right] \\
&= \frac{k-1}{n}\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{k-2}{n}\right), \\
&\qquad\qquad \text{for } k = 2, \ldots, n+1.
\end{aligned}$$

Now try deriving $\Pr\{X = k\}$ directly, for comparison with the "trick" approach. The usefulness of the upper tail probabilities is enhanced by the formula

$$E[X] = \sum_{k=0}^{\infty} \Pr\{X > k\} = \sum_{k=1}^{\infty} \Pr\{X \geq k\}, \tag{1.49}$$

valid for nonnegative integer-valued random variables X. To establish (1.49), abbreviate the notation by using $p(k) = \Pr\{X = k\}$, and rearrange the terms in $E[X] = \Sigma_{k-\geq 0} kp(k)$ as follows:

$$\begin{aligned}
E[X] &= 0p(0) + 1p(1) + 2p(2) + 3p(3) + \cdots \\
&= p(1) + p(2) + p(3) + p(4) + \cdots \\
&\qquad + p(2) + p(3) + p(4) + \cdots \\
&\qquad\qquad + p(3) + p(4) + \cdots \\
&\qquad\qquad\qquad + p(4) + \cdots \\
&\qquad\qquad\qquad\qquad \vdots \\
&= \Pr\{X \geq 1\} + \Pr\{X \geq 2\} + \Pr\{X \geq 3\} + \cdots \\
&= \sum_{k=1}^{\infty} \Pr\{X \geq k\},
\end{aligned}$$

thus establishing (1.49).

For the chip drawing problem, the mean number of draws required is

$$E[X] = \Pr\{X > 0\} + \Pr\{X > 1\} + \cdots + \Pr\{X > n\},$$

since $\Pr\{X > k\} = 0$ for $k > n$. Substituting (1.48) into (1.49) leads directly to

$$E[X] = 2 + \left(1 - \frac{1}{n}\right) + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{n-1}{n}\right).$$

Now let X be a *nonnegative* continuous random variable with density $f(x)$ and distribution function $F(x)$. The analog to (1.49) is

$$E[X] = \int_0^\infty [1 - F(z)]\mathrm{d}z, \tag{1.50}$$

obtained by interchanging an order of integration as follows:

$$E[X] = \int_0^\infty x f(x)\mathrm{d}x = \int_0^\infty \left(\int_0^x \mathrm{d}z\right) f(x)\mathrm{d}x$$
$$= \int_0^\infty \left[\int_z^\infty f(x)\mathrm{d}x\right]\mathrm{d}z = \int_0^\infty [1 - F(z)]\mathrm{d}z.$$

Interchanging the order of integration where the limits are variables often proves difficult for many students. The trick of using indicator functions to make the limits of integration constant may simplify matters. In the preceding interchange, let

$$1_z(x) := 1 \text{ if } 0 \le z < x \text{ and } 1_z(x) := 0 \text{ otherwise.}$$

and then

$$\int_0^\infty \left[\int_0^x \mathrm{d}z\right] f(x)\mathrm{d}x = \int_0^\infty \left[\int_0^\infty \mathbf{1}_z(x) f(x)\mathrm{d}z\right]\mathrm{d}x$$
$$= \int_0^\infty \left[\int_0^\infty \mathbf{1}_z(x) f(x)\mathrm{d}x\right]\mathrm{d}z = \int_0^\infty \left[\int_z^\infty f(x)\mathrm{d}x\right]\mathrm{d}z.$$

As an application of (1.50), let $X_c = \min\{c, X\}$ for some positive constant c. For example, suppose X is the failure time of a certain piece of equipment. A planned

replacement policy is put in use that calls for replacement of the equipment upon its failure or upon its reaching age c, whichever occurs first. Then,

$$X_c = \min\{c, X\} = \begin{cases} X & \text{if } X \le c, \\ c & \text{if } X > c \end{cases}$$

is the time for replacement.

Now

$$\Pr\{X_c > z\} = \begin{cases} 1 - F(z) & \text{if } 0 \le z < c, \\ 0 & \text{if } c \le z, \end{cases}$$

whence we obtain

$$E[X_c] = \int_0^c [1 - F(z)]\mathrm{d}z,$$

which is decidedly shorter than

$$E[X_c] = \int_0^c xf(x)\mathrm{d}x + c[1 - F(c)].$$

Observe that X_c is a random variable whose distribution is partly continuous and partly discrete, thus establishing by example that such distributions do occur in practical applications.

1.5.2 The Exponential Distribution

This exercise is designed to foster intuition about the exponential distribution, as well as to provide practice in algebraic and calculus manipulations relevant to stochastic modeling.

Let X_0 and X_1 be independent exponentially distributed random variables with respective parameters λ_0 and λ_1, so that

$$\Pr\{X_i > t\} = \mathrm{e}^{-\lambda_i t} \quad \text{for } t \ge 0, i = 0, 1.$$

Let

$$N = \begin{cases} 0 & \text{if } X_0 \le X_1, \\ 1 & \text{if } X_1 \le X_0; \end{cases}$$
$$U = \min\{X_0, X_1\} = X_N;$$
$$M = 1 - N;$$
$$V = \max\{X_0, X_1\} = X_M;$$

and

$$W = V - U = |X_0 - X_1|.$$

In this context, we derive the following:

(a) $\Pr\{N = 0 \text{ and } U > t\} = \mathrm{e}^{-(\lambda_0+\lambda_1)t}\left(\dfrac{\lambda_0}{\lambda_0+\lambda_1}\right)$.

The event $\{N = 0 \text{ and } U > t\}$ is exactly the event $\{t < X_0 \le X_1\}$, whence

$$\begin{aligned}
\Pr\{N = 0, U > t\} &= \Pr\{t < X_0 < X_1\} \\
&= \iint\limits_{r<x_0<x_l} \lambda_0 \mathrm{e}^{-\lambda_0 x_0} \lambda_1 \mathrm{e}^{-\lambda_1 x_1} \mathrm{d}x_1 \mathrm{d}x_0 \\
&= \int_t^\infty \left(\int_{x_0}^\infty \lambda_1 \mathrm{e}^{-\lambda_1 X_1} \mathrm{d}x_1 \right) \lambda_0 \mathrm{e}^{-\lambda_0 x_0} \mathrm{d}x_0 \\
&= \int_t^\infty \mathrm{e}^{-\lambda_1 X_0} \lambda_0 \mathrm{e}^{-\lambda_0 X_0} \mathrm{d}x_0 \\
&= \frac{\lambda_0}{\lambda_0+\lambda_1} \int_t^\infty (\lambda_0+\lambda_1) \mathrm{e}^{-(\lambda_0+\lambda_1)x_0} \mathrm{d}x_0 \\
&= \frac{\lambda_0}{\lambda_0+\lambda_1} \mathrm{e}^{-(\lambda_0+\lambda_1)t}.
\end{aligned}$$

(b) $\Pr\{N = 0\} = \dfrac{\lambda_0}{\lambda_0+\lambda_1}$ and $\Pr\{N = 1\} = \dfrac{\lambda_1}{\lambda_0+\lambda_1}$.

We use the result in (a) as follows:

$$\Pr\{N = 0\} = \Pr\{N = 0, U > 0\} = \frac{\lambda_0}{\lambda_0+\lambda_1} \quad \text{from (a).}$$

Obviously, $\Pr\{N = 1\} = 1 - \Pr\{N = 0\} = \lambda_1/(\lambda_0+\lambda_1)$.

(c) $\Pr\{U > t\} = \mathrm{e}^{-(\lambda_0+\lambda_1)t}, \quad t \ge 0.$

Upon adding the result in (a),

$$\Pr\{N = 0 \text{ and } U > t\} = \mathrm{e}^{-(\lambda_0+\lambda_1)t} \frac{\lambda_0}{\lambda_0+\lambda_1},$$

to the corresponding quantity associated with $N = 1$,

$$\Pr\{N = 1 \text{ and } U > t\} = \mathrm{e}^{-(\lambda_0+\lambda_1)t} \frac{\lambda_1}{\lambda_0+\lambda_1},$$

we obtain the desired result via

$$\begin{aligned}
\Pr\{U > t\} &= \Pr\{N = 0, U > t\} + \Pr\{N = 1, U > t\} \\
&= \mathrm{e}^{-(\lambda_0+\lambda_1)t} \left(\frac{\lambda_0}{\lambda_0+\lambda_1} + \frac{\lambda_1}{\lambda_0+\lambda_1} \right) \\
&= \mathrm{e}^{-(\lambda_0+\lambda_1)t}.
\end{aligned}$$

At this point observe that U and N are independent random variables. This follows because (a), (b), and (c) together give

$$\Pr\{N = 0 \text{ and } U > t\} = \Pr\{N = 0\} \times \Pr\{U > t\}.$$

Think about this remarkable result for a moment. Suppose X_0 and X_1 represent lifetimes, and $\lambda_0 = 0.001$, while $\lambda_1 = 1$. The mean lifetimes are $E[X_0] = 1000$ and $E[X_1] = 1$. Suppose we observe that the time of the first death is rather small, say, $U = \min\{X_0, X_1\} = \frac{1}{2}$. In spite of vast disparity between the mean lifetimes, the observation that $U = \frac{1}{2}$ provides no information about which of the two units, 0 or 1, was first to die! This apparent paradox is yet another, more subtle, manifestation of the memoryless property unique to the exponential density.

We continue with the exercise.

(d) $\Pr\{W > t | N = 0\} = \mathrm{e}^{-\lambda_1 t}, \quad t \geq 0.$

The event $\{W > t \text{ and } N = 0\}$ for $t \geq 0$ corresponds exactly to the event $\{t < X_1 - X_0\}$. Thus,

$$\begin{aligned}
\Pr\{W > t \text{ and } N = 0\} &= \Pr\{X_1 - X_0 > t\} \\
&= \iint\limits_{x_1 - x_0 > t} \lambda_0 \mathrm{e}^{-\lambda_0 x_0} \lambda_1 \mathrm{e}^{-\lambda_1 x_1} \mathrm{d}x_0 \mathrm{d}x_1 \\
&= \int_0^\infty \left(\int_{x_0 + t}^\infty \lambda_1 \mathrm{e}^{-\lambda_1 x_1} \mathrm{d}x_1 \right) \lambda_0 \mathrm{e}^{-\lambda_0 x_0} \mathrm{d}x_0 \\
&= \int_0^\infty \mathrm{e}^{-\lambda_1 (x_0 + t)} \lambda_0 \mathrm{e}^{-\lambda_0 x_0} \mathrm{d}x_0 \\
&= \frac{\lambda_0}{\lambda_0 + \lambda_1} \mathrm{e}^{-\lambda_1 t} \int_0^\infty (\lambda_0 + \lambda_1) \mathrm{e}^{-(\lambda_0 + \lambda_1) x_0} \mathrm{d}x_0 \\
&= \frac{\lambda_0}{\lambda_0 + \lambda_1} \mathrm{e}^{-\lambda_1 t} \\
&= \Pr\{N = 0\} \mathrm{e}^{-\lambda_1 t} \quad \text{[from (b)]}.
\end{aligned}$$

Then, using the basic definition of conditional probability (Section 1.2.7), we obtain

$$\Pr\{W > t | N = 0\} = \frac{\Pr\{W > t, N = 0\}}{\Pr\{N = 0\}} = \mathrm{e}^{-\lambda_1 t}, \quad t \geq 0,$$

as desired.

Of course a parallel formula holds conditional on $N = 1$:

$$\Pr\{W > t | N = 1\} = \mathrm{e}^{-\lambda_0 t}, \quad t \geq 0,$$

and using the law of total probability we obtain the distribution of W in the form

$$\begin{aligned}
\Pr\{W > t\} &= \Pr\{W > t, N = 0\} + \Pr\{W > t, N = 1\} \\
&= \frac{\lambda_0}{\lambda_0 + \lambda_1} \mathrm{e}^{-\lambda_1 t} + \frac{\lambda_1}{\lambda_0 + \lambda_1} \mathrm{e}^{-\lambda_0 t}, \quad t \geq 0,
\end{aligned}$$

(e) U and $W = V - U$ are independent random variables.
To establish this final consequence of the memoryless property, it suffices to show that

$$\Pr\{U > u \text{ and } W > w\} = \Pr\{U > u\}\Pr\{W > w\} \quad \text{for all } u \geq 0, w \geq 0.$$

Determining first

$$\begin{aligned}
\Pr\{N = 0, U > u, W > w\} &= \Pr\{u < X_0 < X_1 - w\} \\
&= \iint\limits_{u<x_0<x_1-w} \lambda_0 e^{-\lambda_0 x_0} \lambda_1 e^{-\lambda_1 x_1} dx_0 dx_1 \\
&= \int_u^\infty \left(\int_{x_0+w}^\infty \lambda_1 e^{-\lambda_1 x_1} dx_1 \right) \lambda_0 e^{-\lambda_0 x_0} dx_0 \\
&= \int_u^\infty e^{-\lambda_1 (x_0+w)} \lambda_0 e^{-\lambda_0 x_0} dx_0 \\
&= \left(\frac{\lambda_0}{\lambda_0 + \lambda_1} \right) e^{-\lambda_1 w} \int_u^\infty (\lambda_0 + \lambda_1) e^{-(\lambda_0+\lambda_1)x_0} dx_0 \\
&= \left(\frac{\lambda_0}{\lambda_0 + \lambda_1} \right) e^{-\lambda_1 w} e^{-(\lambda_0+\lambda_1)u},
\end{aligned}$$

and then, by symmetry,

$$\Pr\{N = 1, U > u, W > w\} = \left(\frac{\lambda_1}{\lambda_0 + \lambda_1} \right) e^{-\lambda_0 w} e^{-(\lambda_0+\lambda_1)u},$$

and finally adding the two expressions, we obtain

$$\begin{aligned}
\Pr\{U > u, W > w\} &= \left[\left(\frac{\lambda_0}{\lambda_0 + \lambda_1} \right) e^{-\lambda_1 w} + \left(\frac{\lambda_1}{\lambda_0 + \lambda_1} \right) e^{-\lambda_0 w} \right] e^{-(\lambda_0+\lambda_1)u} \\
&= \Pr\{W > w\}\Pr\{U > u\}, \quad u, w \geq 0.
\end{aligned}$$

The calculation is complete.

Exercises

1.5.1 Let X have a binomial distribution with parameters $n = 4$ and $p = \frac{1}{4}$. Compute the probabilities $\Pr\{X \geq k\}$ for $k = 1, 2, 3, 4$, and sum these to verify that the mean of the distribution is 1.

1.5.2 A jar has four chips colored red, green, blue, and yellow. A person draws a chip, observes its color, and returns it. Chips are now drawn repeatedly, without replacement, until the first chip drawn is selected again. What is the mean number of draws required?

1.5.3 Let X be an exponentially distributed random variable with parameter λ. Determine the mean of X

(a) by integrating by parts in the definition in equation (1.7) with $m = 1$;
(b) by integrating the upper tail probabilities in accordance with equation (1.50).

Which method do you find easier?

1.5.4 A system has two components: A and B. The operating times until failure of the two components are independent and exponentially distributed random variables with parameter 2 for component A, and 3 for B. The system fails at the first component failure.
(a) What is the mean time to failure for component A? For component B?
(b) What is the mean time to system failure?
(c) What is the probability that it is component A that causes system failure?
(d) Suppose that it is component A that fails first. What is the mean remaining operating life of component B?

1.5.5 Consider a post office with two clerks. John, Paul, and Naomi enter simultaneously. John and Paul go directly to the clerks, while Naomi must wait until either John or Paul is finished before she begins service.
(a) If all of the service times are independent exponentially distributed random variables with the same mean $1/\lambda$, what is the probability that Naomi is still in the post office after the other two have left?
(b) How does your answer change if the two clerks have different service rates, say $\lambda_1 = 3$ and $\lambda_2 = 47$?
(c) The mean time that Naomi spends in the post office is less than that for John or Paul provided that $\max\{\lambda_1, \lambda_2\} > c \min\{\lambda_1, \lambda_2\}$ for a certain constant c. What is the value of this constant?

Problems

1.5.1 Let $X_1, X_2, \ldots$ be independent and identically distributed random variables having the cumulative distribution function $F(x) = \Pr\{X \le x\}$. For a fixed number ξ, let N be the first index k for which $X_k > \xi$. That is, $N = 1$ if $X_1 > \xi$; $N = 2$ if $X_1 \le \xi$ and $X_2 > \xi$; etc. Determine the probability mass function for N.

1.5.2 Let $X_1, X_2, \ldots, X_n$ be independent random variables, all exponentially distributed with the same parameter λ. Determine the distribution function for the minimum $Z = \min\{X_1, \ldots, X_n\}$.

1.5.3 Suppose that X is a discrete random variable having the geometric distribution whose probability mass function is

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

(a) Determine the upper tail probabilities $\Pr\{X > k\}$ for $k = 0, 1, \ldots$.
(b) Evaluate the mean via $E[X] = \Sigma_{k \ge 0} \Pr\{X > k\}$.

1.5.4 Let V be a continuous random variable taking both positive and negative values and whose mean exists. Derive the formula

$$E[V] = \int_0^\infty [1 - F_V(v)]\mathrm{d}v - \int_{-\infty}^0 F_V(v)\mathrm{d}v.$$

1.5.5 Show that

$$E[W^2] = \int_0^\infty 2y[1 - F_W(y)]\mathrm{d}y$$

for a nonnegative random variable W.

1.5.6 Determine the upper tail probabilities $\Pr\{V > t\}$ and mean $E[V]$ for a random variable V having the exponential density

$$f_V(v) = \begin{cases} 0 & \text{for } v < 0, \\ \lambda \mathrm{e}^{-\lambda v} & \text{for } v \geq 0, \end{cases}$$

where λ is a fixed positive parameter.

1.5.7 Let $X_1, X_2, \ldots, X_n$ be independent random variables that are exponentially distributed with respective parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$. Identify the distribution of the minimum $V = \min\{X_1, X_2, \ldots, X_n\}$.

Hint: For any real number v, the event $\{V > v\}$ is equivalent to $\{X_1 > v, X_2 > v, \ldots, X_n > v\}$.

1.5.8 Let $U_1, U_2, \ldots, U_n$ be independent uniformly distributed random variables on the unit interval $[0, 1]$. Define the minimum $V_n = \min\{U_1, U_2, \ldots, U_n\}$.

(a) Show that $\Pr\{V_n > v\} = (1 - v)^n$ for $0 \leq v \leq 1$.

(b) Let $W_n = nV_n$. Show that $\Pr\{W_n > w\} = [1 - (w/n)]^n$ for $0 \leq w \leq n$, and thus

$$\lim_{n \to \infty} \Pr\{W_n > w\} = \mathrm{e}^{-w} \quad \text{for } w \geq 0.$$

1.5.9 A flashlight requires two good batteries in order to shine. Suppose, for the sake of this academic exercise, that the lifetimes of batteries in use are independent random variables that are exponentially distributed with parameter $\lambda = 1$. Reserve batteries do not deteriorate. You begin with five fresh batteries. On average, how long can you shine your light?

1.6 Useful Functions, Integrals, and Sums

Collected here for later reference are some calculations and formulas that are especially pertinent in probability modeling.

We begin with several exponential integrals, the first and simplest being

$$\int \mathrm{e}^{-x}\mathrm{d}x = -\mathrm{e}^{-x}. \tag{1.51}$$

When we use integration by parts, the second integral that we introduce reduces to the first in the manner

$$\int x\mathrm{e}^{-x}\mathrm{d}x = -x\mathrm{e}^{-x} + \int \mathrm{e}^{-x}\mathrm{d}x = -\mathrm{e}^{-x}(1 + x). \tag{1.52}$$

Then, (1.51) and (1.52) are the special cases of $\alpha = 1$ and $\alpha = 2$, respectively, in the general formula, valid for any real number α for which the integrals are defined, given by

$$\int x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x = -x^{\alpha-1}\mathrm{e}^{-x} + (\alpha-1)\int x^{\alpha-2}\mathrm{e}^{-x}\mathrm{d}x. \tag{1.53}$$

Fixing the limits of integration leads to the gamma function, defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x, \quad \text{for } \alpha > 0. \tag{1.54}$$

From (1.53), it follows that

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1), \tag{1.55}$$

and therefore, for any integers k,

$$\Gamma(k) = (k-1)(k-2)\cdots 2\cdot\Gamma(1). \tag{1.56}$$

An easy consequence of (1.51) is the evaluation $\Gamma(1) = 1$, which with (1.55) shows that the gamma function at integral arguments is a generalization of the factorial function, and

$$\Gamma(k) = (k-1)! \quad \text{for } k = 1, 2, \ldots. \tag{1.57}$$

A more difficult integration shows that

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \tag{1.58}$$

which with (1.56) provides

$$\Gamma\left(n+\frac{1}{2}\right) = \frac{1\times 3\times 5\times\cdots\times(2n-1)}{2^n}\sqrt{\pi}, \quad \text{for } n = 0, 1, \ldots. \tag{1.59}$$

Stirling's formula is the following important asymptotic evaluation of the factorial function:

$$n! = n^n\mathrm{e}^{-n}(2\pi n)^{1/2}\mathrm{e}^{r(n)/12n}, \tag{1.60}$$

in which

$$1 - \frac{1}{12n+1} < r(n) < 1. \tag{1.61}$$

We sometimes write this in the looser form

$$n! \sim n^n e^{-n} (2\pi n)^{1/2} \quad \text{as } n \to \infty, \tag{1.62}$$

the symbol "$\sim$" signifying that the ratio of the two sides in (1.62) approaches 1 as $n \to \infty$. For the binomial coefficient $\binom{n}{k} = n!/[k!\,(n-k)!\,]$, we then obtain

$$\binom{n}{k} \sim \frac{(n-k)^k}{k!} \quad \text{as } n \to \infty, \tag{1.63}$$

as a consequence of (1.62) and the exponential limit

$$e^{-k} = \lim_{n\to\infty} \left(1 - \frac{k}{n}\right)^n.$$

The integral

$$B(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}\,dx, \tag{1.64}$$

which converges when m and n are positive, defines the *beta* function, related to the gamma function by

$$B(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \quad \text{for } m > 0, n > 0. \tag{1.65}$$

For nonnegative integral values m and n,

$$B(m+1, n+1) = \int_0^1 x^m (1-x)^n dx = \frac{m!\,n!}{(m+n+1)!}. \tag{1.66}$$

For $n = 1, 2, \ldots$, the binomial theorem provides the evaluation

$$(1-x)^n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} x^k, \quad \text{for } -\infty < x < \infty. \tag{1.67}$$

The formula may be generalized for nonintegral n by appropriately generalizing the binomial coefficient, defining for any real number α,

$$\binom{\alpha}{k} = \begin{cases} \dfrac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} & \text{for } k = 1, 2, \ldots, \\ 1 & \text{for } k = 0. \end{cases} \tag{1.68}$$

As a special case, for any positive integer n,

$$\begin{aligned}\binom{-n}{k} &= (-1)^k \frac{n(n+1)\cdots(n+k-1)}{k!} \\ &= (-1)^k \binom{n+k-1}{k}.\end{aligned} \tag{1.69}$$

The general binomial theorem, valid for all real α, is

$$(1-x)^\alpha = \sum_{k=0}^{\infty} (-1)^k \binom{\alpha}{k} x^k \quad \text{for} -1 < x < 1. \tag{1.70}$$

When $\alpha = -n$ for a positive integer n, we obtain a group of formulas useful in dealing with geometric series. For a positive integer n, in view of (1.69) and (1.70), we have

$$(1-x)^{-n} = \sum_{k=0}^{\infty} \binom{n+k-1}{k} x^k \quad \text{for } |x| < 1. \tag{1.71}$$

The familiar formula

$$\sum_{k=0}^{\infty} x^k = 1 + x + x^2 + \cdots = \frac{1}{1-x} \quad \text{for } |x| < 1 \tag{1.72}$$

for the sum of a geometric series results from (1.71) with $n = 1$. The cases $n = 2$ and $n = 3$ yield the formulas

$$\begin{aligned}\sum_{k=0}^{\infty} (k+1)x^k &= 1 + 2x + 3x^2 + \cdots \\ &= \frac{1}{(1-x)^2} \quad \text{for } |x| < 1,\end{aligned} \tag{1.73}$$

$$\sum_{k=0}^{\infty} (k+2)(k+1)x^k = \frac{2}{(1-x)^3} \quad \text{for } |x| < 1. \tag{1.74}$$

Sums of Numbers

The following sums of powers of integers have simple expressions:

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2},$$

$$1 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6},$$

$$1 + 2^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4}.$$