1

INTRODUCTION AND REVIEW

We demand rigidly defined areas of doubt and uncertainty!
–Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

1.1 DETERMINISTIC AND STOCHASTIC MODELS

Probability theory, the mathematical science of uncertainty, plays an ever growing role in how we understand the world around us—whether it is the climate of the planet, the spread of an infectious disease, or the results of the latest news poll.

The word "stochastic" comes from the Greek *stokhazesthai*, which means to aim at, or guess at. A stochastic process, also called a random process, is simply one in which outcomes are uncertain. By contrast, in a deterministic system there is no randomness. In a deterministic system, the same output is always produced from a given input.

Functions and differential equations are typically used to describe deterministic processes. Random variables and probability distributions are the building blocks for stochastic systems.

Consider a simple exponential growth model. The number of bacteria that grows in a culture until its food source is exhausted exhibits exponential growth. A common

Introduction to Stochastic Processes with R, First Edition. Robert P. Dobrow.

deterministic growth model is to assert that the population of bacteria grows at a fixed rate, say 20% per minute. Let $y(t)$ denote the number of bacteria present after t minutes. As the growth rate is proportional to population size, the model is described by the differential equation

$$\frac{dy}{dt} = (0.20)y.$$

The equation is solved by the exponential function

$$y(t) = y_0 e^{(0.20)t},$$

where $y_0 = y(0)$ is the initial size of the population.

As the model is deterministic, bacteria growth is described by a function, and no randomness is involved. For instance, if there are four bacteria present initially, then after 15 minutes, the model asserts that the number of bacteria present is

$$y(15) = 4e^{(0.20)15} = 80.3421 \approx 80.$$

The deterministic model does not address the uncertainty present in the reproduction rate of individual organisms. Such uncertainty can be captured by using a stochastic framework where the times until bacteria reproduce are modeled by random variables. A simple stochastic growth model is to assume that the times until individual bacteria reproduce are independent exponential random variables, in this case with rate parameter 0.20. In many biological processes, the exponential distribution is a common choice for modeling the times of *births* and *deaths*.

In the deterministic model, when the population size is n, the number of bacteria increases by $(0.20)n$ in 1 minute. Similarly, for the stochastic model, after n bacteria arise the time until the next bacteria reproduces has an exponential probability distribution with rate $(0.20)n$ per minute. (The stochastic process here is called a *birth process*, which is introduced in Chapter 7.)

While the outcome of a deterministic system is fixed, the outcome of a stochastic process is uncertain. See Figure 1.1 to compare the graph of the deterministic exponential growth function with several possible outcomes of the stochastic process.

The dynamics of a stochastic process are described by random variables and probability distributions. In the deterministic growth model, one can say with certainty how many bacteria are present after t minutes. For the stochastic model, questions of interest might include:

- What is the *average* number of bacteria present at time t?
- What is the *probability* that the number of bacteria will exceed some threshold after t minutes?
- What is the *distribution* of the time it takes for the number of bacteria to double in size?

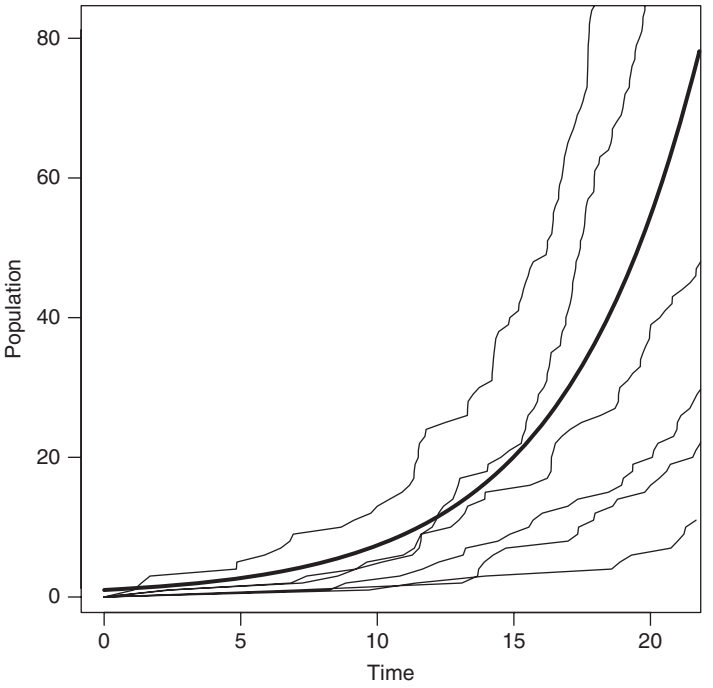

Figure 1.1 Growth of a bacteria population. The deterministic exponential growth curve (dark line) is plotted against six realizations of the stochastic process.

In more sophisticated stochastic growth models, which allow for births and deaths, one might be interested in the likelihood that the population goes extinct, or reaches a long-term equilibrium.

In all cases, conclusions are framed using probability with the goal of quantifying the uncertainty in the system.

Example 1.1 (PageRank) The power of internet search engines lies in their ability to respond to a user's query with an ordered list of web sites ranked by importance and relevance. The heart of Google's search engine is the PageRank algorithm, which assigns an *importance value* to each web page, called its *page rank*. The algorithm is remarkable given the massiveness of the web with over one trillion web pages, and is an impressive achievement of mathematics, particularly linear algebra.

Although the actual PageRank algorithm is complex with many technical (and secret) details, the page rank of a particular web page is easily described by means of a stochastic model. Consider a hypothetical web surfer who travels across the internet moving from page to page at random. When the surfer is on a particular web page, they pick one of the available hypertext links on that page uniformly at random and then move to that page.

The model can be described as a random walk by the web surfer on a giant graph called the *webgraph*. In the webgraph, vertices (nodes) are web pages. Vertex x is joined to vertex y by a directed edge if there is a hypertext link on page x that leads to page y. When the surfer is at vertex x, they choose an edge leading away from x uniformly at random from the set of available edges, and move to the vertex which that edge points to.

The random surfer model is an example of a more general stochastic process called *random walk on a graph*.

Imagine that the web surfer has been randomly walking across the web for a long, long time. What is the probability that the surfer will be at, say, page x? To make this more precise, let p_x^k denote the probability that the surfer is at page x after k steps. The long-term probability of being at page x is defined as $\lim_{k\to\infty} p_x^k$.

This long-term probability is precisely the page rank of page x. Intuitively, the long-term probability of being at a particular page will tend to be higher for pages with more incoming links and smaller for pages with few links, and is a measure of the importance, or popularity, of a page. The PageRank algorithm can be understood as an assignment of probabilities to each site on the web.

Figure 1.2 shows a simplified network of five pages. The numbers under each vertex label are the long-term probabilities of reaching that vertex, and thus the page rank assigned to that page.

Many stochastic processes can be expressed as random walks on graphs in discrete time, or as the limit of such walks in continuous time. These models will play a central role in this book. ■

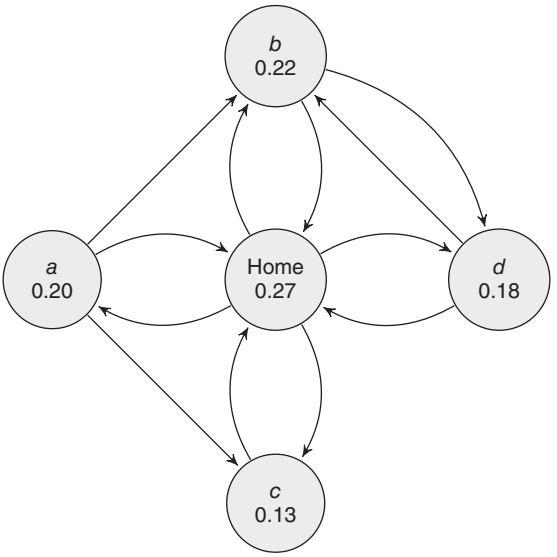

Figure 1.2 Five-page webgraph. Vertex labels show long-term probabilities of reaching each page.

Example 1.2 (Spread of infectious diseases) Models for the spread of infectious diseases and the development of epidemics are of interest to health scientists, epidemiologists, biologists, and public health officials. Stochastic models are relevant because of the randomness inherent in person-to-person contacts and population fluctuations.

The SIR (Susceptible–Infected–Removed) model is a basic framework, which has been applied to the spread of measles and other childhood diseases. At time t, let S_t represent the number of people susceptible to a disease, I_t the number infected, and R_t the number recovered and henceforth immune from infection. Individuals in the population transition from being susceptible to possibly infected to recovered ($S \to I \to R$).

The deterministic SIR model is derived by a system of three nonlinear differential equations, which model interactions and the rate of change in each subgroup.

A stochastic SIR model in discrete time was introduced in the 1920s by medical researchers Lowell Reed and Wade Frost from Johns Hopkins University. In the Reed–Frost model, when a susceptible individual comes in contact with someone who is infected there is a fixed probability z that they will be infected.

Assume that each susceptible person is in contact with all those who are infected. Let p be the probability that a susceptible individual is infected at time t. This is equal to 1 minus the probability that the person is not infected at time t, which occurs if they are not infected by any of the already infected persons, of which there are I_t. This gives

$$p = 1 - (1 - z)^{I_t}.$$

Disease evolution is modeled in discrete time, where one time unit is the incubation period—also the recovery time—of the disease.

The model can be described with a coin-flipping analogy. To find I_{t+1}, the number of individuals infected at time $t + 1$, flip S_t coins (one for each susceptible), where the probability of heads for each coin is the infection probability p. Then, the number of newly infected individuals is the number of coins that land heads.

The number of heads in n independent coin flips with heads probability p has a binomial distribution with parameters n and p. In other words, I_{t+1} has a binomial distribution with $n = S_t$ and $p = 1 - (1 - z)^{I_t}$.

Having found the number of infected individuals at time $t + 1$, the number of susceptible persons decreases by the number of those infected. That is,

$$S_{t+1} = S_t - I_{t+1}.$$

Although the Reed–Frost model is not easy to analyze exactly, it is straightforward to simulate on a computer. The graphs in Figure 1.3 were obtained by simulating the process assuming an initial population of 3 infected and 400 susceptible individuals, with individual infection probability $z = 0.004$. The number of those infected is plotted over 20 time units. ■

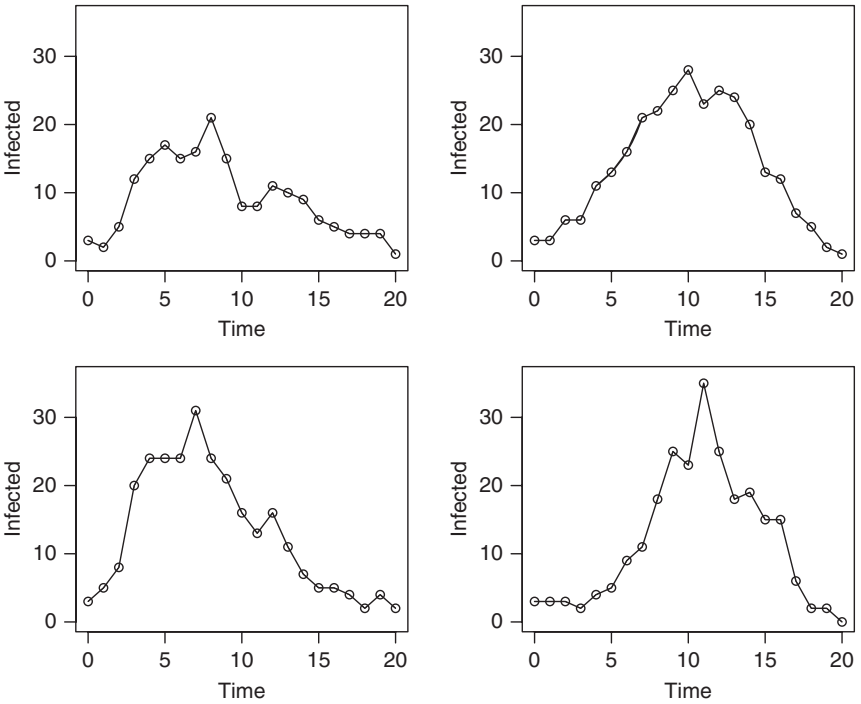

Figure 1.3 Four outcomes of the Reed–Frost epidemic model.

1.2 WHAT IS A STOCHASTIC PROCESS?

In its most general expression, a stochastic process is simply a collection of random variables $\{X_t, t \in I\}$. The index t often represents time, and the set I is the *index set* of the process. The most common index sets are $I = \{0, 1, 2, \ldots\}$, representing discrete time, and $I = [0, \infty)$, representing continuous time. Discrete-time stochastic processes are sequences of random variables. Continuous-time processes are uncountable collections of random variables.

The random variables of a stochastic process take values in a common *state space* $\mathcal{S}$, either discrete or continuous. A stochastic process is specified by its index and state spaces, and by the dependency relations among its random variables.

Stochastic Process

A stochastic process is a collection of random variables $\{X_t, t \in I\}$. The set I is the *index set* of the process. The random variables are defined on a common *state space* $\mathcal{S}$.

Example 1.3 (Monopoly) The popular board game *Monopoly* can be modeled as a stochastic process. Let $X_0, X_1, X_2 \ldots$ represent the successive board positions of an individual player. That is, X_k is the player's board position after k plays.

The state space is $\{1,\ldots,40\}$ denoting the 40 squares of a Monopoly board—from Go to Boardwalk. The index set is $\{0, 1, 2,\ldots\}$ Both the index set and state space are discrete.

An interesting study is to rank the squares of the board in increasing order of probability. Which squares are most likely to be landed on?

Using Markov chain methods (discussed in Chapter 2), Stewart (1996) shows that the most landed-on square is Jail. The next most frequented square is Illinois Avenue, followed by Go, whereas the least frequented location on the board is the third Chance square from Go. ■

Example 1.4 (Discrete time, continuous state space) An air-monitoring station in southern California records oxidant concentration levels every hour in order to monitor smog pollution. If it is assumed that hourly concentration levels are governed by some random mechanism, then the station's data can be considered a realization of a stochastic process $X_0, X_1, X_2, \ldots$, where X_k is the oxidant concentration level at the kth hour. The time variable is discrete. Since concentration levels take a continuum of values, the state space is continuous. ■

Example 1.5 (Continuous time, discrete state space) Danny receives text messages at random times day and night. Let X_t be the number of texts he receives up to time t. Then, $\{X_t, t \in [0, \infty)\}$ is a continuous-time stochastic process with discrete state space $\{0, 1, 2,\ldots\}$.

This is an example of an *arrival process*. If we assume that the times between Danny's texts are independent and identically distributed (i.i.d.) exponential random variables, we obtain a *Poisson process*. The Poisson process arises in myriad settings involving random *arrivals.* Examples include the number of births each day on a maternity ward, the decay of a radioactive substance, and the occurrences of oil spills in a harbor. ■

Example 1.6 (Random walk and gambler's ruin) A random walker starts at the origin on the integer line. At each discrete unit of time the walker moves either right or left, with respective probabilities p and $1 - p$. This describes a *simple random walk* in one dimension.

A stochastic process is built as follows. Let $X_1, X_2,\ldots$ be a sequence of i.i.d. random variables with

$$X_k = \begin{cases} +1, & \text{with probability } p, \\ -1, & \text{with probability } 1 - p, \end{cases}$$

for $k \geq 1$. Set

$$S_n = X_1 + \cdots + X_n, \text{ for } n \geq 1,$$

with $S_0 = 0$. Then, S_n is the random walk's position after n steps. The sequence $S_0, S_1, S_2, \ldots$ is a discrete-time stochastic process whose state space is $\mathbb{Z}$, the set of all integers.

Consider a gambler who has an initial stake of k dollars, and repeatedly wagers \$1 on a game for which the probability of winning is p and the probability of losing is $1 - p$. The gambler's successive fortunes is a simple random walk started at k.

Assume that the gambler decides to stop when their fortune reaches $\$n$ $(n > k)$, or drops to 0, whichever comes first. What is the probability that the gambler is eventually ruined? This is the classic gambler's ruin problem, first discussed by mathematicians Blaise Pascal and Pierre Fermat in 1656.

See Figure 1.4 for simulations of gambler's ruin with $k = 20$, $n = 60$, and $p = 1/2$. Observe that four of the nine outcomes result in the gambler's ruin before 1,000 plays. In the next section, it is shown that the probability of eventual ruin is $(n - k)/n = (60 - 20)/60 = 2/3$. ■

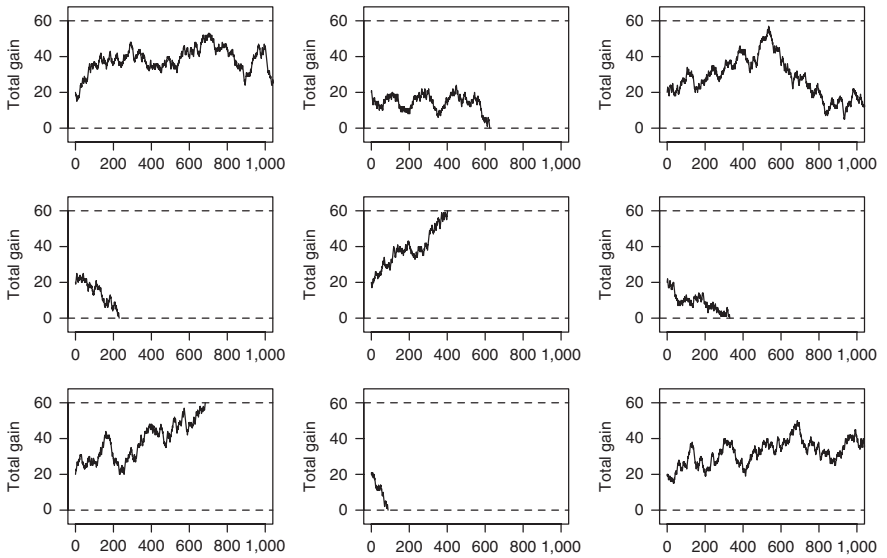

Figure 1.4 Random walk and gambler's ruin.

■ **Example 1.7 (Brownian motion)** Brownian motion is a continuous-time, continuous state space stochastic process. The name also refers to a physical process, first studied by the botanist Robert Brown in 1827. Brown observed the seemingly erratic, zigzag motion of tiny particles ejected from pollen grains suspended in water. He gave a detailed study of the phenomenon but could not explain its cause. In 1905, Albert Einstein showed that the motion was the result of water molecules bombarding the particles.

The mathematical process known as Brownian motion arises as the *limiting process* of a discrete-time random walk. This is obtained by *speeding up* the walk, letting

the interval between discrete steps tend to 0. The process is used as a model for many phenomena that exhibit "erratic, zigzag motion," such as stock prices, the growth of crystals, and signal noise.

Brownian motion has remarkable properties, which are explored in Chapter 8. Paths of the process are continuous everywhere, yet differentiable nowhere. Figure 1.5 shows simulations of two-dimensional Brownian motion. For this case, the index set is $[0, \infty)$ and the state space is $\mathbb{R}^2$. ■

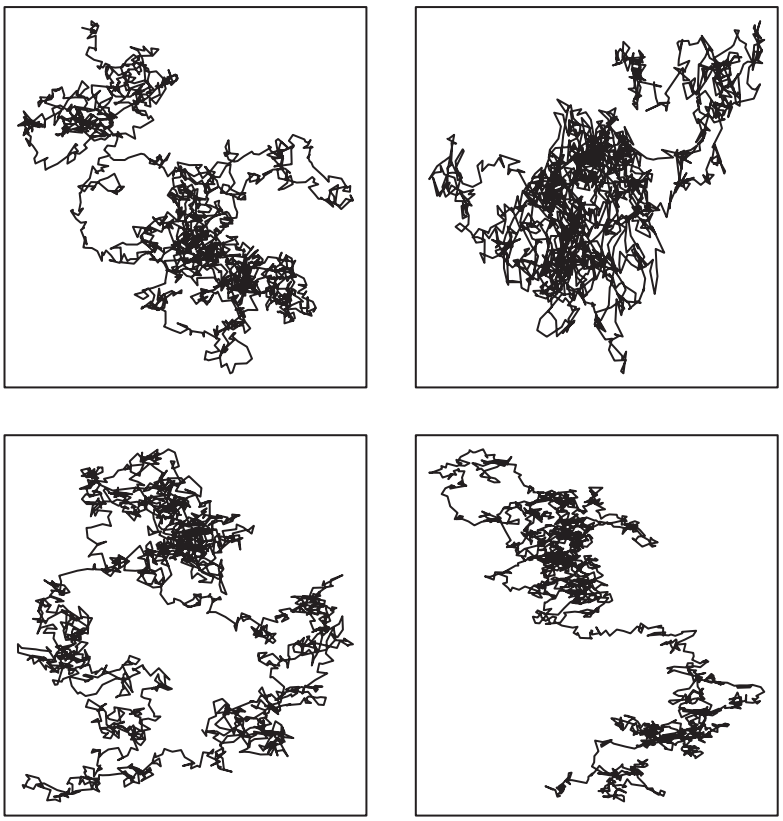

Figure 1.5 Simulations of two-dimensional Brownian motion.

1.3 MONTE CARLO SIMULATION

Advancements in modern computing have revolutionized the study of stochastic systems, allowing for the visualization and simulation of increasingly complex models.

At the heart of the many simulation techniques developed to generate random variables and stochastic processes lies the Monte Carlo method. Given a random experiment and event A, a Monte Carlo estimate of $P(A)$ is obtained by repeating the

random experiment many times and taking the proportion of trials in which A occurs as an approximation for $P(A)$.

The name Monte Carlo evidently has its origins in the fact that the mathematician Stanislaw Ulam, who developed the method in 1946, had an uncle who regularly gambled at the Monte Carlo casino in Monaco.

Monte Carlo simulation is intuitive and matches up with our sense of how probabilities *should* behave. The relative frequency interpretation of probability says that the probability of an event is the long-term proportion of times that the event occurs in repeated trials. It is justified theoretically by the strong law of large numbers.

Consider repeated independent trials of a random experiment. Define the sequence $X_1, X_2, \ldots$, where

$$X_k = \begin{cases} 1, & \text{if } A \text{ occurs on the } k\text{th trial,} \\ 0, & \text{if } A \text{ does not occur on the } k\text{th trial,} \end{cases}$$

for $k \geq 1$. Then, $(X_1 + \cdots + X_n)/n$ is the proportion of n trials in which A occurs. The X_k are identically distributed with common mean $E(X_k) = P(A)$.

By the strong law of large numbers,

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = P(A), \text{ with probability 1.} \tag{1.1}$$

For large n, the Monte Carlo estimate of $P(A)$ is

$$P(A) \approx \frac{X_1 + \cdots + X_n}{n}.$$

In this book, we use the software package `R` for simulation. `R` is a flexible and interactive environment. We often use `R` to illustrate the result of an exact, theoretical calculation with numerical verification. The easy-to-learn software allows the user to see the impact of varying parameters and assumptions of the model. For example, in the Reed–Frost epidemic model of Example 1.2, it is interesting to see how small changes in the infection probability affect the duration and intensity of the epidemic. See the `R` script file **ReedFrost.R** and Exercise 1.36 to explore this question.

If you have not used `R` before, work through the exercises in the introductory tutorial in Appendix A: Getting Started with `R`.

1.4 CONDITIONAL PROBABILITY

The simplest stochastic process is a sequence of i.i.d. random variables. Such sequences are often used to model random samples in statistics. However, most real-world systems exhibit some type of dependency between variables, and an independent sequence is often an unrealistic model.

Thus, the study of stochastic processes really begins with conditional probability—conditional distributions and conditional expectation. These will become essential tools for all that follows.

Starting with a random experiment, the sample space Ω is the set of all possible outcomes. An *event* is a subset of the sample space. For events A and B, the *conditional probability of A given B* is

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

defined for $P(B) > 0$. Events A and B are *independent* if $P(A|B) = P(A)$. Equivalently, A and B are independent if

$$P(A \cap B) = P(A)P(B).$$

Events that are not independent are said to be *dependent*.

For many problems where the goal is to find $P(A)$, partial information and dependencies between events in the sample space are brought to bear. If the sample space can be partitioned into a collection of disjoint events $B_1, \ldots, B_k$, then A can be expressed as the disjoint union

$$A = (A \cap B_1) \cup \cdots \cup (A \cap B_k).$$

If conditional probabilities of the form $P(A|B_i)$ are known, then the law of total probability can be used to find $P(A)$.

Law of Total Probability

Let $B_1, \ldots, B_k$ be a sequence of events that partition the sample space. That is, the B_i are mutually exclusive (disjoint) and their union is equal to Ω. Then, for any event A,

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i).$$

Example 1.8 According to the Howard Hughes Medical Institute, about 7% of men and 0.4% of women are colorblind—either cannot distinguish red from green or see red and green differently from most people. In the United States, about 49% of the population is male and 51% female. A person is selected at random. What is the probability they are colorblind?

Solution Let C, M, and F denote the events that a random person is colorblind, male, and female, respectively. By the law of total probability,

$$\begin{aligned} P(C) &= P(C|M)P(M) + P(C|F)P(F) \\ &= (0.07)(0.49) + (0.004)(0.51) = 0.03634. \end{aligned}$$

■

Using the law of total probability in this way is called *conditioning*. Here, we find the *total probability* of being colorblind by conditioning on sex.

Example 1.9 In a standard deck of cards, the probability that the suit of a random card is hearts is $13/52 = 1/4$. Assume that a standard deck has one card missing. A card is picked from the deck. Find the probability that it is a heart.

Solution Assume that the missing card can be any of the 52 cards picked uniformly at random. Let M denote the event that the missing card is a heart, with the complement M^c the event that the missing card is not a heart. Let H denote the event that the card that is picked from the deck is a heart. By the law of total probability,

$$\begin{aligned} P(H) &= P(H|M)P(M) + P(H|M^c)P(M^c) \\ &= \left(\frac{12}{51}\right)\frac{1}{4} + \left(\frac{13}{51}\right)\frac{3}{4} = \frac{1}{4}. \end{aligned}$$

The result can also be obtained by appealing to symmetry. Since all cards are equally likely, and all four suits are equally likely, the argument by symmetry gives that the desired probability is $1/4$. ■

Example 1.10 (Gambler's ruin) The gambler's ruin problem was introduced in Example 1.6. A gambler starts with k dollars. On each play a fair coin is tossed and the gambler wins \$1 if heads occurs, or loses \$1 if tails occurs. The gambler stops when he reaches \$$n$ ($n > k$) or loses all his money. Find the probability that the gambler will eventually lose.

Solution We make two observations, which are made more precise in later chapters. First, the gambler will eventually stop playing, either by reaching n or by reaching 0. One might argue that the gambler could play forever. However, it can be shown that that event occurs with probability 0. Second, assume that after, say, 100 wagers, the gambler's capital returns to \$$k$. Then, the probability of eventually winning \$$n$ is the same as it was initially. The memoryless character of the process means that the probability of winning \$$n$ or losing all his money only depends on how much capital the gambler has, and not on how many previous wagers the gambler made.

Let p_k denote the probability of reaching n when the gambler's fortune is k. What is the gambler's status if heads is tossed? Their fortune increases to $k+1$ and the probability of winning is the same as it would be if the gambler had started the game with $k+1$. Similarly, if tails is tossed and the gambler's fortune decreases to $k-1$. Hence,

$$p_k = p_{k+1}\left(\frac{1}{2}\right) + p_{k-1}\left(\frac{1}{2}\right),$$

or

$$p_{k+1} - p_k = p_k - p_{k-1}, \quad \text{for} \quad k = 1,\ldots,n-1, \tag{1.2}$$

with $p_0 = 0$ and $p_n = 1$. Unwinding the recurrence gives

$$p_k - p_{k-1} = p_{k-1} - p_{k-2} = p_{k-2} - p_{k-3} = \cdots = p_1 - p_0 = p_1,$$

for $k = 1,\ldots,n$. We have that $p_2 - p_1 = p_1$, giving $p_2 = 2p_1$. Also, $p_3 - p_2 = p_3 - 2p_1 = p_1$, giving $p_3 = 3p_1$. More generally, $p_k = kp_1$, for $k = 1,\ldots,n$.

Sum Equation (1.2) over suitable k to obtain

$$\sum_{k=1}^{n-1}(p_{k+1} - p_k) = \sum_{k=1}^{n-1}(p_k - p_{k-1}).$$

Both sums telescope to

$$p_n - p_1 = p_{n-1} - p_0,$$

which gives $1 - p_1 = p_{n-1} = (n-1)p_1$, so $p_1 = 1/n$. Thus,

$$p_k = kp_1 = \frac{k}{n}, \text{ for } k = 0,\ldots,n.$$

The probability that the gambler eventually wins \$$n$ is k/n. Hence, the probability of the gambler's ruin is $(n-k)/n$. ■

R : Simulating Gambler's Ruin

The file **gamblersruin.R** contains the function `gamble(k,n,p)`, which simulates the gambler's ruin process. At each wager, the gambler wins with probability p, and loses with probability $1 - p$. The gambler's initial stake is \$$k$. The function `gamble` returns 1, if the gambler is eventually ruined, or 0, if the gambler gains \$$n$.

In the simulation the function is called 1,000 times, creating a list of 1,000 ruins and wins, which are represented by 1s and 0s. The mean of the list gives the proportion of 1s, which estimates the probability of eventual ruin.

```
> k <- 20
> n <- 60
> p <- 1/2
> trials <- 1000
> simlist <- replicate(trials, gamble(k,n,p))
> mean(simlist) # Estimate of probability of ruin
[1] 0.664
# Exact probability of ruin is 2/3
```

Sometimes, we need to find a conditional probability of the form $P(B|A)$, but what is given in the problem are *reverse* probabilities of the form $P(A|B)$ and $P(A|B^c)$. Bayes' rule provides a method for *inverting* the conditional probability.