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?

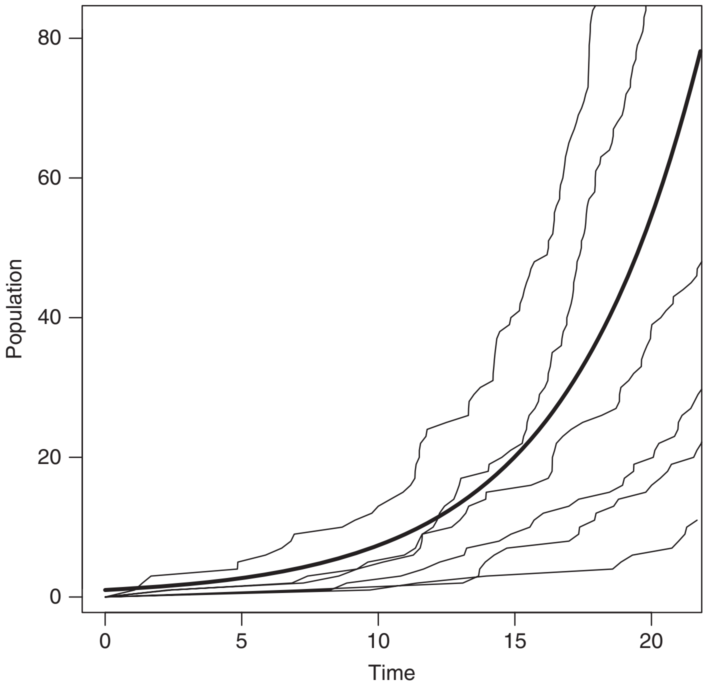

Figure 1.1 Growth of a bacteria population. The deterministic exponential growth curve (dark line) is plotted against six realizations of the stochastic process.

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. ■

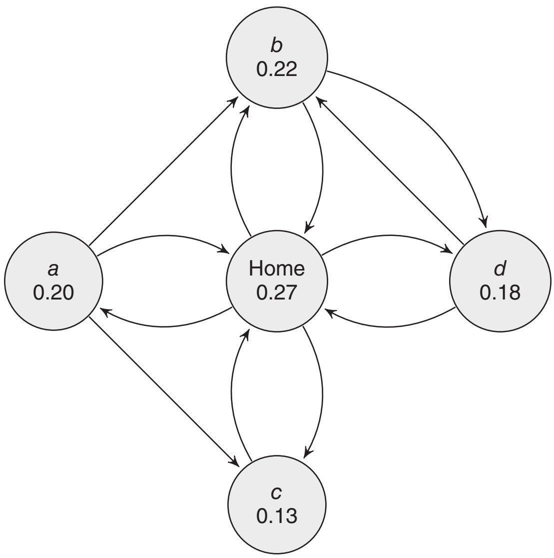

Figure 1.2 Five-page webgraph. Vertex labels show long-term probabilities of reaching each page.

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. ■

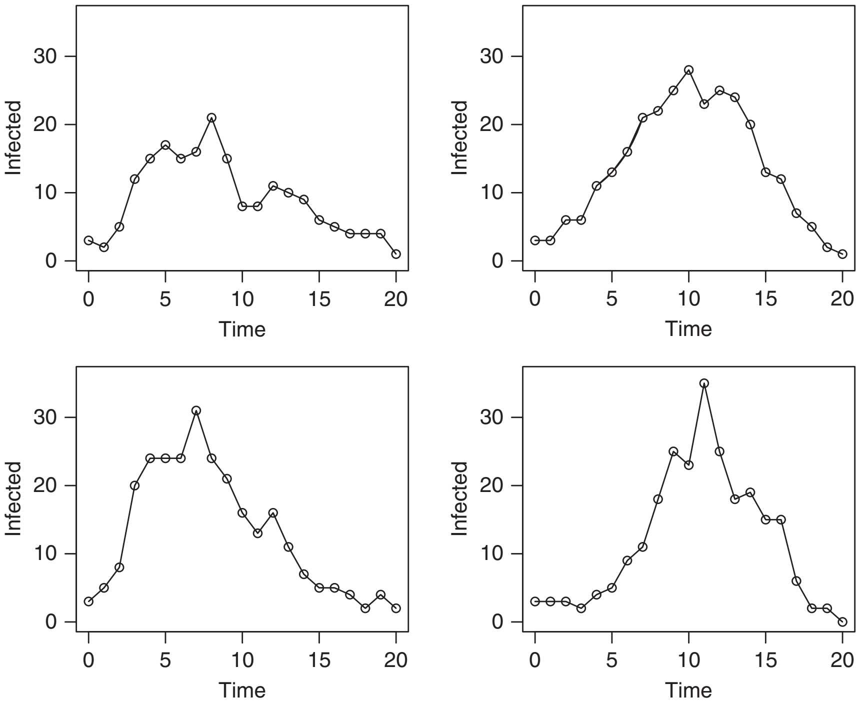

Figure 1.3 Four outcomes of the Reed–Frost epidemic model.

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 $I = \{0, 1, 2, \dots\}$, 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 $X_0, X_1, X_2 \dots$ 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 $\{1,\dots,40\}$ denoting the 40 squares of a Monopoly board—from Go to Boardwalk. The index set is $\{0, 1, 2, \dots\}$ 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 $X_0, X_1, X_2, \dots$, 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 $\{0, 1, 2, \dots\}$.

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 $X_1, X_2, \dots$ 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 $(n - k)/n = (60 - 20)/60 = 2/3$. ■

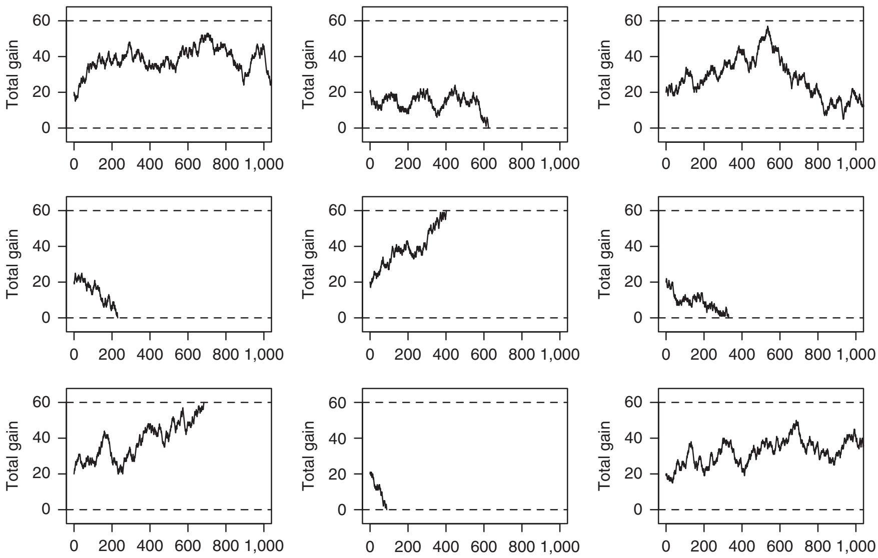

Figure 1.4 Random walk and gambler's ruin.

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 $\mathbb{R}^2$. ■

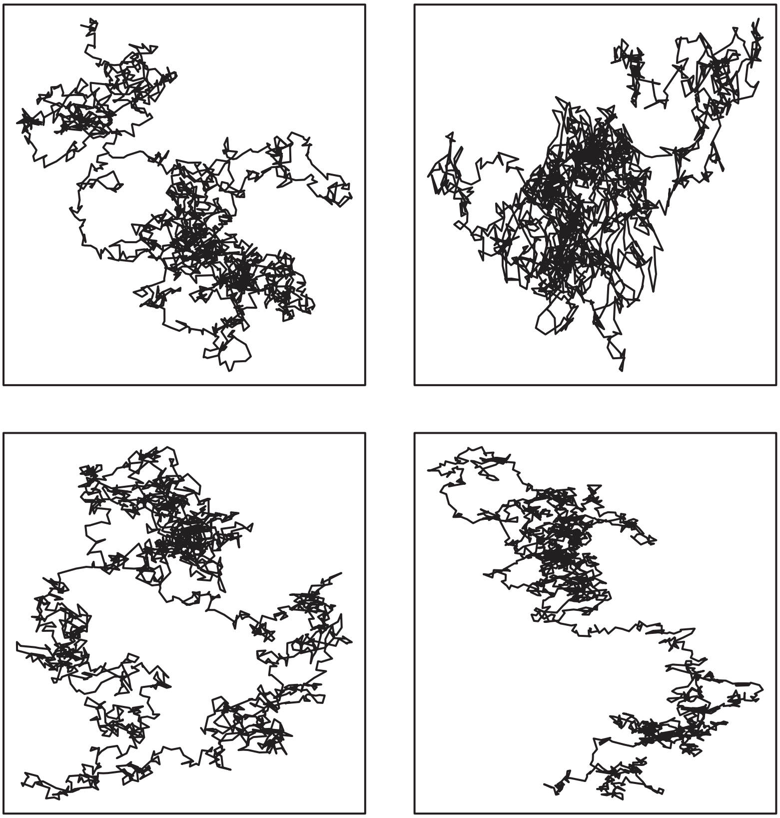

Figure 1.5 Simulations of two-dimensional Brownian motion.

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 $B_1, \dots, B_k$, 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 $B_1, \dots, B_k$ 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.

Bayes' Rule

For events A and B,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

Bayes' rule is a consequence of the law of total probability and the definition of conditional probability, as

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

For any event B, the events B and B^c partition the sample space. Given a countable sequence of events $B_1, B_2, \ldots$, which partition the sample space, a more general form of Bayes' rule is

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}, \quad \text{for} \quad i = 1, 2, \ldots$$

Example 1.11 The use of polygraphs (lie detectors) is controversial, and many scientists feel that they should be banned. On the contrary, some polygraph advocates claim that they are mostly accurate. In 1998, the Supreme Court (*United States v. Sheffer*) supported the right of state and federal governments to bar polygraph evidence in court.

Assume that one person in a company of 100 employees is a thief. To find the thief the company will administer a polygraph test to all its employees. The lie detector has the property that if a subject is a liar, there is a 95% probability that the polygraph will detect that they are lying. However, if the subject is telling the truth, there is a 10% chance the polygraph will report a *false positive* and assert that the subject is lying.

Assume that a random employee is given the polygraph test and asked whether they are the thief. The employee says "no," and the lie detector reports that they are lying. Find the probability that the employee is in fact lying.

Solution Let L denote the event that the employee is a liar. Let D denote the event that the lie detector reports that the employee is a liar. The desired probability is $P(L|D)$. By Bayes' rule,

$$\begin{aligned} P(L|D) &= \frac{P(D|L)P(L)}{P(D|L)P(L) + P(D|L^c)P(L^c)} \\ &= \frac{(0.95)(0.01)}{(0.95)(0.01) + (0.10)(0.99)} = 0.088. \end{aligned}$$

There is less than a 10% chance that the employee is in fact the thief!

Many people, when first given this problem and asked to guess the answer, choose a probability closer to 90%. The mistake is a consequence of confusing the conditional probabilities $P(L|D)$ and $P(D|L)$, the probability that the individual is a liar, given that the polygraph says they are, with the probability that the polygraph says they are lying, given that they are a liar. Since the population of truth tellers is relatively big, the number of false positives—truth tellers whom the lie detector falsely records as being a liar—is also significant. In this case, about 10% of 99, or about 10 employees will be false positives. Assuming that the lie detector correctly identifies the thief as lying, there will be about 11 employees who are identified as liars by the polygraph. The probability that one of them chosen at random is in fact the thief is only about 1/11. ■

Conditional Distribution

The *distribution* of a random variable X refers to the set of values of X and their corresponding probabilities. The distribution of a random variable is specified with either a probability mass function (pmf), if X is discrete, or a probability density function (pdf), if X is continuous.

For more than one random variable, there is a joint distribution, specified by either a joint pmf or a joint pdf.

If X and Y are discrete random variables, their joint pmf is $P(X = x, Y = y)$, considered a function of x and y. If X and Y are continuous, the joint density function $f(x, y)$ satisfies

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\, dt\, ds,$$

for all $x, y \in \mathbb{R}$.

For jointly distributed random variables X and Y, the *conditional distribution of Y given $X = x$* is specified by either a conditional pmf or a conditional pdf.

Discrete Case

The *conditional pmf of Y given $X = x$* is

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)},$$

defined when $P(X = x) > 0$. The conditional pmf is a function of y, where x is treated as fixed.

Example 1.12 Max chooses an integer X uniformly at random between 1 and 100. If $X = x$, Mary then chooses an integer Y uniformly at random between 1 and x. Find the conditional pmf of Y given $X = x$.

Solution By the structure of this two-stage random experiment, the conditional distribution of Y given $X = x$ is uniform on $\{1, \ldots, x\}$. Thus, the conditional pmf is

$$P(Y = y|X = x) = \frac{1}{x}, \text{ for } y = 1, \ldots, x.$$

■

Note that the conditional pmf *is* a probability function. For fixed x, the probabilities $P(Y = y|X = x)$ are nonnegative and sum to 1, as

$$\sum_y P(Y = y|X = x) = \sum_y \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x)}{P(X = x)} = 1.$$

Example 1.13 The joint pmf of X and Y is

$$P(X = x, Y = y) = \frac{x + y}{18}, \quad \text{for } \ x, y = 0, 1, 2.$$

Find the conditional pmf of Y given $X = x$.

Solution The marginal distribution of X is

$$P(X = x) = \sum_{y=0}^{2} P(X = x, Y = y) = \frac{x}{18} + \frac{x + 1}{18} + \frac{x + 2}{18} = \frac{x + 1}{6},$$

for $x = 0, 1, 2$. The conditional pmf is

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{(x + y)/18}{(x + 1)/6} = \frac{x + y}{3(x + 1)},$$

for $y = 0, 1, 2$. ■

Example 1.14 A bag contains 2 red, 3 blue, and 4 white balls. Three balls are picked from the bag (sampling without replacement). Let B be the number of blue balls picked. Let R be the number of red balls picked. Find the conditional pmf of B given $R = 1$.

Solution We have

$$\begin{aligned} P(B = b|R = 1) &= \frac{P(B = b, R = 1)}{P(R = 1)} \\ &= \frac{\binom{3}{b}\binom{2}{1}\binom{4}{3-b-1} \Big/ \binom{9}{3}}{\binom{2}{1}\binom{7}{2} \Big/ \binom{9}{3}} = \frac{2\binom{3}{b}\binom{4}{2-b}}{42} \end{aligned}$$

$$= \frac{\binom{3}{b}\binom{4}{2-b}}{21} = \begin{cases} 2/7, & \text{if } b = 0, \\ 4/7, & \text{if } b = 1, \\ 1/7, & \text{if } b = 2. \end{cases}$$

■

Continuous Case

For continuous random variables X and Y, the *conditional density function of Y given* $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)},$$

where f_X is the marginal density function of X. The conditional density is a function of y, where x is treated as fixed.

Conditional densities are used to compute conditional probabilities. For $R \subseteq \mathbb{R}$,

$$P(Y \in R|X = x) = \int_R f_{Y|X}(y|x)\, dy.$$

Example 1.15 Random variables X and Y have joint density

$$f(x,y) = e^{-x}, \text{for } 0 < y < x < \infty.$$

Find $P(Y < 2|X = 5)$.

Solution The desired probability is

$$P(Y < 2|X = 5) = \int_0^2 f_{Y|X}(y|5)\, dy.$$

To find the conditional density function $f_{Y|X}(y|x)$, find the marginal density

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_0^x e^{-x}\, dy = xe^{-x}, \quad \text{for} \quad x > 0.$$

This gives

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{e^{-x}}{xe^{-x}} = \frac{1}{x}, \quad \text{for} \quad 0 < y < x.$$

The conditional distribution of Y given $X = x$ is uniform on $(0, x)$. Hence,

$$P(Y < 2|X = 5) = \int_0^2 f_{Y|X}(y|5)\, dy = \int_0^2 \frac{1}{5}\, dy = \frac{2}{5}.$$

■

Example 1.16 Tom picks a real number X uniformly distributed on $(0, 1)$. Tom shows his number x to Marisa who then picks a number Y uniformly distributed on $(0, x)$. Find (i) the conditional distribution of Y given $X = x$; (ii) the joint distribution of X and Y; and (iii) the marginal density of Y.

Solution

(i) The conditional distribution of Y given $X = x$ is given directly in the statement of the problem. The distribution is uniform on $(0, x)$. Thus,

$$f_{Y|X}(y|x) = \frac{1}{x}, \quad \text{for} \quad 0 < y < x.$$

(ii) For the joint density,

$$f(x, y) = f_{Y|X}(y|x)f_X(x) = \frac{1}{x}(1) = \frac{1}{x}, \quad \text{for} \quad 0 < y < x < 1.$$

(iii) To find the marginal density of Y, integrate out the x variable in the joint density function. This gives

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_y^1 \frac{1}{x}\, dx = -\ln\ y, \quad \text{for} \quad 0 < y < 1.$$

■

1.5 CONDITIONAL EXPECTATION

A conditional expectation is an expectation computed with respect to a conditional probability distribution. Write $E(Y|X = x)$ for the *conditional expectation of Y given X = x*.

Conditional Expectation of Y given $X = x$

$$E(Y|X = x) = \begin{cases} \sum_y yP(Y = y|X = x), & \text{discrete,} \\ \int_{-\infty}^{\infty} yf_{Y|X}(y|x)\, dy, & \text{continuous.} \end{cases}$$

Most important is that $E(Y|X = x)$ is a function of x.

Example 1.17 A school cafeteria has two registers. Let X and Y denote the number of students in line at the first and second registers, respectively. The joint pmf of X and Y is specified by the following joint distribution table.

		Y 0	1	2	3	4
	0	0.15	0.14	0.03	0	0
	1	0.14	0.12	0.06	0.01	0
X	2	0.03	0.06	0.10	0.03	0.02
	3	0	0.01	0.03	0.02	0.01
	4	0	0	0.02	0.01	0.01

Find the expected number of people in line at the second register if there is one person at the first register.

Solution The problem asks for $E(Y|X = 1)$. We have

$$E(Y|X = 1) = \sum_{y=0}^{4} yP(Y = y|X = 1) = \sum_{y=0}^{4} y\frac{P(X = 1, Y = y)}{P(X = 1)}.$$

The marginal probability $P(X = 1)$ is obtained by summing over the $X = 1$ row of the table. That is, $P(X = 1) = 0.14 + 0.12 + 0.06 + 0.01 + 0 = 0.33$. Hence,

$$\begin{aligned} E(Y|X = 1) &= \frac{1}{0.33}\sum_{y=0}^{4} yP(X = 1, Y = y) \\ &= \frac{1}{0.33}[0(0.14) + 1(0.12) + 2(0.06) + 3(0.01) + 4(0)] \\ &= 0.818. \end{aligned}$$

■

Example 1.18 Let X and Y be independent Poisson random variables with respective parameters λ and μ. Find the conditional expectation of Y given $X + Y = n$.

Solution First find the conditional pmf of Y given $X + Y = n$. We use the fact that the sum of independent Poisson random variables has a Poisson distribution whose parameter is the sum of the individual parameters. That is, $X + Y$ has a Poisson distribution with parameter $\lambda + \mu$. This gives

$$\begin{aligned} P(Y = y|X + Y = n) &= \frac{P(Y = y, X + Y = n)}{P(X + Y = n)} = \frac{P(Y = y, X = n - y)}{P(X + Y = n)} \\ &= \frac{P(Y = y)P(X = n - y)}{P(X + Y = n)} \\ &= \frac{(e^{-\mu}\mu^y/y!)(e^{-\lambda}\lambda^{n-y}/(n - y)!)}{e^{-(\lambda+\mu)}(\lambda + \mu)^n/n!} \end{aligned}$$

$$= \frac{n!}{y!(n-y)!} \left(\frac{\mu^y \lambda^{n-y}}{(\lambda+\mu)^n} \right)$$

$$= \binom{n}{y} \left(\frac{\mu}{\lambda+\mu} \right)^y \left(\frac{\lambda}{\lambda+\mu} \right)^{n-y},$$

for $y = 0, \dots, n$. The form of the conditional pmf shows that the conditional distribution is binomial with parameters n and $p = \mu/(\lambda + \mu)$. The desired conditional expectation is the mean of this binomial distribution. That is,

$$E(Y|X + Y = n) = np = \frac{n\mu}{\lambda + \mu}.$$

■

Example 1.19 Assume that X and Y have joint density

$$f(x, y) = \frac{2}{xy}, \text{ for } 1 < y < x < e.$$

Find $E(Y|X = x)$.

Solution The marginal density of X is

$$f_X(x) = \int_1^x \frac{2}{xy}\, dy = \frac{2 \ln x}{x}, \text{ for } 1 < x < e.$$

The conditional density of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{2/(xy)}{2 \ln x/x} = \frac{1}{y \ln x}, \text{ for } 1 < y < x,$$

with conditional expectation

$$E(Y|X = x) = \int_1^x y f_{Y|X}(y|x)\, dy = \int_1^x \frac{y}{y \ln x}\, dy = \frac{x-1}{\ln x}.$$

■

Key properties of conditional expectation follow.

Properties of Conditional Expectation

1. (Linearity) For constants a and b and random variables X, Y, and Z,

$$E(aY + bZ|X = x) = aE(Y|X = x) + bE(Z|X = x).$$

2. If g is a function,

$$E(g(Y)|X = x) = \begin{cases} \sum_y g(y)P(Y = y|X = x), & \text{discrete,} \\ \int_{-\infty}^{\infty} g(y)f_{Y|X}(y|x)\, dy, & \text{continuous.} \end{cases}$$

3. (Independence) If X and Y are independent,

$$E(Y|X = x) = E(Y).$$

4. If $Y = g(X)$ is a function of X,

$$E(Y|X = x) = g(x).$$

Proof. Properties 1 and 2 are consequences of the fact that conditional expectation *is* an expectation and thus retains all the properties, such as linearity, of the regular expectation. For a proof of property 2, which is sometimes called *the law of the unconscious statistician*, see Dobrow (2013).

For the independence property in the discrete case, if X and Y are independent,

$$E(Y|X = x) = \sum_y yP(Y = y|X = x) = \sum_y yP(Y = y) = E(Y).$$

The continuous case is similar. For property 4,

$$E(Y|X = x) = E(g(X)|X = x) = E(g(x)|X = x) = g(x).$$

■

Example 1.20 Consider random variables X, Y, and U, where U is uniformly distributed on (0, 1). Find the conditional expectation

$$E\left(UX^2 + (1 - U)Y^2|U = u\right).$$

Solution By linearity of conditional expectation,

$$\begin{aligned} E\left(UX^2 + (1 - U)Y^2|U = u\right) &= E\left(uX^2 + (1 - u)Y^2|U = u\right) \\ &= uE\left(X^2|U = u\right) + (1 - u)E\left(Y^2|U = u\right). \end{aligned}$$

If X and Y are also independent of U, the latter expression reduces to

$$uE\left(X^2\right) + (1 - u)E\left(Y^2\right).$$

■

Extending the scope of conditional expectation, we show how to condition on a general event. Given an event A, the *indicator* for A is the 0–1 random variable defined as

$$I_A = \begin{cases} 1, & \text{if } A \text{ occurs,} \\ 0, & \text{if } A \text{ does not occur.} \end{cases}$$

Conditional Expectation Given an Event

Let A be an event such that $P(A) > 0$. The *conditional expectation of Y given A* is

$$E(Y|A) = \frac{E(YI_A)}{P(A)}.$$

For the discrete case, the formula gives

$$E(Y|A) = \frac{1}{P(A)} \sum_y yP(\{Y = y\} \cap A) = \sum_y yP(Y = y|A).$$

Let $A_1, \dots, A_k$ be a sequence of events that partition the sample space. Observe that

$$I_{A_1} + \cdots + I_{A_k} = 1,$$

since every outcome $\omega \in \Omega$ is contained in exactly one of the A_is. It follows that

$$Y = \sum_{i=1}^{k} YI_{A_i}.$$

Taking expectations gives

$$E(Y) = \sum_{i=1}^{k} E\left(YI_{A_i}\right) = \sum_{i=1}^{k} \left(\frac{E\left(YI_{A_i}\right)}{P(A)} \right) P(A) = \sum_{i=1}^{k} E(Y|A_i)P(A_i).$$

The result is known as the law of total expectation. Note the similarity with the law of total probability.

The law of total expectation is often used with partitioning events $\{X = x\}$. This gives

$$E(Y) = \sum_x E(Y|X = x)P(X = x).$$

In summary, here are two forms of the law of total expectation.

Law of Total Expectation

Let Y be a random variable. Let $A_1, \dots, A_k$ be a sequence of events that partition the sample space. Then,

$$E(Y) = \sum_{i=1}^{k} E(Y|A_i)P(A_i).$$

If X and Y are jointly distributed random variables,

$$E(Y) = \sum_x E(Y|X = x)P(X = x).$$

Example 1.21 A fair coin is flipped repeatedly. Find the expected number of flips needed to get two heads in a row.

Solution Let Y be the number of flips needed. Consider three events: (i) T, the first flip is tails; (ii) HT, the first flip is heads and the second flip is tails; and (iii) HH, the first two flips are heads. The events T, HT, HH partition the sample space. By the law of total expectation,

$$\begin{aligned} E(Y) &= E(Y|T)P(T) + E(Y|HT)P(HT) + E(Y|HH)P(HH) \\ &= E(Y|T)\frac{1}{2} + E(Y|HT)\frac{1}{4} + (2)\frac{1}{4}. \end{aligned}$$

Consider $E(Y|T)$. Assume that the first flip is tails. Since successive flips are independent, after the first tails we *start over* waiting for two heads in a row. Since one flip has been used, it follows that $E(Y|T) = 1 + E(Y)$. Similarly, after first heads and then tails we start over again, having used up two coin tosses. Thus, $E(Y|HT) = 2 + E(Y)$. This gives

$$E(Y) = (1 + E(Y))\frac{1}{2} + (2 + E(Y))\frac{1}{4} + (2)\frac{1}{4} = E(Y)\frac{3}{4} + \frac{3}{2}.$$

Solving for $E(Y)$ gives $E(Y)(1/4) = 3/2$, or $E(Y) = 6$. ■

Example 1.22 Every day Bob goes to the pizza shop, orders a slice of pizza, and picks a topping—pepper, pepperoni, pineapple, prawns, or prosciutto—uniformly at random. On the day that Bob first picks pineapple, find the expected number of prior days in which he picked pepperoni.

Solution Let Y be the number of days, before the day Bob first picked pineapple, in which he picks pepperoni. Let X be the number of days needed for Bob to first pick pineapple. Then, X has a geometric distribution with parameter $1/5$.

If $X = x$, then on the first $x - 1$ days pineapple was not picked. And for each of these days, given that pineapple was not picked, there was a $1/4$ chance of picking pepperoni. The conditional distribution of Y given $X = x$ is binomial with parameters $x - 1$ and $1/4$. Thus, $E[Y|X = x] = (x - 1)/4$, and

$$E(Y) = \sum_{x=1}^{\infty} E(Y|X = x)P(X = x)$$

$$= \sum_{x=1}^{\infty} \left(\frac{x-1}{4}\right)\left(\frac{4}{5}\right)^{x-1}\frac{1}{5}$$
$$= \left(\frac{1}{4}\sum_{x=1}^{\infty} x\left(\frac{4}{5}\right)^{x-1}\frac{1}{5}\right) - \left(\frac{1}{4}\sum_{x=1}^{\infty}\left(\frac{4}{5}\right)^{x-1}\frac{1}{5}\right)$$
$$= \frac{1}{4}E(X) - \frac{1}{4} = \frac{5}{4} - \frac{1}{4} = 1.$$

■

R : Simulation of Bob's Pizza Probability

```
> trials <- 10000  # simulation repeated 10,000 times
> simlist <- numeric(trials)
> toppings <- c("pepper","pepperoni","pineapple",
   "prawns","prosciutto")
> for (i in 1:trials) {
> pineapple <- 0
> pepperoni <- 0 # counts pepperonis before pineapple
> while (pineapple == 0) {
     # pick toppings until pineapple is selected
   pick <- sample(toppings,1)
   if (pick == "pepperoni") pepperoni <-pepperoni + 1
   if (pick == "pineapple") pineapple <- 1   }
> simlist[i] <- pepperoni }
> mean(simlist)
[1] 0.9966
```

The next example illustrates conditional expectation given an event in the continuous case.

Example 1.23 The time that Joe spends talking on the phone is exponentially distributed with mean 5 minutes. What is the expected length of his phone call if he talks for more than 2 minutes?

Solution Let Y be the length of Joe's phone call. With $A = \{Y > 2\}$, the desired conditional expectation is

$$E(Y|A) = E(Y|Y > 2) = \frac{1}{P(Y > 2)}\int_2^{\infty} y\frac{1}{5}e^{-y/5}\,dy$$
$$= \left(\frac{1}{e^{-2/5}}\right)7e^{-2/5} = 7\text{ minutes.}$$

Note that the solution can be obtained using the memoryless property of the exponential distribution. The conditional distribution of Y given $Y > 2$ is equal to the distribution of $2 + Z$, where Z has the same distribution has Y. Thus,

$$E(Y|Y > 2) = E(2 + Z) = 2 + E(Z) = 2 + E(Y) = 2 + 5 = 7.$$

■

Conditioning on a Random Variable

From conditioning on an event, we introduce the notion of conditioning on a random variable, a powerful tool for computing conditional expectations and probabilities.

Recall that if X is a random variable and g is a function, then $Y = g(X)$ is a random variable, which is a function of X. The conditional expectation $E(Y|X = x)$ is a function of x. We apply this function to the random variable X and define a new random variable called the *conditional expectation of Y given X*, written $E(Y|X)$. The defining properties of $E(Y|X)$ are given here.

Conditional Expectation of *Y* given *X*

The conditional expectation $E(Y|X)$ has three defining properties.

1. $E(Y|X)$ is a random variable.
2. $E(Y|X)$ is a function of X.
3. $E(Y|X)$ is equal to $E(Y|X = x)$ whenever $X = x$. That is, if

$$E(Y|X = x) = g(x), \text{ for all } x,$$

then $E(Y|X) = g(X)$.

Example 1.24 Let X be uniformly distributed on (0, 1). If $X = x$, a second number Y is picked uniformly on $(0, x)$. Find $E(Y|X)$.

Solution For this two-stage random experiment, the conditional distribution of Y given $X = x$ is uniform on $(0, x)$, for $0 < x < 1$. It follows that $E(Y|X = x) = x/2$. Since this is true for all x, $E(Y|X) = X/2$. ■

It may seem that the difference between $E(Y|X)$ and $E(Y|X = x)$ is just a matter of notation, with capital letters replacing lowercase letters. However, as much as they look the same, there is a fundamental difference. The conditional expectation $E(Y|X = x)$ is a function of x. Its domain is a set of real numbers. The deterministic function can be evaluated and graphed. For instance, in the last example $E(Y|X = x) = x/2$ is a linear function of x with slope 1/2.

On the contrary, $E(Y|X)$ is a random variable. As such, it has a probability distribution. Since $E(Y|X)$ is a random variable with some probability distribution, it makes sense to take *its* expectation with respect to that distribution. The expectation of a conditional expectation may seem pretty far out. But it leads to one of the most important results in probability.

Law of Total Expectation

For random variables X and Y,

$$E(Y) = E(E(Y|X)).$$

We prove this result for the discrete case, and leave the continuous case as an exercise for the reader.

Proof.

$$\begin{aligned}
E(E(Y|X)) &= \sum_x E(Y|X = x)P(X = x) \\
&= \sum_x \left(\sum_y yP(Y = y|X = x) \right) P(X = x) \\
&= \sum_y y \sum_x P(Y = y|X = x)P(X = x) \\
&= \sum_y y \sum_x P(X = x, Y = y) \\
&= \sum_y yP(Y = y) = E(Y).
\end{aligned}$$

■

Example 1.25 Angel will harvest N tomatoes in her garden, where N has a Poisson distribution with parameter λ. Each tomato is checked for defects. The chance that a tomato has defects is p. Defects are independent from tomato to tomato. Find the expected number of tomatoes with defects.

Solution Let X be the number of tomatoes with defects. The conditional distribution of X given $N = n$ is binomial with parameters n and p. This gives $E(X|N = n) = np$. Since this is true for all n, $E(X|N) = Np$. By the law of total expectation,

$$E(X) = E(E(X|N)) = E(Np) = pE(N) = p\lambda.$$

■

Example 1.26 Ellen's insurance will pay for a medical expense subject to a \$100 deductible. Assume that the amount of the expense is exponentially distributed with mean \$500. Find the expectation and standard deviation of the payout.

Solution Let M be the amount of the medical expense and let X be the insurance company's payout. Then,

$$X = \begin{cases} M - 100, & \text{if } M > 100, \\ 0, & \text{if } M \leq 100, \end{cases}$$

where M is exponentially distributed with parameter $1/500$. To find the expected payment, apply the law of total expectation, giving

$$\begin{aligned} E(X) = E(E(X|M)) &= \int_0^\infty E(X|M = m)\lambda e^{-\lambda m}\, dm \\ &= \int_{100}^\infty E(M - 100|M = m)\frac{1}{500}e^{-m/500}\, dm \\ &= \int_{100}^\infty (m - 100)\frac{1}{500}e^{-m/500}\, dm \\ &= 500e^{-100/500} = \$409.365. \end{aligned}$$

For the standard deviation, first find

$$\begin{aligned} E\left(X^2\right) = E\left(E\left(X^2|M\right)\right) &= \int_0^\infty E\left(X^2|M = m\right)\lambda e^{-\lambda m}\, dm \\ &= \int_{100}^\infty E\left((M - 100)^2|M = m\right)\frac{1}{500}e^{-m/500}\, dm \\ &= \int_{100}^\infty (m - 100)^2\frac{1}{500}e^{-m/500}\, dm \\ &= 500000e^{-1/5} = 409365. \end{aligned}$$

This gives

$$\begin{aligned} SD(X) &= \sqrt{Var(X)} = \sqrt{E(X^2) - E(X)^2} \\ &= \sqrt{409365 - (409.365)^2} = \$491.72. \end{aligned}$$

■

R : Simulation of Ellen's Payout

```
> trials <- 100000
> simlist <- numeric(trials)
> for (i in 1:trials) {
> expense <- rexp(1,1/500)
> payout <- max(0, expense-100)
```

```
> simlist[i] <- payout}
> mean(simlist)
[1] 410.0308
> sd(simlist)
[1] 493.5457
```

Example 1.27 (Random sum of random variables) A stochastic model for the cost of damage from traffic accidents is given by Van der Lann and Louter (1986). Let X_k be the amount of damage from an individual's kth traffic accident. Assume $X_1, X_2, \ldots$ is an i.i.d. sequence with mean μ. The number of accidents N for an individual driver is a random variable with mean λ. It is assumed that the number of accidents is independent of the amount of damages for each accident. That is, N is independent of the X_k. For an individual driver, find the total cost of damages.

Solution Let T be the total cost of damages. Then,

$$T = X_1 + \cdots + X_N = \sum_{k=1}^{N} X_k.$$

The number of summands is random. The random variable T is a random sum of random variables. By the law of total expectation $E(T) = E(E(T|N))$. To find $E(T|N)$, consider

$$\begin{aligned} E(T|N = n) &= E\left(\sum_{k=1}^{N} X_k | N = n\right) = E\left(\sum_{k=1}^{n} X_k | N = n\right) \\ &= E\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} E(X_k) = n\mu, \end{aligned}$$

where the third equality is because N is independent of the X_k. Since the final equality holds for all n, $E(T|N) = N\mu$. By the law of total expectation,

$$E(T) = E(E(T|N)) = E(N\mu) = \mu E(N) = \mu\lambda.$$

The result is intuitive. The expected total cost is the product of the expected number of accidents and the expected cost per accident.

Note that it would have been incorrect to write

$$E\left(\sum_{k=1}^{N} X_k\right) = \sum_{k=1}^{N} E(X_k).$$

Linearity of expectation does not apply here because the number of summands is random, not fixed. Indeed, this equation does not even make sense as the left-hand side is a fixed number (the expectation of a random variable), while the right-hand side is a random variable. ■

Computing Probabilities by Conditioning

For an event A, let I_A be the indicator for A. Then,

$$E(I_A) = (1)P(A) + (0)P(A^c) = P(A).$$

From this simple fact, one sees that probabilities can always be expressed as expectations. As such, the law of total expectation can be used when computing probabilities. In particular, if X is a discrete random variable,

$$\begin{aligned}
P(A) = E(I_A) &= E(E(I_A|X)) \\
&= \sum_x E(I_A|X = x)P(X = x) \\
&= \sum_x [(1)P(I_A = 1|X = x)P(X = x) + (0)P(I_A = 0|X = x)P(X = x)] \\
&= \sum_x P(A|X = x)P(X = x).
\end{aligned}$$

We have recaptured the law of total probability, where the conditioning events are $\{X = x\}$ for all x.

If X is continuous with density function f_X,

$$P(A) = \int_{-\infty}^{\infty} E(I_A|X = x)f_X(x)\,dx = \int_{-\infty}^{\infty} P(A|X = x)f_X(x)\,dx,$$

which gives the continuous version of the law of total probability.

Example 1.28 Max arrives to class at time X. Mary arrives at time Y. Assume that X and Y have exponential distributions with respective parameters λ and μ. If arrival times are independent, find the probability that Mary arrives before Max.

Solution Let $A = \{Y < X\}$ be the event that Mary arrives to class before Max. By conditioning on Max's arrival time,

$$\begin{aligned}
P(A) = P(Y < X) &= \int_{-\infty}^{\infty} P(Y < X|X = x)f_X(x)\,dx \\
&= \int_0^{\infty} P(Y < x|X = x)\lambda e^{-\lambda x}\,dx \\
&= \int_0^{\infty} P(Y < x)\lambda e^{-\lambda x}\,dx
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty (1 - e^{-\mu x})\lambda e^{-\lambda x}\, dx \\
&= 1 - \int_0^\infty \lambda e^{-(\lambda+\mu)x}\, dx \\
&= 1 - \frac{\lambda}{\lambda+\mu} = \frac{\mu}{\lambda+\mu}.
\end{aligned}
$$

The fourth equality is by independence of X and Y. ■

Example 1.29 (Sums of independent random variables) Assume that X and Y are independent continuous random variables with density functions f_X and f_Y, respectively. (i) Find the density function of $X + Y$. (ii) Use part (i) to find the density of the sum of two independent standard normal random variables.

Solution

(i) Conditioning on Y,

$$
\begin{aligned}
P(X + Y \le t) &= \int_{-\infty}^{\infty} P(X + Y \le t | Y = y) f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} P(X \le t - y | Y = y) f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} P(X \le t - y) f_Y(y)\, dy.
\end{aligned}
$$

Differentiating with respect to t gives

$$
f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(t - y) f_Y(y)\, dy. \tag{1.3}
$$

Equation (1.3) is known as a *convolution* formula.

(ii) For X and Y independent standard normal random variables, by Equation (1.3), $X + Y$ has density

$$
\begin{aligned}
f_{X+Y}(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(t-y)^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \\
&= \frac{1}{\sqrt{4\pi}} e^{-t^2/4} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1/2)}} e^{-(y-t/2)^2/2(1/2)}\, dy \qquad (1.4) \\
&= \frac{1}{\sqrt{4\pi}} e^{-t^2/4},
\end{aligned}
$$

which is the density of a normal distribution with mean 0 and variance 2. The last equality is because the integrand in Equation (1.4) is the density of a normal distribution with mean $t/2$ and variance 1/2, and thus integrates to 1. ■

Conditional Variance

Analogous to conditional expectation, the conditional variance is a variance taken with respect to a conditional distribution. Given random variables X and Y, let $\mu_x = E(Y|X = x)$. Then, the conditional variance $Var(Y|X = x)$ is defined as

$$Var(Y|X = x) = E\left((Y - \mu_x)^2 | X = x\right)$$
$$= \begin{cases} \sum_y (y - \mu_x)^2 P(Y = y|X = x), & \text{discrete,} \\ \int_{-\infty}^{\infty} (y - \mu_x)^2 f_{Y|X}(y|x)\, dy, & \text{continuous.} \end{cases}$$

Compare with the regular variance formula

$$Var(Y) = E\left((Y - \mu)^2\right),$$

where $\mu = E(Y)$. Observe that the conditional expectation $E(Y|X = x)$ takes the place of the unconditional expectation $E(Y)$ in the regular variance formula.

Example 1.30 Let N be a positive, integer-valued random variable. If $N = n$, flip n coins, each of which has heads probability p. Let Y be the number of coins which come up heads. Find $Var(Y|N = n)$.

Solution The conditional distribution of Y given $N = n$ is binomial with parameters n and p. From the properties of the binomial distribution,

$$Var(Y|N = n) = np(1 - p).$$

■

Properties of the variance transfer to the conditional variance.

Properties of Conditional Variance

1.
$$Var(Y|X = x) = E\left(Y^2|X = x\right) - (E(Y|X = x))^2.$$

2. For constants a and b,

$$Var(aY + b|X = x) = a^2 Var(Y|X = x).$$

As with the development of conditional expectation, define the *conditional variance* $Var(Y|X)$ as the random variable which is a function of X and which takes the value $Var(Y|X = x)$ when $X = x$.

The law of total variance shows how to find the variance of a random variable by conditioning.

Law of Total Variance

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X)).$$

The proof is easier than you might think. We have that

$$\begin{aligned} E(Var(Y|X)) &= E\left(E\left(Y^2|X\right) - (E(Y|X))^2\right) \\ &= E\left(E\left(Y^2|X\right)\right) - E\left((E(Y|X))^2\right) \\ &= E\left(Y^2\right) - E\left((E(Y|X))^2\right). \end{aligned}$$

And

$$\begin{aligned} Var(E(Y|X)) &= E\left((E(Y|X))^2\right) - (E(E(Y|X)))^2 \\ &= E\left((E(Y|X))^2\right) - (E(Y))^2. \end{aligned}$$

Thus,

$$\begin{aligned} &E(Var(Y|X)) + Var(E(Y|X)) \\ &\quad = \left(E(Y^2) - E\left((E(Y|X)^2)\right)\right) + \left(E\left((E(Y|X))^2\right) - (E(Y))^2\right) \\ &\quad = E\left(Y^2\right) - (E(Y))^2 = Var(Y). \end{aligned}$$

Example 1.31 A number X is uniformly distributed on $(0, 1)$. If $X = x$, then Y is picked uniformly on $(0, x)$. Find the variance of Y.

Solution The conditional distribution of Y given $X = x$ is uniform on $(0, x)$. From properties of the uniform distribution,

$$E(Y|X = x) = \frac{x}{2} \quad \text{and} \quad Var(Y|X = x) = \frac{x^2}{12}.$$

This gives $E(Y|X) = X/2$ and $Var(Y|X) = X^2/12$. By the law of total variance,

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X)) = E\left(\frac{X^2}{12}\right) + Var\left(\frac{X}{2}\right)$$

$$= \frac{1}{12}E\left(X^2\right) + \frac{1}{4}Var(X) = \frac{1}{12}\left(\frac{1}{3}\right) + \frac{1}{4}\left(\frac{1}{12}\right)$$
$$= \frac{7}{144} = 0.04861.$$

■

R : Simulation of *Var*(*Y*)

The structure of this two-stage random experiment makes it especially easy to simulate in R .

```
> trials <- 100000
> simlist <- replicate(trials,runif(1,0,runif(1)))
> var(simlist)
[1] 0.04840338
```

Example 1.32 (Variance of a random sum of random variables) Assume that $X_1, X_2, \ldots$ is an i.i.d. sequence with common mean μ_X and variance σ_X^2. Let N be a positive, integer-valued random variable that is independent of the X_i with mean μ_N and variance σ_N^2. Let $T = X_1 + \cdots + X_N$. Find the variance of T.

Solution Random sums of random variables were introduced in Example 1.27 where the expectation $E(T) = \mu_X \mu_N$ was found using the law of total expectation. By the law of total variance,

$$Var(T) = Var\left(\sum_{k=1}^{N} X_k\right) = E\left(Var\left(\sum_{k=1}^{N} X_k|N\right)\right) + Var\left(E\left(\sum_{k=1}^{N} X_k|N\right)\right).$$

We have that

$$\begin{aligned} Var\left(\sum_{k=1}^{N} X_k|N = n\right) &= Var\left(\sum_{k=1}^{n} X_k|N = n\right) \\ &= Var\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} Var(X_k) \\ &= n\sigma_X^2. \end{aligned}$$

The second equality is because N is independent of the X_k. The third equality is because the X_k are independent. This gives

$$Var\left(\sum_{k=1}^{N} X_k|N\right) = N\sigma_X^2.$$

From results for conditional expectation,

$$E\left(\sum_{k=1}^{N} X_k|N\right) = NE(X_1) = N\mu_X.$$

This gives

$$\begin{aligned} Var(T) &= E\left(Var\left(\sum_{k=1}^{N} X_k|N\right)\right) + Var\left(E\left(\sum_{k=1}^{N} X_k|N\right)\right) \\ &= E\left(N\sigma_X^2\right) + Var(N\mu_X) \\ &= \sigma_X^2 E(N) + \mu_X^2 Var(N) \\ &= \sigma_X^2 \mu_N + \mu_X^2 \sigma_N^2. \end{aligned}$$

■

Random sums of independent random variables will arise in several contexts. Results are summarized here.

Random Sums of Random Variables

Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with common mean μ_X and variance σ_X^2. Let N be a positive, integer-valued random variable independent of the X_i with $E(N) = \mu_N$ and $Var(N) = \sigma_N^2$. Let $T = \sum_{i=1}^{N} X_i$. Then,

$$E(T) = \mu_X \mu_N \text{ and } Var(T) = \sigma_X^2 \mu_N + \sigma_N^2 \mu_X^2.$$

EXERCISES

1.1 For the following scenarios identify a stochastic process $\{X_t, t \in I\}$, describing (i) X_t in context, (ii) state space, and (iii) index set. State whether the state space and index set are discrete or continuous.

(a) From day to day the weather in International Falls, Minnesota is either rain, clear, or snow.

Solution: Let X_t denote the weather on day t. Discrete state space is $\mathcal{S} = \{\text{Rain, Clear, Snow}\}$. *Discrete index set is* $I = \{0, 1, 2, \cdots\}$.

(b) At the end of each year, a 4-year college student either advances in grade, repeats their grade, or drops out.

(c) Seismologists record daily earthquake magnitudes in Chile. The largest recorded earthquake in history was the Valdivia, Chile earthquake on May 22, 1960, which had a magnitude of 9.5 on the Richter scale.

(d) Data are kept on the circumferences of trees in an arboretum. The arboretum covers a two square-mile area.

(e) Starting Monday morning at 9 a.m., as students arrive to class, the teacher records student arrival times. The class has 30 students and lasts for 60 minutes.

(f) A card player shuffles a standard deck of cards by the following method: the top card of the deck is placed somewhere in the deck at random. The player does this 100 times to mix up the deck.

1.2 A regional insurance company insures homeowners against flood damage. Half of their policyholders are in Florida, 30% in Louisiana, and 20% in Texas. Company actuaries give the estimates in Table 1.1 for the probability that a policyholder will file a claim for flood damage over the next year.

(a) Find the probability that a random policyholder will file a claim for flood damage next year.

(b) A claim was filed. Find the probability that the policyholder is from Texas.

TABLE 1.1 Probability of Claim for Flood Damage

Florida	Louisiana	Texas
0.03	0.015	0.02

1.3 Let $B_1, \ldots, B_k$ be a partition of the sample space. For events A and C, prove the *law of total probability for conditional probability*

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

1.4 See Exercise 1.2. Among all policyholders who live within five miles of the Atlantic Ocean, 75% live in Florida, 20% live in Louisiana, and 5% live in Texas. For those who live close to the ocean the probabilities of filing a claim increase, as given in Table 1.2.

Assume that a policyholder lives within five miles of the Atlantic coast. Use the law of total probability for conditional probability in Exercise 1.3 to find the chance they will file a claim for flood damage next year.

1.5 Two fair, six-sided dice are rolled. Let X_1, X_2 be the outcomes of the first and second die, respectively.

TABLE 1.2 Probability of Claim for Those Within Five Miles of Atlantic Coast

Florida	Louisiana	Texas
0.10	0.06	0.06

(a) Find the conditional distribution of X_2 given that $X_1 + X_2 = 7$.

(b) Find the conditional distribution of X_2 given that $X_1 + X_2 = 8$.

1.6 Bob has n coins in his pocket. One is two-headed, the rest are fair. A coin is picked at random, flipped, and shows heads. Find the probability that the coin is two-headed.

1.7 A rat is trapped in a maze with three doors and some hidden cheese. If the rat takes door one, he will wander around the maze for 2 minutes and return to where he started. If he takes door two, he will wander around the maze for 3 minutes and return to where he started. If he takes door three, he will find the cheese after 1 minute. If the rat returns to where he started he immediately picks a door to pass through. The rat picks each door uniformly at random. How long, on average, will the rat wander before finding the cheese?

1.8 A bag contains 1 red, 3 green, and 5 yellow balls. A sample of four balls is picked. Let G be the number of green balls in the sample. Let Y be the number of yellow balls in the sample.

(a) Find the conditional probability mass function of G given $Y = 2$ assuming the sample is picked without replacement.

(b) Find the conditional probability mass function of G given $Y = 2$ assuming the sample is picked with replacement.

1.9 Assume that X is uniformly distributed on $\{1, 2, 3, 4\}$. If $X = x$, then Y is uniformly distributed on $\{1, \ldots, x\}$. Find

(a) $P(Y = 2|X = 2)$

(b) $P(Y = 2)$

(c) $P(X = 2|Y = 2)$

(d) $P(X = 2)$

(e) $P(X = 2, Y = 2)$.

1.10 A die is rolled until a 3 occurs. By conditioning on the outcome of the first roll, find the probability that an even number of rolls is needed.

1.11 Consider the gambler's ruin process where at each wager, the gambler wins with probability p and loses with probability $q = 1 - p$. The gambler stops when reaching \$$n$ or losing all their money. If the gambler starts with \$$k$, with $0 < k < n$, find the probability of eventual ruin. See Example 1.10.

1.12 In n rolls of a fair die, let X be the number of times 1 is rolled, and Y the number of times 2 is rolled. Find the conditional distribution of X given $Y = y$.

1.13 Random variables X and Y have joint density

$$f(x, y) = 3y, \quad \text{for } 0 < x < y < 1.$$

(a) Find the conditional density of Y given $X = x$.

(b) Find the conditional density of X given $Y = y$. Describe the conditional distribution.

1.14 Random variables X and Y have joint density function

$$f(x, y) = 4e^{-2x}, \quad \text{for } 0 < y < x < \infty.$$

(a) Find the conditional density of X given $Y = y$.

(b) Find the conditional density of Y given $X = x$. Describe the conditional distribution.

1.15 Let X and Y be uniformly distributed on the disk of radius 1 centered at the origin. Find the conditional distribution of Y given $X = x$.

1.16 A poker hand consists of five cards drawn from a standard 52-card deck. Find the expected number of aces in a poker hand given that the first card drawn is an ace.

1.17 Let X be a Poisson random variable with $\lambda = 3$. Find $E(X|X > 2)$.

1.18 From the definition of conditional expectation given an event, show that

$$E(I_B|A) = P(B|A).$$

1.19 See Example 1.21. Find the variance of the number of flips needed to get two heads in a row.

1.20 A fair coin is flipped repeatedly.

(a) Find the expected number of flips needed to get three heads in a row.

(b) Find the expected number of flips needed to get k heads in a row.

1.21 Let T be a nonnegative, continuous random variable. Show

$$E(T) = \int_0^\infty P(T > t)\, dt.$$

1.22 Find $E(Y|X)$ when (X, Y) is uniformly distributed on the following regions.

(a) The rectangle $[a, b] \times [c, d]$.

(b) The triangle with vertices $(0, 0), (1, 0), (1, 1)$.

(c) The disc of radius 1 centered at the origin.

1.23 Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with common mean μ. Let $S_n = X_1 + \cdots + X_n$, for $n \geq 1$.

(a) Find $E(S_m|S_n)$, for $m \leq n$.

(b) Find $E(S_m|S_n)$ for $m > n$.

1.24 Prove the law of total expectation $E(Y) = E(E(Y|X))$ for the continuous case.

1.25 Let X and Y be independent exponential random variables with respective parameters 1 and 2. Find $P(X/Y < 3)$ by conditioning.

1.26 The density of X is $f(x) = xe^{-x}$, for $x > 0$. Given $X = x$, Y is uniformly distributed on $(0, x)$. Find $P(Y < 2)$ by conditioning on X.

1.27 A restaurant receives N customers per day, where N is a random variable with mean 200 and standard deviation 40. The amount spent by each customer is normally distributed with mean \$15 and standard deviation \$3. The amounts that customers spend are independent of each other and independent of N. Find the mean and standard deviation of the total amount spent at the restaurant per day.

1.28 On any day, the number of accidents on the highway has a Poisson distribution with parameter Λ. The parameter Λ varies from day to day and is itself a random variable. Find the mean and variance of the number of accidents per day when Λ is uniformly distributed on $(0, 3)$.

1.29 If X and Y are independent, does $Var(Y|X) = Var(Y)$?

1.30 Assume that $Y = g(X)$ is a function of X. Find simple expressions for

(a) $E(Y|X)$.

(b) $Var(Y|X)$.

1.31 Consider a sequence of i.i.d. Bernoulli trials with success parameter p. Let X be the number of trials needed until the first success occurs. Then, X has a geometric distribution with parameter p. Find the variance of X by conditioning on the first trial.

1.32 `R`: Simulate flipping three fair coins and counting the number of heads X.

(a) Use your simulation to estimate $P(X = 1)$ and $E(X)$.

(b) Modify the above to allow for a biased coin where $P(\text{Heads}) = 3/4$.

1.33 `R`: Cards are drawn from a standard deck, with replacement, until an ace appears. Simulate the mean and variance of the number of cards required.

1.34 `R`: The time until a bus arrives has an exponential distribution with mean 30 minutes.

(a) Use the command `rexp()` to simulate the probability that the bus arrives in the first 20 minutes.

(b) Use the command `pexp()` to compare with the exact probability.

1.35 `R`: See the script file **gamblersruin.R**. A gambler starts with a $60 initial stake.

(a) The gambler wins, and loses, each round with probability $p = 0.50$. Simulate the probability the gambler wins $100 before he loses everything.

(b) The gambler wins each round with probability $p = 0.51$. Simulate the probability the gambler wins $100 before he loses everything.

1.36 `R`: See Example 1.2 and the script file **ReedFrost.R**. Observe the effect on the course of the disease by changing the initial values for the number of people susceptible and infected. How does increasing the number of infected people affect the duration of the disease?

1.37 `R`: Simulate the results of Exercise 1.28. Estimate the mean and variance of the number of accidents per day.