An Introduction to Stochastic Processes with Applications to Biology

Linda J. S. Allen
Department of Mathematics and Statistics
Texas Tech University

PEARSON EDUCATION, INC., Upper Saddle River, New Jersey 07458

Contents

Chapter 1

Review of Probability Theory and an Introduction to Stochastic Processes

1.1 Introduction

The underlying mathematical theory of stochastic modeling is *stochastic processes*. The theory of stochastic processes is based on probability theory. Therefore, we begin with a brief review of some basic principles from probability theory. An important reference for stochastic processes with applications to biology is the classic textbook by Bailey, *The Elements of Stochastic Processes with Applications to the Natural Sciences*, which has been referenced frequently since its initial publication in 1964. John Wiley & Sons republished this classic textbook in 1990. Other good references for stochastic processes include *Elements of the Theory of Markov Processes and Their Applications*, by Bharucha-Reid (1997, a Dover republication of a 1960 textbook); *Stochastic Models in Biology*, by Goel and Richter-Dyn (1974); *Stochastic Processes and Applications in Biology and Medicine*, Volumes I and II (Theory and Models), by Iosifescu and Tăutu (1973); *A First Course in Stochastic Processes*, by Karlin and Taylor (1975); *A Second Course in Stochastic Processes*, by Karlin and Taylor (1981); *Modelling Fluctuating Populations*, by Nisbet and Gurney (1982); *Modelling Biological Populations in Space and Time*, by Renshaw (1993); *Stochastic Processes*, by Ross (1983); *Classical and Spatial Stochastic Processes*, by Schinazi (1999); and *An Introduction to Stochastic Modeling*, by Taylor and Karlin (1998). The books by Karlin and Taylor have become classics

on stochastic processes; they provide an excellent introduction to general stochastic processes. In addition, Gard's book, *Introduction to Stochastic Differential Equations* (1988), discusses the theory of stochastic differential equations with applications to population dynamics. Additional references on stochastic processes will be given in subsequent chapters.

Numerous stochastic models from biology will be introduced and studied in this book. Often the stochastic models have a deterministic analogue, including models for population growth, competition, predation, and epidemics. For these models, the behavior of the deterministic model will be discussed and compared to that of the corresponding stochastic model. One of the most important differences between deterministic and stochastic models is that deterministic models predict an outcome with absolute certainty, whereas stochastic models provide only the probability of an outcome. For example, in a deterministic model, such as a difference equation or differential equation with initial conditions prescribed at $t = 0$, the solution follows a prescribed path or trajectory in the solution space. The numerical solution of a difference or differential equation gives the value of the solution (to a fixed number of decimal places) at a particular time t. In a stochastic model, the process may still be described by a system of difference equations (transition matrix) or differential equations (forward Kolmogorov equations or stochastic differential equations). However, unlike the deterministic model, the solution to these equations is more complicated, and a single solution trajectory does not describe the entire behavior of the model but represents only a single realization of the process. To understand the behavior of a stochastic model, it is important to know the entire probability distribution of the process over time. When this is not possible, the qualitative behavior of the process is studied by other methods, such as by obtaining the moments (mean, variance, etc.) of the distribution.

In models of populations, where the population size is sufficiently large, a deterministic formulation is often used. However, when population sizes are small, population extinction may occur, and then it is more realistic to model the variation in size by a stochastic formulation. Stochastic models may be used to study the probability of population extinction or the expected duration of time until population extinction. Random variations associated with demography and the environment can be taken into account in stochastic models.

A variety of mathematical techniques will be introduced, and the underlying theory will be developed for stochastic processes. The techniques and theory will be applied to stochastic models in biology. Methods for analyzing the dynamical behavior of the stochastic models will be studied as well as methods for constructing numerical simulations.

In the next section, a review of some basic concepts from probability theory are presented. In sections 1.3 and 1.4, generating functions are introduced and the central limit theorem is stated. Many examples are

given to illustrate and reinforce these important concepts from probability theory. In the last two sections of this chapter, stochastic processes are defined and an example of a simple birth process is given. This chapter is not intended to be a comprehensive review of probability theory but only to be a brief review of some concepts from probability theory that are important to the theory of stochastic processes.

1.2 Brief Review of Probability Theory

In this section, a brief review is given of many important definitions and concepts from probability theory. Important terms and definitions are put in italicized notation. Many books may be consulted for a more extensive review of probability. A classical reference for probability theory is the book by Feller (1968). Ross's 1989 book is another good source of reference for probability theory and probability models. Other references on the basic theory of probability and statistics include Hogg and Craig (1995), Hogg and Tanis (2001), and Hsu (1997).

Let S be a set, any collection of elements, which shall be referred to as the *outcome space* or *sample space*. For example, the sample space could be $S = \{H, T\}$, $S = \{0, 1, 2, \ldots\}$, $S = \{s | s \in [0, \infty)\}$, or any collection of objects or elements. Each element of S is called a *sample point* and each subset of S is referred to as an *event*. For example, suppose a coin is tossed and whether the coin lands with heads or tails showing is recorded. Then the sample space is $\{H, T\}$, a sample point is H or T, and an event may be $\{H\}$, $\{T\}$, $\{H, T\}$, or $\emptyset$. If the coin is fair, then the probability of a head appearing is $1/2$, the probability of a tail is $1/2$, and the probability of any other event is zero. In general, for any experiment, a probability measure is defined on the set of events in S as follows:

Definition 1.1. Let P be a real-valued set function defined on the collection of subsets of the sample space S. The set function $P : S \to [0, 1]$ is called a *probability measure* if it has the following properties:

(1) $0 \leq P(B)$, $B \subset S$.

(2) $P(S) = 1$.

(3) If $B_i \cap B_j = \emptyset$ for $i, j = 1, 2, \ldots, i \neq j$, (pairwise disjoint), where $B_i \subset S$, then $P(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P(B_i)$.

Next, the concepts of conditional probability and independence are defined.

Definition 1.2. Let B_1 and B_2 be two events defined on a sample space S. The *conditional probability of event B_1 given event B_2* (has occurred) is

on stochastic processes; they provide an excellent introduction to general stochastic processes. In addition, Gard's book, *Introduction to Stochastic Differential Equations* (1988), discusses the theory of stochastic differential equations with applications to population dynamics. Additional references on stochastic processes will be given in subsequent chapters.

Numerous stochastic models from biology will be introduced and studied in this book. Often the stochastic models have a deterministic analogue, including models for population growth, competition, predation, and epidemics. For these models, the behavior of the deterministic model will be discussed and compared to that of the corresponding stochastic model. One of the most important differences between deterministic and stochastic models is that deterministic models predict an outcome with absolute certainty, whereas stochastic models provide only the probability of an outcome. For example, in a deterministic model, such as a difference equation or differential equation with initial conditions prescribed at $t = 0$, the solution follows a prescribed path or trajectory in the solution space. The numerical solution of a difference or differential equation gives the value of the solution (to a fixed number of decimal places) at a particular time t. In a stochastic model, the process may still be described by a system of difference equations (transition matrix) or differential equations (forward Kolmogorov equations or stochastic differential equations). However, unlike the deterministic model, the solution to these equations is more complicated, and a single solution trajectory does not describe the entire behavior of the model but represents only a single realization of the process. To understand the behavior of a stochastic model, it is important to know the entire probability distribution of the process over time. When this is not possible, the qualitative behavior of the process is studied by other methods, such as by obtaining the moments (mean, variance, etc.) of the distribution.

In models of populations, where the population size is sufficiently large, a deterministic formulation is often used. However, when population sizes are small, population extinction may occur, and then it is more realistic to model the variation in size by a stochastic formulation. Stochastic models may be used to study the probability of population extinction or the expected duration of time until population extinction. Random variations associated with demography and the environment can be taken into account in stochastic models.

A variety of mathematical techniques will be introduced, and the underlying theory will be developed for stochastic processes. The techniques and theory will be applied to stochastic models in biology. Methods for analyzing the dynamical behavior of the stochastic models will be studied as well as methods for constructing numerical simulations.

In the next section, a review of some basic concepts from probability theory are presented. In sections 1.3 and 1.4, generating functions are introduced and the central limit theorem is stated. Many examples are

given to illustrate and reinforce these important concepts from probability theory. In the last two sections of this chapter, stochastic processes are defined and an example of a simple birth process is given. This chapter is not intended to be a comprehensive review of probability theory but only to be a brief review of some concepts from probability theory that are important to the theory of stochastic processes.

1.2 Brief Review of Probability Theory

In this section, a brief review is given of many important definitions and concepts from probability theory. Important terms and definitions are put in italicized notation. Many books may be consulted for a more extensive review of probability. A classical reference for probability theory is the book by Feller (1968). Ross's 1989 book is another good source of reference for probability theory and probability models. Other references on the basic theory of probability and statistics include Hogg and Craig (1995), Hogg and Tanis (2001), and Hsu (1997).

Let S be a set, any collection of elements, which shall be referred to as the *outcome space* or *sample space*. For example, the sample space could be $S = \{H, T\}$, $S = \{0, 1, 2, \ldots\}$, $S = \{s | s \in [0, \infty)\}$, or any collection of objects or elements. Each element of S is called a *sample point* and each subset of S is referred to as an *event*. For example, suppose a coin is tossed and whether the coin lands with heads or tails showing is recorded. Then the sample space is $\{H, T\}$, a sample point is H or T, and an event may be $\{H\}$, $\{T\}$, $\{H, T\}$, or $\emptyset$. If the coin is fair, then the probability of a head appearing is $1/2$, the probability of a tail is $1/2$, and the probability of any other event is zero. In general, for any experiment, a probability measure is defined on the set of events in S as follows:

Definition 1.1. Let P be a real-valued set function defined on the collection of subsets of the sample space S. The set function $P : S \to [0, 1]$ is called a *probability measure* if it has the following properties:

(1) $0 \leq P(B)$, $B \subset S$.

(2) $P(S) = 1$.

(3) If $B_i \cap B_j = \emptyset$ for $i, j = 1, 2, \ldots, i \neq j$, (pairwise disjoint), where $B_i \subset S$, then $P\left(\cup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i)$.

Next, the concepts of conditional probability and independence are defined.

Definition 1.2. Let B_1 and B_2 be two events defined on a sample space S. The *conditional probability of event B_1 given event B_2* (has occurred) is

denoted as $P(B_1|B_2)$ and defined as

$$P(B_1|B_2) = \frac{P(B_1 \cap B_2)}{P(B_2)},$$

provided that $P(B_2) > 0$. Similarly, the *conditional probability of event* B_2 *given event* B_1 is

$$P(B_2|B_1) = \frac{P(B_1 \cap B_2)}{P(B_1)},$$

provided that $P(B_1) > 0$.

The events B_1 and B_2 are independent if the occurrence of either one of the events does not affect the probability of occurrence of the other. More formally,

Definition 1.3. Let B_1 and B_2 be two events defined on a sample space S. Events B_1 and B_2 are said to be *independent* if and only if

$$P(B_1 \cap B_2) = P(B_1)P(B_2).$$

If the events B_1 and B_2 are not independent, they are said to be *dependent.*

Therefore, B_1 and B_2 are independent if and only if $P(B_2|B_1) = P(B_2)$ or $P(B_1|B_2) = P(B_1)$. In other words, the events B_1 and B_2 are independent if the probability of B_2 does not depend on whether B_1 has occurred or the probability of B_1 does not depend on whether B_2 has occurred.

The concept of a random variable is central to probability theory.

Definition 1.4. A *random variable* X is a real-valued function defined on the sample space S, $X : S \to \mathbf{R} = (-\infty, \infty)$, where there is an associated probability measure P defined on S. Let A be the range of X, $A = \{x|X(s) = x, s \in S\}$. The range A is known as the *space* of X or *state space* of X.

If the range of X is finite or countably infinite, then X is said to be a *discrete random variable*, whereas if the range is an interval (finite or infinite in length), then X is said to be a *continuous random variable.* However, the random variable could be of *mixed type*, having properties of both a discrete and continuous random variable. The distinction between discrete and continuous random variables will be seen in Definitions 1.6 and 1.7. We shall only be concerned with discrete and continuous random variables.

Example 1.1 Suppose two fair coins are tossed sequentially and the outcomes are HH, HT, TH, and TT (i.e., $S = \{HH, HT, TH, TT\}$). Let X be the discrete random variable associated with this experiment having state space $A = \{1, 2, 3, 4\}$, where $X(HH) = 1$, $X(HT) = 2$, and so on. Assume each of the outcomes has an equal probability of 1/4.

$P(\{HH\}) = 1/4$, $P(\{HT\}) = 1/4$, etc. Let B_1 be the event that the first coin is a head, $B_1 = \{HH, HT\}$, and B_2 be the event that the second coin is a head, $B_2 = \{HH, TH\}$. Then $P(B_1) = 1/2$, $P(B_2) = 1/2$, and $P(B_1 \cap B_2) = P(\{HH\}) = 1/4$. Then

$$P(B_2|B_1) = \frac{P(B_1 \cap B_2)}{P(B_1)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

Since $P(B_2) = 1/2$, B_1 and B_2 are independent events. ■

Events associated with a random variable X can be expressed as subsets of $\mathbf{R}$. For example, we shall use the shorthand notation $\{X = x\}$ to mean the event $\{s|X(s) = x, s \in S\}$ and the notation $\{X \leq x\} = (-\infty, x]$ for the event $\{s|X(s) \leq x, s \in S\}$. In Example 1.1, $B_1 = \{HH, HT\}$ expressed in terms of the state space of X is represented by the set $\{1, 2\}$ and $B_2 = \{HH, TH\}$ by the set $\{1, 3\}$. With this convention, the probability measure P associated with the random variable X can be defined on $\mathbf{R}$. Sometimes this measure is denoted as $P_X : \mathbf{R} \to [0, 1]$ and referred to as the *induced probability measure* (Hogg and Craig, 1995). The subscript X is often omitted, but it is clear from the context that the induced probability measure is implied. This notation is used in defining the cumulative distribution function of a random variable X that is defined on the set of real numbers.

Definition 1.5. The *cumulative distribution function* (*c.d.f.*) of the random variable X is the function F defined on $\mathbf{R}$ with values in [0,1], $F : \mathbf{R} \to [0, 1]$, satisfying

$$F(x) = P_X((-\infty, x]).$$

It can be shown that F is nondecreasing, right continuous and satisfies

$$\lim_{x \to -\infty} F(x) = F(-\infty) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1.$$

The cumulative distribution describes how the probabilities accumulate (see Examples 1.2 and 1.3).

Important functions associated with discrete and continuous random variables are the probability mass function and probability density function.

Definition 1.6. Suppose X is a discrete random variable. Then the function $f(x) = P_X(X = x)$ that is defined for each x in the range of X is called the *probability mass function* (*p.m.f.*) of X.

It follows from Definition 1.1 that f has the following two properties:

$$\sum_{x \in A} f(x) = 1 \quad \text{and} \quad P_X(X \in B) = \sum_{x \in B} f(x), \tag{1.1}$$

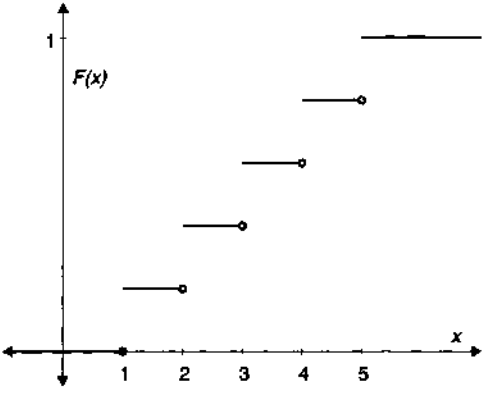

Figure 1.1. Discrete uniform c.d.f.

for any $B \subset A$, where A is the range of X. In addition, the c.d.f. F of a discrete random variable satisfies

$$F(x) = \sum_{a_i \leq x} f(a_i),$$

where $A = \{a_1, a_2, \dots\}$ is the space of X ($F(x) = 0$ if $x < \inf_i\{a_i\}$).

Example 1.2 Let the space of the discrete random variable X be $A = \{1, 2, 3, 4, 5\}$ and $f(x) = 1/5$ for $x \in A$. The c.d.f. $F(x)$ of X satisfies

$$F(x) = \begin{cases} 0, & x < 1, \\ 1/5, & 1 \leq x < 2, \\ 2/5, & 2 \leq x < 3, \\ \vdots & \vdots \\ 1, & 5 \leq x. \end{cases}$$

The graph of F is given in Figure 1.1. This distribution is known as a *discrete uniform distribution.* ■

Definition 1.7. Suppose X is a continuous random variable with c.d.f. F and there exists a nonnegative, integrable function f, $f : \mathbf{R} \to [0, \infty)$, such that

$$F(x) = \int_{-\infty}^{x} f(y)\, dy.$$

Then the function f is called the *probability density function* (*p.d.f.*) of X.

The p.d.f. of a continuous random variable can be used to compute the probability associated with an outcome or event. Suppose A is the space of X and $B \subset A$ is an event. Then

$$P_X(X \in A) = \int_A f(x)\, dx = \int_{-\infty}^{\infty} f(x)\, dx = 1$$

and

$$P_X(X \in B) = \int_B f(x)\,dx. \tag{1.2}$$

In particular,

$$P_X(a \le X \le b) = \int_a^b f(x)\,dx = F(b) - F(a).$$

For a continuous random variable X,

$$P_X(a < X < b) = P_X(a \le X < b) = P_X(a < X \le b) = P_X(a \le X \le b),$$

which follows from (1.2). In addition, if the cumulative distribution function is differentiable, then

$$\frac{dF(x)}{dx} = f(x).$$

Example 1.3 Let the space of a continuous random variable X be $A = [0, 1]$ and the probability density function be $f(x) = 1$ for $x \in A$. The c.d.f. $F(x)$ satisfies

$$F(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x < 1, \\ 1, & 1 \le x. \end{cases}$$

The graph of F is given in Figure 1.2. This distribution is known as a *continuous uniform distribution.* ■

Sometimes we shall use the term *probability density function* to include both the p.d.f of a continuous random variable and the p.m.f. of a discrete random variable. In addition, sometimes the notation Prob$\{\cdot\}$ will be used

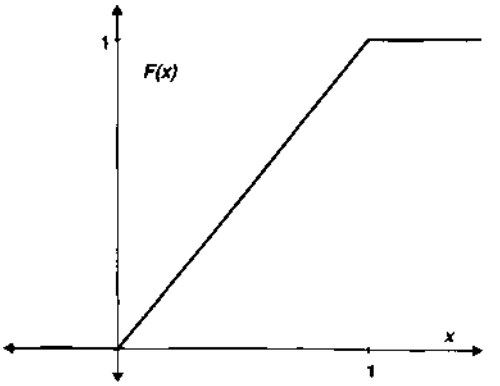

Figure 1.2. Continuous uniform c.d.f.

in place of $P(\cdot)$ or $P_X(\cdot)$ to emphasize the fact that a probability is being computed. For example, for a discrete random variable X,

$$P_X(X = x) = \text{Prob}\{X = x\} = f(x)$$

and for either a continuous or discrete random variable X,

$$P_X(X \leq x) = \text{Prob}\{X \leq x\} = F(x).$$

Some well-known discrete distributions include the uniform, binomial, negative binomial, and Poisson. The probability mass functions for each of these distributions are given below. The binomial distribution will be seen in many of the applications for discrete and continuous time Markov chains. The distribution for the number of deaths in a simple death process will be shown to have a binomial distribution. In addition, the distribution for the number of births in a simple birth process will be shown to have a negative binomial distribution. The Poisson distribution is especially important in the study of continuous time Markov chain models. The Poisson process is discussed in detail in Chapter 5.

Discrete Uniform:

$$f(x) = \begin{cases} \dfrac{1}{n}, & x = 1, 2, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

Binomial:

$$f(x) = \begin{cases} \dbinom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n, \\ 0, & \text{otherwise,} \end{cases}$$

where n is a positive integer and $0 < p < 1$. The notation $\dbinom{n}{x}$ for the binomial coefficient is defined as

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

For example,

$$\binom{5}{1} = 5 \quad \text{and} \quad \binom{5}{3} = 10.$$

It is assumed that $0! = 1$. The binomial probability distribution is denoted as $b(n, p)$. The value of $f(x)$ can be thought of as the probability of x successes in n trials, where p is the probability of success.

Negative Binomial:

$$f(x) = \begin{cases} \binom{x+n-1}{n-1} p^n (1-p)^x, & x = 0, 1, 2, \ldots, \\ 0, & \text{otherwise}, \end{cases} \tag{1.3}$$

where n is a positive integer and $0 < p < 1$. The value of $f(x)$ can be thought of as the probability of n successes in $n + x$ trials, where p is the probability of success.

Poisson:

$$f(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \ldots, \\ 0, & \text{otherwise}, \end{cases}$$

where λ is a positive constant.

Some well-known continuous distributions include the uniform, normal, gamma, and exponential. The uniform distribution is the basis for a random number generator, which is used extensively in numerical simulations of stochastic models. The normal distribution is the underlying distribution for Brownian motion, a diffusion process studied in Chapter 8. The gamma and exponential distributions are associated with waiting time distributions, the time until one or more than one event occurs. The exponential distribution will be seen extensively in the continuous time Markov chain models discussed in Chapters 5, 6, and 7. The probability density functions for each of these distributions are defined below.

Uniform:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise}, \end{cases}$$

where $a < b$ are constants. The uniform distribution is denoted as $U(a, b)$.

Gamma:

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, & x \ge 0, \\ 0, & x < 0, \end{cases}$$

where α and β are positive constants and

$$\Gamma(\alpha) = \int_0^\infty \frac{1}{\beta^\alpha} x^{\alpha-1} e^{-x/\beta}\, dx.$$

For a positive integer n, $\Gamma(n) = (n-1)!$.

Exponential:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0, \end{cases}$$

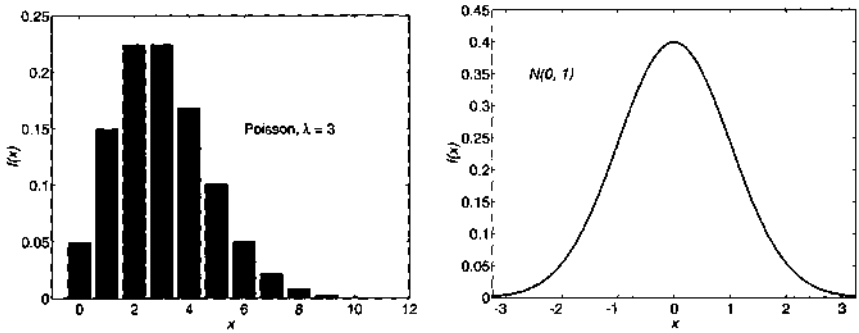

Figure 1.3. Graphs of the Poisson mass function with parameter $\lambda = 3$ and the standard normal density.

where λ is a positive constant.

Normal:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty,$$

where μ and σ are constants. We denote the normal distribution as $N(\mu, \sigma^2)$ and the standard normal as $N(0,1)$, where $\mu = 0$ and $\sigma^2 = 1$.

Note that the exponential distribution is a special case of the gamma distribution, where $\alpha = 1$ and $\beta = 1/\lambda$. Graphs of a Poisson mass function and the standard normal density are given in Figure 1.3. The MATLAB program which generated the Poisson mass function is given in the Appendix for Chapter 1.

An important concept that helps characterize the p.d.f. of a random variable is the expectation.

Definition 1.8. Suppose X is a continuous random variable with p.d.f. f. Then the *expectation of* X, denoted as $E(X)$, is defined as

$$E(X) = \int_{\mathbf{R}} x f(x)\, dx.$$

Suppose X is a discrete random variable with probability function f defined on the space $A = \{a_i\}_{i=1}^{\infty}$. Then the *expectation of* X is defined as

$$E(X) = \sum_{i=1}^{\infty} a_i f(a_i).$$

The expectation of X is a weighted average. The p.d.f. f is weighted by the values of the random variable X.

The definition of expectation of a random variable can be extended to expectation of a function of a random variable. Suppose X is a continuous random variable. Then the *expectation of* $u(X)$ is

$$E(u(X)) = \int_{\mathbf{R}} u(x)f(x)\,dx.$$

If X is a discrete random variable, then the *expectation of* $u(X)$ is

$$E(u(X)) = \sum_{i=1}^{\infty} u(a_i)f(a_i).$$

It can be seen that the expectation is a linear operator defined on the set of functions $u(X)$: If a_1 and a_2 are constants, then it follows from the definition that

$$E(a_1u_1(X) + a_2u_2(X)) = a_1E(u_1(X)) + a_2E(u_2(X)).$$

In addition, for $b \equiv$ constant, $E(b) = b$. The mean, variance, and moments of X, which measure the center and the spread of the p.d.f., are defined in terms of the expectation.

Definition 1.9. The *mean* of the random variable X, denoted as μ or μ_X, is the expectation of X, $\mu_X = E(X)$. The *variance* of X, denoted as σ^2, σ_X^2, or $\mathrm{Var}(X)$, is $\mathrm{Var}(X) = E([X - \mu_X]^2)$. The *standard deviation* of X is $\sigma = \sqrt{\mathrm{Var}(X)}$. The *nth moment of* X *about the point* a is $E([X - a]^n)$.

The subscript X on the mean and variance is used to avoid confusion, especially if more than one random variable is being discussed. The first moment about the origin is the mean, and the second moment about the mean is the variance. An important identity for the variance can be derived from the linearity property of the expectation,

$$\sigma_X^2 = E([X - \mu_X]^2) = E(X^2) - 2\mu_X E(X) + \mu_X^2 = E(X^2) - \mu_X^2.$$

Example 1.4 Suppose X is a random variable with a discrete uniform distribution and Y is a random variable with a continuous uniform distribution [i.e., Y is distributed as $U(0,1)$]. The mean and variance for each of these two random variables are computed. The mean of X is

$$\mu_X = E(X) = \sum_{x=1}^{n} \left(x\frac{1}{n}\right) = \frac{1}{n}\sum_{x=1}^{n} x = \frac{n+1}{2},$$

because $\sum_{x=1}^{n} x = n(n+1)/2$. Then

$$E(X^2) = \sum_{x=1}^{n} \left(x^2\frac{1}{n}\right) = \frac{1}{n}\sum_{x=1}^{n} x^2 = \frac{(n+1)(2n+1)}{6},$$

because $\sum_{x=1}^{n} x^2 = n(n+1)(2n+1)/6$. The variance of X is

$$\sigma_X^2 = E(X^2) - \mu_X^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}.$$

The mean of Y is

$$\mu_Y = E(Y) = \int_0^1 y\,dy = \frac{1}{2}$$

and $E(Y^2) = \int_0^1 y^2\,dy = 1/3$ so that the variance of Y is

$$\sigma_Y^2 = E(Y^2) - \mu_Y^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \qquad \blacksquare$$

Example 1.5 Consider the normal distribution $N(0,\sigma^2)$. The first, second, third, and fourth moments about the origin of this normal distribution are calculated. The normal p.d.f. corresponding to $N(0,\sigma^2)$ satisfies

$$f(x) = \frac{\exp\left(-x^2/2\sigma^2\right)}{\sigma\sqrt{2\pi}}.$$

It is an even function over the interval $(-\infty,\infty)$. Therefore, $xf(x)$ and $x^3f(x)$ are odd functions on $(-\infty,\infty)$. It follows that the first and third moments are zero,

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = 0 = \int_{-\infty}^{\infty} x^3 f(x)\,dx = E(X^3).$$

In addition,

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_{-\infty}^{\infty} \frac{x^2 \exp\left(-x^2/2\sigma^2\right)}{\sigma\sqrt{2\pi}}\,dx \\ &= -\frac{\sigma x \exp(-x^2/2\sigma^2)}{\sqrt{2\pi}}\Bigg|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{\sigma \exp(-x^2/2\sigma^2)}{\sqrt{2\pi}}\,dx \\ &= \sigma^2 \int_{-\infty}^{\infty} f(x)\,dx = \sigma^2, \end{aligned}$$

where integration by parts is used in the first integral, $u = x$, and $dv = xf(x)\,dx$. Therefore, the normal distribution $N(0,\sigma^2)$ has mean and variance, $\mu_X = 0$ and $\sigma_X^2 = \sigma^2$. In a similar manner, it can be shown that the normal distribution $N(\mu,\sigma^2)$ has mean and variance, $\mu_X = \mu$ and $\sigma_X^2 = \sigma^2$.

The fourth moment is computed using the same technique,

$$\begin{aligned} E(X^4) &= \int_{-\infty}^{\infty} x^4 f(x)\,dx = \int_{-\infty}^{\infty} \frac{x^4 \exp\left(-x^2/2\sigma^2\right)}{\sigma\sqrt{2\pi}}\,dx \\ &= -\left.\frac{\sigma x^3 \exp(-x^2/2\sigma^2)}{\sqrt{2\pi}}\right|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{3\sigma x^2 \exp(-x^2/2\sigma^2)}{\sqrt{2\pi}}\,dx \\ &= 3\sigma^2 \int_{-\infty}^{\infty} x^2 f(x)\,dx = 3\sigma^2 E(X^2) = 3(\sigma^2)^2. \end{aligned}$$

■

A simple relationship exists between any normal random variable and the standard normal random variable. If X is distributed as $N(\mu, \sigma^2)$, then it turns out that

$$Z = \frac{X - \mu}{\sigma}$$

is distributed as $N(0, 1)$. This relationship can be verified by showing the c.d.f. of Z corresponds to a standard normal:

$$\begin{aligned} \text{Prob}\{Z \le z\} &= \text{Prob}\left\{\frac{X-\mu}{\sigma} \le z\right\} = \text{Prob}\{X \le z\sigma + \mu\} \\ &= \int_{-\infty}^{z\sigma+\mu} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx. \end{aligned}$$

Make a change of variable in the integral, $y = (x - \mu)/\sigma$ and $dy = dx/\sigma$, so that

$$\text{Prob}\{Z \le z\} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,dy.$$

The latter integral is the c.d.f. of the standard normal distribution.

Example 1.6 The standard normal distribution $N(0, 1)$ has the property that $\text{Prob}\{Z \le 0\} = 0.5$ and thus, for an arbitrary normal distribution, where X is distributed as $N(\mu, \sigma^2)$, we have that

$$\text{Prob}\left\{\frac{X-\mu}{\sigma} \le 0\right\} = \text{Prob}\{X \le \mu\} = 0.5.$$

Values of the c.d.f $F(z)$, $z \in [0, 3]$, for the standard normal distribution can be found in tabular form in many textbooks. In addition, these values can be numerically approximated directly from the integral; for example,

$$\text{Prob}\{-2.1 \le Z < 1\} = \int_{-2.1}^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz \approx 0.8235.$$

If X is distributed as $N(1, 4)$, then

$$\text{Prob}\{X \le 3\} = \text{Prob}\left\{\frac{X-1}{2} \le \frac{3-1}{2}\right\} = \text{Prob}\{Z \le 1\} \approx 0.8413.$$

■

An interesting property of the normal distribution is that approximately two-thirds of the values lie within one standard deviation of the mean, approximately 95% are within two standard deviations, and more than 99% of the values are within three standard deviations. In particular,

$$\text{Prob}\{-k\sigma \leq X - \mu \leq k\sigma\} = \text{Prob}\{|Z| \leq k\} \approx \begin{cases} 0.6827, & k = 1, \\ 0.9545, & k = 2, \\ 0.9973, & k = 3. \end{cases}$$

(Refer to Figure 1.3.)

When several random variables, $X_1, X_2, \ldots,$ and X_n, are associated with the same sample space, a multivariate probability density function or probability mass function $f(x_1, x_2, \ldots, x_n)$ can be defined. Definitions associated with two random variables, X_1 and X_2, are given, that is, for a random vector (X_1, X_2), where $(X_1, X_2) : S \to \mathbf{R}^2$. These definitions can be easily extended to more than two random variables. For each element s in the sample space S, there is associated a unique ordered pair $(X_1(s), X_2(s))$. The set of all ordered pairs

$$A = \{(X_1(s), X_2(s)) | s \in S\} \subset \mathbf{R}^2, \tag{1.4}$$

is known as the *state space* or *space* of the random vector (X_1, X_2).

Definition 1.10. Suppose X_1 and X_2 are two continuous random variables defined on the common sample space S, having probability measure $P : S \to [0, 1]$. If there exists a function $f : \mathbf{R}^2 \to [0, \infty)$ such that

$$\iint_{\mathbf{R}^2} f(x_1, x_2)\, dx_1\, dx_2 = \iint_{A} f(x_1, x_2)\, dx_1\, dx_2 = 1$$

and for $B \subset A$,

$$P_{(X_1, X_2)}(B) = \text{Prob}\,\{(X_1, X_2) \in B\} = \iint_{B} f(x_1, x_2)\, dx_1\, dx_2,$$

then f is called the *joint probability density function* (*joint p.d.f.*) or *joint density function* of the random variables X_1 and X_2. The *marginal p.d.f* of X_1 is defined as

$$f_1(x_1) = \int_{\mathbf{R}} f(x_1, x_2)\, dx_2.$$

The marginal p.d.f. of X_2, $f_2(x_2)$, can be defined in a similar manner. The set A in Definition 1.10 is defined by equation (1.4) and the function $P_{(X_1, X_2)}$ refers to the induced probability measure on $\mathbf{R}^2$.

Definition 1.11. Suppose X_1 and X_2 are two discrete random variables defined on a common sample space S, having probability measure $P : S \to [0, 1]$. If there exists a function $f : A \to [0, 1]$ such that

$$\sum_A f(x_1, x_2) = 1$$

and for $B \subset A$,

$$P_{(X_1,X_2)}(B) = \text{Prob}\{(X_1, X_2) \in B\} = \sum_B f(x_1, x_2),$$

then f is called the *joint probability mass function (p.m.f.)* or *joint mass function* of the random variables X_1 and X_2. The *marginal probability mass function (p.m.f.)* of X_1 is defined as

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2).$$

The marginal p.m.f. of X_2, $f_2(x_2)$, can be defined in a similar manner.

If the set A can be written as the *product space*, $A = A_1 \times A_2$, then the integrals in Definition 1.10 can be expressed as double integrals and the sums in Definition 1.11 as double sums:

$$\iint_A = \int_{A_1}\int_{A_2} \quad \text{and} \quad \sum_A = \sum_{A_1}\sum_{A_2}.$$

In addition, the sum in Definition 1.11 can be expressed as

$$\sum_{x_2} = \sum_{x_2 \in A_2}.$$

The marginal p.d.f.'s or p.m.f.'s $f_1(x_1)$ and $f_2(x_2)$ are indeed p.d.f.'s or p.m.f.'s in their own right, satisfying either (1.2) or (1.1), respectively. The definitions of conditional probability and independence of events are extended to random variables, important concepts in stochastic processes.

Definition 1.12. Let the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2)$ and marginal p.d.f.'s $f_1(x_1)$ and $f_2(x_2)$, respectively. Let $X_1|x_2$ denote the random variable X_1, given that the random variable $X_2 = x_2$, and $X_2|x_1$ denote the random variable X_2, given that the random variable $X_1 = x_1$. The *conditional p.d.f. of the random variable* $X_1|x_2$ is defined as

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \quad f_2(x_2) > 0.$$

The *conditional p.d.f. of* $X_2|x_1$ is

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}, \quad f_1(x_1) > 0.$$

Definition 1.13. Let the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2)$ and marginal p.d.f.'s $f_1(x_1)$ and $f_2(x_2)$, respectively. Assume $A = A_1 \times A_2$. The random variables X_1 and X_2 are said to be *independent* if and only if

$$f(x_1, x_2) = f_1(x_1) f_2(x_2),$$

for all $x_1 \in A_1$ and $x_2 \in A_2$.

Definitions 1.12 and 1.13 apply to discrete and continuous random variables. It follows from these definitions that if X_1 and X_2 are independent, then

$$f(x_1|x_2) = f_1(x_1) \quad \text{and} \quad f(x_2|x_1) = f_2(x_2).$$

An important consequence of the independence of the two random variables X_1 and X_2 is that the expectation of their product is the product of their expectations; that is,

$$E(X_1 X_2) = E(X_1) E(X_2).$$

Example 1.7 Suppose the joint p.d.f of the random variables X_1 and X_2 is $f(x_1, x_2) = 8x_1 x_2$ for $0 < x_1 < x_2 < 1$ and 0 otherwise. The marginal p.d.f.'s of X_1 and X_2 are

$$f_1(x_1) = \int_{x_1}^{1} 8x_1 x_2 \, dx_2 = 4x_1(1 - x_1^2), \quad 0 < x_1 < 1$$

and

$$f_2(x_2) = \int_{0}^{x_2} 8x_1 x_2 \, dx_1 = 4x_2^3, \quad 0 < x_2 < 1.$$

The random variables X_1 and X_2 are dependent. The reason there is not independence, $f_1(x_1) f_2(x_2) \neq f(x_1, x_2)$, is that $A = \{(x_1, x_2) | 0 < x_1 < x_2 < 1\}$ is not a product space, $A \neq A_1 \times A_2$. A necessary condition for independence of the random variables is that the state space A be a product space. ■

Example 1.8 Suppose the joint p.d.f. of the random variables X_1 and X_2 is $f(x_1, x_2) = 4x_1 x_2$ for $0 < x_1 < 1$ and $0 < x_2 < 1$ and 0 otherwise. It is easy to see that the random variables X_1 and X_2 are independent. The marginal p.d.f.'s of X_1 and X_2, respectively, are $f_1(x_1) = 2x_1$ for $0 < x_1 < 1$ and $f_2(x_2) = 2x_2$ for $0 < x_2 < 1$. Hence, to compute $E(X_1 X_2)$ we need only compute

$$E(X_1) = \int_0^1 x f_1(x) \, dx = \int_0^1 2x^2 \, dx = \frac{2}{3} = E(X_2).$$

Then $E(X_1 X_2) = E(X_1) E(X_2) = 4/9$. ■

Linear combinations of independent random variables often occur in applications. Suppose $\{X_i\}_{i=1}^n$ is a set of n independent random variables. Suppose the mean of X_i is μ_i and the variance is σ_i^2, $i = 1, 2, \ldots, n$. Then it is easy to show that the random variable

$$Y = \sum_{i=1}^{n} a_i X_i$$

has mean and variance satisfying

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i \quad \text{and} \quad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2. \tag{1.5}$$

The expectation is a linear operator and, therefore, the identity for the mean μ_Y in (1.5) holds even if the set $\{X_i\}_{i=1}^n$ is not independent. However, the identity for the variance σ_Y^2 in (1.5) requires that the set of random variables be independent. We verify the variance identity in the following example for $n = 2$.

Example 1.9 Suppose X_1 and X_2 are two random variables with respective means μ_1 and μ_2 and respective variances σ_1^2 and σ_2^2. Let $Y = a_1X_1 + a_2X_2$. Then the mean of Y is

$$\mu_Y = E(Y) = E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2) = a_1\mu_1 + a_2\mu_2.$$

Now, suppose the random variables X_1 and X_2 are independent. Then the variance of Y is

$$\begin{aligned}\sigma_Y^2 &= E([Y - \mu_Y]^2) = E([a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2)]^2) \\ &= a_1^2E([X_1 - \mu_1]^2) + 2a_1a_2E([(X_1 - \mu_1)(X_2 - \mu_2)]) + a_2^2E([X_2 - \mu_2]^2) \\ &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2\end{aligned}$$

Because of the independence of X_1 and X_2, $E([(X_1 - \mu_1)(X_2 - \mu_2)]) = E(X_1X_2) - \mu_1\mu_2 = 0$. ■

We shall use the standard notation $\bar{X}$ to denote the *average* of a sum of n independent random variables,

$$\bar{X} = \sum_{i=1}^{n} X_i/n.$$

Suppose $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ so that the independent random variables have the same mean and variance. Then, according to (1.5),

$$\mu_{\bar{X}} = \sum_{i=1}^{n} \frac{1}{n}\mu = \mu \quad \text{and} \quad \sigma_{\bar{X}}^2 = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}.$$

In addition to these basic concepts from probability theory, the concept of generating functions will be helpful when analyzing stochastic models. A short review of some important facts concerning generating functions is given in the next section.

1.3 Generating Functions

Generating functions are first defined in terms of a discrete random variable. Then the definitions are extended to a continuous random variable.

Assume X is a discrete random variable and, for convenience, assume the state space is $\{0, 1, 2, \ldots\}$. Let f denote the probability mass function of X and suppose the probabilities are given by

$$f(j) = \text{Prob}\{X = j\} = p_j, \ j = 0, 1, 2, \ldots, \quad \text{where} \quad \sum_{j=0}^{\infty} p_j = 1.$$

The mean and variance of X satisfy

$$\mu_X = E(X) = \sum_{j=0}^{\infty} j p_j$$

and

$$\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2 = \sum_{j=0}^{\infty} j^2 p_j - \mu_X^2.$$

Definition 1.14. The *probability generating function (p.g.f.)* of the discrete random variable X is a function defined on a subset of the reals, denoted as $\mathcal{P}_X$ and defined by

$$\mathcal{P}_X(t) = E(t^X) = \sum_{j=0}^{\infty} p_j t^j,$$

for some $t \in \mathbf{R}$.

The script notation $\mathcal{P}$ is used for the p.g.f. to distinguish this function from the probability measure P and the subscript X is used to denote its association with the random variable X. This subscript is often omitted when it is clear from the context what is the associated random variable. Because $\sum_{j=0}^{\infty} p_j = 1$, the above sum converges absolutely for $|t| \leq 1$. Thus, $\mathcal{P}(t)$ is well defined for $|t| \leq 1$. As the name implies, the p.g.f. generates the probabilities associated with the distribution

$$\mathcal{P}_X(0) = p_0, \quad \mathcal{P}_X'(0) = p_1, \quad \mathcal{P}_X''(0) = 2! p_2.$$

In general, the kth derivative of the p.g.f. of X satisfies

$$\mathcal{P}_X^{(k)}(0) = k! p_k.$$

(Since the series for the p.g.f. converges absolutely on $|t| < 1$, it is infinitely differentiable inside its interval of convergence.)

The p.g.f. can be used to calculate the mean and variance of a random variable X. Note that $\mathcal{P}'_X(t) = \sum_{j=1}^{\infty} jp_j t^{j-1}$ for $-1 < t < 1$. Let t approach one from the left, $t \to 1^-$ to obtain

$$\mathcal{P}'_X(1) = \sum_{j=1}^{\infty} jp_j = E(X) = \mu_X.$$

The second derivative of $\mathcal{P}_X$ satisfies

$$\mathcal{P}''_X(t) = \sum_{j=1}^{\infty} j(j-1)p_j t^{j-2},$$

so that as $t \to 1^-$,

$$\mathcal{P}''_X(1) = \sum_{j=1}^{\infty} j(j-1)p_j = E(X^2 - X).$$

Suppose the mean is finite. Then the variance of X satisfies

$$\begin{aligned}\sigma_X^2 &= \mathrm{Var}(X) = E(X^2) - E(X) + E(X) - [E(X)]^2 \\ &= \mathcal{P}''_X(1) + \mathcal{P}'_X(1) - [\mathcal{P}'_X(1)]^2.\end{aligned}$$

There are several other generating functions useful to the study of stochastic processes, the moment generating function, the characteristic function, and the cumulant generating function. These are defined in terms of a discrete random variable, then extended to a continuous random variable.

Definition 1.15. The *moment generating function (m.g.f.)* of the discrete random variable X with state space $\{0, 1, 2, \ldots\}$ and probability function $f(j) = p_j$, $j = 0, 1, 2, \ldots$, is denoted as $M_X(i)$ and defined as

$$M_X(t) = E(e^{tX}) = \sum_{j=0}^{\infty} p_j e^{jt}$$

for some $t \in \mathbf{R}$.

For the state space $\{0, 1, 2, \ldots\}$, $M_X(t)$ is defined for $t \leq 0$, but since the series may not converge for $t > 0$, it may not be defined for $t > 0$. The values of t for which the series converges depend on the particular values of the probabilities. The moment generating function generates the moments $E(X^k)$ of the distribution of the random variable X provided the summation converges in some interval about the origin:

$$M_X(0) = 1, \quad M'_X(0) = \mu_X = E(X), \quad M''_X(0) = E(X^2),$$

and, in general,

$$M_X^{(k)}(0) = E(X^k).$$

Definition 1.16. The *characteristic function (ch.f.)* of the discrete random variable X is

$$\phi_X(t) = E(e^{itX}) = \sum_{j=0}^{\infty} p_j e^{ijt}, \quad \text{where } i = \sqrt{-1}.$$

The characteristic function is defined for all real t because the summation converges for all t.

Definition 1.17. The *cumulant generating function (c.g.f)* of the discrete random variable X is the natural logarithm of the moment generating function and is denoted as $K_X(t)$,

$$K_X(t) = \ln[M_X(t)].$$

The generating functions for continuous random variables can be defined in a similar manner.

Definition 1.18. Assume X is a continuous random variable with p.d.f. f. The *probability generating function (p.g.f.)* of X is defined as

$$\mathcal{P}_X(t) = E(t^X) = \int_{\mathbf{R}} f(x) t^x \, dx.$$

The *moment generating function (m.g.f.)* of X is

$$M_X(t) = E(e^{tX}) = \int_{\mathbf{R}} f(x) e^{tx} \, dx$$

and the *characteristic function (ch.f.)* of X is

$$\phi_X(t) = E(e^{itX}) = \int_{\mathbf{R}} f(x) e^{itx} \, dx.$$

Finally, the *cumulant generating function (c.g.f.)* is $K_X(t) = \ln[M_X(t)]$.

The p.g.f. is defined for $|t| < 1$, the ch.f. for all real t and the m.g.f. and the c.g.f. for $t \leq 0$. One generating function can be transformed into another by applying the following identities:

$$\mathcal{P}_X(e^t) = M_X(t) \quad \text{and} \quad M_X(it) = \phi_X(t). \tag{1.6}$$

The same relationships established between the generating functions and the mean and the variance that were shown for discrete random variables hold for continuous random variables as well. In addition, formulas

for the mean and the variance in terms of the cumulant generating function are verified in the Exercises. These formulas are summarized below.

$$\boxed{\mu_X = \mathcal{P}'_X(1) = M'_X(0) = K'_X(0)}$$

and

$$\boxed{\sigma_X^2 = \begin{cases} \mathcal{P}''_X(1) + \mathcal{P}'_X(1) - [\mathcal{P}'_X(1)]^2 \\ M''_X(0) - [M'_X(0)]^2 \\ K''_X(0) \end{cases}}$$

Generating functions for linear combinations of independent random variables can be defined in terms of the generating functions of the individual random variables. Suppose $X_1, X_2, \ldots, X_n$ are n independent random variables and Y is a linear combination of these random variables,

$$Y = \sum_{i=1}^{n} a_i X_i.$$

Then the moment generating function of Y has a simple form; it is the product of the individual moment generating functions:

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E(e^{t(\sum a_i X_i)}) \\ &= E(e^{a_1 t X_1}) E(e^{a_2 t X_2}) \cdots E(e^{a_n t X_n}) \\ &= \prod_{i=1}^{n} M_{X_i}(a_i t). \end{aligned}$$

It is through the generating functions that information is obtained about the distributions of the process over time. In the next example, the p.g.f. and m.g.f. of a binomial distribution are calculated.

Example 1.10 Let X be a binomial random variable, $b(n,p)$, with p.d.f. $f(j) = \binom{n}{j} p^j (1-p)^{n-j}$, for $j = 0, 1, 2, \ldots, n$. The probability and moment generating functions of X are computed using the fact that $\sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} = (p + 1 - p)^n = 1$. The p.g.f. is

$$\mathcal{P}_X(t) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} t^j = \sum_{j=0}^{n} \binom{n}{j} (pt)^j (1-p)^{n-j}$$

so that

$$\mathcal{P}_X(t) = (pt + 1 - p)^n.$$

The m.g.f. can be obtained from the identity (1.6):

$$M_X(t) = (pe^t + 1 - p)^n.$$

Calculation of the derivatives,

$$\mathcal{P}'_X(t) = np(pt + 1 - p)^{n-1} \quad \text{and} \quad \mathcal{P}''_X(t) = n(n-1)p^2(pt + 1 - p)^{n-2},$$

leads to

$$\mu_X = \mathcal{P}'_X(1) = np$$

and

$$\sigma_X^2 = \mathcal{P}''_X(1) + \mathcal{P}'_X(1) - [\mathcal{P}'_X(1)]^2 = n(n-1)p^2 + np - n^2p^2$$

so that

$$\sigma_X^2 = np(1-p). \qquad \blacksquare$$

A very important result concerning generating functions states that the moment generating function *uniquely* defines the probability distribution (provided the m.g.f. exists in an open interval about zero). For example, if the m.g.f. of X equals $M_X(t) = [0.75 + 0.25e^t]^{20}$, then the distribution is binomial with $n = 20$ and $p = 0.25$ [i.e., $b(n,p) = b(20, 0.25)$].

1.4 Central Limit Theorem

An important theorem in probability theory relates the sum of independent random variables to the normal distribution. The central limit theorem states that the mean of n independent and identically distributed random variables, $\bar{X}$, has an approximate normal distribution if n is large. Generally, the expressions *random sample of size n* or *n independent and identically distributed (iid)* random variables are used to denote a collection of n independent random variables with the same distribution. We shall use this standard terminology.

The central limit theorem is an amazing result when one realizes this normal approximation applies to a collection of random variables from any distribution with a finite mean and variance, discrete or continuous. If the distribution is skewed and discrete, the size of the random sample may need to be large to ensure a good approximation.

Recall that the mean and variance of $\bar{X} = \sum_{i=1}^n X_i/n$, where $\{X_i\}_{i=1}^n$ is a random sample satisfying $\mu_{X_i} = \mu$ and $\sigma^2_{X_i} = \sigma^2$, $i = 1, 2, \ldots, n$, is

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma^2_{\bar{X}} = \sigma^2/n.$$

The central limit theorem is stated in terms of the random variable W_n, where

$$W_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Theorem 1.1 (Central Limit Theorem). *Let $X_1, X_2, \ldots, X_n, \ldots$, be a sequence of iid random variables with finite mean, $|\mu| < \infty$, and positive standard deviation, $0 < \sigma < \infty$. Then, as $n \to \infty$, the limiting distribution of*

$$W_n = \frac{\sum_{i=1}^n X_i/n - \mu}{\sigma/\sqrt{n}},$$

is a standard normal distribution.

The sequence $\{W_n\}$ converges in distribution to a standard normal distribution. In particular, it can be shown for each $z \in (-\infty, \infty)$ that

$$\lim_{n\to\infty} \text{Prob}\{W_n \le z\} = F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy,$$

where $F(z)$ is the c.d.f. of the standard normal distribution. In a simple proof of the central limit theorem, based on the assumption that the m.g.f. exists, it is shown that the m.g.f. of W_n, $M(t;n)$, approaches a m.g.f. from a standard normal distribution, $\lim_{n\to\infty} M(t;n) = e^{t^2/2}$ (Hogg and Craig, 1995; Hogg and Tanis, 2001). For a more general proof based on characteristic functions, see Cramér (1945) or Schervish (1995). The central limit theorem is applied in Chapter 8, when deriving stochastic differential equations for interacting populations.

Example 1.11 Suppose $X_1, X_2, \ldots, X_{15}$ are iid random variables from the binomial distribution $b(6, 1/3)$. The mean and variance for each of the X_i are $\mu = 2$ and $\sigma^2 = 4/3$. A graph of the approximate probability density (histogram) of

$$W_{15} = \frac{\sum_{i=1}^{15} X_i/15 - 2}{\sqrt{4/3}/\sqrt{15}}$$

$(n = 15)$ is compared to the standard normal density on the interval $[-3.5, 3.5]$ (Figure 1.4). The histogram is generated from many random samples of size $n = 15$. For each random sample, the value of w_{15} is calculated. It can be seen that the probability histogram is in close agreement with the standard normal p.d.f. ■

A practical application of the central limit theorem is to approximate $\text{Prob}\{a < \bar{X} < b\}$, where $\bar{X} = \sum_{i=1}^n X_i/n$ is the average of a random sample of size n.

Example 1.12 Let $X_1, \ldots, X_n$ be a random sample of size $n = 25$ from the uniform distribution $U(0, 1)$. The mean $E(X_i) = 1/2$ and the variance $Var(X_i) = 1/12$ (see Example 1.4). Thus, by the central limit theorem,

$$W_{25} = \frac{\bar{X} - 1/2}{1/(5\sqrt{12})}$$

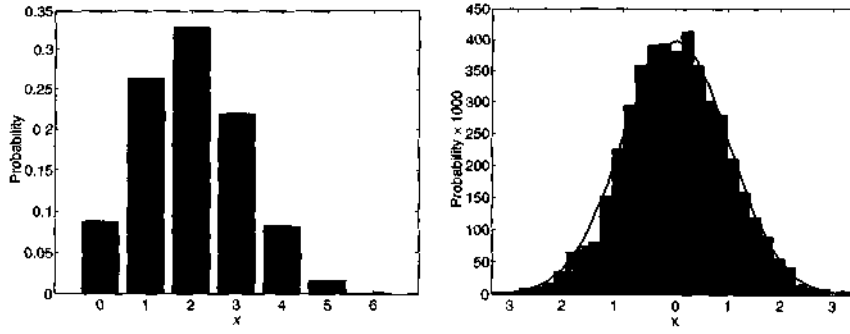

Figure 1.4. Graphs of the binomial mass function $b(6, 1/3)$, and the standard normal density. A probability histogram of p.d.f. W_{15} (defined in the central limit theorem) is graphed against the standard normal density.

has a distribution close to $N(0,1)$. Then

$$\text{Prob}\{\bar{X} < 1/2\} = \text{Prob}\{W_{25} < 0\} \approx 0.5$$

and

$$\begin{aligned}\text{Prob}\{\bar{X} < 1/3\} &= \text{Prob}\left\{\frac{\bar{X} - 1/2}{1/(5\sqrt{12})} < \frac{1/3 - 1/2}{1/(5\sqrt{12})}\right\} \\ &\approx \text{Prob}\{W_{25} < -2.887\} \\ &\approx 0.0019.\end{aligned}$$

■

1.5 Introduction to Stochastic Processes

A stochastic process is just a collection of random variables. More specifically,

Definition 1.19. A *stochastic process* is a collection of random variables $\{X_t(s) : t \in T, s \in S\}$, where T is some index set and S is the *common sample space* of the random variables. For each fixed t, $X_t(s)$ denotes a single random variable defined on S. For each fixed $s \in S$, $X_t(s)$ corresponds to a function defined on T that is called a *sample path* or a *stochastic realization* of the process.

In addition, a stochastic process may be a collection of random vectors. For example, for two random variables, a stochastic process is a collection of random vectors $\{(X_t^1(s), X_t^2(s)) : t \in T, s \in S\}$.

When speaking of a stochastic process, sometimes the variable s is omitted; the random variables are denoted simply as X_t or $X(t)$. We will follow this practice. A stochastic model is based on a stochastic process in which

specific relationships among the set of random variables $\{X_t\}$ are assumed to hold.

There are different methods and techniques for formulating and analyzing stochastic processes that depend on whether the random variables and index set are discrete or continuous. These distinctions between discrete versus continuous random variables and discrete versus continuous index set determine the type of stochastic model and the techniques that can be used to study properties of the model. The set T is often referred to as *time*, and in our stochastic models it will frequently represent time. The first type of stochastic models discussed in the next chapter are those where the index set and the state space are discrete; the models are discrete time Markov chain models. In subsequent chapters, stochastic models are discussed where the index set is continuous but the state space is discrete, referred to as continuous time Markov chain models. These types of models have received the most attention in terms of biological applications: competition, predation, and epidemic processes. It will be easy to see the connection between deterministic differential equation models and stochastic models of this type. Finally, the last type of models that will be discussed are those where the index set and state space are continuous. These types of models are referred to as diffusion processes, and the stochastic realization $X(t)$ satisfies a stochastic differential equation. When the demographic variability in the birth and death rates is included, the model can no longer be expressed in terms of an ordinary differential equation but is expressed as a stochastic differential equation.

The theory of stochastic processes arose from studying biological as well as physical problems. According to Guttorp (1995), one of the first occurrences of a Markov chain may have been in explaining rainfall patterns in Brussels by Quetelet in 1852. The simple branching process was invented by Bienaymé in 1845 to compute the probability of extinction of a family surname. In 1910, Rutherford and Geiger and the mathematician Bateman described the disintegration of radioactive substances using a Poisson process. In 1905, Einstein described Brownian motion of gold particles in solution, and in 1900, Bachelier used this same process to describe bond prices (Guttorp, 1995). The simple birth and death process was introduced by McKendrick in 1914 to describe epidemics, and Gibbs in 1902 used nearest-neighbor models to describe the interactions among large systems of molecules (Guttorp, 1995). Stochastic processes are now used to model many different types of phenomena from a variety of different areas, including biology, physics, chemistry, finance, economics, and engineering.

Four examples of stochastic processes from population biology are described, where the state space or index set are discrete or continuous.

1. X_t is the position of an object at time t, during a 24 hour period, whose directional distance from a particular point 0 is measured in

integer units. In this case, $T = \{0, 1, 2, \ldots, 24\}$ and the state space is $\{0, \pm 1, \pm 2, \ldots\}$; both time and state space are *discrete*.

2. X_t is the number of births in a given population during the time period $[0, t]$. In this case, $T = \mathbf{R}_+ = [0, \infty)$ and the state space is $\{0, 1, 2, \ldots\}$. Time is *continuous* and the state space is *discrete*.

3. X_t is the population density at time $t \in T = \mathbf{R}_+ = [0, \infty)$. The state space is also $\mathbf{R}_+$; both time and state are *continuous*.

4. X_t is the density of an annual plant species in year t, where $T = \{0, 1, 2, \ldots\}$ and the state space is $\mathbf{R}_+$. Time is *discrete* but the state space is *continuous*.

Examples such as these will be studied in more detail in Chapters 2 through 8. In the case of discrete time, assumptions are made regarding the relationship between the state of the system at time t to the state at time $t + 1$. In simple cases, the first example is a random walk model, the second example is a simple birth process, and the third example is a birth and death process with a continuous state space. The fourth example is a stochastic process for which its deterministic analogue is a difference equation. We do not discuss stochastic models of this latter type, where the state space is discrete and time is continuous, because the theory is not as well developed, and there are few biological models of this type.

One of the simplest examples of a stochastic process of the type mentioned in the second example is a simple birth process (the analogue of exponential growth in deterministic theory). Recall that a deterministic model of exponential growth satisfies $y = ae^{bt}$, $a, b > 0$, $t \in [0, \infty)$. Graphs of three stochastic realizations of the simple birth process when $a = 1 = b$ are given in Figure 1.5 and compared to the deterministic exponential growth model. The deterministic model has a single solution, $y = e^t$, whereas the stochastic model has an infinite number of stochastic realizations, three of which are graphed. For a fixed time t, there is, associated with the random variable X_t, a probability function f_t. At a fixed value of t, the stochastic realization equals n, $X_t = n$, for some $n = 0, 1, 2, \ldots$, with probability

$$f_t(n) = \text{Prob}\{X_t = n\}.$$

The stochastic realization may equal any value of n at time t, provided $f_t(n) > 0$. It will be shown in Chapter 6, for this example, that the mean of X_t equals the deterministic solution,

$$\mu_t = E(X_t) = \sum_{n=0}^{\infty} n f_t(n) = e^t.$$

In the simple birth process, time is continuous and the state space is discrete. The simple birth process is discussed briefly in the next section.

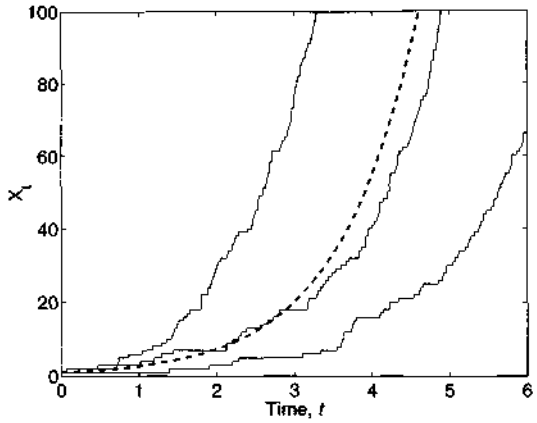

Figure 1.5. Three stochastic realizations of the simple birth process and corresponding deterministic exponential growth model $y = e^t$ (dashed curve).

Some techniques useful to the study of stochastic processes will be introduced and some of the differences between deterministic and stochastic models will be illustrated. The simple birth process will be studied in more detail in Chapter 6.

1.6 An Introductory Example: A Simple Birth Process

Deterministic and stochastic exponential growth models are derived from first principles. The stochastic model is continuous in time but discrete in the state space. The stochastic model is only briefly described, and details are left for discussion in later chapters.

Three assumptions are made in developing this simple birth model. It is assumed that

(i) No individuals die.
(ii) There are no interactions between individuals.
(iii) The birth rate b is the same for all individuals.

See also Renshaw (1993). The term *individual* could mean a cell or some type of organism.

First, a deterministic model is derived. Let $n(t)$ denote the population size at time t. In a small time period Δt, the increase in population size due to a single individual is $b \times \Delta t$ and the increase in size due to all individuals is $b\,\Delta t \times n(t)$. Thus,

$$n(t + \Delta t) = n(t) + b\,\Delta t\, n(t).$$

Rewriting this expression leads to

$$\frac{n(t+\Delta t)-n(t)}{\Delta t} = bn(t).$$

Letting $\Delta t \to 0$, we arrive at the differential equation for exponential growth,

$$\frac{dn(t)}{dt} = bn(t).$$

If the initial population size is $n(0) = a$, then the solution to this differential equation is

$$n(t) = ae^{bt}.$$

The population size is predicted at time t with absolute certainty once the initial size a and the birth rate b are known.

Next, a stochastic model is formulated. In this case, the population size is not known with certainty but with some probability; the population size will be n at time t. It is assumed that the population size is discrete valued but time is continuous. Since the state space is discrete and time is continuous the stochastic process satisfies $X_t \in \{0, 1, 2, \ldots\}$, $t \in [0, \infty)$, where X_t is the discrete random variable for the size of the population at time t. Let the probability mass function associated with the random variable X_t be denoted $\{p_n(t)\}_{n=0}^{\infty}$, where

$$p_n(t) = \text{Prob}\{X_t = n\}.$$

Note that the notation is different from the previous sections. This notation for the probability mass function $p_n(t)$ is consistent with the notation used in later chapters.

The random variables $\{X_t\}$ are related by making the following assumptions. Assume that in a sufficiently small period of time Δt,

1. The probability that a birth occurs is approximately $b\ \Delta t$.
2. The probability of more than one birth in time Δt is negligible.
3. At $t = 0$, $\text{Prob}\{X_0 = a\} = 1$.

That the probability is negligible means it is of order Δt or $o(\Delta t)$; that is, $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$; $o(\Delta t)$ approaches zero faster than Δt. Thus, the first assumption can be stated more precisely as the probability that a birth occurs is $b\ \Delta t + o(\Delta t)$.

Based on assumptions 1 and 2, for the population to be of size n at time $t + \Delta t$, either it is of size n at time t and no birth occurs in $(t, t + \Delta t)$, or else it is of size $n - 1$ at time t and one birth occurs in $(t, t + \Delta t)$. The probability that a population of size n increases to $n + 1$ in $(t, t + \Delta t)$ is approximately $b\ \Delta t \times n$, and the probability that the population fails to increase in that time period is then, approximately $1 - b\ \Delta t \times n$.

The assumptions relate the state of the process at time $t + \Delta t$, $X_{t+\Delta t}$ to the state at time t, X_t. In terms of the probabilities, the probability

that the state equals n at time $t + \Delta t$ depends on whether at time t the population size was $n-1$ and there was a birth or the size was n and there was no birth; that is,

$$p_n(t + \Delta t) = p_{n-1}(t)b(n-1)\,\Delta t + p_n(t)(1 - bn\,\Delta t).$$

There is an inherent assumption about independence in deriving this equation; that is, the state of the process at time $t + \Delta t$ only depends on the state at time t and not the times prior to t (this is known as a *Markov property*). Subtract $p_n(t)$ from both sides of the equation above, divide by Δt, and let $\Delta t \to 0$, to obtain a system of differential equations known as the *forward Kolmogorov differential equations*:

$$\frac{dp_n(t)}{dt} = b(n-1)p_{n-1}(t) - bnp_n(t),$$

where $n = 1, 2, \ldots$. For example,

$$\frac{dp_1(t)}{dt} = -bp_1(t) \quad \text{and} \quad \frac{dp_2(t)}{dt} = bp_1(t) - 2bp_2(t).$$

If the initial population size is zero, then no births can occur and $p_0(t) = 1$ for all time.

The process begins with a known probability distribution for X_0; that is, $\{p_n(0)\}_{n=0}^{\infty}$. Initially, the population size is fixed at a, $X_0 = a$, so that the initial probabilities satisfy $p_a(0) = 1$ and $p_n(0) = 0$, $n \neq a$. The forward Kolmogorov equations can be solved iteratively or by using moment generating function techniques. Methods of solution will be discussed in Chapter 5. It will be shown in Chapter 6 that the solution $p_n(t)$ has the form of a negative binomial distribution for each fixed time t:

$$p_n(t) = \binom{n-1}{a-1} e^{-abt}(1 - e^{-bt})^{n-a}, \quad n = a, a+1, a+2, \ldots.$$

This distribution is a shift of a units to the right of the negative binomial distribution defined in (1.3) (see Exercise 7). The parameter p in the negative binomial distribution is $p = e^{-bt}$. The probability distribution $\{p_n(t)\}_{n=a}^{\infty}$ is graphed in Figure 1.6 when $a = 1$ and $b = 1$ at times $t = 0, 1, 2$, and 3.

Finding the probability distribution solves the stochastic modeling problem, just as finding the solution $n(t)$ solves the deterministic modeling problem. Other information about the stochastic process can be assessed from the probability distribution (e.g., mean and variance). Therefore, the goal in stochastic modeling is to determine the probability distribution associated with the stochastic process. If this is not possible, then information about the mean or variance or other properties of the distribution is sought.

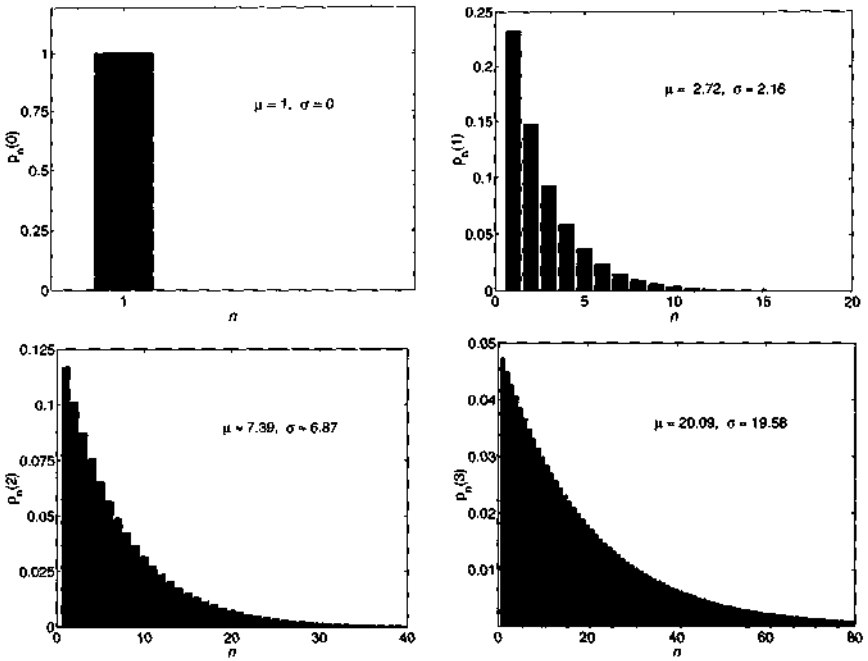

Figure 1.6. Graphs of the probability mass function $p_n(t)$, $n = 0, 1, 2, \ldots$, at times $t = 0, 1, 2, 3$ when $X_0 = a = 1$ and $b = 1$ equals the birth rate.

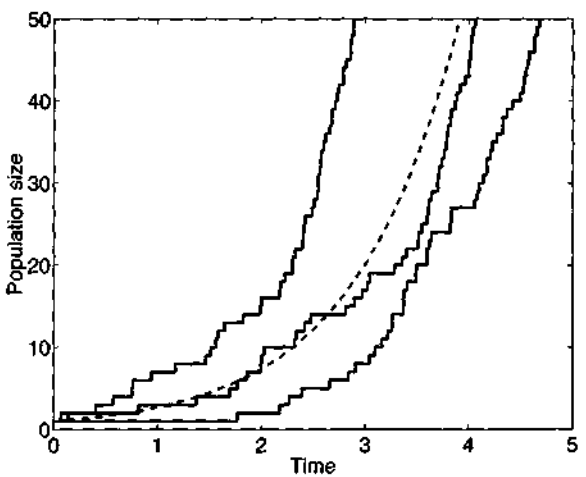

Figure 1.7. Three stochastic realizations of the simple birth process are graphed when $b = 1$, $a = 1$, and $X_0 = 1$. In addition, the deterministic exponential growth model, $n(t) = e^t$, is the dashed curve.

t	μ_t	σ_t^2	σ_t
0	1	0	0
1	2.72	4.67	2.16
2	7.39	47.21	6.87
3	20.09	383.34	19.58

Table 1.1. The mean μ_t, variance σ_t^2, and standard deviation σ_t for the simple birth process at times $t = 0, 1, 2, 3$ when $a = 1$ and $b = 1$

The mean and variance of the simple birth process X_t can be obtained directly from the negative binomial probability distribution:

$$\begin{aligned} \mu_t &= ae^{bt} \\ \sigma_t^2 &= ae^{bt}(e^{bt} - 1). \end{aligned}$$

Note for this example that the mean equals the solution to the deterministic model. However, the variance increases as time increases. Specific values for the mean and variance when $a = 1$ and $b = 1$ are given in Table 1.1.

An important part of modeling is numerical simulation. Many different programming languages can be used to simulate the dynamics of a stochastic model. The output from a MATLAB program for the simple birth process is graphed in Figure 1.7. MATLAB and FORTRAN programs for the simple birth process are given in the Appendix for Chapter 1. Three stochastic realizations of the simple birth process when $b = 1$, $a = 1$, and $p_1(0) = 1$ are graphed in Figure 1.7. The corresponding deterministic exponential growth model, $n(t) = e^t$, is also graphed. Table 1.2 lists the times at which a birth occurs (up to a population size of 50) for two different realizations or sample paths for the simple birth process. Notice that it requires times of 2.675 and 4.573, respectively, for two of the three realizations to reach a population size of 50. Recall that the time to reach a population size of 50 in the deterministic model with $b = 1 = a = n(0)$ is found by solving $50 = \exp(t)$ for t or $t = \ln 50 \approx 3.9120$. In general, the time between births is much longer when the population size is small. As the population size builds up, then births occur more frequently and the interevent time decreases. See the Appendix for Chapter 1 for a brief discussion of interevent time.

The values of the stochastic realizations at a particular time t, $X_t = n$, depend on the probability $p_n(t)$,

$$p_n(t) = \text{Prob}\{X_t = n\}.$$

If $p_n(t) > 0$, then it is possible for a stochastic realization to have the value n at time t. For example, for $t = 0, 1, 2$, and 3, the probability distributions, $p_n(1)$, $p_n(2)$ and $p_n(3)$ graphed in Figure 1.6 show that it is possible for

Realization 1		Realization 2	
Size $n(t)$	Event Time t	Size $n(t)$	Event Time t
1	0	1	0
2	0.1379	2	1.7642
3	0.4065	3	2.1740
4	0.5753	4	2.2688
5	0.7546	5	2.3901
6	0.7664	6	2.6641
7	0.9430	7	2.8388
8	1.1731	8	2.9087
9	1.4634	9	3.0444
10	1.5109	10	3.0947
⋮	⋮	⋮	⋮
50	2.8898	50	4.6784

Table 1.2. For two stochastic realizations, the times at which a birth occurs are given for a simple birth process with $b = 1$ and $a = 1$

the stochastic realization to have any positive value n at $t = 1, 2$, or 3. However, the probability of a large value of n is very small.

In the following chapters discrete time Markov chain models, continuous time Markov chain models, and diffusion processes are studied. In Chapter 2 the theory of discrete time Markov chain models is presented. Discrete time Markov chain models are models that are discrete in time and in state.

1.7 Exercises for Chapter 1

1. The following probability mass function for the discrete random variable X defines a *geometric distribution*:

$$f(j) = p(1-p)^j, \quad j = 0, 1, 2, \ldots, \quad 0 < p < 1.$$

   (a) Show that the probability generating function (p.g.f.) of X satisfies $\mathcal{P}_X(t) = p(1-(1-p)t)^{-1}$.

   (b) Use the p.g.f. to find the mean μ_X and variance σ_X^2.

2. The continuous random variables X_1 and X_2 have the joint probability density function (p.d.f.),

$$f(x_1, x_2) = e^{-x_1 - x_2}, \quad 0 < x_1 < \infty, \quad 0 < x_2 < \infty \qquad (1.7)$$

   and zero otherwise.

(a) Show that X_1 and X_2 are independent and have exponential distributions.

(b) Find the moment generating function (m.g.f.) of X_1 and X_2.

(c) Find $E\left(e^{t(X_1+X_2)}\right)$.

3. Suppose X_1 and X_2 are independent, continuous random variables with joint p.d.f. satisfying $f(x_1, x_2) = f_1(x_1)f_2(x_2)$.

(a) Show that $E(X_1X_2) = E(X_1)E(X_2)$.

(b) For the joint p.d.f. defined by equation (1.7), find $E(X_1X_2)$ and $E(X_1^2X_2^2)$.

4. Assume that the m.g.f. of X, $M_X(t)$, converges in some open interval about the origin. Use the facts that the cumulant generating function (c.g.f) satisfies

$$K_X(t) = \ln M_X(t)$$

and the properties of the m.g.f. to show that the c.g.f. satisfies

$$K_X(0) = 0, \quad K'_X(0) = \mu_X, \quad K''_X(0) = \sigma_X^2.$$

5. Show that the p.g.f. of the Poisson distribution is $\mathcal{P}_X(t) = e^{\lambda(t-1)}$. Then use the p.g.f. to show that the mean and variance satisfy $\mu_X = \lambda = \sigma_X^2$.

6. Show that the p.g.f. of the negative binomial distribution defined in equation (1.3) is

$$\mathcal{P}_X(t) = \frac{p^n}{[1-(1-p)t]^n}.$$

Then use the p.g.f. to find the mean and variance of X. [*Hint*: $\sum_{x=0}^{\infty} \binom{x+n-1}{n-1} p^n(1-p)^x = p^n(p)^{-n} = 1$.]

7. The following p.m.f. is associated with a *negative binomial distribution*:

$$f(y) = \begin{cases} \binom{y-1}{n-1} p^n(1-p)^{y-n}, & y = n, n+1, n+2, \ldots, \\ 0, & \text{otherwise}, \end{cases}$$

where n is a positive integer and $0 < p < 1$ (see Section 1.6). This function of y is a shift of n units to the right of the function $f(x)$ given in (1.3). Show that the p.g.f. of Y satisfies $\mathcal{P}_Y(t) = t^n\mathcal{P}_X(t)$, $\mu_Y = \mu_X + n$, and $\sigma_Y^2 = \sigma_X^2$, where X is the random variable defined by (1.3) (see Exercise 6).

8. For the gamma distribution,

(a) Show that the m.g.f. is $M_X(t) = (1 - \beta t)^{-\alpha}$ for $t < 1/\beta$.

(b) Use the m.g.f. to find the mean and variance of the gamma distribution.

(c) For the special case of the exponential distribution, $\alpha = 1$ and $\beta = 1/\lambda$, find the mean and variance.

9. Suppose that the p.g.f. of a continuous random variable X satisfies $\mathcal{P}(t) = 1/(1 - \theta \ln(t))$, $\theta > 0$.

(a) Find the m.g.f. $M(t)$ and the c.g.f. $K(t)$.

(b) Find the mean and variance of X.

10. Suppose that the random variable X is exponentially distributed. Show that X has the following property:

$$\text{Prob}\{X \geq t + \Delta t | X \geq t\} = \text{Prob}\{X \geq \Delta t\}.$$

This property of the exponential distribution is known as the *memoryless property.*

11. Suppose the discrete random variable X has a geometric distribution (see Exercise 1). Show that X has the memoryless property (see Exercise 10)

$$\text{Prob}\{X \geq j | X \geq i\} = \text{Prob}\{X \geq j - i\}, \quad i \leq j.$$

12. Suppose the continuous random variable X has an exponential distribution with mean $\mu = 10$. Find

(a) $\text{Prob}\{5 < X < 15\}$

(b) $\text{Prob}\{X > 15\}$

(c) $\text{Prob}\{X > 20 | X > 5\}$

13. Suppose the random variable X has a distribution that is $N(2, 1)$. Find

(a) $\text{Prob}\{X \leq 2\}$

(b) $\text{Prob}\{-1 \leq X - 2 \leq 1\}$

(c) $\text{Prob}\{-3 \leq X - 2 \leq 3\}$

(d) $\text{Prob}\{-0.9 \leq X \leq 1.5\}$

14. Show that the m.g.f. of the normal distribution is $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$ and the c.g.f. is $K_X(t) = \mu t + \sigma^2 t^2/2$. Then use the c.g.f. to show that μ and σ^2 are the mean and variance of X, respectively.

15. Suppose that the random variable X has a m.g.f. $M_X(t)$.

(a) Show that the m.g.f. of $Y = X - \mu_X$ is $M_Y(t) = e^{-\mu_X t} M_X(t)$ and $M_Y^{(k)}(0) = E(Y^k) = E[(X - \mu_X)^k]$ gives the kth moment of X about its mean (Bailey, 1990).

(b) Suppose that X is a binomial random variable. Find $M_Y(t)$; then find the first three moments of X about its mean value. [*Hint*: A computer algebra system may be used to calculate the derivatives of M_Y and to evaluate at $t = 0$.]

16. Suppose $X_1, \ldots, X_n$ is a random sample of size $n = 25$ from a Poisson distribution with parameter λ. Apply the central limit theorem to approximate $\text{Prob}\{\bar{X} \leq \lambda\}$ and $\text{Prob}\{\bar{X} \leq \lambda + \sqrt{\lambda}/5\}$.

17. Consider the exponential growth model, $dn/dt = bn$, $n(0) = a$. Find the doubling time, $T_{2a,a}$, the first time the population size is $2a$; then find $T_{ka,a}$.

18. Modify the MATLAB program in the Appendix and graph Poisson probability mass functions for $\lambda = 5$ and $\lambda = 10$. On each of these graphs, superimpose the graph of the normal density, $N(\lambda, \lambda)$. How do these two distributions compare? It can be shown that if X has a Poisson distribution, then the random variable $Z = (X - \lambda)/\sqrt{\lambda}$ approaches a standard normal distribution as $\lambda \to \infty$.

1.8 References for Chapter 1

Anderson, S. L. 1990. Random number generators on vector supercomputers and other advanced architectures. *SIAM Review.* 32: 221–251.

Bailey, N. T. J. 1990. *The Elements of Stochastic Processes with Applications to the Natural Sciences.* John Wiley & Sons, New York.

Bharucha-Reid, A. T. 1997. *Elements of the Theory of Markov Processes and Their Applications.* Dover Pub. Inc., New York.

Cramér, H. 1945. *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications.* Vol 1. 3rd ed. John Wiley & Sons, New York.

Gard, T. C. 1988. *Introduction to Stochastic Differential Equations.* Marcel Dekker, Inc., New York and Basel.

Goel, N. S. and N. Richter-Dyn. 1974. *Stochastic Models in Biology.* Academic Press, New York.

Guttorp, P. 1995. *Stochastic Modeling of Scientific Data.* Chapman & Hall, London.

Hogg, R. V. and A. T. Craig. 1995. *Introduction to Mathematical Statistics.* 5th ed. Prentice Hall, Upper Saddle River, N. J.

Hogg, R. V. and E. A. Tanis 2001. *Probability and Statistical Inference.* 6th ed. Prentice Hall, Upper Saddle River, N. J.

Hsu, H. P. 1997. *Schaum's Outline of Theory and Problems of Probability, Random Variables, and Random Processes.* McGraw-Hill, New York.

Iosifescu, M. and P. Tăutu. 1973. *Stochastic Processes and Applications in Biology and Medicine I. Theory.* Springer-Verlag, Berlin, Heidelberg, New York.

Iosifescu, M. and P. Tăutu. 1973. *Stochastic Processes and Applications in Biology and Medicine II. Models.* Springer-Verlag, Berlin, Heidelberg, New York.

Karlin, S. and H. M. Taylor. 1975. *A First Course in Stochastic Processes.* 2nd ed. Academic Press, New York.

Karlin, S. and H. M. Taylor. 1981. *A Second Course in Stochastic Processes.* Academic Press, New York.

Nisbet, R. M. and W. S. C. Gurney. 1982. *Modelling Fluctuating Populations.* John Wiley & Sons, Chichester and New York.

Renshaw, E. 1993. *Modelling Biological Populations in Space and Time.* Cambridge Studies in Mathematical Biology. Cambridge University Press, Cambridge.

Ross, S. M. 1983. *Stochastic Processes.* John Wiley & Sons, New York.

Ross, S. M. 1989. *Introduction to Probability Models.* 4th ed. Academic Press, New York.

Schervish, M. J. 1995. *Theory of Statistics.* Springer-Verlag, New York, Berlin, Heidelberg.

Schinazi, R. B. 1999. *Classical and Spatial Stochastic Processes.* Birkhäuser, Boston.

Taylor, H. M. and S. Karlin. 1998. *An Introduction to Stochastic Modeling.* 3rd ed. Academic Press, New York.

1.9 Appendix for Chapter 1

1.9.1 MATLAB and FORTRAN Programs

The following MATLAB program can be used to graph the Poisson mass function given in Figure 1.3.

```
% MATLAB Program:
% Poisson function.
clear all % Clears variables and functions from memory.
set(0,'DefaultAxesFontSize',18); % Increases axes labels.
lastx=25; % Truncates the Poisson function at x = 25.
x=linspace(0,lastx,lastx+1);
w(1)=1;
w(2)=1;
lambda=3;
for i=2:lastx
    w(i+1)=i*w(i);
end
y=lambda.^x*exp(-lambda)./w;
bar(x,y,'k') % Graphs a histogram of y.
axis([-1,12,0,0.25]) % Sets the scaling on the axes.
xlabel(x);
ylabel(f(x));
```

The following FORTRAN and MATLAB programs can be used to generate sample paths for the simple birth process.

```
REAL*8 N(50), T(50),Y,B,XX
PRINT *, 'SEED (POSITIVE NUMBER < M)'
READ *, XX
T(1)=0.
N(1)=1.
B=1.
Y=RAND(XX)
DO I=1,49
    Y=RAND(XX)
   T(I+1)=-DLOG(Y)/(B*N(I))+T(I)
   N(I+1)=N(I)+1
   PRINT *, 'T', T(I+1), 'N', N(I+1)
ENDDO
STOP
END

FUNCTION RAND(XX)
REAL*8 XX,A,M,D
A=16807.
```

```
M=2147483647.
ID=A*XX/M
D=ID
XX=A*XX-D*M
RAND=XX/M
RETURN
END
```

```
% MATLAB Program:
% Sample paths for the simple birth process.
clear all % Clears variables and functions from memory.
set(0,'DefaultAxesFontSize',18); % Increases axes labels.
b=1;
x=linspace(0,50,51); % Defines the vector [0,1,2,...,50].
y=exp(x);
n=linspace(1,50,50); % Defines the population vector.
for j=1:3; % Three sample paths.
    t(1)=0;
    for i=1:49;
       t(i+1)=t(i)-log(rand)/(b*n(i));
    end % End of i loop.
    s= stairs(t,n); % Draws stairstep graph of n.
    set(s,'LineWidth',2); % Thickens the line width.
    hold on % Holds the current plot.
end % end of j loop
plot(x,y,'k--','LineWidth',2); % Plots the exponential.
axis([0,5,0,50]); % Sets scaling for the x- and y-axes.
xlabel('Time'); % Label for the x-axis.
ylabel('Population Size'); % Label for the y-axis.
hold off  % Erases previous plots before drawing new ones.
```

Note: A statement following % explains the command; these statements are not executable.

1.9.2 Interevent Time

To simulate the simple birth process, it is necessary to know the random variable for the time between births or *interevent time*. It is shown in Chapter 5 that the random variable for the interevent time is exponentially distributed; if the population is of size N, then the time H to the next event (or interevent time) has a distribution satisfying

$$P(H \geq h) = \exp(-bNh).$$

To simulate a value $h \in H$, a uniformly distributed random number Y is selected in the range $0 \leq Y \leq 1$ [i.e., from the uniform distribution $U(0,1)$].

Then

$$Y = \exp(-bNh),$$

which yields

$$h = -\frac{\ln(Y)}{bN}$$

(Renshaw, 1993). Notice in the simple birth process that as N increases, $h = -\ln(Y)/bN$ decreases; the interevent time decreases as the population size increases. To simulate an interevent time, it is necessary to apply a random number generator that generates uniformly distributed numbers in $[0, 1]$. The function subroutine RAND in the FORTRAN program is a pseudo–random number generator and is based on the recursion relation $y_{n+1} = (Ay_n) \mathrm{mod} M$, where $RAND = y_{n+1}/M \in [0, 1)$ and the modulus $M = 2^{31} - 1$ is a Mersenne prime (Anderson, 1990). The term "rand" in the MATLAB program is a built-in MATLAB function for a uniform random number generator on $[0, 1]$.

Chapter 2

Discrete Time Markov Chains

2.1 Introduction

In this chapter, discrete time Markov chains are introduced. Both time and state space are discrete. The theory and application of Markov chains is probably one of the most well-developed theories of stochastic processes. A classic textbook on finite Markov chains (where the state space is finite) is the textbook by Kemeny and Snell (1960). Some additional references on the theory and numerical methods for discrete time Markov chains include *A First Course in Stochastic Processes*, by Karlin and Taylor (1975); *An Introduction to Stochastic Modeling*, by Taylor and Karlin (1998); *Classical and Spatial Stochastic Processes*, by Schinazi (1999); *Markov Chains*, by Norris (1997); and *Introduction to the Numerical Solution of Markov Chains*, by Stewart (1994).

We introduce some basic notation and theory for discrete time Markov chains in this chapter. A discrete time chain can be classified as irreducible or reducible, periodic or aperiodic, and recurrent or transient. These classifications help in determining the behavior of the Markov chain. Basic theorems concerning the asymptotic behavior are stated that apply to particular types of Markov chains. For example, it is shown that a stationary limiting distribution exists for an aperiodic, irreducible, and recurrent Markov chain. A stationary distribution is analogous to a stable equilibrium in a deterministic model. However, in a stochastic model, the "equilibrium" is defined by a probability distribution, known as the stationary probability distribution. Some well-known examples of discrete time Markov chains are discussed in this chapter, including the random walk model in one, two, and three dimensions. In addition, a problem related to genetics inbreeding is discussed.

2.2 Definitions and Notation

Consider a discrete time stochastic process, $\{X_n\}$, $n = 0, 1, 2, \ldots$, where the random variable X_n is a discrete random variable defined on a finite or countably infinite *state space*. For convenience, we denote the state space as $\{1, 2, \ldots\}$. However, the set could be finite and could include nonpositive integer values. Also, the variable n is used instead of t to denote an element of the index set; this notation is used frequently in discrete time processes. The index set is defined as $\{0, 1, 2, \ldots\}$, since it often represents the progression of time, which is also the reason for the terminology, discrete time processes. Therefore, the index n shall be referred to as "time" n.

A Markov stochastic process is a stochastic process in which the future behavior of the system depends only on the present and not on its past history. More formally,

Definition 2.1. A discrete time stochastic process $\{X_n\}_{n=0}^{\infty}$ is said to have the *Markov property* if

$$\text{Prob}\{X_n = i_n | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} = \text{Prob}\{X_n = i_n | X_{n-1} = i_{n-1}\},$$

where the values of $i_k \in \{1, 2, \ldots\}$ for $k = 0, 1, 2, \ldots, n$. The stochastic process is then called a *Markov chain* or, more specifically, a *discrete time Markov chain*. It is called a *finite state Markov chain* or a *finite Markov chain* if the state space is finite.

The stochastic process is referred to as a *chain* when the state space is discrete. The name *Markov* refers to Andrei A. Markov, a Russian probabilist (1856–1922), whose work in Markov chains contributed much to the theory of stochastic processes.

The notation Prob is used to denote the induced probability measure, $\text{Prob}\{\cdot\} = P_{X_n}(\cdot)$, because P will refer to the transition matrix that is defined below. Denote the probability mass function associated with the random variable X_n by $\{p_i(n)\}_{i=0}^{\infty}$, where

$$p_i(n) = \text{Prob}\{X_n = i\}. \tag{2.1}$$

The state of the process at time n, X_n, is related to the process at time $n + 1$ through what is known as the transition probabilities. If the process is in state i at time n, at the next time step $n + 1$, it will either stay in state i or move or transfer to another state j. The probabilities for these changes in state are defined by the one-step transition probabilities.

Definition 2.2. The *one-step transition probability*, denoted as $p_{ji}(n)$, is defined as the following conditional probability:

$$p_{ji}(n) = \text{Prob}\{X_{n+1} = j | X_n = i\},$$

the probability that the process is in state j at time $n + 1$ given that the process was in state i at the previous time n, for $i, j = 1, 2, \ldots$.

Definition 2.3. If the transition probabilities $p_{ji}(n)$ in a Markov chain do not depend on time n, they are said to be *stationary* or *time homogeneous* or simply *homogeneous*. In this case, we shall use the notation p_{ji}. If the transition probabilities are time dependent, $p_{ji}(n)$, then they are said to be *nonstationary* or *nonhomogeneous*.

A Markov chain can have either stationary or nonstationary transition probabilities. Unless stated otherwise, it shall be assumed that the transition probabilities of the Markov chain are stationary. For each state, the one-step transition probabilities satisfy

$$\sum_{j=1}^{\infty} p_{ji} = 1, \quad \text{for } i = 1, 2, \ldots \text{ and } p_{ji} \geq 0,$$

meaning that, with probability one, the process in any state i must move or transfer to some other state j, $j \neq i$ or stay in state i at the next time interval. This identity also states for a fixed i, $\{p_{ji}\}$ is a probability distribution.

The one-step transition probabilities can be expressed in matrix form, which is referred to as the transition matrix.

Definition 2.4. The *transition matrix* of the discrete time Markov chain $\{X_n\}_{n=0}^{\infty}$ with state space $\{1, 2, \ldots\}$ and one-step transition probabilities, $\{p_{ij}\}_{i,j=1}^{\infty}$, is denoted as $P = (p_{ij})$, where

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

If the set of states is finite, $\{1, 2, \ldots, N\}$, then P is an $N \times N$ matrix. Note that the column elements sum to one since $\sum_{j=1}^{N} p_{ji} = 1$. A nonnegative matrix with the property that each column sum equals one is called a *stochastic matrix*. The transition matrix P is a stochastic matrix. It is left as an Exercise to show that if P is a stochastic matrix, then P^n is a stochastic matrix, for n any positive integer. If the row sums also equal one, then the matrix is called *doubly stochastic*.

The notation used here differs from that used in some textbooks in two respects. First, the transition matrix is sometimes defined as the transpose of P, P^T. Then the definition of a stochastic matrix is defined as a nonnegative matrix whose row sums equal one (rather than column sums equal one) (Kemeny and Snell, 1960; Norris, 1997; Stewart, 1994). Second, generally, the one-step transition probability p_{ij} is defined as the probability of a transition from state i to state j rather than a transition from j to i as in our notation (Bailey, 1990; Karlin and Taylor, 1975; Kemeny and Snell,

1960; Norris, 1997; Schinazi, 1999; Stewart, 1994; Taylor and Karlin, 1998). We prefer this notation because it closely resembles the notation used in deterministic models and will allow us to more easily relate deterministic models to stochastic models (see Tuljapurkar, 1997). For example, suppose $Y_{n+1} = AY_n$ represents the dynamics of a deterministic system that changes over time. Matrix $A = (a_{ij})$ and $Y = (y_1, y_2, \ldots, y_k)^T$. The term a_{ij} in matrix A represents the effect variable y_j has on y_i, $j \to i$ during the time interval $[n, n+1]$. In addition, using our notation, the element p_{ij} is in the ith row and jth column of the transition matrix P, which is the standard notation used to define matrix elements. As an aid in setting up and understanding how the elements of P are related, note that the nonzero elements in the ith row of P represent all those states j (column j) that can transfer into state i in one time step. Next, we define the n-step transition probabilities.

Definition 2.5. The *n-step transition probability*, denoted $p_{ji}^{(n)}$, is the probability of moving or transferring from state i to state j in n time steps,

$$p_{ji}^{(n)} = \text{Prob}\{X_n = j | X_0 = i\}.$$

The *n-step transition matrix* is denoted as $P^{(n)} = \left(p_{ji}^{(n)}\right)$. For the cases $n = 0$ and $n = 1$, $p_{ji}^{(1)} = p_{ji}$ and

$$p_{ji}^{(0)} = \delta_{ji} = \begin{cases} 1, & j = i, \\ 0, & j \neq i, \end{cases}$$

where δ_{ji} represents the Kronecker delta symbol. Then $P^{(1)} = P$ and $P^{(0)} = I$, where I is the identity matrix.

Relationships exist between the n-step transition probabilities and s-step and $(n-s)$-step transition probabilities. These relationships are known as the *Chapman-Kolmogorov equations*:

$$p_{ji}^{(n)} = \sum_{k=1}^{\infty} p_{jk}^{(n-s)} p_{ki}^{(s)}, \quad 0 < s < n.$$

Verification of the Chapman-Kolmogorov equations can be shown as follows (Stewart, 1994):

$$\begin{aligned} p_{ji}^{(n)} &= \text{Prob}\{X_n = j | X_0 = i\}, \\ &= \sum_{k=1}^{\infty} \text{Prob}\{X_n = j, X_s = k | X_0 = i\}, \qquad (2.2) \\ &= \sum_{k=1}^{\infty} \text{Prob}\{X_n = j | X_s = k, X_0 = i\} \text{Prob}\{X_s = k | X_0 = i\}, \qquad (2.3) \end{aligned}$$

$$= \sum_{k=1}^{\infty} \text{Prob}\{X_n = j | X_s = k\} \text{Prob}\{X_s = k | X_0 = i\}, \tag{2.4}$$

$$= \sum_{k=1}^{\infty} p_{jk}^{(n-s)} p_{ki}^{(s)}, \tag{2.5}$$

where equations (2.2)–(2.5) hold for $0 < s < n$. The relationship (2.3) follows from conditional probabilities (see Exercise 2). The relationship (2.4) follows from the Markov property. The preceding identity written in terms of matrix notation yields

$$P^{(n)} = P^{(n-s)} P^{(s)}. \tag{2.6}$$

However, because $P^{(1)} = P$, it follows from the Chapman-Kolmogorov equations (2.6) that $P^{(2)} = P^2$ and, in general, $P^{(n)} = P^n$. The n-step transition matrix $P^{(n)}$ is just the nth power of P. The elements of P^n are the n-step transition probabilities, $p_{ij}^{(n)}$. Be careful not to confuse the notation p_{ij}^n with $p_{ij}^{(n)}$, $p_{ij}^{(n)} \neq p_{ij}^n$. The notation p_{ij}^n is the nth power of the element p_{ij}, whereas $p_{ij}^{(n)}$ is the ij element in the nth power of P.

Let $p(n)$ denote the vector form of the probability mass function associated with X_n; that is, $p(n) = (p_1(n), p_2(n), \ldots)^T$, where $p_i(n)$ is defined in (2.1) and the states are arranged in increasing order in the column vector $p(n)$. The probabilities satisfy

$$\sum_{i=1}^{\infty} p_i(n) = 1.$$

Given the probability distribution associated with X_n, the probability distribution associated with X_{n+1} can be found by multiplying the transition matrix P by $p(n)$; that is,

$$p_i(n+1) = \sum_{j=1}^{\infty} p_{ij} p_j(n)$$

or

$$p(n+1) = Pp(n).$$

In general,

$$p(n+m) = P^{n+m} p(0) = P^n \left(P^m p(0)\right) = P^n p(m).$$

2.3 Classification of States

Relationships between the states of a Markov chain lead to a classification scheme for the states and ultimately classification for Markov chains.

Figure 2.1. In the directed graph, $i \to j$ ($p_{ji} > 0$) and $i \to k$ ($p_{ki}^{(2)} > 0$), but it is not the case that $k \to i$.

Definition 2.6. The state j can be *reached* from the state i (or state j is *accessible* from state i) if there is a nonzero probability, $p_{ji}^{(n)} > 0$, for some $n \geq 0$. This relationship is denoted as $i \to j$. If i can be reached from j, $j \to i$, and if j can be reached from i, $i \to j$, then i and j are said to *communicate*, or to be in the *same class*, denoted $i \leftrightarrow j$; that is, there exists nonnegative integers n and n' such that

$$p_{ji}^{(n)} > 0 \quad \text{and} \quad p_{ij}^{(n')} > 0.$$

The relation $i \to j$ can be represented in graph theory as a directed graph (see Figure 2.1).

The relation $i \leftrightarrow j$ is an *equivalence relation* on the state space $\{1, 2, \ldots\}$. The relation satisfies the following three properties (Karlin and Taylor, 1975):

(1) reflexivity: $i \leftrightarrow i$, because $p_{ii}^{(0)} = 1$. Beginning in state i, the system stays in state i if there is no time change.

(2) symmetry: $i \leftrightarrow j$ implies $j \leftrightarrow i$ follows from the definition.

(3) transitivity: $i \leftrightarrow j$, $j \leftrightarrow k$ implies $i \leftrightarrow k$. To verify this last property, note that the the first two properties imply there exist nonnegative integers n and m such that $p_{ji}^{(n)} > 0$ and $p_{kj}^{(m)} > 0$. Thus,

$$\begin{aligned} p_{ki}^{(n+m)} &= \text{Prob}\{X_{n+m} = k | X_0 = i\}, \\ &\geq \text{Prob}\{X_{n+m} = k, X_n = j | X_0 = i\}, \\ &= \text{Prob}\{X_{n+m} = k | X_n = j\} \text{Prob}\{X_n = j | X_0 = i\}, \qquad (2.7) \\ &= p_{kj}^{(m)} p_{ji}^{(n)}, \end{aligned}$$

where probability (2.7) follows from conditional probabilities and the Markov property. Thus, $p_{ki}^{(n+m)} > 0$ and $i \to k$. Similarly, it can be shown that $p_{ik}^{(n+m)} > 0$, which implies $k \to i$.

The equivalence relation on the states of the Markov chain define a set of equivalence classes. These equivalence classes are known as classes of the Markov chain.

Definition 2.7. The set of equivalence classes in a discrete time Markov chain are called the *communication classes* or, more simply, the *classes* of the Markov chain.

If every state in the Markov chain can be reached from every other state, then there is only one communication class (all the states are in the same class).

Definition 2.8. If there is only one communication class, then the Markov chain is said to be *irreducible*, but if there is more than one communication class, then the Markov chain is said to be *reducible*.

A communication class may have the additional property that it is closed.

Definition 2.9. A set of states C is said to be *closed* if it is impossible to reach any state outside of C from any state in C by one-step transitions; $p_{ji} = 0$ if $i \in C$ and $j \notin C$.

A sufficient condition that shows that a Markov chain is irreducible is the existence of a positive integer n such that $p_{ji}^{(n)} > 0$ for all i and j; that is, every element in P^n is positive, $P^n > 0$, for some positive integer n. For a finite Markov chain, irreducibility can be checked from the directed graph for that chain. A finite Markov chain with states $\{1, 2, \ldots, N\}$ is irreducible if there is a directed path from i to j for every $i, j \in \{1, 2, \ldots, N\}$.

The definitions of *irreducible* and *reducible* apply more generally to $N \times N$ matrices, $A = (a_{ij})$. A *directed graph* or *digraph* with N nodes can be constructed from an $N \times N$ matrix. There is a single directed path from node i to node j if $a_{ji} \neq 0$. Then node j can be reached from node i in one step. A more general signed digraph can be constructed, where the sign of a_{ji} is associated with each directed path. Node j can be reached from node i in n steps if $a_{ji}^{(n)} \neq 0$, where $a_{ji}^{(n)}$ is the element in the jth row and ith column of A^n. A directed graph with N nodes constructed from a matrix A is said to be *strongly connected* if there exists a series of directed paths from i to j for every $i, j \in \{1, 2 \ldots, N\}$ $(i \leftrightarrow j)$. Then a directed graph is strongly connected if it is possible to start from any node i and reach any other node j in a finite number of steps. Matrix irreducibility is defined as a strongly connected digraph (Ortega, 1987).

Definition 2.10. Matrix A is said to be *irreducible* if and only if its directed graph is strongly connected. Alternately, matrix A is is said to be *reducible* if and only if its directed graph is not strongly connected.

Example 2.1 A discrete time Markov chain with four states $\{1, 2, 3, 4\}$ has the following transition matrix:

$$P = \begin{pmatrix} 0 & 0 & p_{13} & 0 \\ p_{21} & 0 & p_{23} & p_{24} \\ 0 & 0 & 0 & 0 \\ 0 & p_{42} & 0 & 0 \end{pmatrix},$$

where p_{ij} denotes a positive element. Then it is easy to see that $4 \leftrightarrow 2 \leftarrow 1 \leftarrow 3$ and $4 \leftrightarrow 2 \leftarrow 3$ (see Figure 2.2).

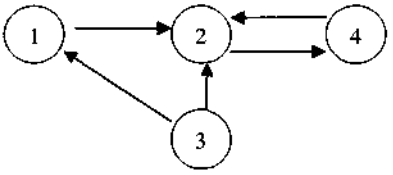

Figure 2.2. Digraph for Example 2.1.

Since it is impossible to return to states 1 or 3 after having left them, each of these states forms a single communication class, $\{1\}$, $\{3\}$. The set $\{2,4\}$ is a third communication class. The Markov chain is reducible. In addition, the set $\{2,4\}$ is closed, but the sets $\{1\}$ and $\{3\}$ are not closed. If one of the elements, either p_{12} or p_{14}, is positive, then the communication classes consist of $\{1,2,4\}$ and $\{3\}$. If one of the elements, either p_{32} or p_{34}, is positive, then there is a single communication class $\{1,2,3,4\}$; the directed graph is strongly connected and matrix P is irreducible. Also, the discrete time Markov chain is irreducible. ■

The following example illustrates the classical gambler's ruin problem.

Example 2.2 The state space is the set $\{0,1,2\ldots,N\}$. The states represent the amount of money of one of the players (gambler). The gambler bets \$1 per game and either wins or loses each game. The gambler is ruined if he/she reaches state 0. The probability of winning (moving to the right) is $p > 0$ and the probability of losing (moving to the left) is $q > 0$, $p + q = 1$ (i.e., $p_{i,i+1} = q$ and $p_{i+1,i} = p$, $i = 1,\ldots,N-1$). In addition, $p_{00} = 1$ and $p_{NN} = 1$, which are referred to as *absorbing boundaries*. All other elements of the transition matrix are zero. In general, a state i is called *absorbing* if $p_{ii} = 1$. See the directed graph in Figure 2.3 and the corresponding $(N+1) \times (N+1)$ transition matrix:

$$P = \begin{pmatrix} 1 & q & 0 & \cdots & 0 & 0 \\ 0 & 0 & q & \cdots & 0 & 0 \\ 0 & p & 0 & \cdots & 0 & 0 \\ 0 & 0 & p & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & p & 1 \end{pmatrix}.$$

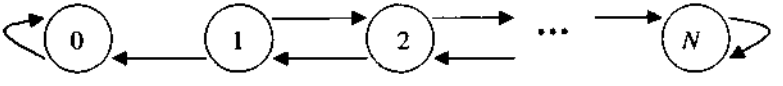

Figure 2.3. The probability of winning is p and losing is q. The boundaries or end states, 0 and N, are absorbing, $p_{00} = 1 = p_{NN}$.

There are three communication classes for the Markov chain graphed in Figure 2.3: $\{0\}$, $\{1, 2, \ldots, N-1\}$, and $\{N\}$. The Markov chain is reducible. The sets $\{0\}$ and $\{N\}$ are closed, but the set $\{1, 2, \ldots, N-1\}$ is not closed. Also, states 0 and N are absorbing; the remaining states are *transient*. A transient state is defined more formally later. This example is also an illustration of a random walk with absorbing boundaries at 0 and N. ■

Example 2.3 In an infinite-dimensional random walk or unrestricted random walk, the states are the integers, $0, \pm 1, \pm 2, \ldots$. Let $p > 0$ be the probability of moving to the right and $q > 0$ be the probability of moving to the left, $p + q = 1$. There are no absorbing boundaries, $p_{i,i+1} = q$ and $p_{i+1,i} = p$ for $i \in \{0, \pm 1, \pm 2, \ldots\}$. From the directed graph in Figure 2.4 it is easy to see that the Markov chain is irreducible. Every state in the system communicates with every other state. The set of states forms a closed set. In this case, the transition matrix P is infinite dimensional. If the states are ordered such that $\ldots, -1, 0, 1, \ldots$, then matrix P is an extension of the matrix in Example 2.2 with q along the superdiagonal and p along the subdiagonal. ■

Example 2.4 Suppose the states of the system are $\{1, 2, 3, 4, 5\}$ with directed graph in Figure 2.5 and transition matrix P given as follows:

$$P = \begin{pmatrix} 1/2 & 1/3 & 0 & 0 & 0 \\ 1/2 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 \\ 0 & 0 & 1 & 1/2 & 1 \\ 0 & 0 & 0 & 1/4 & 0 \end{pmatrix}.$$

There are two communication classes, $\{1, 2\}$ and $\{3, 4, 5\}$. Both classes are closed. The Markov chain is reducible. ■

Figure 2.4. Unrestricted random walk; the probability of moving right is p and left is q.

Figure 2.5. Directed graph for Example 2.4.

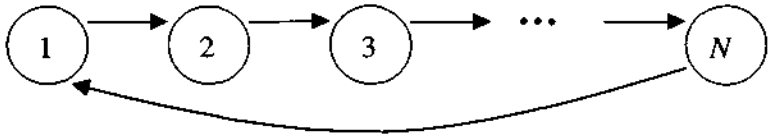

Figure 2.6. Directed graph for Example 2.5.

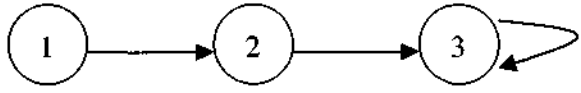

Figure 2.7. Directed graph for Example 2.6.

Example 2.5 Suppose the states of the system are $\{1, 2, \ldots, N\}$ with transition matrix given by P and directed graph in Figure 2.6. For this example, the Markov chain is irreducible. The set $\{1, 2, \ldots, N\}$ is closed.

$$P = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

■

The chain in Example 2.5 has the property that beginning in state i it takes exactly N time steps to return to state i. In addition, $P^N = I$. The chain is periodic with period equal to N.

Definition 2.11. The *period of state* i, denoted as $d(i)$, is the greatest common divisor of all integers $n \geq 1$ for which $p_{ii}^{(n)} > 0$; that is,

$$d(i) = \text{g.c.d}\{n | p_{ii}^{(n)} > 0 \text{ and } n \geq 1\}.$$

If a state i has period $d(i) > 1$, it is said to be *periodic of period* $d(i)$. If the period of a state equals one, it is said to be *aperiodic*. If $p_{ii}^{(n)} = 0$ for all $n \geq 1$, we define $d(i) = 0$.

It follows from the definition that $d(i)$ is a nonnegative integer.

Example 2.6 The directed graph of a Markov chain with three states $\{1, 2, 3\}$ is given in Figure 2.7. The corresponding transition matrix is

$$P = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

It is easy to see that there are three communication classes, $\{1\}$, $\{2\}$, and $\{3\}$. The value of $d(i) = 0$ for $i = 1, 2$ because $p_{ii}^{(n)} = 0$ for $i = 1, 2$ and $n = 1, 2, \ldots$. Also, $d(3) = 1$; state 3 is aperiodic. ■

In the special case $d(i) = 0$, it can be shown that the set $\{i\}$ forms a communication class (see Exercise 5). Also, if $p_{ii} > 0$, then $d(i) = 1$. Generally, the term periodic is reserved for the case $d(i) > 1$.

Example 2.7 In Example 2.1, the classes are $\{1\}$, $\{3\}$ and $\{2,4\}$. States 1 and 3 satisfy $d(1) = 0 = d(3)$. States 2 and 4 are periodic with period 2, $d(2) = 2 = d(4)$; $p_{ii}^{(2n)} = 1$ and $p_{ii}^{(2n+1)} = 0$ for $i = 2, 4$ and $n = 1, 2, \ldots$. ■

Periodicity is a class property; that is, if $i \leftrightarrow j$, then $d(i) = d(j)$. All states in one class have the same period. Thus, we can speak of a periodic class or a periodic chain. This result is verified in the next theorem.

Theorem 2.1. *If $i \leftrightarrow j$, then $d(i) = d(j)$.*

Proof. The case $d(i) = 0$ is trivial. Suppose $d(i) \geq 1$ and $p_{ii}^{(s)} > 0$ for some $s > 0$. Then $d(i)$ divides s. Since $i \leftrightarrow j$, there exists m and n such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$. Then

$$p_{jj}^{(n+s+m)} \geq p_{ji}^{(n)} p_{ii}^{(s)} p_{ij}^{(m)} > 0.$$

Also, since $p_{ii}^{(2s)} > 0$, $p_{jj}^{(n+2s+m)} > 0$. Thus, $d(j)$ divides $n + s + m$ and $n + 2s + m$ and must divide $(n + 2s + m) - (n + s + m) = s$. Since s was arbitrary, $d(j) \leq d(i)$.

Reverse the argument by assuming $p_{jj}^{(r)} > 0$. Then it can be shown that $d(i) \leq d(j)$. Combining these two inequalities gives the desired result, $d(i) = d(j)$. □

In the random walk model with absorbing boundaries, Example 2.2, the classes $\{0\}$ and $\{N\}$ are aperiodic. The class $\{1, 2, \ldots, N-1\}$ has period 2. In the unrestricted random walk model, Example 2.3, the entire chain is periodic of period 2. In this case, we shall use the notation $d = 2$ rather than stating that $d(i) = 2$ for each of the i states. The two classes in Example 2.4 are both aperiodic.

Some additional definitions and notation are needed to define a transient state. This is done in the next section.

2.4 First Passage Time

Assume the process begins in state i, $X_0 = i$. Then we define a first return to state i and a first passage to state j for $j \neq i$.

Definition 2.12. Let $f_{ii}^{(n)}$ denote the probability that, starting from state i, $X_0 = i$, the *first return* to state i is at the nth time step, $n \geq 1$; that is,

$$f_{ii}^{(n)} = \text{Prob}\{X_n = i, X_m \neq i, m = 1, 2, \ldots, n-1 | X_0 = i\}.$$

The probabilities $f_{ii}^{(n)}$ are known as *first return probabilities*. Define $f_{ii}^{(0)} = 0$.

Note that $f_{ii}^{(1)} = p_{ii}$ but, in general, $f_{ii}^{(n)}$ is not equal to $p_{ii}^{(n)}$. The first return probabilities, $f_{ii}^{(n)}$, represent the *first time* the chain returns to state i; thus,

$$0 \leq \sum_{n=1}^{\infty} f_{ii}^{(n)} \leq 1.$$

A transient state is defined in terms of these first return probabilities.

Definition 2.13. State i is said to be *transient* if $\sum_{n=1}^{\infty} f_{ii}^{(n)} < 1$. State i is said to be *recurrent* if $\sum_{n=1}^{\infty} f_{ii}^{(n)} = 1$.

The term *persistent* is sometimes used instead of recurrent (Bailey, 1990). If state i is recurrent, then the set $\{f_{ii}^{(n)}\}_{n=0}^{\infty}$ defines a probability distribution for the random variable representing the first return time, which is

$$T_{ii} = \inf_{m \geq 1} \{m | X_m = i \text{ and } X_0 = i\};$$

that is, $T_{ii} = n$ with probability $f_{ii}^{(n)}$, $n = 0, 1, 2, \ldots$. When state i is transient, $\{f_{ii}^{(n)}\}_{n=0}^{\infty}$ does not define a complete set of probabilities necessary to define a probability distribution. However, if $f_{ii} = \sum_{n=0}^{\infty} f_{ii}^{(n)} < 1$, then we can define $1 - f_{ii}$ as the probability of never returning to state i. The random variable T_{ii} may be thought of as a "waiting time" until the chain returns to state i.

Definition 2.14. The mean of the distribution of T_{ii} is referred to as the *mean recurrence time* or *mean first return time* for state i and is denoted as $\mu_{ii} = E(T_{ii})$. For a recurrent state i,

$$\mu_{ii} = \sum_{n=1}^{\infty} n f_{ii}^{(n)}. \tag{2.8}$$

Although T_{ii} is not defined for a transient state, we shall assume the mean recurrence time for a transient state is always infinite (formally $T_{ii} = \infty$ with probability $1 - f_{ii}$). The mean recurrence time for a recurrent state can be either finite or infinite.

Definition 2.15. If a recurrent state i satisfies $\mu_{ii} < \infty$, then it is said to be *positive recurrent*, and if it satisfies $\mu_{ii} = \infty$, then it is said to be *null recurrent*.

Sometimes the term *nonnull recurrent* is used instead of positive recurrent (Bailey, 1990).

Example 2.8 A simple example of a positive recurrent state is an absorbing state. If i is an absorbing state, then $p_{ii} = 1$, so that $f_{ii}^{(1)} = p_{ii} = 1$ and $f_{ii}^{(n)} = 0$ for $n \neq 1$. The mean recurrence time of an absorbing state is $\mu_{ii} = 1$. ■

Example 2.9 Suppose the transition matrix of a two-state Markov chain satisfies

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix},$$

where $0 < p_{ii} < 1$ for $i = 1, 2$. Then all of the elements of matrix P are positive, $p_{ij} > 0$, $i, j = 1, 2$. Hence, $f_{11}^{(1)} = p_{11}$, $f_{11}^{(2)} = p_{12}p_{21}$, $f_{11}^{(3)} = p_{12}p_{22}p_{21}$, and, in general,

$$f_{11}^{(n)} = p_{12}p_{22}^{n-2}p_{21}, \quad n \geq 3.$$

This can be verified easily from the directed graph. Because $p_{22} < 1$, it follows that $\lim_{n\to\infty} f_{11}^{(n)} = 0$ and that

$$\sum_{n=1}^{\infty} f_{11}^{(n)} = p_{11} + p_{12}p_{21} \sum_{n=0}^{\infty} p_{22}^{n} = p_{11} + \frac{p_{12}p_{21}}{1 - p_{22}}.$$

Next, applying the definition of a stochastic matrix, $p_{11} + p_{21} = 1$ and $p_{12} + p_{22} = 1$, it follows that

$$\sum_{n=1}^{\infty} f_{11}^{(n)} = p_{11} + p_{21} = 1,$$

which implies that state 1 is recurrent. Similarly, it can be shown that state 2 is recurrent. In addition, it can be shown that the mean recurrence times are finite; for example,

$$\mu_{11} = p_{11} + p_{12}p_{21} \sum_{n=0}^{\infty} (n+2)p_{22}^{n} < \infty$$

(see Exercise 6). Therefore, the Markov chain is positive recurrent. ■

Note that in the definitions of first return probabilities and mean recurrence time, the Markov property was not assumed. These concepts do not require the Markov assumption and are sometimes discussed in the context of renewal processes. These definitions are extended to first passage time probabilities and mean first passage time. Then they are related to Markov chains.

Define the probability $f_{ji}^{(n)}$ for $j \neq i$ in a manner analogous to $f_{ii}^{(n)}$.

Definition 2.16. Let $f_{ji}^{(n)}$ denote the probability that, starting from state i, $X_0 = i$, the *first* return to state j, $j \neq i$ is at the nth time step, $n \geq 1$,

$$f_{ji}^{(n)} = \text{Prob}\{X_n = j, X_m \neq j, m = 1, 2, \ldots, n-1 | X_0 = i\}, \quad j \neq i.$$

The probabilities $f_{ji}^{(n)}$ are known as *first passage time probabilities.* Define $f_{ji}^{(0)} = 0$.

It follows from the definition that $0 \leq \sum_{n=0}^{\infty} f_{ji}^{(n)} \leq 1$. If $\sum_{n=0}^{\infty} f_{ji}^{(n)} = 1$, $\{f_{ji}^{(n)}\}_{n=0}^{\infty}$ defines a probability distribution for a random variable $T_{ji} = \inf_{m \geq 1}\{m | X_m = j \text{ and } X_0 = i\}$ known as the first passage to state j from state i. If $i = j$, then Definition 2.16 is the same as Definition 2.12.

Definition 2.17. If $X_0 = i$, then the *mean first passage time* to state j is denoted as $\mu_{ji} = E(T_{ji})$ and defined as

$$\mu_{ji} = \sum_{n=1}^{\infty} n f_{ji}^{(n)}, \quad j \neq i.$$

This definition can be extended to include the case $f_{ji} = \sum_{n=0}^{\infty} f_{ji}^{(n)} < 1$ by defining the probability of never reaching state j from state i as $1 - f_{ji}$. If $f_{ji} < 1$, then the mean first passage time is infinite.

There exists relationships between the n-step transition probabilities of a Markov chain and the first return probabilities. The transition from state i to i at the nth step, $p_{ii}^{(n)}$, may have its first return to state i at any of the steps $j = 1, 2, \ldots, n$. It is easy to see that

$$\begin{aligned} p_{ii}^{(n)} &= f_{ii}^{(0)} p_{ii}^{(n)} + f_{ii}^{(1)} p_{ii}^{(n-1)} + \cdots + f_{ii}^{(n)} p_{ii}^{(0)} \\ &= \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)}, \end{aligned} \tag{2.9}$$

since $f_{ii}^{(0)} = 0$ and $p_{ii}^{(0)} = 1$. A similar relationship exists for $f_{ji}^{(n)}$ and $p_{ji}^{(n)}$:

$$p_{ji}^{(n)} = \sum_{k=1}^{n} f_{ji}^{(k)} p_{jj}^{(n-k)}, \quad j \neq i. \tag{2.10}$$

Let the generating function for the sequence $\{f_{ji}^{(n)}\}$ be

$$F_{ji}(s) = \sum_{n=0}^{\infty} f_{ji}^{(n)} s^n, \quad |s| < 1$$

and the generating function for the sequence $\{p_{ji}^{(n)}\}$ be

$$P_{ji}(s) = \sum_{n=0}^{\infty} p_{ji}^{(n)} s^n, \quad |s| < 1$$

for all states i, j of the Markov chain. Note that these functions may not be probability generating functions since the set of probabilities $\{f_{ji}^{(n)}\}_{n=0}^{\infty}$ and $\{p_{ji}^{(n)}\}_{n=0}^{\infty}$ may not represent a probability distribution (the sum may be less than one). However, relationships between these two generating functions are shown and these relationships are used to prove results about Markov chains.

Multiply $F_{ii}(s)$ by $P_{ii}(s)$ using the definition for the product of two series. The product $C(s)$ of two series $A(s)$ and $B(s)$, where $A(s) = \sum_0^{\infty} a_k s^k$ and $B(s) = \sum_0^{\infty} b_l s^l$, is

$$C(s) = A(s)B(s) = \sum_{r=0}^{\infty} c_r s^r,$$

where

$$c_r = a_0 b_r + a_1 b_{r-1} + \cdots + a_r b_0 = \sum_{k=0}^{r} a_k b_{r-k}.$$

If $A(s)$ and $B(s)$ converge on the interval $(-1, 1)$, then $C(s)$ also converges on $(-1, 1)$ (Wade, 2000). Identify the coefficient a_k of $A(s)$ with $f_{ii}^{(k)}$ of $F_{ii}(s)$ and the coefficient b_l of $B(s)$ with $p_{ii}^{(l)}$ of $P_{ii}(s)$ and apply equation (2.9) so that $c_r = p_{ii}^{(r)}$. The following relationship between the generating functions is obtained:

$$F_{ii}(s)P_{ii}(s) = \sum_{r=1}^{\infty} p_{ii}^{(r)} s^r = P_{ii}(s) - 1,$$

where $p_{ii}^{(r)} = \sum_{k=1}^{r} f_{ii}^{(k)} p_{ii}^{(r-k)}$ and $|s| < 1$. Note that the number one is subtracted from $P_{ii}(s)$ since the first term $c_0 = f_{ii}^{(0)} p_{ii}^{(0)}$ in the product of $F_{ii}(s)P_{ii}(s)$ is zero but the first term in $P_{ii}(s)$ is $p_{ii}^{(0)} = 1$. Hence,

$$P_{ii}(s) = \frac{1}{1 - F_{ii}(s)}. \tag{2.11}$$

A similar relationship exists between $P_{ji}(s)$ and $F_{ji}(s)$ that follows from (2.10):

$$F_{ji}(s)P_{jj}(s) = P_{ji}(s), \quad i \neq j. \tag{2.12}$$

For equation (2.12), the number one is not subtracted from $P_{ji}(s)$ since the first term in its series representation is $p_{ji}^{(0)} = 0$, $i \neq j$, which equals the first term in the series representation of $F_{ji}(s)P_{jj}(s)$; that is, $c_0 = f_{ji}^{(0)} p_{ji}^{(0)} = 0$. The above identities are used to verify some theoretical results on Markov chains.

2.5 Basic Theorems for Markov Chains

The relationships between the generating functions are used to relate a recurrent state i to the n-step transition probabilities $p_{ii}^{(n)}$. In addition, an important result referred to as the basic limit theorem for Markov chains is stated that gives conditions for a Markov chain to have a limiting distribution. First, a lemma is needed.

Lemma 2.1 (Abel's Convergence Theorem).

(i) If $\sum_{k=0}^{\infty} a_k$ *converges, then* $\lim_{s\to 1^-} \sum_{k=0}^{\infty} a_k s^k = \sum_{k=0}^{\infty} a_k = a.$

(ii) If $a_k \geq 0$ *and* $\lim_{s\to 1^-} \sum_{k=0}^{\infty} a_k s^k = a \leq \infty$, *then* $\sum_{k=0}^{\infty} a_k = a.$

For a proof of Abel's convergence theorem, consult Karlin and Taylor (1975, pp. 64–65). The lemma is straightforward if the series converges absolutely. Lemma 2.1 is used to verify the following theorem.

Theorem 2.2. *A state i is recurrent (transient) if and only if* $\sum_{n=0}^{\infty} p_{ii}^{(n)}$ *diverges (converges); that is,*

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty \ (< \infty).$$

Proof. We prove the theorem in the case of a recurrent state. The proof in the case of a transient state follows as a direct consequence because if a state i is not recurrent it is transient. Assume state i is recurrent; that is,

$$\sum_{n=1}^{\infty} f_{ii}^{(n)} = 1.$$

By part (i) of Lemma 2.1,

$$\lim_{s\to 1^-} \sum_{n=1}^{\infty} f_{ii}^{(n)} s^n = \lim_{s\to 1^-} F_{ii}(s) = 1.$$

From the identity (2.11), it follows that

$$\lim_{s\to 1^-} P_{ii}(s) = \lim_{s\to 1^-} \frac{1}{1 - F_{ii}(s)} = \infty.$$

Because $P_{ii}(s) = \sum_{n=0}^{\infty} p_{ii}^{(n)} s^n$, it follows from Lemma 2.1 part (ii) that

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty.$$

The converse of the theorem is proved by contradiction. Assume that $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$ and state i is transient; that is,

$$\sum_{n=1}^{\infty} f_{ii}^{(n)} < 1.$$

Applying Lemma 2.1 part (i),

$$\lim_{s\to 1^-} F_{ii}(s) = \lim_{s\to 1^-} \sum_{n=0}^{\infty} f_{ii}^{(n)} s^n = \sum_{n=0}^{\infty} f_{ii}^{(n)} = \sum_{n=1}^{\infty} f_{ii}^{(n)} < 1.$$

Now, applying the identity (2.11), it follows that

$$\lim_{s\to 1^-} P_{ii}(s) = \lim_{s\to 1^-} \frac{1}{1 - F_{ii}(s)} < \infty.$$

Finally, Lemma 2.1 part (ii) and the above inequality yield

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} < \infty,$$

which contradicts the original assumption. The theorem is verified. □

Next it is shown that if state i is recurrent, then all states in the same communicating class are recurrent. Thus, recurrence and transience are class properties. If the chain is irreducible, then the chain is either recurrent or transient.

Corollary 2.1. *Assume $i \leftrightarrow j$. State i is recurrent (transient) if and only if state j is recurrent (transient).*

Proof. Suppose $i \leftrightarrow j$ and state i is recurrent. Then there exists $n, m \geq 1$ such that

$$p_{ij}^{(n)} > 0 \quad \text{and} \quad p_{ji}^{(m)} > 0.$$

Let k be a nonnegative integer,

$$p_{jj}^{(m+n+k)} \geq p_{ji}^{(m)} p_{ii}^{(k)} p_{ij}^{(n)}.$$

Summing on k,

$$\sum_{k=0}^{\infty} p_{jj}^{(k)} \geq \sum_{k=0}^{\infty} p_{jj}^{(n+m+k)} \geq \sum_{k=0}^{\infty} p_{ji}^{(m)} p_{ii}^{(k)} p_{ij}^{(n)} = p_{ji}^{(m)} p_{ij}^{(n)} \sum_{k=0}^{\infty} p_{ii}^{(k)}.$$

The right-hand side is infinite by Theorem 2.2 because state i is recurrent. Thus, $\sum_{k=0}^{\infty} p_{jj}^{(k)}$ is divergent and state j is recurrent. The theorem also holds for transient states because if a state is not recurrent it is transient. □

An important property about recurrent classes follows from the definition of a recurrent state. The next corollary shows that a recurrent class forms a closed set.

Corollary 2.2. *Every recurrent class in a discrete time Markov chain is a closed set.*

Proof. Let C be a recurrent class. Suppose C is not closed. Then for some $i \in C$ and $j \notin C$, $p_{ji} > 0$. Because $j \notin C$, it is impossible to return to the set C from state j (otherwise $i \leftrightarrow j$). Thus, beginning from state i, the probability of never returning to C is at least p_{ji} or $\sum_n f_{ii}^{(n)} \leq 1 - p_{ji} < 1$, a contradiction to the fact that i is a recurrent state. Hence, C must be closed. □

Example 2.10 Suppose the transition matrix of a Markov chain with states $\{1, 2, 3, \ldots\}$ satisfies

$$P = \begin{pmatrix} a_1 & 0 & 0 & \cdots \\ a_2 & a_1 & 0 & \cdots \\ a_3 & a_2 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix},$$

where $a_i > 0$ and $\sum_{i=1}^{\infty} a_i = 1$. The communication classes consist of $\{1\}$, $\{2\}$, $\{3\}$, and so on. Each state represents a communication class. In addition, none of the classes are closed. Hence, by Corollary 2.2, it follows that none of the classes are recurrent, they must all be transient. In fact, each class is aperiodic and transient. ■

We will use Example 2.3, the one-dimensional unrestricted random walk, to illustrate Theorem 2.2. It will be shown that the Markov chain for this process is recurrent if and only if (iff) the probabilities of moving right or left are equal, $p = 1/2 = q$, which means it is a symmetric random walk.

Example 2.11 Consider the one-dimensional, unrestricted random walk in Example 2.3. The chain is irreducible and periodic of period 2. Let p be the probability of moving to the right and q be the probability of moving left, $p + q = 1$. We verify that the state 0 or the origin is recurrent iff $p = 1/2 = q$. However, if the origin is recurrent, then all states are recurrent because the chain is irreducible. Notice that starting from the origin, it is impossible to return to the origin in an odd number of steps,

$$p_{00}^{(2n+1)} = 0 \quad \text{for} \quad n = 0, 1, 2, \ldots.$$

The chain has period 2 because only in an even numbers of steps is the transition probability positive. In $2n$ steps, there are a total of n steps to the right and a total of n steps to the left, and the n steps to the left must

be the reverse of those steps taken to the right in order to return to the origin. In particular, in $2n$ steps, there are

$$\binom{2n}{n} = \frac{(2n)!}{n!n!}$$

different paths (combinations) that begin and end at the origin. Also, the probability of occurrence of each one of these paths is $p^n q^n$. Thus,

$$\sum_{n=0}^{\infty} p_{00}^{(n)} = \sum_{n=0}^{\infty} p_{00}^{(2n)} = \sum_{n=0}^{\infty} \binom{2n}{n} p^n q^n.$$

We need an asymptotic formula for $n!$ known as Stirling's formula to verify recurrence. The notation $f(n) \sim g(n)$ means

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1.$$

Both quantities, $f(n)$ and $g(n)$, grow at the same rate as $n \to \infty$. In Stirling's formula,

$$\lim_{n\to\infty} \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = 1.$$

Thus, *Stirling's formula* is the following asymptotic relation:

$$\boxed{n! \sim n^n e^{-n}\sqrt{2\pi n}.}$$

Verification of Stirling's formula can be found in Feller (1968) or Norris (1997).

Stirling's formula gives the following approximation:

$$\begin{aligned} p_{00}^{(2n)} &= \frac{(2n)!}{n!n!} p^n q^n \\ &\sim \frac{\sqrt{4\pi n}(2n)^{2n} e^{-2n}}{2\pi n^{2n+1} e^{-2n}} p^n q^n \\ &= \frac{(4pq)^n}{\sqrt{\pi n}}. \end{aligned} \tag{2.13}$$

Thus, there exists a positive integer N such that for $n \geq N$,

$$\frac{(4pq)^n}{2\sqrt{\pi n}} < p_{00}^{(2n)} < \frac{2(4pq)^n}{\sqrt{\pi n}}.$$

Considered as a function of p, the expression $4pq = 4p(1-p)$ has a maximum at $p = 1/2$. If $p = 1/2$, then $4pq = 1$ and if $p \neq 1/2$, then $4pq < 1$. When $pq \neq 1/4$, then

$$\sum_{n=0}^{\infty} p_{00}^{(2n)} < N + \sum_{n=N}^{\infty} \frac{2(4pq)^n}{\sqrt{\pi n}} < \infty.$$

The latter series converges by the ratio test. When $pq = 1/4$, we have

$$\sum_{n=0}^{\infty} p_{00}^{(2n)} > \sum_{n=N}^{\infty} \frac{(4pq)^n}{2\sqrt{\pi n}} = \frac{1}{2\sqrt{\pi}} \sum_{n=N}^{\infty} \frac{1}{\sqrt{n}} = \infty.$$

The latter series diverges because it is just a multiple of a divergent p-series. Therefore, the series $\sum_{n=0}^{\infty} p_{00}^{(2n)}$ diverges iff $p = 1/2 = q$, which means the one-dimensional random walk is recurrent iff it is a symmetric random walk. If $p \neq q$, then all states are transient; there is a positive probability that an object starting from the origin will never return to the origin. What happens if an object never returns to the origin? It can be shown that either the object tends to $+\infty$ or to $-\infty$. ■

Before giving additional examples, some important results concerning irreducible and recurrent Markov chains are stated. Verification of these results are quite lengthy and the proofs are not given. They depend on a result from renewal theory known as the discrete renewal theorem. A statement of the discrete renewal theorem, its proof, and proofs of the following theorems can be found in Karlin and Taylor (1975) (also consult Norris, 1997). The first result is known as the basic limit theorem for aperiodic Markov chains, which gives conditions for recurrent, irreducible, and aperiodic Markov chains to have a limiting probability distribution. The second result applies to periodic Markov chains.

Theorem 2.3 (Basic limit theorem for aperiodic Markov chains). *Consider a recurrent, irreducible, and aperiodic Markov chain. Then*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_{ii}},$$

where μ_{ii} is the mean recurrence time for state i defined by (2.8) and i and j are any states of the chain. [If $\mu_{ii} = \infty$, then $\lim_{n\to\infty} p_{ij}^{(n)} = 0$.]

Theorem 2.4 (Basic limit theorem for periodic Markov chains). *Consider a recurrent, irreducible, and d-periodic Markov chain. Then*

$$\lim_{n\to\infty} p_{ii}^{(nd)} = \frac{d}{\mu_{ii}}$$

and $p_{ii}^{(m)} = 0$ if m is not a multiple of d, where μ_{ii} is the mean recurrence time for state i defined by (2.8). [If $\mu_{ii} = \infty$, then $\lim_{n\to\infty} p_{ii}^{(nd)} = 0$.]

In Theorem 2.4, $d \geq 1$, since the chain is irreducible (see Exercise 5).

Example 2.12 The Markov chain in Example 2.5 is periodic with period $d = N$. The Markov chain is also irreducible and recurrent. Applying Theorem 2.4, $\lim_{n\to\infty} p_{ii}^{(nN)} = N/\mu_{ii}$. However, $P^N = I$, so that $p_{ii}^{(nN)} = 1$,

and if $i \neq j$, $p_{ij}^{(nN)} = 0$ for $n = 1, 2, \ldots$. Therefore, $1 = N/\mu_{ii}$, which implies that the mean recurrence time is N. This result is obvious if one notices that $f_{ii}^{(N)} = 1$ and $f_{ii}^{(n)} = 0$ for $n \neq N$. Then $\mu_{ii} = N f_{ii}^{(N)} = N$. ■

Theorems 2.3 and 2.4 apply to recurrent classes as well as to recurrent chains. Suppose C is a recurrent communication class. Then since C is closed, $p_{ki}^{(n)} = 0$ for $i \in C$ and $k \notin C$ for $n \geq 1$. Therefore, the submatrix P_C of P given by $P_C = (p_{ij})_{i,j \in C}$ is a transition matrix for C. By Corollary 2.1, the associated Markov chain for C is irreducible and recurrent. Therefore, Theorems 2.3 and 2.4 can be applied to any aperiodic, recurrent class (rather than a chain) or to any periodic, recurrent class. We state the extension of Theorem 2.3 as a corollary.

Corollary 2.3. *If i and j are any states in a recurrent and aperiodic class of a Markov chain, then*

$$\lim_{n \to \infty} p_{ij}^{(n)} = \frac{1}{\mu_{ii}},$$

where μ_{ii} is defined in (2.8).

Example 2.13 The Markov chain in Example 2.4 has two aperiodic, recurrent classes, $\{1, 2\}$ and $\{3, 4, 5\}$. Then $\lim_{n \to \infty} p_{ij}^{(n)} = 1/\mu_{ii}$ for $i, j = 1, 2$ and for $i, j = 3, 4, 5$. Note that $\lim_{n \to \infty} p_{ij}^{(n)} = 0$ when $i \in \{1, 2\}$ and $j \in \{3, 4, 5\}$ or when $i \in \{3, 4, 5\}$ and $j \in \{1, 2\}$. We shall show in the next section how to compute the limit, $1/\mu_{ii}$. ■

If $\mu_{ii} = \infty$, then state i is null recurrent and if $0 < \mu_{ii} < \infty$, then state i is positive recurrent. It can be shown that if one state is positive recurrent in a communication class, then all states in that class are positive recurrent. In this case, the entire class is positive recurrent. In addition, it follows that if one state is null recurrent in a communication class, then all states are null recurrent. Verification of these results is left as an exercise (see Exercise 11). Hence, null recurrence and positive recurrence are class properties. Therefore, it follows from the previous results that every irreducible Markov chain can be classified as either periodic or aperiodic and as either transient, null recurrent, or positive recurrent:

(1) periodic or (2) aperiodic.

(i) transient or (ii) null recurrent or (iii) positive recurrent.

The classifications (1) and (2) are disjoint, and the three classifications, (i), (ii), and (iii), are disjoint. This classification scheme can be applied to communication classes as well, provided the period $d \geq 1$. The special case where $d = 0$ consists of a class with only a single element and the class must be transient (see Example 2.6 and Exercise 5). The term ergodic is used to classify states or irreducible chains that are aperiodic and positive recurrent.

Definition 2.18. A state is *ergodic* if it is both aperiodic and positive recurrent. An *ergodic chain* is a Markov chain that is irreducible, aperiodic, and positive recurrent.

When the entire class or chain is ergodic, it is also referred to as *strongly ergodic* (Karlin and Taylor, 1975). If the ergodic class or chain is null recurrent rather than positive recurrent, then it is said to be *weakly ergodic* (Karlin and Taylor, 1975).

The next example reconsiders the unrestricted random walk model. It has already been shown that this Markov chain is irreducible and periodic. In the case of a symmetric random walk, it is shown that the chain is null recurrent.

Example 2.14 The unrestricted random walk model is irreducible and periodic with period $d = 2$. The chain is recurrent iff it is a symmetric random walk, $p = 1/2 = q$ (Example 2.11). Recall that the $2n$-step transition probability satisfies

$$p_{00}^{(2n)} \sim \frac{1}{\sqrt{\pi n}}$$

[see equation (2.13)] and hence, $\lim_{n\to\infty} p_{00}^{(2n)} = 0$. It follows from the basic limit theorem for periodic Markov chains that $d/\mu_{00} = 0$. Thus, $\mu_{00} = \infty$; the zero state is null recurrent. Since the chain is irreducible, all states are null recurrent. Thus, when $p = 1/2 = q$, the chain is periodic and null recurrent and when $p \neq 1/2$, the chain is periodic and transient. ■

2.6 Stationary Probability Distribution

A stationary probability distribution represents an "equilibrium" of the Markov chain; that is, a probability distribution that remains fixed in time. For instance, if the chain is initially at a stationary probability distribution, $p(0) = \pi$, then $p(n) = P^n\pi = \pi$ for all time n.

Definition 2.19. A *stationary probability distribution* of a Markov chain with states $\{1, 2, \dots\}$ is a nonnegative vector $\pi = (\pi_1, \pi_2, \dots)^T$ that satisfies $P\pi = \pi$ and whose elements sum to one (i.e., $\sum_{i=1}^{\infty} \pi_i = 1$).

Definition 2.19 also applies to a finite Markov chain, where the vector $\pi = (\pi_1, \pi_2, \dots, \pi_N)^T$ and $\sum_{i=1}^{N} \pi_i = 1$. In the finite case, π is a right eigenvector of P corresponding to the eigenvalue $\lambda = 1$, $P\pi = \lambda\pi$. There may be one or more than one linearly independent eigenvector corresponding to the eigenvalue $\lambda = 1$. In fact, there may be at most N linearly independent eigenvectors. If there is more than one linearly independent eigenvector corresponding to $\lambda = 1$, then the stationary probability distribution of the finite Markov chain is not unique.

Example 2.15 Suppose we have the transition matrix

$$P = \begin{pmatrix} p & 0 & q \\ q & p & 0 \\ 0 & q & p \end{pmatrix},$$

where $p > 0$, $q > 0$ and $p + q = 1$. To determine π, we solve $P\pi = \pi$ or

$$(P - I)\pi = \mathbf{0},$$

where I is the 3×3 identity matrix and $\mathbf{0}$ is the zero vector. There is only one linearly independent eigenvector corresponding to the eigenvalue $\lambda = 1$. The unique stationary probability distribution satisfies $\pi_1 = \pi_2 = \pi_3$ so that $\pi = (1/3, 1/3, 1/3)^T$. ■

Example 2.16 If the transition matrix P is the $N \times N$ identity matrix, then there exist N linearly independent eigenvectors, $e_1 = (1, 0, \ldots, 0)^T, \ldots, e_N = (0, 0, \ldots, 1)^T$, corresponding to the eigenvalue $\lambda = 1$. Hence, there is an infinite number of stationary probability distributions. Any vector $\pi = (\pi_1, \pi_2, \ldots, \pi_N)^T$, where $\pi_i \geq 0$ for $i = 1, 2, \ldots, N$, and $\sum_{i=1}^{N} \pi_i = 1$ is a stationary probability distribution. ■

A stationary probability distribution for a finite state Markov chain always exists; although it may not be unique. This is due to the fact that a finite stochastic matrix always has an eigenvalue $\lambda = 1$. However, if the state space is infinite, a stationary probability distribution may not exist.

Example 2.17 Consider the transition matrix in Example 2.10,

$$P = \begin{pmatrix} a_1 & 0 & 0 & \cdots \\ a_2 & a_1 & 0 & \cdots \\ a_3 & a_2 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix},$$

where $a_i > 0$ and $\sum_{i=1}^{\infty} a_i = 1$. There exists no stationary probability distribution because $P\pi = \pi$ implies $\pi = \mathbf{0}$, the zero vector. It is impossible for the sum of the elements of π to equal one. ■

As illustrated in the previous examples, nonexistence of a stationary distribution only applies to infinite Markov chains. It can be shown that every finite Markov chain has at least one stationary probability distribution (Gantmacher, 1964). In addition, if the finite Markov chain is irreducible, it has a unique stationary probability distribution (Ortega, 1987). The Markov chain in Example 2.15 is irreducible, but the one in Example 2.16 is reducible.

If a Markov chain is irreducible, positive recurrent, and aperiodic, then the next theorem shows that there exists a unique stationary probability

distribution and, in addition, this distribution is the limiting distribution of the Markov chain. It is the property of aperiodicity that is needed for convergence to the stationary probability distribution. A proof of this result can be found in Karlin and Taylor (1975).

Theorem 2.5. *Suppose a discrete time Markov chain is irreducible, positive recurrent, and aperiodic (strongly ergodic) with states* $\{1, 2, \ldots\}$ *and transition matrix* P. *Then there exists a unique positive stationary probability distribution* $\pi = (\pi_1, \pi_2, \ldots)^T$, $P\pi = \pi$, *such that*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_i \quad for \quad i, j = 1, 2, \ldots.$$

Theorem 2.5 gives sufficient conditions on the chain for existence and uniqueness of the limiting probability distribution. The transition matrix of a strongly ergodic chain satisfies

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi_1 & \pi_1 & \pi_1 & \cdots \\ \pi_2 & \pi_2 & \pi_2 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

Thus,

$$\lim_{n\to\infty} P^n p(0) = \pi. \tag{2.14}$$

The limit is independent of the initial distribution and equals the stationary probability distribution. The convergence to a stationary probability distribution is similar to convergence to a stable equilibrium in a deterministic model. Theorem 2.5 applies to finite and infinite Markov chains. The following example shows that a unique stationary probability distribution may exist but that the Markov chain may not converge to that distribution.

Example 2.18 Suppose the transition matrix of Markov chain satisfies

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This chain is irreducible, positive recurrent, and periodic of period 2. There exists a unique stationary distribution, $\pi = (1/2, 1/2)^T$, but there is no limiting distribution. For any initial distribution $p(0)$, $P^{2n}p(0) = p(0)$ and $P^{2n+1}p(0) = p(1)$. This example shows the necessity of aperiodicity in Theorem 2.5. ■

Comparing Theorem 2.5 with the basic limit theorem for aperiodic Markov chains, it follows that

$$\pi_i = \frac{1}{\mu_{ii}} > 0,$$

where μ_{ii} is the mean recurrence time for state i. The mean recurrence time of a positive recurrent, irreducible, and aperiodic chain can be computed from the stationary probability distribution, $\mu_{ii} = 1/\pi_i$.

Example 2.19 Suppose the transition matrices for two Markov chains are

$$P_1 = \begin{pmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{pmatrix} \quad \text{and} \quad P_2 = \begin{pmatrix} 0 & 1/4 & 0 \\ 1 & 1/2 & 1 \\ 0 & 1/4 & 0 \end{pmatrix}.$$

These Markov chains are strongly ergodic. (In the next section, it is shown that all irreducible finite Markov chains are positive recurrent.) These two Markov chains represent the two recurrent classes in the Markov chain discussed in Examples 2.4 and 2.13.

There exist unique limiting stationary probability distributions π for each matrix P_i. The stationary distribution corresponding to P_1 satisfies $\pi_1 = (2/3)\pi_2$ and $1 = \pi_1 + \pi_2 = (2/3)\pi_2 + \pi_2$, so that $\pi = (2/5, 3/5)^T$. In addition, the mean recurrence times are $\mu_{11} = 5/2$ and $\mu_{22} = 5/3$. For matrix P_2, it can be shown that the stationary probability distribution is $\pi = (1/6, 2/3, 1/6)^T$. Hence, the mean recurrence times are $\mu_{11} = 6$, $\mu_{22} = 3/2$, and $\mu_{33} = 6$.

It follows from Theorem 2.5 that the columns of P_i^n approach the stationary probability distribution,

$$\lim_{n\to\infty} P_1^n = \begin{pmatrix} 2/5 & 2/5 \\ 3/5 & 3/5 \end{pmatrix} \quad \text{and} \quad \lim_{n\to\infty} P_2^n = \begin{pmatrix} 1/6 & 1/6 & 1/6 \\ 2/3 & 2/3 & 2/3 \\ 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

For example, in the first Markov chain, eventually, 40% of the time is spent in state 1 and 60% of the time is spent in state 2. After leaving state 1, it takes on the average about 2.5 time steps until there is a return to state 1, and, after leaving state 2, it takes about 1.67 time steps until there is a return to state 2. ■

2.7 Finite Markov Chains

An important property of finite Markov chains is that there are no null recurrent states and not all states can be transient. Therefore, an irreducible, finite Markov chain is positive recurrent. The assumption of recurrence is not needed when the basic limit theorems are applied to finite Markov chains. To verify these results, we begin with a lemma.

Lemma 2.2. *If j is a transient state of a Markov chain and i is any state in the Markov chain, then* $\lim_{n\to\infty} p_{ji}^{(n)} = 0$.

Actually, Lemma 2.2 applies to finite and infinite Markov chains. The proof is left as an exercise.

Theorem 2.6. *In a finite Markov chain, not all states can be transient and no states can be null recurrent. In particular, an irreducible finite Markov chain is positive recurrent.*

Proof. From Lemma 2.2, it follows that if j is transient, then

$$\lim_{n\to\infty} p_{ji}^{(n)} = 0, \quad \text{for } i = 1, 2, \ldots, N, \tag{2.15}$$

where N is the number of states. Suppose all states are transient. Then the identity (2.15) holds for all i and j, $i, j = 1, 2, \ldots, N$ and

$$\lim_{n\to\infty} P^n = \mathbf{0},$$

the zero matrix. Matrix P^n is a stochastic matrix (Exercise 1). Hence, $\sum_{j=1}^{N} p_{ji}^{(n)} = 1$; the column sums of P^n are one. Taking the limit as $n \to \infty$ and interchanging the limit and summation (possible because of the finite sum) leads to $\sum_{j=1}^{N} \lim_{n\to\infty} p_{ji}^{(n)} = \mathbf{1}$, a contradiction to the above limit. Thus, not all states can be transient.

Suppose there exists a null recurrent state i and $i \in C$, where C is a class of states. The class C is closed by Corollary 2.2 and all states in C are null recurrent. (See the remarks following Corollary 2.3 and Exercise 11.)

Suppose the class C is aperiodic and null recurrent. Then according to the basic limit theorem for aperiodic Markov chains, $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ for all $i, j \in C$. The submatrix P_C of P consisting of all states in C is a stochastic matrix ($p_{kj}^{(n)} = 0$ for $k \notin C$). But $\lim_{n\to\infty} P_C^n = \mathbf{0}$, an impossibility. Thus, all states are positive recurrent.

Suppose the class C is periodic and null recurrent. Then according to the basic limit theorem for periodic Markov chains, $\lim_{n\to\infty} p_{ii}^{(n)} = 0$ for any $i \in C$. Furthermore, for any state $j \in C$, since $i \leftrightarrow j$, there exists positive integers m and n such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$. Therefore,

$$p_{ii}^{(m+n)} \geq p_{ij}^{(n)} p_{ji}^{(m)} > 0.$$

Fix n and let $m \to \infty$. Then it follows that $\lim_{m\to\infty} p_{ji}^{(m)} = 0$. Also, fix m and let $n \to \infty$. Then $\lim_{n\to\infty} p_{ij}^{(n)} = 0$. Thus, $\lim_{n\to\infty} P_C^n = \mathbf{0}$, where P_C is the submatrix of P consisting of states in C. This is an impossibility since P_C^n is stochastic. Thus, all states are positive recurrent.

In the case that the finite Markov chain is irreducible, there is only one class and all states in that class must be either positive recurrent, null recurrent, or transient. Since they cannot all be transient and there are no null recurrent states, they all must be positive recurrent. □

Since there are no null recurrent states in finite Markov chains, there are only four different types of classification schemes based on periodicity and recurrence. The states of a finite Markov chain can be classified as either periodic or aperiodic and either transient or positive recurrent. Recurrence in a finite Markov chain will always mean positive recurrence.

Example 2.20 Let P be the transition matrix of a Markov chain:

$$P = \begin{pmatrix} 1/2 & 0 & 0 & 1/2 \\ 1/2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

There are three communication classes, $\{1\}$, $\{2\}$, and $\{3,4\}$. Class $\{1\}$ is aperiodic and transient. Class $\{2\}$ is aperiodic and recurrent. Class $\{3,4\}$ is periodic and transient. State 2 is an absorbing state. Matrix P can be partitioned according to the classes,

$$P = \left(\begin{array}{c|c|cc} 1/2 & 0 & 0 & 1/2 \\ \hline 1/2 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{array}\right) = \begin{pmatrix} T_1 & 0 & A \\ B & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & T_3 \end{pmatrix}.$$

The diagonal matrices T_1 and T_3 corresponding to the transient classes have the property that

$$\lim_{n\to\infty} T_i^n = \mathbf{0}.$$

In addition, the entire first, third, and fourth rows all tend to zero as $n \to \infty$ (Lemma 2.2). Eventually, all transient classes are absorbed into the recurrent classes. In this example, there is eventual absorption into state 2. ■

An additional property of finite Markov chains is that a communication class that is closed is recurrent. This is verified in the next theorem. See Schinazi (1999).

Theorem 2.7. *In a finite Markov chain, a class is recurrent iff it is closed.*

Proof. We have already shown that if a class is recurrent, it is closed. The reverse implication is verified by contradiction. Assume a class of states C is closed, but C is not recurrent. Then C is transient. By Lemma 2.2, if j is a transient state, then $\lim_{n\to\infty} p_{ji}^{(n)} = 0$ for all states i. In particular, for state $i \in C$, we have

$$\sum_{j\in C} \lim_{n\to\infty} p_{ji}^{(n)} = 0. \tag{2.16}$$

But because C is closed, the submatrix P_C consisting of all states in C is a stochastic matrix. In addition, $P_C^{(n)}$ is a stochastic matrix (Exercise 1). The column sums of $P_C^{(n)}$ equal one and must equal one in the limit as $n \to \infty$, a contradiction to (2.16). Hence, C cannot be transient; C must be recurrent. □

Example 2.21 The finite Markov chain in Examples 2.4 and 2.13 has two recurrent classes, $\{1,2\}$ and $\{3,4,5\}$. The transition matrix can be partitioned according to these classes,

$$P = \left(\begin{array}{cc|ccc} 1/2 & 1/3 & 0 & 0 & 0 \\ 1/2 & 2/3 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1/4 & 0 \\ 0 & 0 & 1 & 1/2 & 1 \\ 0 & 0 & 0 & 1/4 & 0 \end{array}\right).$$

Since both classes are closed, they are both recurrent. They are also aperiodic. In addition, from the partition, it is easy to see that each of the diagonal submatrices forms a stochastic matrix. ■

In finite Markov chain theory, a stochastic matrix with the property $p_{ji}^{(n)} > 0$, for some $n > 0$ and all $i, j = 1, 2, \ldots, N$ ($P^n > 0$), is often referred to as a *regular* matrix. If the transition matrix is regular, then the Markov chain is irreducible and aperiodic (Why?). The Markov chain, in this case, is also referred to as regular. Therefore, a regular Markov chain is positive recurrent (strongly ergodic). The theorems of Perron and Frobenius from linear algebra state that a regular matrix P has a positive dominant eigenvalue λ that is simple and satisfies $\lambda > |\lambda_i|$, where λ_i is any other eigenvalue of P (see Gantmacher, 1964, or Ortega, 1987). In addition, the dominant eigenvalue λ has an associated positive eigenvector (see Gantmacher, 1964). The eigenvalue λ of a regular stochastic matrix P is $\lambda = 1$, and the associated eigenvector π satisfying $\sum \pi_j = 1$ defines a stationary probability distribution, $P\pi = \pi$. If the assumption of regularity is weakened, so that the stochastic matrix P is irreducible, then the theorems of Perron and Frobenius still imply that $\lambda = 1$ is a simple eigenvalue satisfying $\lambda \geq |\lambda_i|$ with associated positive eigenvector π (Ortega, 1987). Therefore, all that is required for existence of a unique stationary probability distribution is that P be irreducible. However, the additional property of aperiodicity (or regularity) is needed to show convergence to the stationary probability distribution. The following result is a corollary of Theorems 2.5 and 2.6.

Corollary 2.4. *Suppose a finite Markov chain is irreducible and aperiodic. Then there exists a unique stationary probability distribution* $\pi = (\pi_1, \ldots, \pi_N)^T$ *such that*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_i, \quad i, j = 1, 2, \ldots, N.$$

The next example is a finite Markov chain that models the dynamics of two squirrel populations.

Example 2.22 The introduction of a new or alien species into an environment will often disrupt the dynamics of native species (Hengeveld, 1989; Shigesada and Kawasaki, 1997; Williamson, 1996). For example, the gray squirrel, *Sciurus carolinensis*, was introduced into Great Britain in the late nineteenth century and it quickly invaded many regions previously occupied by the native red squirrel, *Sciurus vulgaris* (Reynolds, 1985). Data were collected in various regions in Great Britain as to the presence of red squirrels only (R), gray squirrels only (G), both squirrels (B), or absence of both squirrels (A). The data were summarized and a Markov chain model was developed with the four states, $\{R, G, B, A\}$. The model is reported in Mooney and Swift (1999). Each region was classified as being in one of these states, and the transitions between states over a period of one year were estimated (e.g., $p_{RR} = 0.8797$, $p_{RG} = 0.0382$). If the states 1, 2, 3, and 4 are ordered as R, G, B, and A, respectively, then the transition matrix has the form

$$P = \begin{pmatrix} 0.8797 & 0.0382 & 0.0527 & 0.0008 \\ 0.0212 & 0.8002 & 0.0041 & 0.0143 \\ 0.0981 & 0.0273 & 0.8802 & 0.0527 \\ 0.0010 & 0.1343 & 0.0630 & 0.9322 \end{pmatrix}.$$

It is easy to check that the corresponding Markov chain is irreducible, positive recurrent, and aperiodic (P is regular). There exists a unique, limiting stationary distribution π. We calculate this distribution by finding the eigenvector of P corresponding to the eigenvalue 1,

$$\pi = (0.1705, 0.0560, 0.3421, 0.4314)^T.$$

Over the long term, the model predicts that 17.05% of the region will be populated by red squirrels, 5.6% by gray squirrels, 34.21% by both species, and 43.14% by neither species. The mean recurrence times, $\mu_{ii} = 1/\pi_i$, $i = 1, 2, 3, 4$, are given by the vector

$$\mu = (5.865, 17.857, 2.923, 2.318)^T.$$

For example, a region populated by red squirrels (R) may change to other states (G, B, or A) but, on the average, it will again be populated by red squirrels after a period of about 5.9 years. ■

2.7.1 Mean Recurrence Time and Mean First Passage Time

A method is derived for calculating the mean recurrence times and mean first passages in irreducible finite Markov chains. Denote the matrix of

mean recurrence times and mean first passage times by

$$M = (\mu_{ij}) = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1N} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{N1} & \mu_{N2} & \cdots & \mu_{NN} \end{pmatrix}.$$

Instead of calculating the matrix elements via their definitions, using $\{f_{ii}^{(n)}\}$ and $\{f_{ji}^{(n)}\}$, an alternate method is applied. A relationship between the mean recurrence and the mean first passage times is derived that defines a linear system whose solution is M.

Consider what happens at the first time step. Either state j can be reached in one time step with probability p_{ji} or it takes more than one time step. If it takes more than one time step to reach j, then in one step another state k is reached, $k \neq j$, with probability p_{ki}. Then the time it takes to reach state j is $1 + \mu_{jk}$, one time step plus the mean time it takes to reach state j from state k. This relationship is given by

$$\mu_{ji} = p_{ji} + \sum_{k=1, k\neq j}^{N} p_{ki}(1 + \mu_{jk}) = 1 + \sum_{k=1, k\neq j}^{N} p_{ki}\mu_{jk}. \tag{2.17}$$

The relationship (2.17) assumes matrix P is irreducible; every state j can be reached from any other state i.

The equations in (2.17) can be expressed in matrix form,

$$M = E + (M - \text{diag}(M))P, \tag{2.18}$$

where E is an $N \times N$ matrix of ones. Since P is irreducible, the Markov chain is irreducible, which means it is positive recurrent, $1 \leq \mu_{ii} < \infty$, $i = 1, 2, \ldots, N$. It follows that $1 \leq \mu_{ji} < \infty$ for $j \neq i$. The system (2.18) can be written as a linear system of equations, N^2 equations and N^2 unknowns (the μ_{ji}'s). It can be shown that the linear system has a unique solution given by the μ_{ji}'s. Stewart (1994) discusses an iterative method based on equation (2.18) to estimate the mean recurrence times and mean first passage times.

Example 2.23 Suppose

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then equation (2.18) can be expressed as

$$\begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix} = \begin{pmatrix} 1 + \mu_{12} & 1 \\ 1 & 1 + \mu_{21} \end{pmatrix}.$$

Hence, $\mu_{12} = 1 = \mu_{21}$ and $\mu_{11} = 2$ and $\mu_{22} = 2$. This result is obvious once we recognize that the chain is periodic of period 2. It takes two time steps

to return to states 1 or 2 and only one time step to go from state 1 to state 2 or from state 2 to state 1. ■

The next section illustrates a method for determining the n-step transition matrix P^n in the case of a finite Markov chain.

2.8 The n-Step Transition Matrix

In the case of a finite Markov chain, a general form for the n-step transition matrix can be derived. A particularly simple form for P^n can be generated if P can be expressed as

$$P = UDU^{-1},$$

where D is a diagonal matrix and U is a nonsingular matrix. In this case, matrix P^n satisfies

$$P^n = UD^nU^{-1}.$$

An important theorem in linear algebra states that P can be expressed as $P = UDU^{-1}$ iff P is diagonalizable iff P has n linearly independent eigenvectors (Ortega, 1987). Hence, it shall be assumed that P has n linearly independent eigenvectors.

We show how the matrices U and D can be formed. Assume P is an $N \times N$ matrix with N eigenvalues, $\lambda_1, \lambda_2, \cdots, \lambda_N$. Let x_j be a right eigenvector (column vector) corresponding to λ_j and y_j be a left eigenvector (column vector):

$$Px_j = \lambda_j x_j \quad \text{and} \quad y_j^T P = \lambda_j y_j^T. \tag{2.19}$$

Define $N \times N$ matrices

$$H = (x_1, x_2, \ldots, x_N) \quad \text{and} \quad K = (y_1, y_2, \ldots, y_N),$$

where the columns of H are the right eigenvectors and the columns of K are the left eigenvectors. These matrices are nonsingular because the vectors are linearly independent. Because of the identities in (2.19),

$$PH = HD \quad \text{and} \quad K^T P = DK^T,$$

where $D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$. Thus,

$$P = HDH^{-1} \quad \text{or} \quad P = (K^T)^{-1} D K^T;$$

$U = H$ or $U = (K^T)^{-1}$. The n-step transition matrix satisfies

$$P^n = HD^nH^{-1} \quad \text{and} \quad P^n = (K^T)^{-1} D^n K^T. \tag{2.20}$$

The identities in (2.20) demonstrate one method that can be used to calculate P^n. Another method is also demonstrated below (see Bailey, 1990).

Note that

$$y_j^T P x_i = y_j^T \lambda_i x_i = y_j^T \lambda_j x_i.$$

If $\lambda_i \neq \lambda_j$, then $y_j^T x_i = 0$; the left and right eigenvectors are orthogonal. Thus, in this method, it is assumed that the eigenvalues are distinct (distinct eigenvalues imply the corresponding eigenvectors are linearly independent). Suppose y_j and x_j are chosen to be orthonormal:

$$y_j^T x_i = \begin{cases} 0 & i \neq j \\ 1 & i = j. \end{cases}$$

Then $K^T H = I$ (identity matrix) or $K^T = H^{-1}$ and

$$\begin{aligned} P &= HDH^{-1} = HDK^T \\ &= (\lambda_1 x_1, \lambda_2 x_2, \ldots, \lambda_N x_N)(y_1, y_2, \ldots, y_N)^T \\ &= \lambda_1 x_1 y_1^T + \lambda_2 x_2 y_2^T + \cdots + \lambda_N x_N y_N^T \end{aligned}$$

Therefore,

$$P = \sum_{i=1}^{N} \lambda_i x_i y_i^T.$$

Because the matrix $x_i y_i^T x_j y_j^T$ is the zero matrix for $i \neq j$ and the sum $\sum_{i=1}^{N} x_i y_i^T = HK^T = I$, it follows that P^2 can be expressed in terms of the matrices $x_i y_i^T$:

$$P^2 = \left(\sum_{i=1}^{N} \lambda_i x_i y_i^T \right) \left(\sum_{k=1}^{N} \lambda_k x_k y_k^T \right) = \sum_{i=1}^{N} \lambda_i^2 x_i y_i^T.$$

In general, the n-step transition matrix satisfies

$$P^n = \sum_{i=1}^{N} \lambda_i^n x_i y_i^T. \tag{2.21}$$

In the case where the Markov chain is regular (or ergodic), which means it is irreducible and aperiodic, then P^n has a limiting distribution. The limiting distribution is the stationary distribution corresponding to the eigenvalue $\lambda_1 = 1$. In this case,

$$\lim_{n \to \infty} P^n = x_1 y_1^T = \begin{pmatrix} \pi_1 & \pi_1 & \cdots & \pi_1 \\ \pi_2 & \pi_2 & \cdots & \pi_2 \\ \vdots & \vdots & \cdots & \vdots \\ \pi_N & \pi_N & \cdots & \pi_N \end{pmatrix},$$

where $x_1 = \pi$ and $y_1^T = (1, 1, \ldots, 1)$.

Note that both methods apply to any finite matrix with distinct eigenvalues. The two methods of computing P^n, given in (2.20) and (2.21), are illustrated in the next example.

Example 2.24 Consider a Markov chain with two states $\{1,2\}$ and transition matrix

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix},$$

where $0 < p < 1$. Note that this is a doubly stochastic matrix and all states are positive recurrent. Thus, $\lim_{n\to\infty} P^n p(0) = \pi = (\pi_1, \pi_1)^T$, where $\pi_2 = \pi_1$ (see Exercise 14).

The eigenvalues and corresponding eigenvectors of P are $\lambda_1 = 1$, $\lambda_2 = 1-2p$, $x_1^T = (1,1)$, $x_2^T = (1,-1)$, $y_1^T = (1,1)$, and $y_2^T = (1,-1)$. Note that $x_1/2 = \pi$ is the stationary probability distribution and $|\lambda_2| = |1-2p| < 1$. Using the first identity in (2.20), an expression for P^n is given by

$$P^n = H\Lambda^n H^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (1-2p)^n \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{2}.$$

Multiplication of the three matrices above yields

$$\begin{aligned} P^n &= \frac{1}{2} \begin{pmatrix} 1+(1-2p)^n & 1-(1-2p)^n \\ 1-(1-2p)^n & 1+(1-2p)^n \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{(1-2p)^n}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \end{aligned}$$

For example, the probability $p_{11}^{(n)}$ is $1/2 + (1-2p)^n/2$. Note that $P^n p(0)$ approaches the stationary probability distribution given by $\pi = (1/2, 1/2)^T$.

For the second method (2.21), the eigenvectors are normalized. Since $y_1^T x_1 = y_2^T x_2 = 2$, we divide by 2. Thus,

$$\begin{aligned} P^n &= \lambda_1^n \frac{x_1 y_1^T}{2} + \lambda_2^n \frac{x_2 y_2^T}{2} \\ &= \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{(1-2p)^n}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \end{aligned}$$

The two methods give the same expression for P^n. ■

There are other methods for computing P^n (see Elaydi, 1999; Elaydi and Harris, 1998; Kwapisz, 1998). We shall discuss one additional method for computing P^n, where it is not necessary that P be diagonalizable. This method is based on the Cayley-Hamilton theorem from linear algebra. Verification of this method is given in the Appendix for Chapter 2.

Suppose the characteristic polynomial of an $N \times N$ matrix P is given by

$$\det(\lambda I - P) = \lambda^N + a_{N-1}\lambda^{N-1} + \cdots + a_0 = 0.$$

This polynomial equation is also the characteristic polynomial of an Nth-order scalar difference equation of the form

$$x(N+n) + a_{N-1}x(N+n-1) + \cdots + a_0 x(n) = 0.$$

To find a general formula for P^n, it is necessary to find N linearly independent solutions to this Nth-order scalar difference equation, $x_1(n)$, $x_2(n), \ldots, x_N(n)$, with initial conditions

$$\left.\begin{array}{c} x_1(0)=1 \\ x_1(1)=0 \\ \vdots \\ x_1(N-1)=0 \end{array}\right\}, \quad \left.\begin{array}{c} x_2(0)=0 \\ x_2(1)=1 \\ \vdots \\ x_2(N-1)=0 \end{array}\right\}, \quad \cdots, \quad \left.\begin{array}{c} x_N(0)=0 \\ x_N(1)=0 \\ \vdots \\ x_N(N-1)=1 \end{array}\right\}.$$

Then

$$P^n = x_1(n)I + x_2(n)P + \cdots + x_N(n)P^{N-1}, \quad n = 0, 1, 2, \ldots. \tag{2.22}$$

Example 2.25 The nth power of matrix $P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$, given in Example 2.24, is computed using equation (2.22). The characteristic polynomial of P is

$$\lambda^2 - (2-2p)\lambda + 1 - 2p = (\lambda - 1)(\lambda - 1 + 2p) = 0.$$

The second-order linear difference equation,

$$x(n+2) - (2-2p)x(n+1) + (1-2p)x(n) = 0,$$

has two linearly independent solutions, 1 and $(1-2p)^n$, for $p \neq 0$. The general solution is a linear combination of these two solutions, $x(n) = c_1 + c_2(1-2p)^n$. Next, we find the constants c_1 and c_2 so that the two solutions x_1 and x_2 satisfy the required initial conditions. For the first solution, $x_1(0) = c_1 + c_2 = 1$ and $x_1(1) = c_1 + c_2(1-2p) = 0$. Solving for c_1 and c_2 we obtain the first solution:

$$x_1(n) = \frac{2p-1}{2p} + \frac{(1-2p)^n}{2p}.$$

For the second solution, $x_2(0) = c_1 + c_2 = 0$ and $x_2(1) = c_1 + c_2(1-2p) = 1$. Solving for c_1 and c_2 we obtain the second solution:

$$x_2(n) = \frac{1}{2p} - \frac{(1-2p)^n}{2p}.$$

Then applying the identity (2.22),

$$P^n = x_1(n)I + x_2(n)P = \frac{1}{2}\begin{pmatrix} 1+(1-2p)^n & 1-(1-2p)^n \\ 1-(1-2p)^n & 1+(1-2p)^n \end{pmatrix}.$$

This latter formula agrees with the one given in Example 2.24. ■

2.9 An Example: Genetics Inbreeding Problem

Inheritance depends on the information contained in the chromosomes that are passed down from generation to generation. Humans have two sets of chromosomes (diploid), one obtained from each parent. Certain locations along the chromosomes contain the instructions for some physical characteristic. The collections of chemicals at these locations are called *genes* and their locations are called *loci* (see, e.g., Hoppensteadt, 1975). At each locus, the gene may take one of several forms referred to as an *allele*.

Suppose there are only two types of alleles for a given gene, denoted a and A. A diploid individual could then have one of three different combinations of alleles: AA, Aa, or aa, known as the *genotypes* of the locus. The combinations AA and aa are called *homozygous*, whereas Aa is called *heterozygous*.

Bailey (1990) and Feller (1968) discuss a problem on the genetics of inbreeding and formulate a Markov chain model. We discuss this problem. Assume two individuals are randomly mated. Then, in the next generation, two of their offspring of opposite sex are randomly mated. The process of brother and sister mating or inbreeding continues each year. This process can be formulated as a finite, discrete time Markov chain whose states consist of the six mating types,

1. $AA \times AA$, 2. $AA \times Aa$, 3. $Aa \times Aa$, 4. $Aa \times aa$, 5. $AA \times aa$, 6. $aa \times aa$.

Suppose the parents are of type 1, $AA \times AA$. Then the next generation of offspring from these parents will be AA individuals, so that crossing of brother and sister will give only type 1, $p_{11} = 1$. Now, suppose the parents are of type 2, $AA \times Aa$. Offspring of type $AA \times Aa$ will occur in the following proportions, 1/2 AA and 1/2 Aa, so that crossing of brother and sister will give 1/4 type 1 ($AA \times AA$), 1/2 type 2 ($AA \times Aa$), and 1/4 type 3-($Aa \times Aa$). If the parents are of type 3, $Aa \times Aa$, offspring are in the proportions 1/4 AA, 1/2 Aa, and 1/4 aa, so that brother and sister mating will give 1/16 type 1, 1/4 type 2, 1/4 type 3, 1/4 type 4, 1/8 type 5, and 1/16 type 6. Continuing in this manner, we can complete the transition probability matrix P:

$$P = \begin{pmatrix} 1 & 1/4 & 1/16 & 0 & 0 & 0 \\ 0 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1 & 0 \\ 0 & 0 & 1/4 & 1/2 & 0 & 0 \\ 0 & 0 & 1/8 & 0 & 0 & 0 \\ 0 & 0 & 1/16 & 1/4 & 0 & 1 \end{pmatrix}$$

$$= \left(\begin{array}{c|cc|cc|c} 1 & 1/4 & 1/16 & 0 & 0 & 0 \\ \hline 0 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1 & 0 \\ 0 & 0 & 1/4 & 1/2 & 0 & 0 \\ 0 & 0 & 1/8 & 0 & 0 & 0 \\ \hline 0 & 0 & 1/16 & 1/4 & 0 & 1 \end{array}\right)$$

$$= \begin{pmatrix} 1 & A & 0 \\ \mathbf{0} & T & \mathbf{0} \\ 0 & B & 1 \end{pmatrix}.$$

The Markov chain is reducible and has three communicating classes: $\{1\}$, $\{6\}$, and $\{2,3,4,5\}$. The first two classes are positive recurrent and the third class is transient. States 1 and 6 are absorbing states, $p_{ii} = 1$, $i = 1, 6$.

Note that

$$P^n = \begin{pmatrix} 1 & A_n & 0 \\ \mathbf{0} & T^n & \mathbf{0} \\ 0 & B_n & 1 \end{pmatrix},$$

where A_n and B_n are functions of T, A, and B, $A_n = A\sum_{i=0}^{n-1} T^i$, and $B_n = B\sum_{i=0}^{n-1} T^i$. Thus, to determine P^n, we first determine T^n. Because T corresponds to a transient class, $\lim_{n\to\infty} T^n = \mathbf{0}$.

A general formula can be found for T^n. The eigenvalues of T are $\lambda_i = 1/2, 1/4, \frac{1}{4}(1+\sqrt{5}), \frac{1}{4}(1-\sqrt{5})$, $i = 1, 2, 3, 4$. For example, applying (2.20) or (2.21),

$$T^n = HD^nH^{-1} \quad \text{or} \quad T^n = \sum_{i=1}^{4} \lambda_i^n x_i y_i^T.$$

In addition, it can be seen that

$$\lim_{n\to\infty} B_n = B(I-T)^{-1} \quad \text{and} \quad \lim_{n\to\infty} A_n = A(I-T)^{-1}.$$

Once T^n is calculated, various questions can be addressed about the dynamics of the model at the nth time step. For example, what is the probability of absorption and the proportion of heterozygotes in the population in the nth generation? Absorption into states 1 or 6 can be calculated as follows. Absorption at step n into state 1 implies that at the $(n-1)$st step state 2 or 3 is entered. Then state 1 is entered at the next step. Thus, absorption into state 1 at the nth step is

$$p_{12}p_{2i}^{(n-1)} + p_{13}p_{3i}^{(n-1)} = \frac{1}{4}p_{2i}^{(n-1)} + \frac{1}{16}p_{3i}^{(n-1)}.$$

The values of $p_{2i}^{(n-1)}$ and $p_{3i}^{(n-1)}$ can be calculated from T^{n-1}. Absorption into state 6 at the nth step is

$$p_{63}p_{3i}^{(n-1)} + p_{64}p_{4i}^{(n-1)} = \frac{1}{16}p_{3i}^{(n-1)} + \frac{1}{4}p_{4i}^{(n-1)}.$$

The proportion of heterozygous individuals, Aa, at the nth time step satisfies

$$h_n = \frac{1}{2}p_2(n) + p_3(n) + \frac{1}{2}p_4(n),$$

where $p_i(n)$ is the proportion of the population in state i at time n. Because states 2, 3, and 4 are transient, $\lim_{n\to\infty} h_n = 0$. A simpler method of calculating h_n is discussed in the Appendix for Chapter 2.

2.10 Unrestricted Random Walks in Two and Three Dimensions

The random walk model can be extended to two and three dimensions. It was shown for the unrestricted random walk in one dimension that the chain is null recurrent if and only if $p = 1/2 = q$. For two and three dimensions, it is assumed that the probability of moving in any one direction is the same. Thus, for two dimensions, the probability is 1/4 of moving in any of the four directions: up, down, right, or left. For three dimensions, the probability is 1/6 of moving in any of the six directions: up, down, right, left, forward, or backward. It is shown for two dimensions that the chain is null recurrent but for three dimension it is transient. These examples illustrate the distinctly different behavior between one and two dimensions and dimensions greater than two. These examples were first studied by Polya and are discussed in many textbooks (see, e.g., Bailey 1990; Karlin and Taylor, 1975; Norris, 1997; Schinazi, 1999). The verifications are quite lengthy.

The Markov chain represented by this unrestricted random walk is irreducible and periodic of period 2. Therefore, recurrence and transience can be verified by checking recurrence or transience at the origin. Let the origin be denoted as 0 and $p_{00}^{(n)}$ be the probability of returning to the origin after n steps. Note that $p_{00}^{(2n)} > 0$, but $p_{00}^{(2n+1)} = 0$ for $n = 0, 1, 2, \ldots$. It is impossible to begin at the origin and return to the origin in an odd number of steps.

2.10.1 Two Dimensions

In two dimensions, for a path length of $2n$ beginning and ending at 0, if k steps are taken to the right, then k steps must be also taken to the left, and if $n - k$ steps are taken in the upward direction, then $n - k$ steps must be taken downward, $k + k + n - k + n - k = 2n$. There are

$$\sum_{k=0}^{n} \frac{(2n)!}{k!k!(n-k)!(n-k)!}$$

different paths of length $2n$ that begin and end at the origin. Each of these paths is equally likely and has a probability of occurring equal to $(1/4)^{2n}$. Thus,

$$\begin{aligned} p_{00}^{(2n)} &= \sum_{k=0}^{n} \frac{(2n)!}{k!k!(n-k)!(n-k)!} \left(\frac{1}{4}\right)^{2n} \\ &= \frac{(2n)!}{(n!)^2} \sum_{k=0}^{n} \left(\frac{n!}{k!(n-k)!}\right)^2 \left(\frac{1}{4}\right)^{2n} \\ &= \frac{(2n)!}{(n!)^2} \sum_{k=0}^{n} \binom{n}{k}^2 \left(\frac{1}{4}\right)^{2n}. \end{aligned}$$

It can be shown that

$$\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$$

(see, e.g., Bailey, 1990). Hence, $p_{00}^{(2n)}$ can be simplified to

$$p_{00}^{(2n)} = \frac{(2n)!}{(n!)^2} \binom{2n}{n} \left(\frac{1}{4}\right)^{2n} = \left[\frac{(2n)!}{n!n!}\right]^2 \frac{1}{4^{2n}}.$$

Stirling's formula can be applied to the right side of the above expression ($n! \sim n^n\sqrt{2\pi n}e^{-n}$) so that

$$\begin{aligned} p_{00}^{(2n)} &\sim \left[\frac{(2n)^{2n}\sqrt{4\pi n}e^{-2n}}{n^{2n}2\pi n e^{-2n}}\right]^2 \frac{1}{4^{2n}} \\ &= \left[\frac{4^n}{\sqrt{\pi n}}\right]^2 \frac{1}{4^{2n}} = \frac{1}{\pi n}. \end{aligned}$$

By comparing $\sum p_{00}^{(2n)}$ with the divergent harmonic series $\sum 1/[\pi n]$, it follows that $\sum p_{00}^{(2n)}$ also diverges. Thus, by Theorem 2.2, the origin is recurrent and all states must be recurrent. In addition, by applying the basic limit theorem for periodic Markov chains, $\lim_{n\to\infty} p_{00}^{(2n)} = 2/\mu_{00}$. But this limit is zero; thus, $\mu_{00} = \infty$. The zero state is null recurrent and, hence, the Markov chain for the symmetric, two-dimensional random walk is null recurrent.

2.10.2 Three Dimensions

In three dimensions, in a path of length $2n$ beginning and ending at the origin, if k steps are taken to the right, then k must be taken to the left; if j steps are taken upward, then j steps must be taken downward; and if $n - k - j$ steps are taken forward, then $n - k - j$ steps must be taken

backward, $k+k+j+j+n-k-j+n-k-j=2n$. The total number of paths of length $2n$ is

$$\sum_{j+k\le n}\frac{(2n)!}{(k!)^2(j!)^2[(n-k-j)!]^2},$$

where the sum is over all of the j and k, $j+k\le n$. Because each path has probability $(1/6)^{2n}$, it follows that

$$\begin{aligned} p_{00}^{(2n)} &= \sum_{j+k\le n}\frac{(2n)!}{(k!)^2(j!)^2[(n-k-j)!]^2}\left(\frac{1}{6}\right)^{2n} \\ &= \frac{(2n)!}{2^{2n}(n!)^2}\sum_{j+k\le n}\left(\frac{n!}{j!k!(n-j-k)!}\right)^2\left(\frac{1}{3}\right)^{2n}. \end{aligned}$$

We use the fact that the trinomial distribution satisfies

$$\sum_{j+k\le n}\frac{n!}{j!k!(n-j-k)!}\frac{1}{3^n}=1.$$

For convenience, denote the trinomial coefficient as

$$\frac{n!}{j!k!(n-j-k)!}=\binom{n}{j\ k}.$$

The maximum value of the trinomial distribution can be shown to occur when $j\approx n/3$ and $k\approx n/3$ and is approximately equal to $M_n\approx n!(1/3)^n/[(n/3)!]^3$ when n is large. This can be seen as follows. Suppose the maximum value occurs at j' and k'. Then

$$\begin{aligned} \binom{n}{j'\ (k'-1)} &\le \binom{n}{j'\ k'} \\ \binom{n}{j'\ (k'+1)} &\le \binom{n}{j'\ k'} \\ \binom{n}{(j'-1)\ k'} &\le \binom{n}{j'\ k'} \\ \binom{n}{(j'+1)\ k'} &\le \binom{n}{j'\ k'} \end{aligned}$$

so that

$$n-k'-1\le 2j'\le n-k'+1,\quad n-j'-1\le 2k'\le n-j'+1$$

or

$$\frac{n-1}{n}\le\frac{2j'+k'}{n}\le\frac{n+1}{n},\quad \frac{n-1}{n}\le\frac{2k'+j'}{n}\le\frac{n-1}{n}.$$

Letting $n \to \infty$, then $2j' + k' \sim n$ and $2k' + j' \sim n$, from which it follows that $j' \sim n/3$ and $k' \sim n/3$.

We use the above facts to get an upper bound on $p_{00}^{(2n)}$. First,

$$\begin{aligned} p_{00}^{(2n)} &\leq \frac{1}{2^{2n}} \frac{(2n)!}{(n!)^2} M_n \left[\sum_{j+k \leq n} \frac{n!}{j!k!(n-j-k)!} \frac{1}{3^n} \right] \\ &= \frac{1}{2^{2n}} \frac{(2n)!}{(n!)^2} \frac{n!}{[(n/3)!]^3} \frac{1}{3^n}, \end{aligned}$$

because the expression in the square brackets is a trinomial distribution whose sum equals one. Next, Stirling's formula is used to approximate the right-hand side of the above inequality for large n:

$$\begin{aligned} \frac{1}{2^{2n}} \frac{(2n)!}{(n!)^2} \frac{n!}{[(n/3)!]^3} \frac{1}{3^n} &\sim \frac{1}{2^{2n}} \frac{(2n)^{2n}\sqrt{4\pi n}e^{-2n}}{n^n\sqrt{2\pi n}e^{-n}(n/3)^n(\sqrt{2\pi n/3})^3 e^{-n}} \frac{1}{3^n} \\ &= \frac{c}{n^{3/2}}, \end{aligned}$$

where $c = (1/2)(3/\pi)^{3/2}$. Thus, for large n, $p_{00}^{(2n)} \leq c/n^{3/2}$. Because $\sum_n c/n^{3/2}$ is a convergent p-series, it follows by comparison and Theorem 2.2 that the origin is a transient state. Hence, because the Markov chain is irreducible, all states are transient. The discrete time Markov chain for the symmetric, three-dimensional random walk is transient.

The distinctly different behavior of discrete time Markov chains in three dimensions as opposed to one or two dimensions is not unusual. A path along a line or in a plane is much more restricted than a path in space. This difference in behavior is demonstrated in other models as well (e.g., systems of autonomous differential equations), where the behavior in three or higher dimensions is much more complicated and harder to predict than in one or two dimensions.

2.11 Exercises for Chapter 2

1. Suppose P is an $N \times N$ stochastic matrix (column sums equal one),

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix}.$$

   (a) Show that P^2 is a stochastic matrix. Then show that P^n is a stochastic matrix for all positive integers n.

(b) Suppose P is a doubly stochastic matrix (row and column sums equal one). Show that P^n is a doubly stochastic matrix for all positive integers n.

2. Show that the relation (2.3) follows from conditional probabilities. In particular, show that

$$\text{Prob}\{A \cap B|C\} = \text{Prob}\{A|B \cap C\}\text{Prob}\{B|C\}.$$

3. If j is a transient state of a Markov chain with states $\{1, 2, \ldots\}$, prove that for all states $i = 1, 2, \ldots$,

$$\sum_{n=1}^{\infty} p_{ji}^{(n)} < \infty$$

and $\lim_{n\to\infty} p_{ji}^{(n)} = 0$. [*Hint*: Use the identity $P_{ji}(s) = F_{ji}(s)P_{jj}(s)$ when s equals 1.]

4. Suppose a finite Markov chain has N states. State 1 is absorbing and the remaining states are transient. Use Exercises 3 and 1 to show that

$$\lim_{n\to\infty} P^n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Then for any initial probability distribution corresponding to X_0, $p(0) = (p_1(0), p_2(0), \ldots, p_N(0))^T$, it follows that

$$\lim_{n\to\infty} P^n p(0) = (1, 0, \ldots, 0)^T.$$

5. Verify the following two statements.

(a) Assume the period of state i in a discrete time Markov chain model satisfies $d(i) = 0$. Then the set $\{i\}$ is a communication class in the Markov chain.

(b) In an irreducible, discrete time Markov chain, the period $d \geq 1$.

6. Refer to Example 2.9. Show that the mean recurrence times for this example are finite, $\mu_{ii} < \infty$ for $i = 1, 2$.

7. A Markov chain has the following transition matrix:

$$P = \begin{pmatrix} 0 & 1/2 & 0 \\ 1 & 0 & 1 \\ 0 & 1/2 & 0 \end{pmatrix}.$$

(a) Draw a directed graph for the chain.

(b) Identify the communicating classes and classify them as periodic or aperiodic, transient or recurrent.

(c) Calculate the probability of the first return to state i at the nth step, $f_{ii}^{(n)}$, for each state $i = 1, 2, 3$ and for each time step $n = 1, 2, \ldots$.

(d) Use (c) to calculate the mean recurrence times for each state, μ_{ii}, $i = 1, 2, 3$.

8. Three different Markov chains are defined by the following transition matrices:

$$\text{(i)} \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}, \quad \text{(ii)} \begin{pmatrix} 1 & 0 & 1/3 \\ 0 & 0 & 1/3 \\ 0 & 1 & 1/3 \end{pmatrix}, \text{(iii)} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

(a) Draw a directed graph for each chain. Is the Markov chain irreducible?

(b) Identify the communicating classes and classify them as periodic or aperiodic, transient or recurrent.

9. Three different Markov chains are defined by the following transition matrices:

$$\text{(i)} \begin{pmatrix} 1/3 & 1/4 & 0 & 1/2 \\ 1/3 & 1/4 & 0 & 0 \\ 0 & 1/4 & 1 & 0 \\ 1/3 & 1/4 & 0 & 1/2 \end{pmatrix}, \quad \text{(ii)} \begin{pmatrix} 1/2 & 1/3 & 0 & 0 & 1 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 1/3 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1 & 0 \\ 1/2 & 0 & 1/3 & 0 & 0 \end{pmatrix},$$

$$\text{(iii)} \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 2/3 & 0 \\ 1/3 & 0 & 2/3 \end{pmatrix}.$$

(a) Draw a directed graph for each chain. Is the Markov chain irreducible?

(b) Identify the communicating classes and classify them as periodic or aperiodic, transient or recurrent.

10. Suppose the states of three different Markov chains are $\{1, 2, \ldots\}$ and their corresponding transition matrices are

$$P_1 = \begin{pmatrix} a_1 & 0 & 0 & \cdots \\ a_2 & a_1 & 0 & \cdots \\ a_3 & a_2 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}, \quad P_2 = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ a_1 & 0 & 0 & \cdots \\ a_2 & a_1 & 0 & \cdots \\ a_3 & a_2 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix},$$

and

$$P_3 = \begin{pmatrix} 1 & 1/2 & 1/3 & \cdots \\ 0 & 1/2 & 1/3 & \cdots \\ 0 & 0 & 1/3 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

The elements a_i of P_1 and P_2 are positive and $\sum_{i=1}^{\infty} a_i = 1$.

(a) Draw a directed graph for each chain. Is the Markov chain irreducible?

(b) Identify the communicating classes, find their period, then classify them as transient, null recurrent, or positive recurrent.

11. Assume $i \leftrightarrow j$ for states i and j in a Markov chain. Prove the following: State i is positive recurrent if and only if state j is positive recurrent. Show that the same result holds if positive recurrent is replaced by null recurrent. (*Hint*: Apply the basic limit theorems for periodic and aperiodic Markov chains and use the relation $p_{jj}^{(m+n)} \geq p_{ji}^{(m)} p_{ij}^{(n)} > 0$ for some m and n.)

12. The transition matrix for a three-state Markov chain is

$$P = \begin{pmatrix} 1 & q & 0 \\ 0 & r & q \\ 0 & p & p+r \end{pmatrix},$$

$p, q > 0$, $r \geq 0$, and $p + q + r = 1$.

(a) Draw the directed graph of the chain.

(b) Is the set $\{2, 3\}$ closed? Why or why not?

(c) Find an expression for $p_{11}^{(n)}$. Then verify that state 1 is positive recurrent.

(d) Show that the process has a unique stationary probability distribution, $\pi = (\pi_1, \pi_2, \pi_3)^T$.

13. The transition matrix for a four-state Markov chain is

$$P = \begin{pmatrix} 0 & 1/4 & 0 & 1/2 \\ 1/2 & 0 & 3/4 & 0 \\ 0 & 3/4 & 0 & 1/2 \\ 1/2 & 0 & 1/4 & 0 \end{pmatrix}.$$

(a) Show that the chain is irreducible, positive recurrent, and periodic. What is the period?

(b) Find the unique stationary probability distribution.

14. Suppose the transition matrix P of a finite Markov chain is doubly stochastic; that is, row and column sums equal one, $p_{ij} \geq 0$,

$$\sum_{i=1}^{N} p_{ij} = 1, \quad \text{and} \quad \sum_{j=1}^{N} p_{ij} = 1.$$

Prove the following: If an irreducible, aperiodic finite Markov chain (ergodic chain) has a doubly stochastic transition matrix, then all stationary probabilities are equal, $\pi_1 = \pi_2 = \cdots = \pi_N$.

15. The transition matrix for a three-state Markov chain is

$$P = \begin{pmatrix} 0 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1 & 1 & 0 \end{pmatrix}.$$

(a) Draw a directed graph of the chain and show that P is irreducible.

(b) Show that P is periodic of period 2 and find P^{2n}, $n = 1, 2, \ldots$.

(c) Use the identity (2.18) to find the mean recurrence times and mean first passage times. Show that the mean recurrence times agree with the formula given in the basic limit theorem for periodic Markov chains.

16. Suppose that the transition matrix of a two-state Markov chain is

$$P = \begin{pmatrix} 1-a & b \\ a & 1-b \end{pmatrix}, \tag{2.23}$$

where $0 < a < 1$ and $0 < b < 1$. Use the identity (2.18) to find a general formula for the mean recurrence times and mean first passage times.

17. Suppose that the transition matrix of a two-state Markov chain is given by equation (2.23) in Exercise 16. Use the identity $P^n = UD^nU^{-1}$ to show that P^n can be expressed as follows:

$$P^n = \frac{1}{a+b}\begin{pmatrix} b & b \\ a & a \end{pmatrix} + \frac{(1-a-b)^n}{a+b}\begin{pmatrix} a & -b \\ -a & b \end{pmatrix}.$$

18. Let $P = \begin{pmatrix} 1 & 1/4 & 1/2 \\ 0 & 1/2 & 0 \\ 0 & 1/4 & 1/2 \end{pmatrix}$. A general formula for P^n will be derived using the method of Example 2.25.

(a) Show that the characteristic polynomial of P is $\lambda^3 - 2\lambda^2 + (5/4)\lambda - 1/4 = (\lambda - 1)(\lambda - 1/2)^2 = 0$. Therefore, three linearly independent solutions of the third order linear difference equation, $x(n+3) - 2x(n+2) + (5/4)x(n+1) - (1/4)x(n) = 0$, are $y_1(n) = 1$, $y_2(n) = 1/2^n$, and $y_3(n) = n/2^n$.

(b) Use the three linearly independent solutions, $y_i(n)$, $i = 1, 2, 3$, to find three solutions $x_i(n)$, $i = 1, 2, 3$ that satisfy the initial conditions.

(c) Use the identity (2.22) to find a general expression for P^n.

19. Suppose that two unbiased coins are tossed repeatedly and after each toss the accumulated number of heads and tails that have appeared on each coin is recorded. Let the random variable X_n denote the difference in the accumulated number of heads on coin 1 and coin 2 after the nth toss [e.g, (Total # Heads Coin 1) – (Total # Heads Coin 2)]. Thus, the state space is $\{0, \pm 1, \pm 2, \ldots\}$. Show that the zero state, where the total number of heads is equal on each coin, is null recurrent. *Hint*: Show that

$$p_{00}^{(n)} = \frac{1}{2^{2n}} \sum_{k=0}^{n} \binom{n}{k}^2 \sim \frac{1}{\sqrt{n\pi}}$$

(Bailey, 1990).

20. Consider the genetics inbreeding problem. Let

$$p(0) = (0, 1/4, 1/4, 1/4, 1/4, 0)^T.$$

(a) Find a general formula for the proportion of heterozygotes h_n in terms of the eigenvalues:

$$h_n = a\lambda_1^n + b\lambda_2^n + c\lambda_3^n + d\lambda_4^n.$$

(b) Find h_{20} and h_{40} (see the Appendix).

21. A Markov chain model for the growth and replacement of trees assumes that there are three stages of growth based on the size of the tree: young tree, mature tree, and old tree. When an old tree dies, it is replaced by a young tree with probability $1 - p$. Order the states numerically, 1, 2, and 3. State 1 is a young tree, state 2 is a mature tree, and state 3 is an old tree. A Markov chain model for the transitions between each of the states over a period of eight years has the following transition matrix:

$$P = \begin{pmatrix} 1/4 & 0 & 1-p \\ 3/4 & 1/2 & 0 \\ 0 & 1/2 & p \end{pmatrix}.$$

Transitions occur over an eight-year period. For example, after a period of eight years the probability that a young tree becomes a mature tree is $3/4$ and the probability it remains a young tree is $1/4$.

(a) Suppose $0 \leq p < 1$. Show that the Markov chain is irreducible and aperiodic. Find the unique limiting stationary distribution.

(b) Suppose $p = 7/10$. Find the mean recurrence time for $i = 1, 2, 3$ (i.e., the mean number of eight-year periods it will take a tree in stage i to be replaced by another tree of stage i).

22. A Markov chain model for the growth and replacement of trees assumes that there are four stages of growth based on the size of the tree: seedling, young tree, mature tree, and old tree. When an old tree dies, it is replaced by a seedling. Order the states numerically, $1, 2, 3$, and 4. State 1 is a seedling, state 2 is a young tree, and so on. A Markov chain model for the transitions between each state over a period of five years has the following transition matrix:

$$P = \begin{pmatrix} p_{11} & 0 & 0 & 1-p_{44} \\ 1-p_{11} & p_{22} & 0 & 0 \\ 0 & 1-p_{22} & p_{33} & 0 \\ 0 & 0 & 1-p_{33} & p_{44} \end{pmatrix}.$$

Transition p_{ii} is the probability that a tree remains in the same state for five years and $1 - p_{ii}$ is the probability a tree is at the next stage after five years of growth.

(a) Suppose $0 < p_{ii} < 1$ for $i = 1, 2, 3, 4$. Show that the Markov chain is irreducible and aperiodic. Find the unique limiting stationary distribution.

(b) Suppose $p_{44} = 1$ and $0 < p_{ii} < 1$ for $i = 1, 2, 3$. What do these assumptions imply about the growth and replacement of trees? Show that $\lim_{n\to\infty} p_{ii}^{(n)} = p_{ii}^{n}$. Identify the communicating classes and determine if they are transient or recurrent.

2.12 References for Chapter 2

Bailey, N. T. J. 1990. *The Elements of Stochastic Processes with Applications to the Natural Sciences.* John Wiley & Sons, New York.

Elaydi, S. N. 1999. *An Introduction to Difference Equations.* 2nd ed. Springer-Verlag, New York.

Elaydi, S. and W. Harris. 1998. On the computation of A^n. *SIAM Review* 40: 965–971.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications.* Vol 1. 3rd ed. John Wiley & Sons, New York.

Gantmacher, F. R. 1964. *The Theory of Matrices.* Vol. II. Chelsea Pub. Co., New York.

Hengeveld, R. 1989. *Dynamics of Biological Invasions.* Chapman and Hall, London and New York.

Hoppensteadt, F. 1975. *Mathematical Methods of Population Biology.* Cambridge University Press, Cambridge.

Karlin, S. and H. Taylor. 1975. *A First Course in Stochastic Processes.* 2nd ed. Academic Press, New York.

Kemeny, J. G. and J. L. Snell. 1960. *Finite Markov Chains.* Van Nostrand, Princeton, N. J.

Kwapisz, M. 1998. The power of a matrix. *SIAM Review* 40: 703–705.

Leonard, I. E. 1996. The matrix exponential. *SIAM Review* 38: 507–512.

Mooney, D. and R. Swift. 1999. *A Course in Mathematical Modeling.* The Mathematical Association of America, Washington, D. C.

Norris, J. R. 1997. *Markov Chains.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Ortega, J. M. 1987. *Matrix Theory A Second Course.* Plenum Press, New York.

Reynolds, J. C. 1985. Details of the geographic replacement of the red squirrel (*Sciurus vulgaris*) by the grey squirrel (*Sciurus carolinensis*) in eastern England. *J. Anim. Ecol.* 54: 149–162.

Schinazi, R. B. 1999. *Classical and Spatial Stochastic Processes.* Birkhäuser, Boston.

Shigesada N. and K. Kawasaki. 1997. *Biological Invasions: Theory and Practice.* Oxford University Press, Oxford, New York, and Tokyo.

Stewart, W. J. 1994. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, Princeton, N. J.

Taylor, H. M. and S. Karlin. 1998. *An Introduction to Stochastic Modeling.* 3rd ed. Academic Press, New York.

Tuljapurkar, S. 1997. Stochastic matrix models. In *Structured-Population Models in Marine, Terrestrial, and Freshwater Systems.* S. Tuljapurkar and H. Caswell (eds.) pp. 59–87. Chapman & Hall, New York.

Wade, W. R. 2000. *An Introduction to Analysis.* 2nd ed. Prentice Hall, Upper Saddle River, N. J.

Williamson, M. 1996. *Biological Invasions.* Chapman & Hall, London.

2.13 Appendix for Chapter 2

2.13.1 Power of a Matrix

Suppose P is an $N \times N$ matrix with characteristic polynomial

$$c(\lambda) = \det(\lambda I - P) = \lambda^N + a_{N-1}\lambda^{N-1} + \cdots + a_0 = 0.$$

Note that matrix $Z(n) = P^n$ is the unique solution to the matrix difference equation,

$$Z(N+n) + a_{N-1}Z(N+n-1) + \cdots + a_0 Z(n) = \mathbf{0}, \tag{2.24}$$

with initial conditions $Z(0) = I$, $Z(1) = P, \ldots,$ and $Z(N-1) = P^{N-1}$. This follows from the Cayley-Hamilton theorem from linear algebra that states a matrix P satisfies its characteristic polynomial, $c(P) = \mathbf{0}$. The following theorem is due to Kwapisz (1998). It is based on a similar theorem for matrix exponentials by Leonard (1996).

Theorem 2.8. *Suppose $x_1(n), x_2(n), \ldots, x_N(n)$ are solutions of the Nth-order scalar difference equation,*

$$x(N+n) + a_{N-1}x(N+n-1) + \cdots + a_0 x(n) = 0,$$

with initial conditions

$$\left.\begin{array}{c} x_1(0) = 1 \\ x_1(1) = 0 \\ \vdots \\ x_1(N-1) = 0 \end{array}\right\}, \quad \left.\begin{array}{c} x_2(0) = 0 \\ x_2(1) = 1 \\ \vdots \\ x_2(N-1) = 0 \end{array}\right\}, \quad \cdots, \quad \left.\begin{array}{c} x_N(0) = 0 \\ x_N(1) = 0 \\ \vdots \\ x_N(N-1) = 1 \end{array}\right\}.$$

Then

$$P^n = x_1(n)I + x_2(n)P + \cdots + x_N(n)P^{N-1}, \quad n = 0, 1, 2, \ldots.$$

Proof. Let $Z(n) = x_1(n)I + x_2(n)P + \cdots + x_N(n)P^{N-1}$ for $n = 0, 1, 2, \ldots$. Then substitution of $Z(n)$ into the difference equation (2.24) shows that $Z(n)$ satisfies the equation. In addition, $Z(0) = I$, $Z(1) = P, \ldots,$ and $Z(N-1) = P^{N-1}$. Because the solution of (2.24) is unique it follows that $Z(n) = P^n$ for $n = 0, 1, 2, \ldots$. □

2.13.2 Genetics Inbreeding Problem

In the genetics inbreeding problem, h_n is the proportion of heterozygotes at time n,

$$h_n = \frac{1}{2}p_2(n) + p_3(n) + \frac{1}{2}p_4(n),$$

where $p_i(n)$ is the proportion of the population in state i at time n (Bailey, 1990). The three states are elements of the matrix T^n. Let $p(0) = (p_2(0), p_3(0), p_4(0), p_5(0))^T$. Then $T^n p(0) = p(n) = \sum_{i=1}^4 \lambda_i^n x_i y_i^T p(0)$. It follows that the $p_i(n)$ satisfy

$$p_i(n) = c_{i1}\lambda_1^n + c_{i2}\lambda_2^n + c_{i3}\lambda_3^n + c_{i4}\lambda_4^n, \quad i = 2, 3, 4, 5.$$

Hence,

$$h_n = a\lambda_1^n + b\lambda_2^n + c\lambda_3^n + d\lambda_4^n,$$

where a, b, c, d are combinations of the c_{ij}. The coefficients a, b, c, d can be found by solving the following four linear equations (linear in a, b, c, d):

$$h_i = a\lambda_1^i + b\lambda_2^i + c\lambda_3^i + d\lambda_4^i, \quad i = 0, 1, 2, 3.$$

Suppose, initially, the entire population is of type 2, $AA \times Aa$, $p^{(0)} = (0, 1, 0, 0, 0, 0)^T$. Then $h_0 = 1/2$ and $Pp^{(0)} = (1/4, 1/2, 1/4, 0, 0, 0)^T$ so that $h_1 = (1/2)(1/2) + (1/2)(1/2) + (1/2)(0) = 1/2$. By computing $P^2 p^{(0)}$ and $P^3 p^{(0)}$, values for h_2 and h_3 can be calculated, $h_2 = 3/8$ and $h_3 = 5/16$. The following *Maple* program was used to calculate h_{20} and h_{30} and a general formula was obtained for h_n

$$h_{20} = 0.008445, \quad h_{30} = 0.001014,$$

and

$$h_n = (1/4 + 3\sqrt{5}/20)\lambda_3^n + (1/4 - 3\sqrt{5}/20)\lambda_4^n.$$

```
> with(linalg):
> P:=matrix(6,6,[1,1/4,1/16,0,0,0,0,1/2,1/4,0,0,0,0,1/4,1/4,
1/4,1,0,0,0,1/4,1/2,0,0,0,0,1/8,0,0,0, 0,0,1/16,1/4,0,1]):
> T:=matrix(4,4,[1/2,1/4,0,0,1/4,1/4,1,0,0,1/4,1/2,0,0,1/8,
0,0]):
> p0:=vector([0,1,0,0,0,0]):
> h0:=1/2*p0[2]+p0[3]+1/2*p0[4]:
> p:=n->evalm(P ^ n&*p0):
> h:=n->1/2*p(n)[2]+p(n)[3]+1/2*p(n)[4]:
> evalm(p0); h0;
```

$$[0, 1, 0, 0, 0, 0]$$

```
> evalm(p(1)); h(1);
```

$$[1/4, 1/2, 1/4, 0, 0, 0]$$

$$1/2$$

```
> evalm(p(2)); h(2);
```

$$[25/64, 5/16, 3/16, 1/16, 1/32, 1/4]$$

$$3/8$$

```
> evalm(p(3)); h(3);
```

$$[123/256, 13/64, 11/64, 5/64, 3/128, 11/256]$$

$$5/16$$

```
> ll:=[eigenvals(P)];
```

$$ll := \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{4} + \frac{1}{4}\sqrt{5}, \frac{1}{4} - \frac{1}{4}\sqrt{5}, 1, 1\right]$$

```
> f:=n->a*ll[1] ^ n+b*ll[2] ^ n+c*ll[3] ^ n+d*ll[4] ^ n:
> solve({f(0)=q0,f(1)=q(1),f(2)=q(2),f(3)=q(3)},{a,b,c,d});
```

$$\left\{a = 0, b = 0, c = \frac{1}{4} + \frac{3}{20}\sqrt{5}, d = \frac{1}{4} - \frac{3}{20}\sqrt{5}\right\}$$

```
> f(20):=evalf(subs({a = 0, b = 0, c = 1/4+3/20*sqrt(5),
d = 1/4-3/20*sqrt(5)},f(20)));
```

$$f(20) := 0.008445262939$$

```
> f(30):=evalf(subs({a = 0, b = 0, c = 1/4+3/20*sqrt(5),
d = 1/4-3/20*sqrt(5)},f(30)));
```

$$f(30) := 0.001014354178$$

```
> evalf(h(20));
```

$$0.008445262909$$

```
> evalf(h(30));
```

$$0.001014354173$$

Chapter 3

Biological Applications of Discrete Time Markov Chains

3.1 Introduction

Several classical and biological applications of discrete time Markov chain models are discussed in this chapter. The first application of Markov chains is a random walk on the finite set $\{0, 1, 2 \ldots, N\}$, with absorbing boundaries at $x = 0$ and $x = N$. This first application is often referred to as the *gambler's ruin problem*, a classical discrete time Markov chain model, discussed briefly in Example 2.2. An expression is derived for the probability of absorption using techniques from difference equations. Then an expression for the expected duration until absorption is derived. Finally, the entire probability distribution for absorption at the nth time step is derived using generating functions and difference equations. Some of these results are extended to a random walk on a semi-infinite domain, where the states include $\{0, 1, 2 \ldots\}$.

The second application of discrete time Markov chain models is to birth and death processes. A general discrete time birth and death process is described. The general birth and death process is applied to a logistic birth and death process, where the birth and death probabilities are nonlinear functions of the population size. In this model, it is assumed there is a maximal population size, so that the processes are finite Markov chains. A transition matrix can be defined. The theory developed from random walk models will be useful to the analysis of the birth and death processes. The probability of absorption or population extinction and the expected time until population extinction are discussed. Further, the distribution conditioned on nonextinction, known as the quasistationary distribution, is studied.

The final application of discrete time Markov chain models is to epidemic models. A stochastic Susceptible-Infected-Susceptible (SIS) epidemic model is studied. In an SIS epidemic model, susceptible individuals become infected but do not develop immunity; they immediately become susceptible again. In this model, there are N states, where the states correspond to the number of infected individuals. It is shown that the process is equivalent to a logistic growth model. In addition to the SIS epidemic model, some other epidemic models, known as chain binomial models, are studied. The chain binomial epidemic models were first developed in the 1920s and 1930s by Reed, Frost, and Greenwood and are appropriately named Reed-Frost and Greenwood models. For these models, the duration and size of the epidemic are computed.

3.2 Restricted Random Walk Models

A *restricted random walk* is a random walk model with at least one boundary, so that either the state space of the process is finite, $\{0, 1, 2 \ldots, N\}$, with two boundaries at 0 and N, or semi-infinite, $\{0, 1, 2, \ldots\}$, with one boundary at 0. In a random walk model, the states are positions and will be denoted by the variable x, $x = 0, 1, 2, \ldots$. The variable n will denote time, where $n \in \{0, 1, 2 \ldots\}$. In the simplest random walk model, it is assumed that p is the probability of moving to the right, x to $x + 1$, and q is the probability of moving to the left, x to $x - 1$.

Assumptions about movement at the boundary, at $x = 0$ or $x = N$, differ from movement at other positions. We shall discuss three types of boundary behavior: absorbing, reflecting, and elastic. An *absorbing boundary* at $x = 0$ assumes the one-step transition probability

$$p_{00} = 1.$$

A *reflecting boundary* at $x = 0$ assumes the transition probabilities

$$p_{00} = 1 - p \quad \text{and} \quad p_{10} = p, \quad 0 < p < 1.$$

An *elastic boundary* at $x = 0$ assumes the transition probabilities

$$p_{21} = p, \quad p_{11} = sq, \quad p_{01} = (1 - s)q, \quad p + q = 1, \quad \text{and} \quad p_{00} = 1,$$

for $0 < p, s < 1$. An elastic boundary is intermediate in relation to absorbing and reflecting boundaries. If $s = 0$, then $x = 0$ is an absorbing boundary, and if $s = 1$, then $x = 1$ is a reflecting boundary. When $0 < s < 1$, an object moving toward the boundary from position $x = 1$ will either reach $x = 0$ with probability $(1 - s)q$ or return to $x = 1$ with probability sq (elastic property).

In the next several sections, the restricted random walk with absorbing boundaries at $x = 0$ and $x = N$ is studied. This process is also known

as the *gambler's ruin problem*, discussed briefly in Example 2.2. It will be shown in Section 3.5 that the theory from the gambler's ruin problem can be applied to a simple birth and death process.

3.3 Gambler's Ruin Problem

The position x in the gambler's ruin problem represents the gambler's capital, and each time step represents one game where the gambler may either increase his/her capital to $x+1$ or decrease it to $x-1$. After each game there is either a gain or loss, never a tie. If the gambler's capital reaches zero, he/she is ruined, the opponent has won, and the games stop, whereas if the capital reaches N, he/she has won all of the capital (the opponent is ruined), and the games stop.

First, we define the transition matrix for the gambler's ruin problem. Let p be the probability of moving to the right (winning a game), q be the probability of moving to the left (losing a game), and $p+q=1$; that is,

$$p_{ji} = \text{Prob}\{X_{n+1} = j | X_n = i\} = \begin{cases} p, & \text{if } j = i+1, \\ q, & \text{if } j = i-1, \\ 0, & \text{if } j \neq i+1, i-1, \end{cases}$$

for $i = 1, 2, \ldots, N-1$. The state space is $\{0, 1, 2, \ldots, N\}$. The boundaries 0 and N are absorbing,

$$p_{00} = 1 \quad \text{and} \quad p_{NN} = 1.$$

The transition matrix P is an $(N+1) \times (N+1)$ matrix of the following form:

$$P = \begin{pmatrix} 1 & q & 0 & \cdots & 0 & 0 \\ 0 & 0 & q & \cdots & 0 & 0 \\ 0 & p & 0 & \cdots & 0 & 0 \\ 0 & 0 & p & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & p & 1 \end{pmatrix}.$$

There are three communication classes, $\{0\}$, $\{1, 2, \ldots, N-1\}$, and $\{N\}$. States 0 and N are absorbing (recurrent) and the other states are transient.

To understand the dynamics of the gambler's ruin problem, we investigate the probability that the gambler either loses or wins all of his/her money (probability of absorption) and the mean number of games until the gambler either wins all of the money or loses all of the money (expected duration of the games). Absorption occurs at either $x = 0$ (ruin) or at $x = N$ (jackpot). Beginning with a capital of k, the expected duration of

the games (expected duration until absorption) is the sum of the following mean first passage times, $\mu_{0k} + \mu_{Nk}$.

Let

a_{kn} = probability of absorption at $x = 0$ on the nth game beginning with a capital of k, Prob$\{X_n = 0 | X_0 = k\}$. The gambler has lost everything on the nth game.

b_{kn} = probability of absorption at $x = N$ on the nth game beginning with a capital of k, Prob$\{X_n = N | X_0 = k\}$. The gambler has won everything on the nth game.

Assume that the beginning capital is restricted to $k = 1, \ldots, N-1$. Note that $a_{0n} = 1$, $a_{Nn} = 0$, $b_{0n} = 0$, and $b_{Nn} = 1$ because the games stop when the capital is either zero or N. As a mnemonic device, associate a with the left endpoint ($x = 0$) and b with the right endpoint ($x = N$). Note that $a_{kn} + b_{kn}$ is the probability of absorption at the nth game. Because absorption occurs at either $x = 0$ or $x = N$, $\{a_{kn} + b_{kn}\}_{n=0}^{\infty}$ represents the probability distribution associated with absorption,

$$\sum_{n=0}^{\infty} (a_{kn} + b_{kn}) = 1, \quad 1 \le k \le N-1.$$

Let A_k and B_k be the generating functions of the sequences $\{a_{kn}\}_{n=0}^{\infty}$ and $\{b_{kn}\}_{n=0}^{\infty}$

$$A_k(t) = \sum_{n=0}^{\infty} a_{kn} t^n \quad \text{and} \quad B_k(t) = \sum_{n=0}^{\infty} b_{kn} t^n, \quad |t| \le 1.$$

The functions $A_k(t)$ and $B_k(t)$ by themselves are *not* probability generating functions. However, their sum $A_k(t) + B_k(t) = \sum_{n=0}^{\infty} (a_{kn} + b_{kn}) t^n$ is a *probability generating function.* Define

$$a_k = A_k(1) = \sum_{n=0}^{\infty} a_{kn},$$

$$b_k = B_k(1) = \sum_{n=0}^{\infty} b_{kn},$$

and

$$\tau_k = A'(1) + B'(1) = \sum_{n=0}^{\infty} n(a_{kn} + b_{kn}).$$

Then a_k is the probability of absorption at $x = 0$ or the probability of ruin beginning with a capital of k, and b_k is the probability of absorption at $x = N$ or probability of winning all of the capital beginning with a

capital of k. Finally, τ_k is the expected or mean duration of the games until absorption occurs either at $x = 0$ or at $x = N$, $\tau_k = \mu_{0k} + \mu_{Nk}$. In particular, if T_k denotes the random variable for the time until absorption, then $\tau_k = E(T_k)$. Note that

$$a_k + b_k = 1, \tag{3.1}$$

so that $b_k = 1 - a_k$. Expressions for a_k, τ_k, and a_{kn} will be derived in the next three sections.

3.3.1 Probability of Absorption

An expression for a_k is derived, the probability of absorption at $x = 0$ beginning with a capital of k. A difference equation relating a_{k-1}, a_k, and a_{k+1} is formulated. When boundary conditions are assigned at $k = 0$ and $k = N$, the difference equation can be solved for a_k.

An expression for the probability of ruin a_k on the interval $k \in [0, N]$ can be derived as follows. With a capital of k, the gambler may either win or lose the next game with probabilities p or q, respectively. If the gambler wins, the capital is $k+1$ and the probability of ruin is a_{k+1}. If the gambler loses, the capital is $k-1$ and the probability of ruin is a_{k-1}. This relationship is given by the following difference equation:

$$a_k = pa_{k+1} + qa_{k-1} \tag{3.2}$$

for $1 \leq k \leq N-1$. Equation (3.2) is a second-order difference equation in a_k. Expressed in standard form, the difference equation is

$$pa_{k+1} - a_k + qa_{k-1} = 0. \tag{3.3}$$

This method of deriving equation (3.2) is referred to as a *first-step analysis* (Taylor and Karlin, 1998). In the derivation, we only consider what happens in the next step, then apply the Markov property.

To solve the difference equation (3.3), we need the boundary conditions:

$$a_0 = 1 \quad \text{and} \quad a_N = 0.$$

If the capital is zero, then the probability of ruin equals one, and if the the capital is N, then the probability of ruin equals zero. The difference equation is linear and homogeneous, and the coefficients are constants. This type of difference equation can be solved easily (Elaydi, 1999). We review the method of solution below.

To solve the difference equation, let $a_k = \lambda^k \neq 0$, and substitute this value for a_k into the difference equation. The result is the *characteristic equation*,

$$p\lambda^{k+1} - \lambda^k + q\lambda^{k-1} = 0.$$

Since $\lambda \neq 0$, the characteristic equation simplifies to

$$p\lambda^2 - \lambda + q = 0.$$

The roots of the characteristic equation are the *eigenvalues*

$$\lambda_{1,2} = \frac{1 \pm \sqrt{1 - 4pq}}{2p}.$$

The expression for the eigenvalues can be simplified by noticing that $(p + q)^2 = 1$. Expanding and rearranging terms,

$$\begin{aligned} 1 &= p^2 + 2pq + q^2 \\ 1 - 4pq &= p^2 - 2pq + q^2 = (p - q)^2. \end{aligned}$$

The radical in the expression for $\lambda_{1,2}$ simplifies to $\sqrt{1 - 4pq} = |p - q|$.

The solution to the difference equation (3.3) must be divided into two cases, depending on whether $p \neq q$ or $p = 1/2 = q$. In the first case, $p \neq q$, $\lambda_{1,2} = (1 \pm (p - q))/(2p)$, so that $\lambda_1 = 1$ and $\lambda_2 = q/p$. The general solution is

$$a_k = c_1 + c_2(q/p)^k.$$

The constants c_1 and c_2 are found by applying the boundary conditions: $a_0 = 1 = c_1 + c_2$ and $a_N = 0 = c_1 + c_2(q/p)^N$. Solving for c_1 and c_2 and substituting these values into the general solution yields the following particular solution to the difference equation:

$$a_k = \frac{(q/p)^N - (q/p)^k}{(q/p)^N - 1}, \quad p \neq q. \tag{3.4}$$

Since $a_k + b_k = 1$, the solution for b_k is

$$b_k = \frac{(q/p)^k - 1}{(q/p)^N - 1}, \quad p \neq q.$$

For the second case, $p = 1/2 = q$, note that $1 - 4pq = 0$, so that the characteristic equation has a root of multiplicity two, $\lambda_{1,2} = 1$. The general solution to the difference equation (3.3) is $a_k = c_1 + c_2 k$. Again applying the boundary conditions gives $a_0 = 1 = c_1$ and $a_N = 0 = c_1 + c_2 N$ so that the particular solution is

$$a_k = \frac{N - k}{N}, \quad p = 1/2 = q. \tag{3.5}$$

The solution to b_k is

$$b_k = \frac{k}{N}, \quad p = 1/2 = q.$$

Prob.	a_{50}	b_{50}	τ_{50}	$A'_{50}(1)$	$B'_{50}(1)$
$q = 0.50$	0.5	0.5	2500	1250	1250
$q = 0.51$	0.880825	0.119175	1904	1677	227
$q = 0.55$	0.999956	0.000044	500	499.93	0.07
$q = 0.60$	1.00000	0.00000	250	250	0

Table 3.1. Gambler's ruin problem with a beginning capital of $k = 50$ and a total capital of $N = 100$.

Example 3.1 Suppose that the total capital is $N = 100$ and a gambler has $k = 50$ dollars. Table 3.1 gives values for the probability of losing all of the money, a_{50}, and winning all of the money, b_{50}, for different values of p and q. (Recall that $p = 1 - q$, where p is the probability of winning \$1 in each game.) Values for the expected duration, $\tau_{50} = A'(50) + B'(50)$, are also given in Table 3.1. Their derivation is discussed in Section 3.3.2. ■

Note that as the probability of losing a single game increases (q increases), the probability of ruin a_{50} also increases but a_{50} increases at a much faster rate than q. When $q = 0.6$, the probability of ruin starting with a capital of $k = 50$ is very close to one and the expected number of games until ruin equals 250.

We have applied the theory of difference equations to find general solutions for a_k and b_k. Numerical methods can also be used to find solutions for a_k and b_k. To apply a numerical method, the system of equations is first expressed as a matrix equation. Equations (3.3) can be expressed as the following single matrix equation, $D\mathbf{a} = \mathbf{c}$, where $\mathbf{a} = (a_0, a_1, \ldots, a_N)^T$, $\mathbf{c} = (1, 0, \ldots, 0)^T$ and

$$D = \left(\begin{array}{c|cccccccc|c} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \hline q & -1 & p & 0 & \cdots & 0 & 0 & 0 \\ 0 & q & -1 & p & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & q & -1 & p \\ \hline 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{array}\right) = \begin{pmatrix} 1 & \mathbf{0} & 0 \\ D_1 & D_{N-1} & D_2 \\ 0 & \mathbf{0} & 1 \end{pmatrix}. \tag{3.6}$$

Matrix D is an $(N+1) \times (N+1)$ matrix that can be partitioned into block form according to the three communication classes, $\{0\}$, $\{1, 2, \ldots, N-1\}$, and $\{N\}$.

Note that the first row and last row of $D\mathbf{a} = \mathbf{c}$ give the boundary conditions $a_0 = 1$ and $a_N = 0$. The second row gives $qa_0 - a_1 + pa_2 = 0$, which is equation (3.3) when $k = 1$ and, in general, the kth row of $D\mathbf{a} = \mathbf{c}$ gives the general equation (3.3). Matrix D is nonsingular and a sparse matrix. Therefore, many efficient numerical methods can be applied to solve for $\mathbf{a}$.

A matrix with a form such as D will be seen in many other problems. Therefore, we shall digress briefly to show the properties of D that make it nonsingular.

Definition 3.1. An $n \times n$ matrix $A = (a_{ij})$ is said to be *diagonally dominant* if

$$|a_{ii}| \geq \sum_{j=1, i \neq j}^{n} |a_{ij}|, \quad \text{for } i = 1, 2, \ldots, n. \tag{3.7}$$

Matrix A is said to be *strictly diagonally dominant* if the inequality (3.7) is strict for all i.

Inequality (3.7) states that the absolute value of the diagonal element dominates the sum of the absolute values of all of the off diagonal elements in that row.

Recall the definition of an irreducible matrix from the discussion in Chapter 2, Section 2.3. A square matrix A is irreducible iff its directed graph is strongly connected (Ortega, 1987). The properties of irreducibility and diagonal dominance lead to the following definition.

Definition 3.2. If matrix A is irreducible, diagonally dominant, and the inequality in (3.7) is strict for at least one i, then A is said to be *irreducibly diagonally dominant.*

A result from linear algebra shows that irreducibly diagonally dominant matrices are nonsingular. For a proof of this result, please consult Ortega (1987).

Theorem 3.1. *If an $n \times n$ matrix A is strictly diagonally dominant or irreducibly diagonally dominant, then matrix A is nonsingular.*

In addition, if matrix A is nonsingular, then matrix A^T is nonsingular. The $(N-1) \times (N-1)$ submatrix D_{N-1} of D is irreducibly diagonally dominant. By Theorem 3.1, $\det(D_{N-1}) \neq 0$. But $\det(D) = \det(D_{N-1})$; therefore, D is nonsingular. Theorem 3.1 will be applied to many other matrices that appear in problems associated with the probability of absorption or mean time until absorption.

3.3.2 Expected Time until Absorption

The duration of the games, beginning with a capital of k, lasts until all of the capital N is either gained or lost. The expected duration of the games is

denoted as $\tau_k = E(T_k)$. As was done for the probability of ruin, a difference equation for τ_k can also be derived. We use a first-step analysis. Beginning with a capital of k, the gambler may either win or lose the next game with probabilities p or q, respectively. If the gambler wins, then the capital is $k+1$ and the duration of the games is $1+\tau_{k+1}$ (counting the game just played), and if the gambler loses, then the capital is $k-1$ and the duration of the game is $1+\tau_{k-1}$. The difference equation for τ_k has the following form:

$$\tau_k = p(1+\tau_{k+1}) + q(1+\tau_{k-1}),$$

for $k = 1, 2, \ldots, N-1$. Using the fact that $p+q=1$, the difference equation can be expressed as follows:

$$p\tau_{k+1} - \tau_k + q\tau_{k-1} = -1, \tag{3.8}$$

a second-order linear, nonhomogeneous difference equation with constant coefficients. The boundary conditions satisfy

$$\tau_0 = 0 = \tau_N$$

because if the capital is either zero or N, there can be no more games (absorption has occurred). The difference equation (3.8) can be easily solved in a manner similar to the difference equation for a_k. First, the general solution to the homogeneous difference equation is found, and then a particular solution to the nonhomogeneous is added to the homogeneous solution.

To solve for the homogeneous solution, let $\tau_k = \lambda^k \neq 0$ and substitute this value into the difference equation. The following characteristic equation is obtained:

$$p\lambda^2 - \lambda + q = 0.$$

The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = q/p$, if $p \neq q$ and $\lambda_1 = 1 = \lambda_2$, if $p = 1/2 = q$.

In the first case, $p \neq q$, the general solution to the homogeneous difference equation is $\tau_k = c_1 + c_2(q/p)^k$. To find a particular solution to the nonhomogeneous equation, let $\tau_k = ck$ for an arbitrary constant c and solve for c. Substituting $\tau_k = ck$ into the difference equation gives the following solution for c: $c = 1/(q-p)$. Thus, the general solution to the nonhomogeneous difference equation (3.8) is

$$\tau_k = c_1 + c_2(q/p)^k + \frac{k}{q-p}.$$

Next, applying the boundary conditions, the constants c_1 and c_2 can be found, $\tau_0 = 0 = c_1 + c_2$ and $\tau_N = 0 = c_1 + c_2(q/p)^N + N/(q-p)$; then $c_1 = -N/(q-p)\left[1/(1-(q/p)^N)\right]$ and $c_2 = -c_1$ so that

$$\begin{aligned}\tau_k &= \frac{k}{q-p} - \frac{N}{q-p}\left[\frac{1-(q/p)^k}{1-(q/p)^N}\right] \\ &= \frac{1}{q-p}\left[k - N\left(\frac{1-(q/p)^k}{1-(q/p)^N}\right)\right], \quad p \neq q. \end{aligned} \tag{3.9}$$

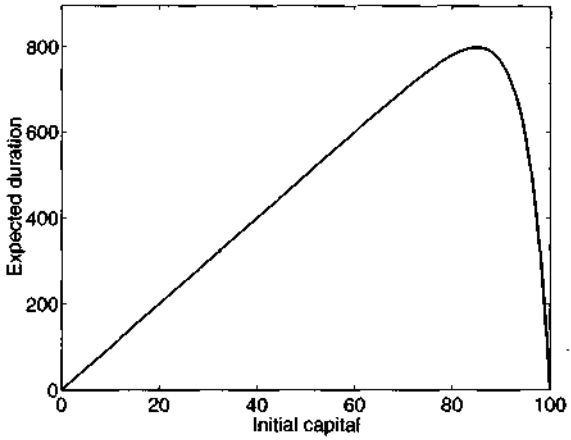

Figure 3.1. Expected duration of the games, τ_k for $k = 0, 1, 2, \ldots, 100$, when $q = 0.55$.

A similar method is applied in the second case $p = 1/2 = q$. The general solution to the homogeneous difference equation (3.8) is $\tau_k = c_1 + c_2 k$ and a particular solution has the form ck^2. Substituting ck^2 into the difference yields $c = -1$, so that the general solution to the nonhomogeneous equation is $\tau_k = c_1 + c_2 k - k^2$. Solving for c_1 and c_2 gives $c_1 = 0$ and $c_2 = N$. The solution to (3.8) in the case $p = 1/2 = q$ is

$$\tau_k = k(N - k), \quad p = 1/2 = q. \tag{3.10}$$

Example 3.2 Suppose $N = 100$ and $p = 1/2 = q$. Then, applying the formula (3.10), $\tau_{50} = 2500$ (see Table 3.1). But when $p = 0.45$ and $q = 0.55$, $\tau_{50} \approx 500$. If the capital is increased to $N = 1000$, then for $p = 1/2 = q$, $\tau_{500} = 250,000$. For this same capital, but $p = 0.45$ and $q = 0.55$, $\tau_{500} = 5000$. The duration of the game increases significantly when the capital increases. ■

Figure 3.1 graphs the expected duration, τ_k, $k = 0, 1, 2, \ldots, N$, when $N = 100$ and $q = 0.55$. It can be seen from Figure 3.1 that although $\tau_{50} \approx 500$, the maximum duration is approximately 800 games at an initial capital of $k = 85$ dollars.

As was demonstrated for the probabilities of absorption, a numerical method can also be applied to find the expected duration of the games. The equations (3.8) can be expressed as a single matrix equation,

$$D\tau = \mathbf{d},$$

where D is defined by (3.6) and $d = (0, -1, -1, \ldots, -1, 0)^T$. The solution τ, the expected duration of the games is given by $\tau = D^{-1}\mathbf{d}$ (see the MATLAB program in the Appendix for Chapter 3).

3.3.3 Probability Distribution for Absorption

In the previous section, a formula for the expected duration until absorption was derived. To calculate the entire probability distribution until absorption at either $x = 0$ or at $x = N$, expressions for a_{kn} and b_{kn} are required. The probability distribution until absorption is

$$\{a_{kn} + b_{kn}\}_{n=0}^{\infty}$$

with generating function

$$A_k(t) + B_k(t) = \sum_{n=0}^{\infty}(a_{kn} + b_{kn})t^n, \quad |t| \leq 1.$$

Closed form expressions can be derived for $A_k(t)$ and $B_k(t)$, but they do not give explicit expressions for a_{kn} and b_{kn}. However, assuming $A_k(t)$ and $B_k(t)$ have Maclaurin series expansions in t, these coefficients can be obtained by repeated differentiation; for example,

$$a_{kn} = \frac{1}{n!}\frac{d^n A_k(t)}{dt^n}\bigg|_{t=0}. \tag{3.11}$$

A difference equation for A_k is derived using a first-step analysis similar to the analysis used to derive expressions for a_k and τ_k in the previous sections. First, a difference equation is derived for a_{kn}. If absorption at $x = 0$ occurs in $n + 1$ steps from an initial capital of k, then in the next game if the gambler wins, the capital is $k+1$ and absorption will occur with n more games. Otherwise, if the gambler loses the next game, the capital is $k - 1$ and absorption will occur with n more games; that is,

$$a_{k,n+1} = pa_{k+1,n} + qa_{k-1,n}, \quad k \geq 1, \ n \geq 0.$$

The above equation is a partial difference equation, the difference equation version of a partial differential equation, since a_{kn} is a function of two variables, k and n. The boundary and initial conditions for this system of difference equations are

$$a_{0n} = 0 = a_{Nn}, \quad n = 1, 2, \ldots, \quad a_{00} = 1, \quad \text{and} \quad a_{k0} = 0,$$

for $k = 1, 2, \ldots, N - 1$. These conditions follow because if the beginning capital is zero, absorption has already occurred and no games are required, $a_{00} = 1$, and absorption cannot occur in $n > 0$ steps, $a_{0n} = 0$. In addition, if the gambler has all of the capital, his/her opponent has already lost and no games are required $a_{N0} = 0$ and absorption cannot occur in $n > 0$ steps, a_{Nn}. Finally, it takes at least k games to be ruined beginning with a capital of k, $a_{kn} = 0$ for $n < k$. These conditions give rise to simple expressions for the generating functions, $A_0(t)$ and $A_N(t)$:

$$A_0(t) = a_{00} + a_{01}t + a_{02}t^2 + \cdots = 1$$

and

$$A_N(t) = a_{N0} + a_{N1}t + a_{N2}t^2 + \cdots = 0.$$

To obtain a difference equation for A_k, the equation in a_{kn} is multiplied by t^{n+1} and summed from $n = 0$ to $n = \infty$. For $k \geq 1$, it follows that

$$\sum_{n=0}^{\infty} a_{k,n+1}t^{n+1} = \sum_{n=0}^{\infty} pa_{k+1,n}t^{n+1} + \sum_{n=0}^{\infty} qa_{k-1,n}t^{n+1}.$$

Because $a_{k0} = 0$ for $k \geq 1$, the above equation can be simplified:

$$\begin{aligned} \sum_{n=0}^{\infty} a_{kn}t^n &= pt\sum_{n=0}^{\infty} a_{k+1,n}t^n + qt\sum_{n=0}^{\infty} a_{k-1,n}t^n \\ A_k(t) &= ptA_{k+1}(t) + qtA_{k-1}(t). \end{aligned}$$

For t fixed, $0 < t < 1$, the difference equation can be solved subject to the following boundary conditions:

$$A_0(t) = 1 \quad \text{and} \quad A_N(t) = 0.$$

Let $A_k(t) = \lambda^k \neq 0$ so that the characteristic equation is

$$pt\lambda^2 - \lambda + qt = 0.$$

The eigenvalues satisfy

$$\lambda_{1,2} = \frac{1 \pm \sqrt{1 - 4pqt^2}}{2pt},$$

where $\lambda_1 > \lambda_2 > 0$. The two roots are real and distinct. The general solution is $A_k(t) = c_1\lambda_1^k + c_2\lambda_2^k$. The constants c_1 and c_2 are found by applying the boundary conditions, $c_1 + c_2 = 1$ and $c_1\lambda_1^N + c_2\lambda_2^N = 0$, so that

$$A_k(t) = \frac{\lambda_1^N\lambda_2^k - \lambda_2^N\lambda_1^k}{\lambda_1^N - \lambda_2^N}. \tag{3.12}$$

A closed form expression for $B_k(t)$ can be obtained in a similar manner. For example, $B_k(t)$ satisfies the same difference equation as $A_k(t)$,

$$B_k(t) = ptB_{k+1}(t) + qtB_{k-1}(t)$$

for $k = 1, 2, \ldots, N-1$, but the boundary conditions differ,

$$B_0(t) = 0 \quad \text{and} \quad B_N(t) = 1.$$

The solution $B_k(t)$ is

$$B_k(t) = \frac{\lambda_1^k - \lambda_2^k}{\lambda_1^N - \lambda_2^N}.$$

Thus, the p.g.f. for the duration of the games is given by $A_k(t) + B_k(t)$. Although the formula was derived for t restricted to $0 < t < 1$, if the p.g.f. is expressed as a Maclaurin series over this region,

$$A_k(t) + B_k(t) = \sum_{n=0}^{\infty} (a_{kn} + b_{kn}) t^n,$$

then the series is absolutely convergent for $|t| < 1$ and all of its derivatives exist for $|t| < 1$ (Wade, 1995). Abel's convergence theorem (Chapter 2) can be applied to the series and the derivatives of the series to extend the domain to $t = 1$. For example, the derivative of the series satisfies

$$A_k'(t) + B_k'(t) = \sum_{n=1}^{\infty} n(a_{kn} + b_{kn}) t^{n-1},$$

which is finite for $|t| < 1$ and $n(a_{kn} + b_{kn}) \geq 0$. Therefore,

$$\lim_{t \to 1^-} \left[\sum_{n=1}^{\infty} n(a_{kn} + b_{kn}) t^{n-1} \right] = L \leq \infty.$$

By Abel's convergence theorem, $\sum_{n=1}^{\infty} n(a_{kn} + b_{kn}) = L$. Therefore, the expected duration of the games, τ_k, can be derived from the generating function,

$$\tau_k = A_k'(1) + B_k'(1) = \lim_{t \to 1^-} [A_k'(t) + B_k'(t)].$$

Fortunately, the derivation in the previous section has already provided a solution for τ_k. Bailey (1990) derives a general formula for a_{kn} using a partial fraction expansion. For specific values of N, k, and n, it is an easy task to use a computer algebra system such as *Maple* to compute a_{kn}. This is demonstrated in the next example.

Example 3.3 Suppose $N = 10$ and $k = 5$. Then $A_5(t)$ can be computed from (3.12) so that

$$A_5(t) = \frac{q^5 t}{1 - 5pqt^2 + 5p^2q^2t^4}.$$

Through differentiation and application of (3.11), the values of a_{5n} for $n = 5, 6, \ldots$, can be easily computed. The computer algebra system *Maple* was used to compute a_{5n} for $n = 5, 7, 9, 11$,

$$a_{55} = q^5, \; a_{57} = 5q^6p, \; a_{59} = 20q^7p^2, \; a_{5,11} = 75q^8p^3.$$

It can be easily seen that $a_{5n} = 0$ for $n < 5$ and for n even. ■

```
> N:=10:  k:=5:  dAk:=n->simplify(diff(Ak,t$n)):
> ak:=n->limit(dAk(n),t=0)/n!:
> ak(5);
```

$$q^5$$

```
> ak(6);
```

$$0$$

```
> ak(7);
```

$$5q^6p$$

```
> ak(9);
```

$$20q^7p^2$$

```
> ak(11);
```

$$75q^8p^3$$

Example 3.4 Assume $N = 100$, $k = 50$, and $q = 0.55$. The expected duration of the games is $\tau_{50} \approx 500$ and the standard deviation is $\sigma_{50} \approx 222.4$. Approximations to these values were calculated from 1000 sample paths,

$$\text{mean} \approx 482.74, \quad \text{standard deviation} \approx 203.27.$$

The mean and variance of the distribution for the duration of the games, τ_{50} and σ_{50}^2, can be calculated directly from the p.g.f. $S_k(t) = A_k(t) + B_k(t)$ when $k = 50$,

$$\tau_{50} = S'_{50}(1) \quad \text{and} \quad \sigma_{50}^2 = S''_{50}(1) + S'_{50}(1) - \left[S'_{50}(1)\right]^2.$$

Three sample paths are graphed for this stochastic process in Figure 3.2. ■

3.4 Gambler's Ruin Problem on a Semi-Infinite Domain

The gambler's ruin problem can be extended to a semi-infinite domain $\{0, 1, 2 \ldots\}$, where it is assumed that there is an unlimited amount of capital, $N \to \infty$. This problem is related to a birth and death process, where the population size is unlimited. The general solution to $A_k(t)$ has the same form as in the finite domain:

$$A_k(t) = c_1\lambda_1^k + c_2\lambda_2^k.$$

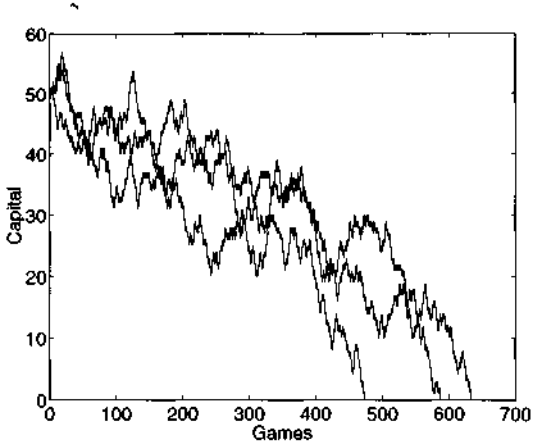

Figure 3.2. Three sample paths for the gambler's ruin problem when $N = 100$, $k = 50$, and $q = 0.55$

However, one of the boundary conditions changes,

$$A_0(t) = 1 \quad \text{and} \quad \lim_{k\to\infty} |A_k(t)| < \infty.$$

The boundedness condition must be satisfied because

$$|A_k(t)| \le \sum_{n=0}^{\infty} a_{kn} = a_k \le 1$$

for $|t| \le 1$ and all k.

Recall that the two eigenvalues λ_1 and λ_2 satisfy

$$|\lambda_1| = \frac{1 + \sqrt{1 - 4pqt^2}}{2p|t|} \quad \text{and} \quad |\lambda_2| = \frac{1 - \sqrt{1 - 4pqt^2}}{2p|t|} \quad \text{for} \quad |t| \le 1.$$

In addition, the absolute values of the eigenvalues $|\lambda_i|$, $i = 1, 2$, are roots of the characteristic equation, $f(\lambda) = p|t|\lambda^2 - \lambda + q|t| = 0$. Because the graph of f is a parabola satisfying $f(0) = q|t| > 0$, $f(1) = |t| - 1 < 0$ for $0 < |t| < 1$, and $\lim_{\lambda\to\infty} f(\lambda) = \infty$, it follows that the two roots satisfy $0 < |\lambda_2| < 1 < |\lambda_1|$. Thus, the coefficient $c_1 = 0$, $c_2 = 1$, and

$$A_k(t) = \lambda_2^k = \left(\frac{1 - \sqrt{1 - 4pqt^2}}{2pt}\right)^k.$$

The probability of ruin or of absorption at $x = 0$, a_k, can be calculated directly from $A_k(t)$:

$$a_k = A_k(1) = \begin{cases} 1, & \text{if } p \le q, \\ \left(\dfrac{q}{p}\right)^k, & \text{if } p > q. \end{cases} \tag{3.13}$$

Also, for $p \leq q$, the sequence $\{a_{kn}\}_{n=0}^{\infty}$ is a probability distribution and the expected duration until absorption at $x = 0$ can be computed. It is easy to see that for $p > q$ the expected duration is infinite. In the other cases, τ_k is computed from $A_k'(1)$:

$$\tau_k = \begin{cases} \dfrac{k}{q-p}, & \text{if } p < q, \\ \infty, & \text{if } p \geq q. \end{cases} \tag{3.14}$$

The probability of absorption in n time steps, a_{kn}, can be calculated by applying formula (3.11).

3.5 General Birth and Death Process

A general birth and death process is formulated as a discrete time Markov chain. The Markov chain model is related to the gambler's ruin problem, but the probability of a birth (or winning) is not constant but depends on the size of the population and the probability of a death (or losing) also depends on the population size. To define a birth and death process, let X_n, $n = 0, 1, 2, \ldots$, denote the size of the population, where the state space may be either finite or infinite, $\{0, 1, 2, \ldots, N\}$ or $\{0, 1, 2, \ldots\}$, and N is the maximal population size. The birth and death probabilities are b_i and d_i, respectively. In addition, $b_0 = 0 = d_0$, $b_i > 0$ and $d_i > 0$ for $i = 1, 2, \ldots$, except in the finite case, where $b_N = 0$. It is assumed that the time interval, $n \to n + 1$, is sufficiently small such that during this time interval at most one event occurs, either a birth or a death. Assume the transition probabilities satisfy

$$\begin{aligned} p_{ji} &= \text{Prob}\{X_{n+1} = j | X_n = i\} \\ &= \begin{cases} b_i, & \text{if } j = i + 1, \\ d_i, & \text{if } j = i - 1, \\ 1 - (b_i + d_i), & \text{if } j = i, \\ 0, & \text{if } j \neq i - 1, i, i + 1, \end{cases} \end{aligned}$$

for $i = 1, 2, \ldots$, $p_{00} = 1$, and $p_{j0} = 0$ for $j \neq 0$. In the case of a finite state space, where N is the maximal population size, $p_{N+1,N} = b_N = 0$.

The transition matrix P for the finite Markov chain has the following form:

$$P = \begin{pmatrix} 1 & d_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1-(b_1+d_1) & d_2 & 0 & \cdots & 0 & 0 \\ 0 & b_1 & 1-(b_2+d_2) & d_3 & \cdots & 0 & 0 \\ 0 & 0 & b_2 & 1-(b_3+d_3) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1-(b_{N-1}+d_{N-1}) & d_N \\ 0 & 0 & 0 & 0 & \cdots & b_{N-1} & 1-d_N \end{pmatrix}. \tag{3.15}$$

To ensure P is a stochastic matrix, it is assumed that

$$\sup_{i\in\{1,2,\ldots\}} \{b_i + d_i\} \leq 1.$$

During each time interval, n to $n+1$, either the population size increases by one, decreases by one, or stays the same size. This is a reasonable assumption only if the time step is sufficiently small.

There are two communication classes, $\{0\}$ and $\{0, 1, \ldots, N\}$ in the finite case. It is easy to see that zero is positive recurrent; all other states are transient. There exists a unique stationary probability distribution π, $P\pi = \pi$, where $\pi_0 = 1$ and $\pi_i = 0$ for $i = 1, 2, \ldots$. In the case of a finite Markov chain, it can be easily shown that eventually population extinction occurs from any initial state,

$$\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = \lim_{n\to\infty} p_0(n) = 1.$$

In the notation of Section 3.3.1, $a_k = 1$ for $k = 0, 1, 2, \ldots, N$.

3.5.1 Expected Time to Extinction

The techniques from Section 3.3.2 are used to find expressions for the expected time until population extinction. It is assumed that $\lim_{n\to\infty} p_0(n) = 1$. Let τ_k denote the expected time until extinction for a population with initial size k. Then $\tau_0 = 0$ and the following relationship holds for τ_k, $k = 1, 2, \ldots$:

$$\tau_k = b_k(1 + \tau_{k+1}) + d_k(1 + \tau_{k-1}) + (1 - (b_k + d_k))(1 + \tau_k). \tag{3.16}$$

If the maximal population size is finite, then for $k = N$, $\tau_N = d_N(1 + \tau_{N-1}) + (1 - d_N)(1 + \tau_N)$. The difference equation can be simplified as follows:

$$d_k\tau_{k-1} - (b_k + d_k)\tau_k + b_k\tau_{k+1} = -1, \tag{3.17}$$

$k = 1, 2, \ldots$. If $k = N$, then $d_N\tau_{N-1} - d_N\tau_N = -1$. Because the coefficients for these difference equations are not constant, the same techniques cannot be employed as in the previous sections. However, when the maximal population size is finite, then these difference equations can be expressed as a matrix equation, $D\tau = \mathbf{c}$, where $\tau = (\tau_0, \tau_1, \ldots, \tau_N)^T$, $\mathbf{c} = (0, -1, \ldots, -1)^T$ and

$$\begin{aligned} D &= \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ d_1 & -b_1 - d_1 & b_1 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & -b_2 - d_2 & b_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & d_N & -d_N \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0} \\ D_1 & D_N \end{pmatrix}. \end{aligned}$$

Submatrix D_N of D is irreducibly diagonally dominant, and by Theorem 3.1 $\det(D_N) \neq 0$. But $\det(D) = \det(D_N)$. Thus, D is nonsingular and the solution to the expected time until extinction is

$$\tau = D^{-1}\mathbf{c}. \tag{3.18}$$

The solution τ can be found using numerical methods. However, because the matrix D is tridiagonal, simple recursion relations can be applied to obtain explicit formulas for the τ_k, $k = 1, 2, \ldots, N$. This formula is given in the next theorem. See Nisbet and Gurney (1982).

Theorem 3.2. *Suppose* $\{X_n\}$, $n = 0, 1, 2, \ldots, N$, *is a general birth and death process with* $X_0 = m \geq 1$ *satisfying* $b_0 = 0 = d_0$, $b_i > 0$ *for* $i = 1, 2, \ldots, N-1$, *and* $d_i > 0$ *for* $i = 1, 2, \ldots, N$. *The expected time until population extinction satisfies*

$$\tau_m = \begin{cases} \dfrac{1}{d_1} + \displaystyle\sum_{i=2}^{N} \frac{b_1 \cdots b_{i-1}}{d_1 \cdots d_i}, & m = 1 \\ \tau_1 + \displaystyle\sum_{s=1}^{m-1} \left[\frac{d_1 \cdots d_s}{b_1 \cdots b_s} \sum_{i=s+1}^{N} \frac{b_1 \cdots b_{i-1}}{d_1 \cdots d_i} \right], & m = 2, \ldots, N. \end{cases} \tag{3.19}$$

Proof. For $k = 1, 2, \ldots, N-1$, the equations (3.17) are solved recursively for $\tau_2, \ldots, \tau_N$ to obtain the formulas

$$\tau_m = \tau_1 + \sum_{k=1}^{m-1} \frac{d_1 \cdots d_k}{b_1 \cdots d_k} \left[\tau_1 - \frac{1}{d_1} - \sum_{i=2}^{k} \frac{b_1 \cdots b_{i-1}}{d_1 \cdots d_i} \right] \tag{3.20}$$

for $m = 2, \ldots, N$. The second summation is zero when $k < 2$. Then applying the relation for $k = N$,

$$\tau_N = \frac{1}{d_N} + \tau_{N-1}$$

and equating the two values for τ_N, the following formula for τ_1 is obtained:

$$\tau_1 = \frac{1}{d_1} + \sum_{i=2}^{N} \frac{b_1 \cdots b_{i-1}}{d_1 \cdots d_i}.$$

Substituting τ_1 into (3.20), the formula for (3.19) follows. □

Example 3.5 Suppose the maximal population size is $N = 20$ in a birth and death process, where $b_i \equiv bi$, for $i = 1, 2, \ldots, 19$, $d_i \equiv di$, for $i = 1, 2, \ldots, 20$, b and d are constants. This process is often referred to as a *simple* birth and death process. When $b > d$, there is population growth, and when $b < d$, there is population decline. Three cases are considered: (i) $b = 0.02 < 0.03 = d$, (ii) $b = 0.025 = d$, and (iii) $b = 0.03 > 0.02 = d$. The expected time until population extinction $\tau = (\tau_0, \ldots, \tau_{20})^T$ is plotted

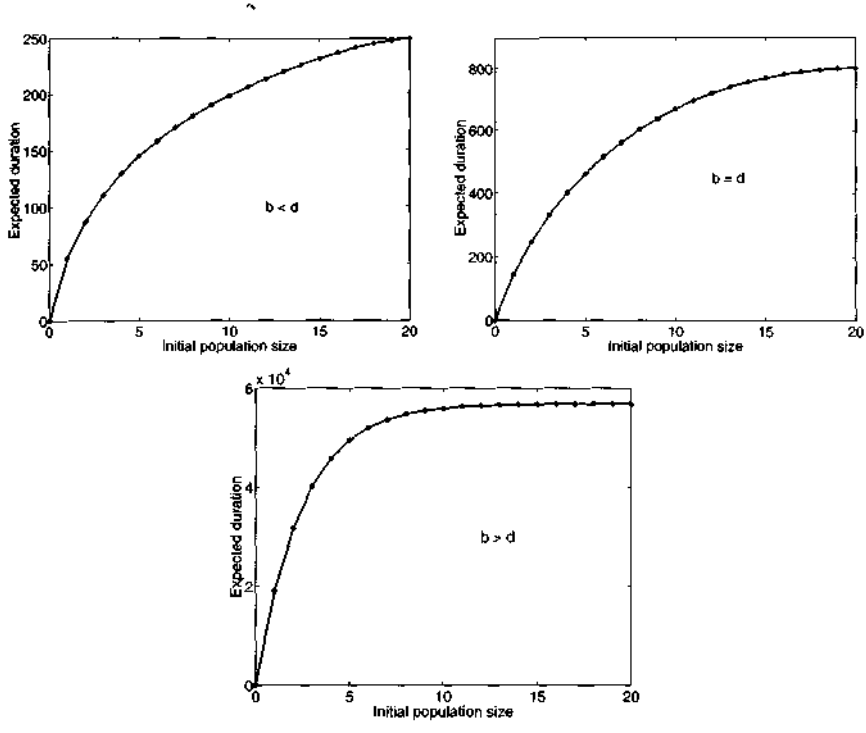

Figure 3.3. Expected time until population extinction τ when the maximal population size is $N = 20$ and $b = 0.02 < 0.03 = d$, $b = 0.025 = d$, or $b = 0.03 > 0.02 = d$.

in each of these three cases in Figure 3.3. The formula in Theorem 3.2 can be applied or τ can be found using a numerical method, $\tau = D^{-1}\mathbf{c}$. Note how much greater the expected duration is for $b > d$ than for $b < d$. Three sample paths of the simple birth and death process in Example 3.5 are graphed in the case $b = 0.025 = d$ in Figure 3.4. ■

Birth and death processes that are continuous in time will be discussed in more detail in Chapter 6. For continuous time processes, the assumption of at most one event occurring during a time period t to $t + \Delta t$ is more reasonable than in the discrete time process, because in the continuous time process, we let $\Delta t \to 0$.

3.6 Logistic Growth Process

In this section, assumptions are made on the general birth and death probabilities b_i and d_i so that the process has a logistic form. Recall that in the deterministic logistic model, if $y(t)$ is the population size at time t, then

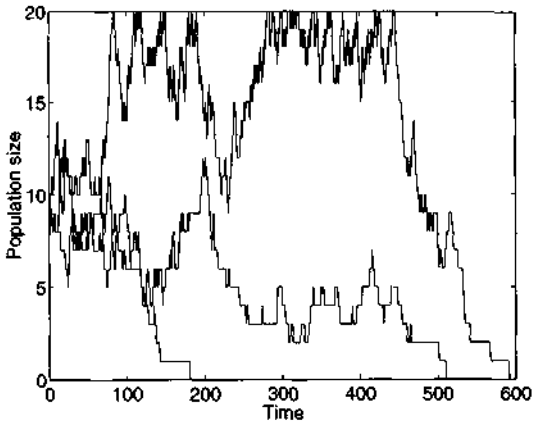

Figure 3.4. Three sample paths for the simple birth and death process in the case $b = 0.025 = d$ with maximal population size $N = 20$ and $X_0 = 10$.

the rate of change of $y(t)$ satisfies

$$\frac{dy}{dt} = ry\left(1 - \frac{y}{K}\right), \quad y(0) = y_0 > 0.$$

The right-hand side of the differential equation is a quadratic function of y and equals the birth rate minus death rate. The parameter r is the intrinsic growth rate and K is the carrying capacity. It is well known that the unique solution $y(t)$ to this differential equation satisfies $\lim_{t\to\infty} y(t) = K$. The population size approaches the carrying capacity.

For a logistic growth process, we make the assumption that

$$b_i - d_i = ri(1 - i/K), \tag{3.21}$$

for $i = 0, 1, 2, \ldots, N$, where $N > K$. Note that the birth probability equals the death probability when the population size is zero or when the population size is at carrying capacity K. This agrees with the logistic deterministic model. Due to the relationship (3.21), it is reasonable to assume that b_i and d_i are either linear or quadratic functions of i.

As demonstrated in the previous section, the expected time until population extinction can be calculated analytically (Theorem 3.2) or numerically $\tau = D^{-1}\mathbf{c}$, where $\tau = (\tau_0, \tau_1, \ldots, \tau_N)^T$ and τ_k is the expected time until extinction for a population with initial size k.

Two cases for the birth and death probabilities b_i and d_i are considered:

(a) $b_i = r\left(i - \dfrac{i^2}{2K}\right)$ and $d_i = r\dfrac{i^2}{2K}$, $\quad i = 0, 1, 2, \ldots, 2K$

(b) $b_i = \begin{cases} ri, & i = 0, 1, 2, \ldots, N-1 \\ 0, & i \geq N \end{cases}$ and $d_i = r\dfrac{i^2}{K}$, $\quad i = 0, 1, \ldots, N$

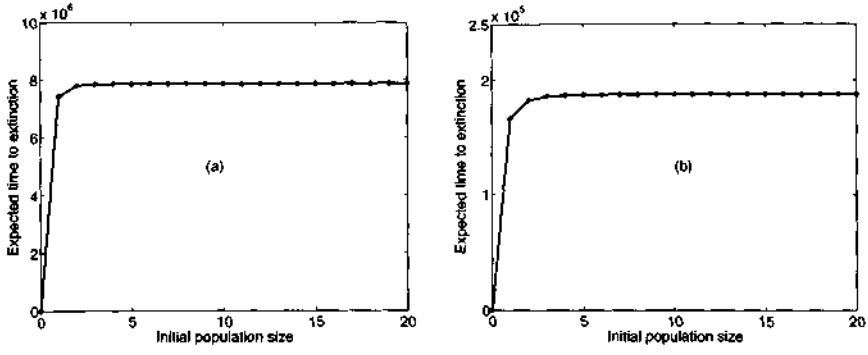

Figure 3.5. Expected time until population extinction when the birth and death rates satisfy (a) and (b) and the parameters are $r = 0.015$, $K = 10$, and $N = 20$.

In case (a), the maximal population size is $N = 2K$. Also, the birth probability increases when the population size is less than K but decreases when the population size is greater than K so that it equals zero at the maximal population size $N = 2K$. In case (a), the death probability is an increasing function of the population size. In case (b), both birth and death probabilities increase as the population size increases.

Example 3.6 For the two cases defined above, the expected time until population extinction is calculated. Let $r = 0.015$, $K = 10$, and $N = 20$. A MATLAB program was used to calculate D^{-1} and $\tau = D^{-1}\mathbf{c}$. Graphs of the expected time to extinction in cases (a) and (b) are given in Figure 3.5. Note how much longer the population persists in case (a). ■

At this point, we should wonder why the logistic growth process appears to behave so much differently from the deterministic model. Why doesn't the population size approach the carrying capacity K? If, in the previous example, the maximal population size is doubled, $N = 40$ and $K = 20$ ($r = 0.0075$), we can check that the expected time to extinction increases to 7.7×10^{12} in case (a) and 3.6×10^9 in case (b). For large N, the stochastic and deterministic logistic models have better agreement because prior to extinction (which may take a long time), the distribution is approximately stationary for a long period of time (quasistationary distribution). This can be seen in Figure 3.6, where the probability distribution $p(n) = (p_0(n), p_1(n), \ldots, p_{20}(n))^T$ is graphed as a function of time $n = 0, 1, 2, \ldots, 1000$. An approximate stationary distribution has been reached by $n = 500$. The MATLAB program that generated Figure 3.6(a) is given in the Appendix for Chapter 3.

The logistic growth process has no positive stationary distribution, $\pi > 0$ (it is not irreducible). The unique stationary distribution corresponding to the simple birth and death process and the logistic growth process is $\pi = (1, 0, 0, \ldots, 0)^T$. However, if the time to extinction is sufficiently long,

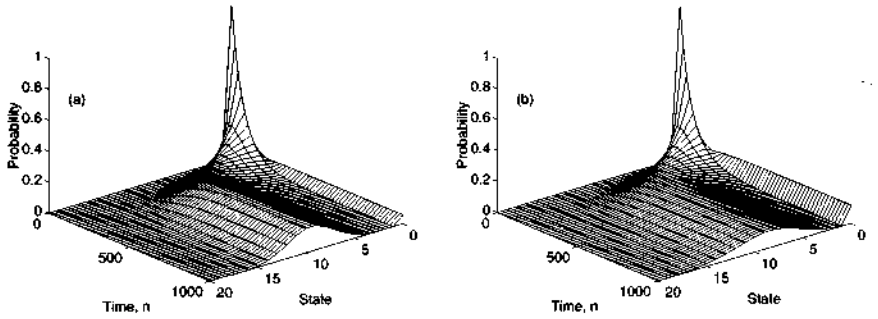

Figure 3.6. Probability distribution for the stochastic logistic model in cases (a) and (b) when $r = 0.015$, $K = 10$, $N = 20$, and $X_0 = 1$.

the process approaches a quasistationary probability distribution, the distribution conditioned on nonextinction. For large K and N, it will be seen that the mean of this quasistationary distribution is close to K (see Figure 3.6).

3.7 Quasistationary Probability Distribution

When the expected duration until absorption is large, it is reasonable to examine the dynamics of the process prior to absorption. Let $\{X_n\}$ for $n = 0, 1, 2, \ldots$ denote a general birth and death process with $p_i(n) = \text{Prob}\{X_n = i\}$, $i = 0, 1, 2, \ldots, N$. Define the conditional probability,

$$\begin{aligned} q_i(n) &= \text{Prob}\{X_n = i | X_j \neq 0, j = 0, 1, 2, \ldots, n-1\} \\ &= \frac{p_i(n)}{1 - p_0(n)} \end{aligned}$$

for $i = 1, 2, \ldots, N$. The distribution $q(n) = (q_1(n), q_2(n), \ldots, q_N(n))^T$ defines a probability distribution because

$$\sum_{i=1}^{N} q_i(n) = \frac{\sum_{i=1}^{N} p_i(n)}{1 - p_0(n)} = \frac{1 - p_0(n)}{1 - p_0(n)} = 1.$$

It is a conditional probability distribution. The probability $q_i(n)$ is conditioned on the population size not hitting zero by time n (i.e., conditional on nonextinction). Let this conditional discrete time Markov chain be denoted as $\{Q_n\}$, where Q_n is the random variable for the population size at time n conditional on nonextinction; $q_i(n) = \text{Prob}\{Q_n = i\}$. The stationary probability distribution for this process is denoted as q^*; q^* is referred to as the *quasistationary probability distribution* or *quasiequilibrium probability distribution.*

Difference equations satisfied by $q_i(n)$ can be derived based on those for $p_i(n)$ [i.e., $p(n+1) = Pp(n)$]. From these difference equations the quasistationary probability distribution q^* can be determined. It will be seen that q^* cannot be calculated by a direct method but by an indirect method, an iterative scheme. An approximation to the process $\{Q_n\}$ yields an irreducible, positive recurrent, aperiodic Markov chain, $\{\tilde{Q}_n\}$, with associated probability distribution $\tilde{q}(n)$. For this new process, a transition matrix, $\tilde{P}$, and the limiting positive stationary probability distribution $\tilde{q}^*$ can be defined. The stationary probability distribution $\tilde{q}^*$ is an approximation for the quasistationary probability distribution q^*.

Difference equations for $q_i(n+1)$ are derived from the identity $p(n+1) = Pp(n)$, where transition matrix P is defined in (3.15). Note that

$$\begin{aligned} q_i(n+1) &= \frac{p_i(n+1)}{1-p_0(n+1)} \\ &= \left(\frac{p_i(n+1)}{1-p_0(n)}\right)\left(\frac{1-p_0(n)}{1-p_0(n+1)}\right) \\ &= \left(\frac{p_i(n+1)}{1-p_0(n)}\right)\left(\frac{1-p_0(n)}{1-p_0(n)-d_1p_1(n)}\right) \end{aligned}$$

or

$$q_i(n+1)(1-d_1q_1(n)) = \left(\frac{p_i(n+1)}{1-p_0(n)}\right).$$

Using the identity for $p_i(n+1)$, the following relation is obtained:

$$q_i(n+1)[1-d_1q_1(n)] = b_{i-1}q_{i-1}(n)+(1-b_i-d_i)q_i(n)+d_{i+1}q_{i+1}(n) \quad (3.22)$$

for $i = 1, 2, \ldots, N$, $b_0 = 0$, and $q_i(n) = 0$ for $i \notin \{1, 2, \ldots, N\}$. It is similar to the difference equation satisfied by $p_i(n)$ except for an additional factor multiplying $q_i(n+1)$. An analytical solution to the stationary solution q^* cannot be found directly from these equations since the coefficients depend on n, but q^* can be found via a numerical iterative method (Nåsell, 1999, 2001).

To approximate the quasistationary probability distribution, q^*, the process $\{Q_n\}$ is approximated by assuming $d_1 = 0$. Equivalently, when the population size reduces to one, it is assumed that the probability of dying is zero. This is a reasonable assumption when $d_1 \approx 0$. With this assumption, equation (3.22) simplifies to

$$\tilde{q}_i(n+1) = b_{i-1}\tilde{q}_{i-1}(n) + (1-b_i-d_i)\tilde{q}_i(n) + d_{i+1}\tilde{q}_{i+1}(n),$$

$i = 2, \ldots, N-1$, $\tilde{q}_1(n+1) = (1-b_1)\tilde{q}_1(n) + d_2\tilde{q}_2(n)$, and $\tilde{q}_N(n+1) = b_{N-1}\tilde{q}_{N-1}(n) + (1-d_N)\tilde{q}_N(n)$. The new transition matrix corresponding

to this approximation satisfies

$$\tilde{P} = \begin{pmatrix} 1-b_1 & d_2 & \cdots & 0 & 0 \\ b_1 & 1-(b_2+d_2) & \cdots & 0 & 0 \\ 0 & b_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1-(b_{N-1}+d_{N-1}) & d_N \\ 0 & 0 & \cdots & b_{N-1} & 1-d_N \end{pmatrix}.$$

Note that $\tilde{P}$ is a submatrix of matrix P given in (3.15), where the first column and first row are deleted and $d_1 = 0$. The discrete time Markov chain $\{\tilde{Q}_n\}$, described by $\tilde{q}(n+1) = \tilde{P}\tilde{q}(n)$, is ergodic (irreducible, positive recurrent, and aperiodic) and has a unique stationary probability distribution, $\tilde{q}^*$, $\tilde{P}\tilde{q}^* = \tilde{q}^*$. It can be shown that $\tilde{q}^* = (\tilde{q}_1^*, \tilde{q}_2^*, \ldots, \tilde{q}_N^*)^T$ satisfies

$$\tilde{q}_{i+1}^* = \frac{b_i \cdots b_1}{d_{i+1} \cdots d_2} \tilde{q}_1^* \quad \text{and} \quad \sum_{i=1}^{N} \tilde{q}_i^* = 1. \tag{3.23}$$

Example 3.7 The approximate quasistationary probability distribution, $\tilde{q}^*$, is compared to the quasistationary probability distribution q^* when $r = 0.015$, $K = 10$, and $N = 20$ in cases (a) and (b) in Figure 3.7. Both distributions have good agreement for $N = 20$, but when $N = 10$ and $K = 5$, then the two distributions differ, especially for values near zero [Figure 3.7(c)].

The means and standard deviations for the quasistationary distribution q^*, which are graphed in Figures 3.7(a) and (b), are

(a) $\mu_{q^*} = 9.435$ and $\sigma_{q^*} = 2.309$

(b) $\mu_{q^*} = 8.848$ and $\sigma_{q^*} = 3.171$. ■

Some of the differences and similarities between the deterministic and stochastic models have been illustrated in the previous examples. The estimate for the expected time until population extinction, τ_k, and the probability of population extinction, $\lim_{n\to\infty} p_0(n) = 1$, are important in stochastic theory but they have no counterpart in deterministic theory. For large population sizes, K and N large, and initial conditions sufficiently large, the deterministic model agrees much better with the stochastic model than for K and N small. This can be seen in the stochastic logistic model graphed in Figure 3.8. The shape of the probability distribution over time is similar to that of the solution to the logistic differential equation, $dy/dt = ry(1 - y/K)$. The quasistationary probability distribution shows a mean close to $K = 50$. The mean of the stochastic process and the solution to the logistic differential equation show close agreement in Figure 3.8. Formulation and analysis of deterministic and stochastic models provide us with a greater understanding of and appreciation for the modeling process and of the underlying biological phenomena.

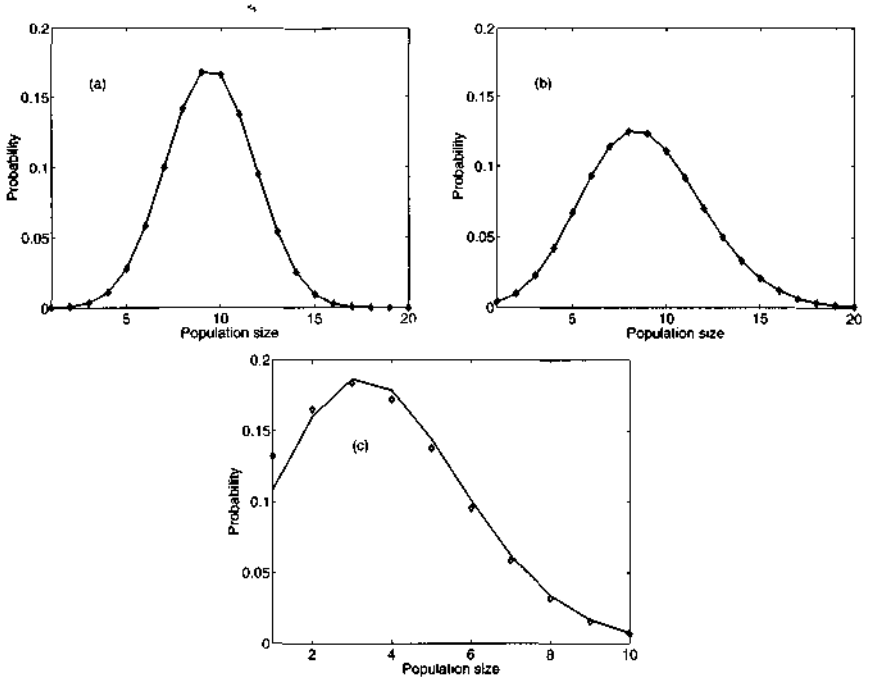

Figure 3.7. Quasistationary probability distribution, q^* (solid curve), and the approximate quasistationary probability distribution, $\tilde{q}^*$ (diamond marks), when $r = 0.015$, $K = 10$, and $N = 20$ in cases (a) and (b). In (c), $r = 0.015$, $K = 5$, $N = 10$, where $b_i = ri$ and $d_i = ri^2/K$.

3.8 SIS Epidemic Model

In this section, a stochastic epidemic model is formulated. The model is referred to as an SIS epidemic model because susceptible individuals (S) become infected (I) but do not develop immunity after they recover. They can immediately become infected again, $S \rightarrow I \rightarrow S$. No latent period is included in the model; therefore, individuals that become infected are also infectious (i.e., they can pass on the infection to others). It is also assumed that there is no vertical transmission of the disease; that is, the disease is not passed from the mother to her offspring. In our simple model, having no vertical transmission means no individuals are born infected; newborns enter the susceptible class. The total population size remains constant for all time since the number of births equals the number of deaths, $S+I = N$.

The compartmental diagram in Figure 3.9 represents the transitions between the two states, S and I.

Let the time period Δt from time n to $n+1$ be sufficiently small so that at most one event occurs. In the time interval Δt, either there is a suscep-

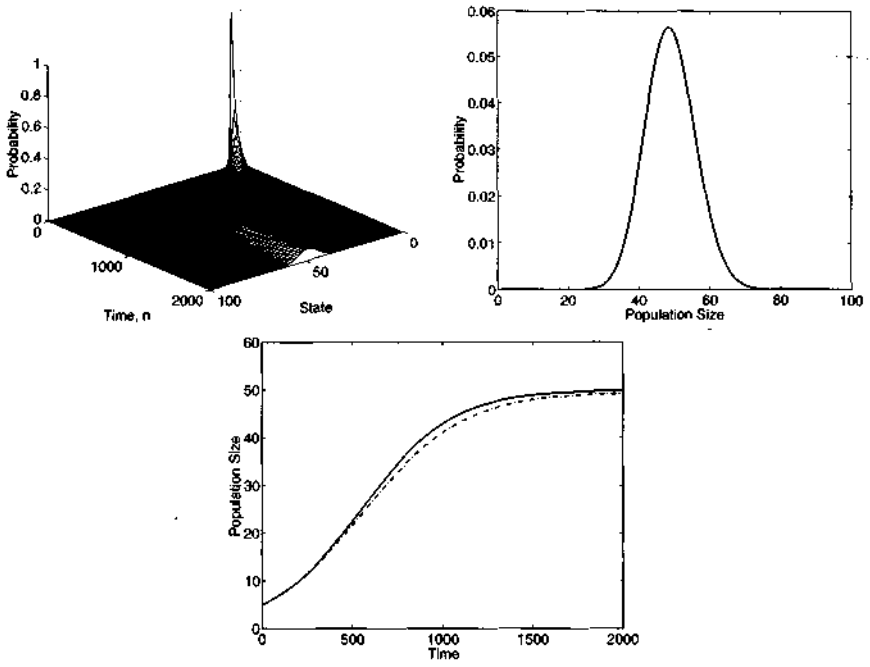

Figure 3.8. The stochastic logistic probability distribution, $p(n)$, and the solution to the logistic differential equation are compared, $dy/dt = ry(1 - y/K)$, $r = 0.004$, $K = 50$, $N = 100$, $y(0) = 5 = X_0$. In the top two figures, the probability distribution $p(n)$ for $n = 0, \ldots, 2000$ and the quasistationary distribution are graphed. In the bottom figure, the mean of the stochastic process (dashed and dotted curve) and the solution to the deterministic model (solid curve) are compared.

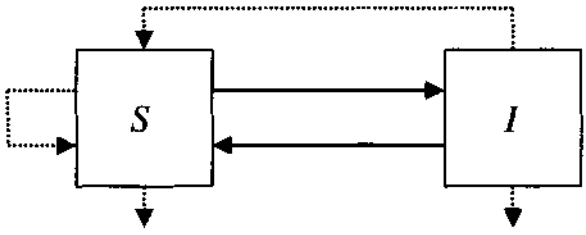

Figure 3.9. Compartmental diagram of the SIS epidemic model, a susceptible individual becomes infected with probability $\beta I/N$ and an infected individuals recovers with probability γ (solid lines). Birth and death probabilities of susceptible or infected individuals equal b (dotted lines).

tible individual that becomes infected, a birth of a susceptible individual (and corresponding death of either a susceptible or infected individual), or an infected individual recovers. A susceptible individual becomes infected with probability $\beta I/N$. The constant β is the number of contacts made by one infected (and infectious) individual that results in infection during the time interval Δt; only $\beta S/N$ of these contacts can result in a new infection, and the total number of new infections by the entire class of infected individuals is $\beta SI/N$. Susceptible and infected individuals die or are born with probability b, during the time interval Δt. Also, infected individuals recover with probability γ.

3.8.1 Deterministic SIS Epidemic Model

First, the dynamics of the deterministic SIS epidemic model are reviewed; then an analogous stochastic model is formulated and analyzed. Let S_n and I_n represent the number of susceptible and infected individuals at time n. The change in the states S_n and I_n during the time interval Δt can be represented by the system of difference equations:

$$\begin{aligned} S_{n+1} &= S_n - \beta S_n I_n/N + I_n(b+\gamma) \\ I_{n+1} &= \beta I_n S_n/N + I_n\,(1-b-\gamma)\,, \end{aligned}$$

where $n = 0, 1, 2, \ldots$, $S_0 > 0$, $I_0 > 0$, and $S_0 + I_0 = N$. For example, the number of new susceptible individuals at time $n+1$ equals those individuals that did not become infected, $S_n[1-\beta I_n/N]$, plus infected individuals that recovered, γI_n, plus newborns from the infected class, bI_n. The number of newborns from the susceptible class equals the number of susceptible individuals who die, bS_n, because the total population size is assumed to be constant.

It is assumed that the parameters are positive and satisfy

$$0 < \beta \leq 1, \quad 0 < b+\gamma \leq 1.$$

It can be seen that $S_n + I_n = N$. Therefore, it is sufficient to consider only the difference equation for I_n. Replacing S_n by $N - I_n$,

$$\begin{aligned} I_{n+1} &= I_n\left(\beta\frac{N-I_n}{N} + 1 - b - \gamma\right) \\ &= I_n\left(1 + \beta - b - \gamma - \beta\frac{I_n}{N}\right). \end{aligned} \tag{3.24}$$

Because of the assumptions on the parameters and the initial conditions, the solution I_n satisfies $0 \leq I_n \leq N$ for all time. There exists two constant or equilibrium solutions $I_n = \bar{I}$ to (3.24):

$$\bar{I} = 0 \quad \text{and} \quad \bar{I} = N\left(1 - \frac{b+\gamma}{\beta}\right). \tag{3.25}$$

It can be shown for model (3.24) that the dynamics depend on the following parameter $\mathcal{R}_0$, known as the *basic reproduction number*,

$$\mathcal{R}_0 = \frac{\beta}{b+\gamma}$$

(Allen, 1994). The parameter $\mathcal{R}_0$ has a biological interpretation. When the entire population is susceptible, $\mathcal{R}_0$ represents the average number of successful contacts (β) by one infected individual during the period of infectivity ($1/[b+\gamma]$) (Anderson and May, 1992). If $\mathcal{R}_0 > 1$, then one infected individual gives rise to more than one new infection, and if $\mathcal{R}_0 < 1$, then one infected individual gives rise to less than one new infection. Note that the second equilibrium in (3.25) is positive iff $\mathcal{R}_0 > 1$. It can be shown that if $\mathcal{R}_0 \leq 1$, then $\lim_{n\to\infty} I_n = 0$ and if $\mathcal{R}_0 > 1$, then $\lim_{n\to\infty} I_n = N(1-1/\mathcal{R}_0)$, where this limit is the second equilibrium given in (3.25) (Allen, 1994; Allen and Burgin, 2000). The magnitude of $\mathcal{R}_0$ determines whether the epidemic persists in the population, that is, whether it becomes an endemic infection.

3.8.2 Stochastic SIS Epidemic Model

Now, the stochastic SIS epidemic model is formulated. Let the random variable I_n represent the number of infected individuals at time n. The state space of the random variable I_n is the set $\{0, 1, 2, \ldots, N\}$ and the index set of the stochastic process $\{I_n\}$ is $n = 0, 1, 2, \ldots$. It will be shown that the stochastic process $\{I_n\}$ is a discrete time Markov chain. A transition matrix P is defined, an expression for the expected duration of the epidemic, τ_k, is derived, and an approximation to the probability of absorption, a_k (probability the epidemic dies out), is given for a large population size N.

Assume that Δt is sufficiently small such that during this time interval there is at most one change in the random variable I_n. If $I_n = i$, then I_{n+1} may change to only one of the following states, $i+1$, $i-1$ or i. The one-step transition probabilities satisfy

$$\begin{aligned}
p_{i+1,i} &= \text{Prob}\{I_{n+1} = i+1 | I_n = i\} = \beta i(N-i)/N = \Pi_i \\
p_{i-1,i} &= \text{Prob}\{I_{n+1} = i-1 | I_n = i\} = (b+\gamma)i \\
p_{ii} &= \text{Prob}\{I_{n+1} = i | I_n = i\} = 1 - \beta i(N-i)/N - (b+\gamma)i \\
&= 1 - \Pi_i - (b+\gamma)i,
\end{aligned}$$

for $i = 1, 2, \ldots, N-1$ and $p_{ji} = 0$ if $j \neq i-1, i, i+1$. Also, $p_{00} = 1$, the zero state is an absorbing state. The transition matrix P has the following

form:

$$\begin{pmatrix} 1 & (b+\gamma) & 0 & \cdots & 0 \\ 0 & 1-\Pi_1-(b+\gamma) & 2(b+\gamma) & \cdots & 0 \\ 0 & \Pi_1 & 1-\Pi_2-2(b+\gamma) & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & N(b+\gamma) \\ 0 & 0 & 0 & \cdots & 1-N(b+\gamma) \end{pmatrix},$$

where $\max_i\{\Pi_i + i(b+\gamma)\} \leq 1$.

It can be seen from the transition matrix that there are two classes, $\{0\}$ and $\{1, 2, \ldots, N\}$. The zero class is absorbing and the remaining states are all transient. Thus, $\lim_{n\to\infty} P^n p(0) = (1, 0, \ldots, 0)^T$. Eventually absorption occurs into the zero state, where there are zero infected individuals.

This model is similar to a logistic growth process if $\mathcal{R}_0 > 1$. Let $b_i = \Pi_i = \beta i(1 - i/N)$ and $d_i = (b+\gamma)i$. Then

$$b_i - d_i = i\left[\beta - (b+\gamma) - \beta i/N\right] = ri\left[1 - i/K\right],$$

where the intrinsic growth rate is represented by $r = \beta - (b+\gamma) > 0$ and the carrying capacity by $K = Nr/\beta = N(1 - 1/\mathcal{R}_0)$. Note that this value of K is the stable equilibrium of the deterministic model, equation (3.25). If $\mathcal{R}_0 \leq 1$, then $b_i - d_i \leq 0$ and only at $i = 0$ does the birth rate equal the death rate. The transition matrix can be used to calculate the probability distribution $p(n)$ and set up a linear system as in (3.18), $D\tau = \mathbf{c}$, to solve for the expected duration of the epidemic, τ.

Example 3.8 Suppose the population size $N = 100$, $\beta = 0.01$, $b = .0025 = \gamma$, and $\mathcal{R}_0 = 2$. Figure 3.10 shows the graphs of the probability distribution $p(n)$ when $I_0 = 1$ and the expected duration of the epidemic. ■

As can be seen in Figure 3.10, it may take a long time until the epidemic ends, especially if N is large and k is large. If N is sufficiently large and k is sufficiently small, the SIS epidemic model may behave similarly to a semi-infinite random walk model; either there is absorption with probability a_k or the size of the epidemic gets very large prior to ultimate absorption. The probability of absorption, a_k, can be used to estimate the probability that the epidemic ends quickly [i.e., the value of $p_0(n) \approx$constant is seen in Figure 3.10]. The probability of absorption at $x = 0$ for the semi-infinite random walk model is given by (3.13):

$$(q/p)^k, \quad \text{if } q < p \quad \text{and} \quad 1, \quad \text{if } q \geq p,$$

where q is the probability of moving to the left ($x \to x - 1$), p is the probability of moving to the right ($x \to x + 1$), and k is the initial position. In the epidemic model, the probability of moving to the left is $(b + \gamma)k$

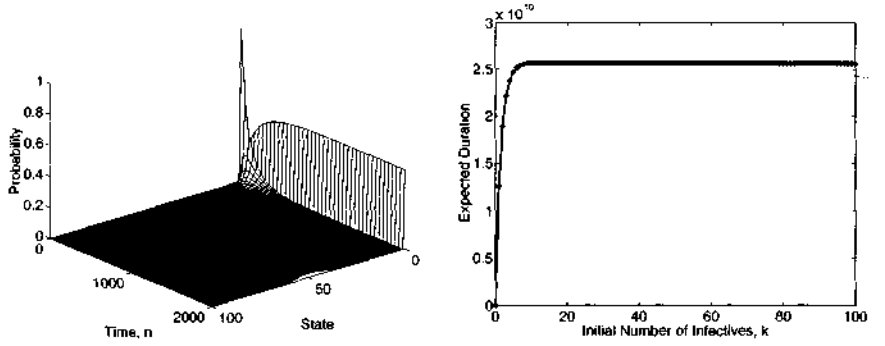

Figure 3.10. The parameters for the SIS epidemic model satisfy $N = 100$, $\beta = 0.01$, $b = .0025 = \gamma$, and $\mathcal{R}_0 = 2$. The probability distribution $p(n)$ for $n = 0, 1, \ldots, 2000$ is graphed when $I_0 = 1$ and the expected duration of the epidemic τ_k as a function of the initial number of infected individuals $k = I_0$.

and the probability of moving to the right is $\beta k(N-k)/N$. For N large, $q/p \approx (b+\gamma)/\beta = 1/\mathcal{R}_0$. Therefore, by analogy with the semi-infinite random walk model, the probability that the epidemic dies out quickly, given initially k infected individuals, is $(q/p)^k \approx (1/\mathcal{R}_0)^k$. An estimate of $p_0(n)$ at the outset of an epidemic is

$$p_0(n) \approx (1/\mathcal{R}_0)^k, \quad \text{if} \quad \mathcal{R}_0 > 1 \quad \text{and} \quad p_0(n) \approx 1, \quad \text{if} \quad \mathcal{R}_0 \leq 1$$

(Allen and Burgin, 2000; Jacquez and Simon, 1993). In Figure 3.10, $I_0 = 1 = k$ and $\mathcal{R}_0 = 2$, so that the probability the epidemic dies out at the outset of the epidemic is approximately $1/\mathcal{R}_0 = 1/2$. We can see in Figure 3.10 that $p_0(n)$ rises rapidly to $1/2$ and stays approximately constant. The increase in $p_0(n)$ after reaching $1/2$ is very slow; the average number of times steps until absorption is on the order of 10^{10} (see Figure 3.10).

Difference equations for the conditional distribution $q(n)$ can also be derived in a manner similar to the stochastic logistic model. As before,

$$q_i(n+1) = \frac{p_i(n+1)}{1 - p_0(n+1)}.$$

Using the relations satisfied by $p_i(n)$, it follows that

$$q_i(n+1)(1 - (b+\gamma)q_1(n)) = \left(\frac{p_i(n+1)}{1-p_0(n)}\right).$$

An approximation to the stationary distribution of $q(n)$ can be found by assuming that when there is one infected individual, that individual does not recover or give birth. The approximate quasistationary distribution $\tilde{q}^*(n)$ satisfies (3.23).

Consult the references for further information about stochastic SIS and other stochastic epidemic models (e.g., Allen and Burgin, 2000; Bailey,

1975; Bartlett, 1956; Daley and Gani, 1999; Gabriel, Lefevre, and Picard, 1990; Jacquez and Simon, 1993; Nåsell, 1996, 1999). Models such as an Susceptible-Infected-Removed (SIR) epidemic model, where there are immune or removed individuals and the population size is constant $S + I + R = N$, require a bivariate Markov process that includes two random variables, $\{S_n, I_n\}$. The epidemic models studied in the next section are bivariate Markov processes. They are known as chain binomial epidemic models.

3.9 Chain Binomial Epidemic Models

Let S_n and I_n be discrete random variables for the number of susceptible and infected individuals at time n, respectively. The time interval n to $n+1$ is of length Δt and represents the latent period, the time period until individuals become infectious, $n = 0, 1, 2, \ldots$. The infectious period is contracted to a point. In other words, the number of infected individuals I_n represents new infected individuals who were latent during the time interval $n-1$ to n. These new infected individuals are infectious. They will contact susceptible individuals at time n, who may then become infected at time $n+1$. There are no births nor deaths; the number of susceptible individuals is nonincreasing over time. The newly infected individuals at time $n+1$ and the susceptible individuals at time $n+1$ represent all those individuals who were susceptible at time n:

$$S_{n+1} + I_{n+1} = S_n.$$

The epidemic ends when the number of infected individuals equals zero, $I_n = 0$, because in the next time interval no more individuals can become infected, $I_{n+1} = 0$. Thus, $S_{n+1} = S_n$. Consult Daley and Gani (1999) for further details about this model.

Two models based on the above assumptions are formulated. They are known as the Greenwood and Reed-Frost models, named after the individuals who developed the models. These models were developed in the 1931 and 1928, respectively (Abbey, 1952; Daley and Gani, 1999; Greenwood, 1931). Lowell Reed and Wade Hampton Frost, two medical researchers at John's Hopkins University, developed their model for the purpose of showing medical students the variability in the epidemic process. However, neither Reed nor Frost thought their model was worthy of publication; it was Abbey who published their results in 1952. Primarily, these two models have been applied to small epidemics, or to epidemics within a household, where an initial infected individual spreads the infection to other members of the household (e.g., Bailey, 1975; Daley and Gani, 1999; Gani and Mansouri, 1987). Both models are *bivariate* Markov chain models because they depend on two random variables, the number of susceptible individuals, S_n, and the number of infected individuals, I_n. The bivariate Markov process is

denoted as $\{S_n, I_n\} = \{(S, I)_n\}$ for $n = 0, 1, 2, \ldots$. The state of the system at time $n + 1$ is determined only by the state of the system at the previous time n. The transition probability $p_{(s,i)_{n+1},(s,i)_n}$ specifies the one-step transition probability for moving between the two states, $(s, i)_n \rightarrow (s, i)_{n+1}$. The lower case letters s and i or s_n and i_n represent values of the random variables, S_n and I_n, respectively, at time n.

Let α be the probability of a contact between a susceptible individual and an infected individual and β be the probability that the susceptible individual is infected after contact. Then the probability that a susceptible individual does not become infected is

$$1 - \alpha + \alpha(1 - \beta) = 1 - \alpha\beta = p.$$

The probability p is an important parameter in the Greenwood and Reed-Frost models. Please consult Daley and Gani (1999) and Bailey (1975) for additional information about the mathematical properties of these models and Ackerman, Elveback, and Fox (1984) for a discussion of numerical simulations based on the Reed-Frost model.

3.9.1 Greenwood Model

The Greenwood model assumes that the transition probability $p_{(s,i)_{n+1},(s,i)_n}$ is a binomial probability. The probability of a successful contact resulting in infection is $1 - p$, and the probability of a contact not resulting in infection (not successful) is p. At time $n+1$, if there are s_{n+1} susceptible individuals, s_{n+1} contacts were not successful and $i_{n+1} = s_n - s_{n+1}$ contacts were successful, so that

$$p_{(s,i)_{n+1},(s,i)_n} = \binom{s_n}{s_{n+1}} p^{s_{n+1}}(1 - p)^{s_n - s_{n+1}}. \tag{3.26}$$

As shown in the expression above, the transition probability is independent of i_n. Because the transition probability can be expressed in terms of s_n and s_{n+1}, we shall denote it as p_{s_{n+1},s_n}. To initiate an epidemic, $I_0 = i_0 > 0$. The state space for S_n and I_n is $\{0, 1, 2, \ldots, s_0\}$, where $S_0 = s_0 > 0$. The maximal number of infected individuals is s_0.

A particular *realization* or *sample path* of the process can be denoted as $\{s_0, s_1, \ldots, s_{t-1}, s_t\}$, where $i_t = 0$, or, alternately, $s_t - s_{t-1} = 0$. The value t is the length of the sample path or the *duration* of the epidemic. Also, the *size* of the epidemic is the number of susceptible individuals who become infected during the epidemic or $s_0 - s_t$. It can be seen from the identity (3.26) that the random variable S_{n+1} has a binomial distribution, $b(S_n, p)$. It is for this reason that the Greenwood model is referred to as a *chain binomial model*. Using the facts that S_{n+1} has a binomial distribution and that $I_{n+1} = S_n - S_{n+1}$, the conditional expectation can be shown to satisfy

$$E(S_{n+1}|S_n = s_n) = ps_n,$$

and

$$E(I_{n+1}|S_n = s_n) = s_n - ps_n = (1-p)s_n.$$

[Recall that the mean of a binomial distribution $b(s_n, p)$ is $\mu = ps_n$.]

A transition matrix for the Greenwood model can be expressed in terms of the initial condition s_0. It is a matrix of size $(s_0 + 1) \times (s_0 + 1)$. The transition matrix is given by

$$P = \begin{pmatrix} 1 & (1-p) & (1-p)^2 & \cdots & (1-p)^{s_0} \\ 0 & p & 2p(1-p) & \cdots & \binom{s_0}{1} p(1-p)^{s_0-1} \\ 0 & 0 & p^2 & \cdots & \binom{s_0}{2} p^2(1-p)^{s_0-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & p^{s_0} \end{pmatrix}.$$

The transition matrix does not tell the whole story. Once there has been a transition of the type $s_n \to s_n, p_{s_n,s_n}$, the epidemic ends because $s_n = s_{n+1}$ and $i_n = i_{n+1}$.

Example 3.9 Suppose that, initially, there are three susceptible individuals and one infective. The epidemic with sample path $\{s_0, s_1\} = \{3, 3\}$ has probability $p_{33} = p^3$. The duration of the epidemic in this case is $T = 1$ and the size of the epidemic is $W = s_0 - s_1 = 0$. The epidemic with sample path $\{s_0, s_1, s_2\} = \{3, 1, 1\}$ has probability $p_{13}p_{11} = [3p(1-p)^2]p = 3(1-p)^2p^2$ with duration $T = 2$ and size $W = s_0 - s_2 = 1$. The other sample paths and their corresponding probabilities are given in Table 3.2 (see Daley and Gani, 1999). ■

Sample Path $\{s_0, s_1, \ldots, s_{t-1}, s_t\}$	Duration T	Size W	Greenwood	Reed-Frost
3 3	1	0	p^3	p^3
3 2 2	2	1	$3(1-p)p^4$	$3(1-p)p^4$
3 2 1 1	3	2	$6(1-p)^2p^4$	$6(1-p)^2p^4$
3 1 1	2	2	$3(1-p)^2p^2$	$3(1-p)^2p^3$
3 2 1 0 0	4	3	$6(1-p)^3p^3$	$6(1-p)^3p^3$
3 2 0 0	3	3	$3(1-p)^3p^2$	$3(1-p)^3p^2$
3 1 0 0	3	3	$3(1-p)^3p$	$3(1-p)^3p(1+p)$
3 0 0	2	3	$(1-p)^3$	$(1-p)^3$

Table 3.2. All of the sample paths, their duration, and size are computed for the Greenwood and Reed-Frost models when $s_0 = 3$ and $i_0 = 1$.

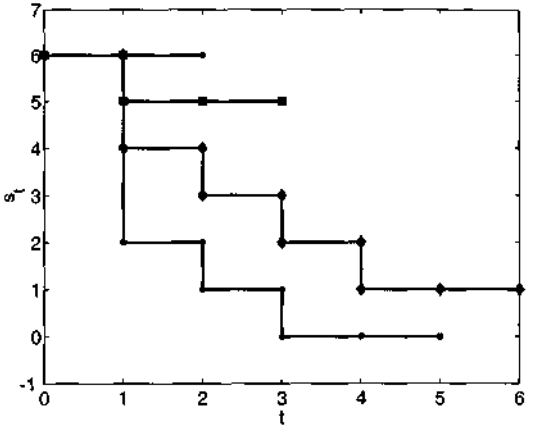

Figure 3.11. Four sample paths for the Greenwood chain binomial model when $s_0 = 6$ and $i_0 = 1$, $\{6,6\}$, $\{6,5,5\}$, $\{6,4,3,2,1,1\}$, and $\{6,2,1,0,0\}$.

Figure 3.11 illustrates four sample paths for the Greenwood model when $s_0 = 6$ and $i_0 = 1$.

3.9.2 Reed-Frost Model

In the Reed-Frost model, a susceptible individual at time n will still be susceptible at time $n+1$ if there is no contact with an infected individual. If the number of infected individuals at time n is i_n, it is assumed that the probability that there is no successful contact of a susceptible individual with any of the i_n infected individuals is p^{i_n}. The Reed-Frost model has the form of the Greenwood model except that p is replaced by p^{i_n}. The transition probabilities in the Reed-Frost model are binomial probabilities satisfying

$$p_{(s,i)_{n+1},(s,i)_n} = \binom{s_n}{s_{n+1}} (p^{i_n})^{s_{n+1}} (1 - p^{i_n})^{s_n - s_{n+1}}.$$

The one-step transition probability depends on i_n, s_n, and s_{n+1}, and, therefore, it cannot be expressed just in terms of the values of s_n and s_{n+1} as was done in the Greenwood model. Recall that $s_n + i_n = s_{n-1}$ or $i_n = s_{n-1} - s_n$. Although the transition probability depends on i_n, s_n and s_{n+1}, for simplicity of notation, we denote the transition probability for the Reed-Frost model as p_{s_{n+1},s_n}. It follows from the form of the transition probability that the random variable S_{n+1} has a binomial distribution, $b(S_n, p^{I_n})$. Hence, the Reed-Frost model is also referred to as a *chain binomial model.* Using the facts that S_{n+1} has a binomial distribution and $I_{n+1} = S_n - S_{n+1}$, it

can be shown that the conditional expectation

$$E(S_{n+1}|(S,I)_n = (s_n, i_n)) = s_n p^{i_n}$$

and

$$E(I_{n+1}|(S,I)_n = (s_n, i_n)) = s_n - s_n p^{i_n} = s_n(1 - p^{i_n}).$$

Example 3.10 Suppose that, initially, there are three susceptible individuals and one infected individual, $s_0 = 3$, $i_0 = 1$. In the Reed-Frost epidemic model, the sample path $\{3,3\}$ has probability $p_{33} = p^3$, the sample path $\{3,2,2\}$ has probability $p_{23}p_{22} = [3p^2(1-p)]p^2 = 3(1-p)p^4$, and the sample path $\{3,1,1\}$ has probability $p_{13}p_{11} = [3p(1-p)^2]p^2 = 3(1-p)^2p^3$. This last sample path has a probability that is different from the Greenwood model. In general, if there is more than one infected individual in a time interval, the Greenwood and Reed-Frost model will differ. ■

3.9.3 Duration and Size of the Epidemic

Let T denote the duration of an epidemic and let W denote the size of an epidemic or the total number of susceptible individuals who become infected. For example, for a sample path $\{s_0, s_1, \ldots, s_{t-1}, s_t\}$, $T = t$ and $W = s_0 - s_t$. For a given number of initial susceptible and infected individuals, $s_0 > 0$ and $i_0 > 0$, the maximum value of T is $s_0 + 1$, $T \in \{1, 2, \ldots, s_0+1\}$ and the maximum value of W is s_0, $W \in \{0, 1, \ldots, s_0\}$. The epidemic may end in one time step if no one gets infected, $S_1 = s_0$ and $I_1 = 0$ ($T = 1$ and $W = 0$), or it may end after $s_0 + 1$ time steps when one individual gets infected each time step ($T = s_0 + 1$ and $W = s_0$). The variables T and W are random variables whose probability distributions can be computed from the probabilities of the sample paths (see Table 3.2).

Example 3.11 The probability distribution corresponding to the duration an epidemic, T, in the Greenwood model for $s_0 = 3$ and $i_0 = 1$ can be computed from Table 3.2:

$$\begin{aligned}
\text{Prob}\{T = 1\} &= p^3 \\
\text{Prob}\{T = 2\} &= (1-p)^3 + 3p^2(1-p)^2 + 3p^4(1-p) \\
\text{Prob}\{T = 3\} &= 3p(1+p)(1-p)^3 + 6p^4(1-p)^2 \\
\text{Prob}\{T = 4\} &= 6p^3(1-p)^3.
\end{aligned}$$

■

Another method can be applied to find the probability distributions for T and W in the Greenwood model. This method is described by Daley and Gani (1999) and is briefly presented here. First, partition the transition matrix P of the Greenwood model into two matrices, $P = U + D$, where U

is a strictly upper triangular matrix with zeros along the diagonal and D is a diagonal matrix–that is, $D = \text{diag}(1, p, p^2, \ldots, p^{s_0})$ and

$$U = \begin{pmatrix} 0 & (1-p) & (1-p)^2 & \cdots & (1-p)^{s_0} \\ 0 & 0 & 2p(1-p) & \cdots & \binom{s_0}{1} p(1-p)^{s_0-1} \\ 0 & 0 & 0 & \cdots & \binom{s_0}{2} p^2(1-p)^{s_0-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Note that the matrix U represents those transitions that do not return to the same state in one time step with probability p_{ij} for $i \neq j$, whereas D represents those transitions that do return to the same state in one time step with probability p_{ii}. When $s_t = s_{t-1}$ or $p^{(n)}_{s_t, s_t} > 0$, there is a positive probability that the epidemic ends at time n. The elements of the matrix U^{n-1} represent the probability of transition between states i and j in $n-1$ time steps, $p^{(n-1)}_{ij}$, where $j \to i$, $j \neq i$. Let

$$p(n) = (p_0(n), p_1(n), \ldots, p_{s_0}(n))^T,$$

denote the probability distribution for the state of susceptible individuals at time n, then $U^{n-1}p(0)$ represents the probability distribution $p(n-1)$ given that the epidemic has not ended at time $n-1$. Multiplying by D, $DU^{n-1}p(0)$ gives the probability distribution vector that the epidemic has ended exactly at time n. The sum of the elements of the probability distribution vector $DU^{n-1}p(0)$ is the probability that the epidemic has ended at time n, $\text{Prob}\{T = n\}$. Let $E = (1, 1, 1, \ldots, 1)$ be a row vector of ones. Then

$$\text{Prob}\{T = n\} = EDU^{n-1}p(0).$$

Since the epidemic could end at states $0, 1, 2, \ldots, s_0$, the p.g.f. for the random variable T, the duration of the epidemic, is given by

$$\sum_{n=1}^{s_0+1} EDU^{n-1}p(0)t^n = \sum_{n=1}^{s_0+1} \text{Prob}\{T = n\}t^n.$$

The computer algebra system *Maple* can be used to calculate the probability distribution T with the matrix formulas above (see the Appendix for Chapter 3).

In a similar manner, the probability generating function for the random variable W, the size of the epidemic, can be derived (Daley and Gani, 1999).

For the Greenwood model, let $U(t)$ be defined as follows:

$$U(t) = \begin{pmatrix} 0 & (1-p)t & [(1-p)t]^2 & \cdots & [(1-p)t]^{s_0} \\ 0 & 0 & 2p(1-p)t & \cdots & \binom{s_0}{1} p[(1-p)t]^{s_0-1} \\ 0 & 0 & 0 & \cdots & \binom{s_0}{2} p^2[(1-p)t]^{s_0-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

The matrix U defined previously is then $U \equiv U(1)$. The elements $p_{ij}(t)$ of $U(t)$ satisfy $p_{ij}(t) = p_{ij}t^{j-i}$. The elements of $U(t)$, $p_{ij}(t)$, $i \neq j$, represent generating elements for the probability the epidemic has not ended in one-step transitions. Since the number of susceptible individuals has gone from j to i, the size of the epidemic is $j - i$. In addition, the elements in $U^2(t)$ satisfy $p_{ij}^{(2)}(t) = p_{ij}^{(2)}t^{j-i}$. If, in two time steps, the number of susceptible individuals has gone from j to i, then the size of the epidemic is $j - i$. Thus, $EDU(t)^{n-1}p(0)$ represents the generating function for the size of the epidemic when it has ended in n time steps. Since the epidemic can end in $1, 2, \ldots, s_0 + 1$ time steps, the *probability generating function* for W satisfies

$$\sum_{n=1}^{s_0+1} EDU(t)^{n-1}p(0) = \sum_{k=0}^{s_0} \text{Prob}\{W = k\}t^k.$$

The coefficient of t^k in the expansion of the left-hand side is $\text{Prob}\{W = k\}$ (*Daley and Gani, 1999*).

Example 3.12 It can be shown using the probability generating function for W (see the Appendix for Chapter 3) or directly from Table 3.2 that if $s_0 = 3$ and $i_0 = 1$, then, in the Greenwood model,

$$\begin{aligned} \text{Prob}\{W = 0\} &= p^3 \\ \text{Prob}\{W = 1\} &= 3p^4(1-p) \\ \text{Prob}\{W = 2\} &= 3p^2(1+2p^2)(1-p)^2 \\ \text{Prob}\{W = 3\} &= (1-p)^3(1+3p+3p^2+6p^3) \end{aligned}$$

The probability distributions for W and T are graphed in Figure 3.12 in the case $p = 0.5$, $s_0 = 3$, and $i_0 = 1$, $E(W) \approx 2.156$ and $E(T) \approx 2.438$. ■

3.10 Exercises for Chapter 3

1. Modify the gambler's ruin problem so that the probability of winning is p, losing is q and a tie is r, $p + q + r = 1$.

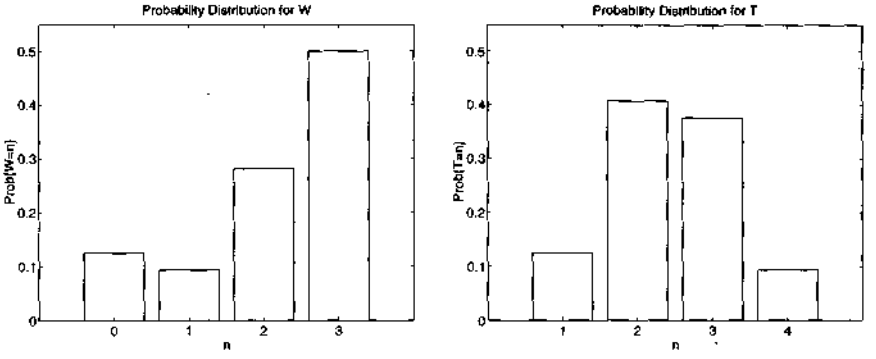

Figure 3.12. Probability distributions for the size of the epidemic, W, and the duration of the epidemic, T, when $p = 0.5$, $s_0 = 3$, and $i_0 = 1$.

(a) Show that the probability of ruin (or absorption at $x = 0$), beginning with a capital of k, satisfies

$$a_k = \frac{(q/p)^N - (q/p)^k}{(q/p)^N - 1}, \quad p \neq q.$$

(b) Show that the expected duration of the games is

$$\tau_k = \frac{1}{(q-p)} \left[k - N \frac{(q/p)^k - 1}{(q/p)^N - 1} \right], \quad p \neq q. \tag{3.27}$$

Note that this expression is the same as the one for $r = 0$, but for $r > 0$ the values of p and q are smaller, so the expected duration will have a larger value.

(c) Find the probability of ruin a_k and expected duration τ_k when $p = q$.

2. For the semi-infinite random walk model in one dimension, derive the formulas for the probability of absorption, a_k, equation (3.13), and the expected duration until absorption, τ_k, equation (3.14), directly from the random walk model with absorbing boundaries. Let the right-hand endpoint of the domain approach infinity, $N \to \infty$ in (3.4), (3.5), (3.9), and (3.10).

3. Consider a restricted random walk in one dimension with an absorbing barrier at $x = 0$ and an elastic barrier at $x = N$. Assume p is the probability of moving to the right and q is the probability of moving to the left, $p + q = 1$ and $p \neq q$. Find the probability of absorption at $x = 0$.

4. Verify the following statements for the gambler's ruin problem.

   (a) If $n < k$, then $a_{kn} = 0$, and if $n = k$, then $a_{kk} = q^k$.

   (b) The probability $a_{k,k+2i+1} = 0$ for $i = 0, 1, 2, \ldots$.

   (c) The probability $a_{k,k+2} = kq^{k+1}p$.

   (d) What are the values of b_{kn}, if $n < N - k$? What are the values of $b_{k,N-k}$ and $b_{k,N-k+2i+1}$ for $i = 0, 1, 2, \ldots$?

5. Consider the random walk model on $\{0, 1, 2, \ldots, N\}$. Suppose state N is absorbing, $p_{N,N} = 1$, and state zero is reflecting, $p_{00} = 1 - p$ and $p_{10} = p$. Let b_k be the probability of absorption at $x = N$ beginning at state $x = k$.

   (a) Show that $b_N = 1$ and $b_0 = b_1$.

   (b) Derive a difference equation for the probability of absorption, b_k; then show that $b_k = 1$ (i.e., the probability of absorption is one).

6. In a simple birth and death process the transition matrix is

$$P = \begin{pmatrix} 1 & d & 0 & \cdots & 0 \\ 0 & 1-b-d & 2d & \cdots & 0 \\ 0 & b & 1-2(b+d) & \cdots & 0 \\ 0 & 0 & 2b & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & dN \\ 0 & 0 & 0 & \cdots & 1-dN \end{pmatrix}.$$

   (See Example 3.5.)

   (a) Show that the unique stationary distribution for this process is $\pi = (1, 0, 0, \ldots, 0)^T$.

   (b) Assume $N = 20$ and $b = 0.025 = d$. Suppose the initial probability distribution for X_0 is $X_0 = 5$ [i.e., $p_5(0) = 1$]. Find $p(1), p(2), \ldots, p(100)$. Then graph the probability distribution over time, $n = 0, 5, 10, \ldots, 500$. (*Hint*: Modify the MATLAB program in the Appendix for Chapter 3 that was applied to the logistic process.)

   (c) Graph $p_0(n)$ for $n = 0, 1, 2, \ldots, 500$. What is $\lim_{n\to\infty} p_0(n)$?

7. The formula for the expected duration of the game, given by equation (3.27), can be checked numerically by performing some numerical simulations. Let $N = 100$, $k = 50$, $q = 0.5$, $r = 0.2$, and $p = 0.3$. Write a computer program for this gambler's ruin problem and simulate a sufficient number of sample paths (total sample paths$\geq$

1000) to find the time until the game ends. Then compute the mean duration for all of the sample paths. Compare the computer generated mean duration with formula (3.27).

8. For the gambler's ruin problem, calculate the mean τ_k and standard deviation σ_k of the the duration of the games. Use the p.g.f. $S_k(t) = A_k(t) + B_k(t)$ when $N = 100$, $k = 50$ and for values of $p = 2/5$, $9/20$ and $1/2$. (*Hint*: The mean and standard deviation can be computed using *Maple* or MATLAB, if the limit is taken from the left at $t = 1$. When using *Maple*, it is important to use exact values for the parameters, e.g., $p = 2/5$ and not $p = 0.4$.)

9. In a simple birth and death process, the birth and death rates satisfy $b_i = bi$ for $i = 1, \ldots, N-1$, $d_i = di$ for $i = 1, \ldots, N$ and zero elsewhere (see Example 3.5 and Exercise 6). The parameters b and d are positive and satisfy $(b+d)N \leq 1$. The mean of the population size at time n, denoted as $\mu(n)$, satisfies

$$\mu(n) = \sum_{i=0}^{N} i p_i(n).$$

   (a) Use the transition matrix P to compute $p_i(n+1)$, $p(n+1) = Pp(n)$. Then show that $\mu(n)$ satisfies the following first-order difference equation:

$$\mu(n+1) = (1 + b - d)\mu(n) - bNp_N(n).$$

   (b) Use the fact that $0 \leq \mu(n+1) \leq (1+b-d)\mu(n)$ to show

$$\mu(n) \leq (1 + b - d)^n \mu(0).$$

   If $b < d$, find $\lim_{n\to\infty} \mu(n)$.

10. For the logistic growth process,

   (a) Show that the approximate quasistationary probability distribution, $\tilde{q}^*$, satisfies the relation given in equation (3.23).

   (b) Use this relation to compute $\tilde{q}^*$ for $N = 50$ and $r = 0.01$ when $b_i = r(i - i^2/N)$ and $d_i = ri^2/N$ ($K = 25$) for $i = 1, 2, \ldots, 50$. Graph $\tilde{q}_i^*$ for $i = 1, 2, \ldots, 50$.

11. For the deterministic SIS epidemic model (3.24), verify the following.

   (a) If $\mathcal{R}_0 \leq 1$, then $\lim_{n\to\infty} I_n = 0$. (*Hint*: $I_{n+1} < I_n$.)

   (b) If $\mathcal{R}_0 > 1$, then $\lim_{n\to\infty} I_n = N(1 - 1/\mathcal{R}_0)$. [*Hint*: Make a change of variable and compare the model to the discrete logistic equation $x_{n+1} = rx_n(1 - x_n)$. The discrete logistic equation has a stable equilibrium at $(r-1)/r$ iff $1 < r \leq 3$ (Elaydi, 2000).]

12. Consider the stochastic SIS epidemic model with the following parameters.

    (a) Let $N = 20$, $\beta = 0.01$, $b = 0.0025 = \gamma$, and $\mathcal{R}_0 = 2$. Calculate the expected duration of an epidemic $\tau_k = E(T_k)$; then sketch a graph of τ_k for $k = 0, 1, \ldots, 20$. *Maple* or MATLAB may be used to solve the linear system.

    (b) Let $N = 20$, $\beta = 0.005$, $b = 0.0025 = \gamma$, and $\mathcal{R}_0 = 1$. Calculate the expected duration of an epidemic $\tau_k = E(T_k)$; then sketch a graph of τ_k.

13. Find the approximate quasistationary distribution $\tilde{q}^*$ for the stochastic SIS epidemic model when $N = 20$, $\beta = 0.01$, and $b = 0.0025 = \gamma$ (i.e., the solution to $\tilde{P}\tilde{q}^* = \tilde{q}^*$, where $\sum \tilde{q}_i^* = 1$). Sketch $\tilde{q}_i^*$ for $i = 1, 2, \ldots, 20$.

14. In the SIS epidemic model, an estimate was obtained for the probability that the epidemic ends quickly. It was found that the probability of extinction, $p_0(n)$, reaches a plateau or constant value that is less than one prior to ultimate extinction, $p_0(n) \approx \text{constant} = \hat{p}_0 < 1$. In particular,

$$p_0(n) \approx \hat{p}_0 \approx (1/\mathcal{R}_0)^k$$

    for $\mathcal{R}_0 > 1$ and $I_0 = k$. This latter estimate was obtained from the formula for the probability of absorption at $x = 0$ in the semi-infinite random walk model, $a_k = (q/p)^k$. This estimate can also be obtained from the product $\Pi_{i=1}^{k}(d_i/b_i)$. We shall use this latter formula to estimate the probability of population extinction $\hat{p}_0$ in the stochastic logistic model.

    (a) Consider the stochastic logistic model with $b_i = ri(1 - i/(2K))$ and $d_i = ri^2/(2K)$, for $i = 0, 1, 2, \ldots, 2K$. Suppose $r = 0.015$ and $K = 20$. Use the formula $\Pi_{i=1}^{k}(d_i/b_i)$ to estimate the probability of population extinction, $\hat{p}_0$, when $I_0 = 1$ and $I_0 = 2$. Compare these estimates with the values obtained from the probability distribution $p_0(n)$ in Figure 3.13. For $1000 < n \leq 2000$, $p_0(n)$ is approximately constant, $p_0(n) \approx \hat{p}_0$. For $I_0 = 1$, $\hat{p} \approx 0.0264$, and for $I_0 = 2$, $\hat{p}_0 \approx 0.00143$.

    (b) Use the formula $\Pi_{i=1}^{k}(d_i/b_i)$ to estimate $\hat{p}_0$ when $b_i = ri$ and $d_i = ri^2/K$ for $i = 0, 1, 2, \ldots, N$, $N = 2K$, $b_N = 0$, and $r = 0.005$ and compare the estimates with those in Table 3.3. The values in Table 3.3 are the values of $p_0(n)$ for $2000 < n \leq 6000$.

15. Consider a chain binomial epidemic model with initially one infective and two susceptible individuals, $s_0 = 2$ and $i_0 = 1$.

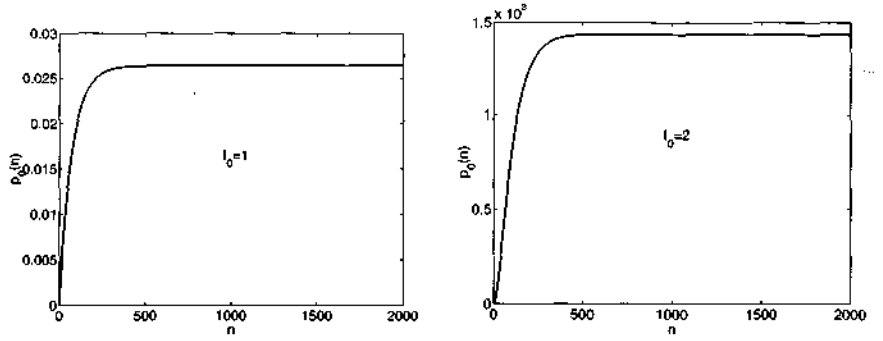

Figure 3.13. Probability of population extinction for the stochastic logistic model when $b_i = ri(1 - i/(2K))$ and $d_i = ri^2/(2K)$, for $i = 0, 1, 2, \ldots, 2K$, $r = 0.015$ and $K = 20$.

K	I_0	$\hat{p}_0$
20	1	0.0530
20	2	0.0056
30	1	0.0346
30	2	0.0024

Table 3.3. Estimates of the probability of rapid population extinction for the stochastic logistic model when $b_i = ri$ and $d_i = ri^2/K$ for $i = 0, 1, 2, \ldots, N$, $N = 2K$, $b_N = 0$, and $r = 0.005$

(a) Calculate the probabilities for the different types of chains in the Reed-Frost and Greenwood chain binomial models and show that both models have the same probabilities.

(b) Find the probability distribution for the duration time, T; that is, Prob$\{T = n\}$, $n = 1, 2, 3$. Graph this distribution for $p = 0.2$, 0.5 and 0.8. Find $E(T)$.

(c) Find the probability distribution for the size of the epidemic, W: $Prob\{W = n\}$, $n = 0, 1, 2$. Graph this distribution for $p = 0.2$, 0.5 and 0.8. Find $E(W)$.

16. Consider the Reed-Frost chain binomial epidemic model with initially one infective and three susceptible individuals, $s_0 = 3$ and $i_0 = 1$. Use Table 3.2 for the different types of chains.

(a) Find the probability distribution for the duration time, T.

(b) Find the probability distribution for the size of the epidemic, W.

(c) Sketch the probability distributions in (a) and (b) when $p = 0.2$ and $p = 0.8$.

17. Suppose the size of a population remains constant from generation to generation; the size equals N. The dynamics of a particular gene in this population is modeled. Suppose the gene has two alleles, A and a. Therefore, individual genotypes are either AA, Aa, or aa. Let the random variable X_n denote the number of A alleles in the population in the nth generation, $n = 0, 1, 2, \ldots$. Then $X_n \in \{0, 1, 2, \ldots, 2N\}$. Assume random mating of individuals so that the genes in generation $n + 1$ are found by sampling with replacement from the genes in generation n (Ewens, 1979). Then the one step transition probability has a binomial probability distribution with the probability of success $X_n/(2N)$, i.e., if $X_n = i$, then the one step transition probability is the binomial p.d.f $b(2N, i/2N)$,

$$p_{ji} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j},$$

$i, j = 0, 1, 2, \ldots, 2N$ (Ewens, 1979; Schinazi, 1999). This model is known as the Wright-Fisher model.

(a) Given $X_n = k$, show that the mean of X_{n+1} satisfies $\mu_{X_{n+1}} = E(X_{n+1}|X_n = k) = k$. A discrete-time Markov process with the property $E(X_{n+1}|X_n = k) = k$ is called a *martingale.*

(b) Show that in the gambler's ruin problem, with $1 \leq k \leq N - 1$, $p_{k+1,k} = p$ and $p_{k-1,k} = q$, $E(X_{n+1}|X_n = k) = k$ iff $p = q$. In game theory, a martingale is a "fair game". On the average, there is no gain nor loss with each game that is played.

3.11 References for Chapter 3

Abbey, H. 1952. An examination of the Reed-Frost theory of epidemics. *Hum. Biology.* 24: 201–233.

Ackerman, E., L. R. Elveback, and J. P. Fox. 1984. *Simulation of Infectious Disease Epidemics.* Charles C. Thomas Publ., Springfield, Ill.

Allen, L. J. S. 1994. Some discrete-time SI, SIR and SIS epidemic models. *Math. Biosci.* 124: 83–105.

Allen, L. J. S. and A. Burgin. 2000. Comparison of deterministic and stochastic SIS and SIR models in discrete time. *Math. Biosci.* 163: 1–33.

Anderson, R. M. and R. M. May. 1992. *Infectious Diseases of Humans Dynamics and Control.* Oxford University Press, Oxford.

Bailey, N. T. J. 1975. *The Mathematical Theory of Infectious Diseases and Its Applications.* Charles Griffin, London.

Bailey, N. T. J. 1990. *The Elements of Stochastic Processes with Applications to the Natural Sciences.* John Wiley & Sons, New York.

Bartlett, M. S. 1956. Deterministic and stochastic models for recurrent epidemics. *Proc. Third Berkeley Symp. Math. Stat. and Prob.* 4: 81–109.

Daley, D. J. and J. Gani. 1999. *Epidemic Modelling: An Introduction.* Cambridge Studies in Mathematical Biology: 15, Cambridge University Press, Cambridge.

Elaydi, S. N. 1999. *An Introduction to Difference Equations.* 2nd ed. Springer-Verlag, New York.

Elaydi, S. N. 2000. *Discrete Chaos.* Chapman & Hall/CRC, Boca Raton, Fla.

Ewens, W. J. 1979. *Mathematical Population Genetics.* Springer-Verlag, Berlin, Heidelberg, New York.

Gabriel, J. -P., C. Lefèvre, and P. Picard (eds.) 1990. *Stochastic Processes in Epidemic Theory.* Lecture Notes in Biomathematics, Springer-Verlag, New York.

Gani, J. and H. Mansouri. 1987. Fitting chain binomial models to the common cold. *Math. Scientist* 12: 31–37.

Greenwood, M. 1931. On the statistical measure of infectiousness. *J. Hyg. Cambridge* 31: 336–351.

Jacquez, J. A. and C. P. Simon. 1993. The stochastic SI model with recruitment and deaths I. Comparison with the closed SIS model. *Math. Biosci.* 117: 77–125.

Nåsell, I. 1996. The quasi-stationary distribution of the closed endemic SIS model. *Adv. Appl. Rrob.* 28: 895–932.

Nåsell, I. 1999. On the quasi-stationary distribution of the stochastic logistic epidemic. *Math. Biosci.* 156: 21–40.

Nåsell, I. 2001. Extinction and quasi-stationarity in the Verhulst logistic model. *J. Theor. Biol.* 211: 11–27.

Nisbet, R. M. and W. S. C. Gurney. 1982. *Modelling Fluctuating Populations.* John Wiley & Sons, Chichester and New York.

Ortega, J. M. 1987. *Matrix Theory A Second Course.* Plenum Press, New York.

Schinazi, R. B. 1999. *Classical and Spatial Stochastic Processes.* Birkhäuser, Boston.

Taylor, H. M. and S. Karlin. 1998. *An Introduction to Stochastic Modeling.* 3rd ed. Academic Press, New York.

Wade, W. R. 1995. *An Introduction to Analysis.* Prentice Hall, Upper Saddle River, N. J.

3.12 Appendix for Chapter 3

3.12.1 Matlab Programs

The following three MATLAB programs can be used to compute the expected duration of the gambler's ruin problem, sample paths for the gambler's ruin problem, and the probability distribution for logistic growth.

```
% MATLAB Program:
% Expected duration for the
% gambler's ruin problem, Figure 3.1.
clear all
set(0,'DefaultAxesFontSize',18);
N=100;
q=0.55; % Probability of losing.
p=1-q;
C=sparse(2:N,2:N,-ones(1,N-1),N+1,N+1); % Diagonal entries.
C(1,1)=1;
C(N+1,N+1)=1;
L=sparse(2:N,1:N-1,q*ones(1,N-1),N+1,N+1); % Subdiagonal.
U=sparse(2:N,3:N+1,p*ones(1,N-1),N+1,N+1); % Superdiagonal.
D=L+C+U; % Matrix D for the expected duration.
d=-ones(N+1,1);
d(1)=0;
d(N+1)=0;
t=D; % Expected Duration is the solution t = D^{-1}d.
plot([0:N],t,'k-','LineWidth',2); % Graphs the expected
xlabel('Initial capital');        % duration.
ylabel('Expected duration');
set(gca,'ytick',[0,200,400,600,800]); % Sets the tick marks.
max(t) % Maximum value of the expected duration.
```

Notes: The command "sparse" is used to make the computations more efficient. A statement following % explains the Matlab command. This statement is not executed. If a semicolon is left off an executable command, then the value generated by the command prints to the computer screen.

```
% MATLAB Program:
% Sample paths and expected duration
% for the gambler's ruin problem.
clear all
set(0,'DefaultAxesFontSize',18) % Increases axes labels.
sim=1000; % Number of simulations.
q=0.55;
for j=1:sim
    j % Simulation number; prints on the computer screen.
    clear r
    r(1)=50;
    i=1;
    while r(i)>0 & r(i)<100
       y=rand;
       if y<=q
          r(i+1)=r(i)-1;
       else
          r(i+1)=r(i)+1;
       end % End of while loop.
       i=i+1;
    end % End of while loop.
    t(j)=i; % Time until absorption.
    if j<=3 % Plots three sample paths.
       l1=stairs([0:1:i-1],r) % Draws stairstep graph.
       set(l1,'LineWidth',2) % Thickens the line width.
       hold on % Holds the current plot.
    end % End of sample path loop.
end % end of j loop
meandur=mean(t) % Mean duration; printed on screen.
stdevdur=std(t) % Standard deviation; printed on screen.
xlabel('Games')
ylabel('Capital')
hold off % Erases previous plots before drawing new ones.
```

```
% MATLAB Program:
% Probability distribution for logistic growth.
clear all
set(0,'DefaultAxesFontSize',18);
time=2000;
K=50;
N=2*K;
r=0.004;
en=25; % Plot every enth time interval.
T=zeros(N+1,N+1); % T is the transition matrix (3.15).
p=zeros(time+1,N+1);
```

```
p(1,6)=1;
v=linspace(0,N,N+1);
b1=r*v.*(1-v/(2*K)); % Defines the probabilities.
d1=r*v.^ 2 /(2*K);
b2=r*v;
d2=r*v.*v/K;
b2(N+1)=0;
for i=2:N  % Define the transition matrix.
    T(i,i)=1-b1(i)-d1(i);
    T(i,i+1)=d1(i+1);
    T(i+1,i)=b1(i);
end
T(1,1)=1;
T(1,2)=d1(2);
T(N+1,N+1)=1-d1(N+1);
for t=1:time
    y=T*p(t,:)';
    p(t+1,:)=y';
end
pm(1,:)=p(1,:);
for t=1:time/en;
    pm(t+1,:)=p(en*t,:);
end
mesh([0:1:N],[0:en:time],pm); % Three dimensional plot.
xlabel('State');
ylabel('Time, n');
zlabel('Probability');
view(140,30)
```

3.12.2 Maple Program

The following *Maple* program can be used to find p.g.f.'s for the duration time and size of the epidemic in the Greenwood chain binomial epidemic model.

```
>> with(linalg):
>> P:=t->matrix(4,4,[1,(1-p)*t,((1-p)*t)^2,((1-p)*t)^3,0,p,
2*p*(1-p)*t,3*p*((1-p)*t)^2,0,0,p^2,3*p^2*(1-p)*t,0,0,0,p^3])
>> Di:=diag(1,p,p^2,p^3):
>> U:=evalm(P(1)-Di):
>> E:=vector([1,1,1,1]):
>> p0:=vector([0,0,0,1]):
>> T:=t*dotprod(evalm(E&*Di),p0):
>> for k from 2 to 4 do
T:=T+factor(dotprod(evalm(E&*Di&*U^(k-1)),p0))*t^k;
```

Chapter 4

Discrete Time Branching Processes

4.1 Introduction

Discrete time branching processes are a special type of discrete time Markov chain. The study of branching processes has a long history. The subject of branching processes began in 1845 with Bienaymé and was advanced in the 1870s with the work of Reverend Henry William Watson, a clergyman and mathematician, and the biometrician Francis Galton (Mode, 1971). These individuals were interested in studying the survival of family names. Galton in 1873 submitted a problem to the *Educational Times* (Mode, 1971) stating the following: Suppose adult males (N in number) in a population each have different surnames. Suppose in each generation, a_0 percent of the adult males have no male children who survive to adulthood; a_1 have one such child; a_2 have two, and so on up to a_5, who have five. Then Galton posed two questions (Mode, 1971):

(1) Find what proportion of the surnames become extinct after r generations.

(2) Find how many instances there are of the same surname being held by m persons.

Galton did not receive satisfactory solutions to his problems and sought help from Watson. Watson used probability generating functions to study the problems. Even Galton and Watson did not completely solve the problems, and it wasn't until the 1930s that complete solutions were found. Fisher, Haldane, Erlang, and Steffenson contributed to the solution of the problems (Mode, 1971). Thus, appropriately, these discrete time processes are known as Galton-Watson branching processes. Schinazi (1999)

also notes Bienaymé's contributions and refers to the theory as Bienaymé-Galton-Watson branching processes. Branching processes have been applied to electron multipliers, neutron chain reactions, population growth, and the survival of mutant genes.

In the next section, some notation and preliminary results will be given. The main result regarding extinction will be given in Section 4.4. Some extensions to multitype branching processes will be introduced in Section 4.6, and in Section 4.7 they will be applied to a discrete age-structured model, a Leslie matrix model. Excellent references for branching processes with applications to biological problems include the books by Harris (1963), Jagers (1975), Kimmel and Axelrod (2002), and Mode (1971).

4.2 Definitions and Notation

Discrete time branching processes are discrete time Markov chains; the time variable and state space are discrete and the state of the system at time $n+1$ depends only on the state of the system at time n. Frequently, branching processes are studied separately from Markov chains. One reason for this separate study is the wide variety of applications in branching processes. Another reason is that different techniques other than the transition matrix are used to analyze their behavior. Techniques that employ probability generating functions are important in the study of branching processes.

The following assumptions are made in studying Galton-Watson branching processes. Let X_0 denote the total size of the population at the zeroth generation and X_n the size of the population at the nth generation. The process $\{X_n\}_{n=0}^{\infty}$ has state space $\{0, 1, 2, \ldots\}$. Assume that each individual in generation n gives birth to Y offspring in the next generation, where Y is a random variable that takes values in $\{0, 1, 2 \ldots\}$ whose offspring distribution is $\{p_k\}_{k=0}^{\infty}$,

$$\text{Prob}\{Y = k\} = p_k, \quad k = 0, 1, 2, \ldots .$$

In addition, assume that each individual gives birth *independently* from all other individuals. The process $\{X_n\}_{n=0}^{\infty}$ is referred to as a *branching process*. Figure 4.1 indicates why the process is referred to as a branching process; one sample path is graphed in the case $X_0 = 1$.

If, in any generation n, the population size is zero, $X_n = 0$, then the process stops, $X_{n+k} = 0$ for $k = 1, 2, \ldots$. Thus, the zero state is absorbing (i.e., the one-step transition probability $p_{00} = 1$). The zero state is positive recurrent. We verify later that all of the other states are transient.

We do not study branching processes in terms of the transition matrix. In fact, it would be very difficult to write down each of the one-step transition probabilities for the states $\{2, 3 \ldots\}$. Instead, we study branching processes via probability generating functions (p.g.f.'s).

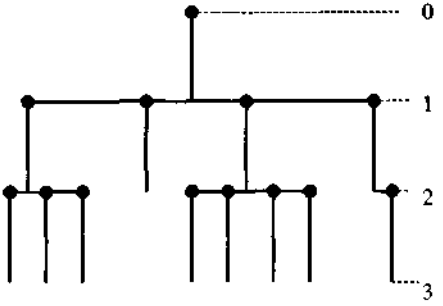

Figure 4.1. A sample path or stochastic realization of a branching process $\{X_n\}_{n=0}^{\infty}$. In the first generation, four individuals are born, $X_1 = 4$. The four individuals in generation 1 give birth to three, zero, four, and one individuals, respectively, making a total of eight individuals in generation 2, $X_2 = 8$.

Some preliminary results are needed before we define the p.g.f. for the random variable X_n. Suppose $S_2 = Y_1 + Y_2$, where Y_1 and Y_2 are independent discrete random variables with values in $\{0, 1, 2 \ldots\}$. Suppose the p.g.f. of S_2 is denoted as H and the p.g.f. of Y_i is denoted as F_i for $i = 1, 2$. Then the p.g.f. of S_2 is

$$\begin{aligned} H(t) &= \sum_{j=0}^{\infty} \text{Prob}\{S_2 = j\} t^j \\ &= E(t^{S_2}) \\ &= E(t^{Y_1+Y_2}) \\ &= E(t^{Y_1}) E(t^{Y_2}) \\ &= F_1(t) F_2(t), \end{aligned}$$

which follows because Y_1 and Y_2 are independent. The p.g.f. of S_2 is the product of the p.g.f.'s of Y_i. If the distributions of the Y_i are equal, then $F_i = F$ and $H(t) = [F(t)]^2$. In general, if $S_k = \sum_{j=1}^{k} Y_j$, where the Y_j are independent and identically distributed (iid) and k is fixed, then

$$H(t) = [F(t)]^k. \tag{4.1}$$

Now, suppose

$$S_M = \sum_{i=1}^{M} Y_i,$$

where M is not a fixed number but a random variable. Also, suppose the random variables Y_i are iid. Suppose the p.g.f.'s for the random variables

Y_i, M, and S_M are F, G, and H, respectively, and defined as follows:

$$\begin{aligned} F(t) &= \sum_{j=0}^{\infty} f_j t^j, \quad f_j = \text{Prob}\{Y_i = j\} \\ G(t) &= \sum_{j=0}^{\infty} g_j t^j, \quad g_j = \text{Prob}\{M = j\} \\ H(t) &= \sum_{j=0}^{\infty} h_j t^j, \quad h_j = \text{Prob}\{S_M = j\} \end{aligned}$$

(see Bailey, 1990). Because M is a random variable, the p.g.f. of S_M does not have a simple form given by equation (4.1). The coefficient $h_j = \text{Prob}\{S_M = j\}$ can be expressed as follows:

$$\begin{aligned} h_j &= \text{Prob}\{S_M = j\} \\ &= \sum_{m=0}^{\infty} \text{Prob}\{S_m = j | M = m\} \text{Prob}\{M = m\} \\ &= \sum_{m=0}^{\infty} g_m \text{Prob}\{S_m = j | M = m\}. \end{aligned}$$

If $M = m$ is a constant, it follows from equation (4.1) that

$$[F(t)]^m = \sum_{j=0}^{\infty} \text{Prob}\{S_m = j | M = m\} t^j. \tag{4.2}$$

Now, we use equation (4.2) to calculate the p.g.f. H of S_M, where M is a random variable (Bailey, 1990):

$$\begin{aligned} H(t) &= \sum_{j=0}^{\infty} h_j t^j = \sum_{j=0}^{\infty} t^j \sum_{m=0}^{\infty} g_m \text{Prob}\{S_m = j | M = m\} \\ &= \sum_{m=0}^{\infty} g_m \sum_{j=0}^{\infty} \text{Prob}\{S_m = j | M = m\} t^j = \sum_{m=0}^{\infty} g_m [F(t)]^m, \end{aligned}$$

where the summations can be interchanged if they converge absolutely. Thus,

$$H(t) = G(F(t)),$$

the p.g.f. of S_M is the composition of the generating function of M and the generating function of the Y's. Notice that if $M = m$ is a constant, the p.g.f. of $G(t) = t^m$, so that $G(F(t))$ simplifies to $[F(t)]^m$. We use these results to derive the p.g.f. of the branching process $\{X_n\}_{n=0}^{\infty}$.

4.3 Probability Generating Function of X_n

Suppose $X_0 = 1$ and denote the p.g.f. of X_n as h_n. The p.g.f. of X_0 is $h_0(t) = t$. In the next generation, each individual gives birth to k individuals with probability p_k. The p.g.f. of X_1 is

$$h_1(t) = \sum_{k=0}^{\infty} p_k t^k, \tag{4.3}$$

(i.e., $X_1 = Y_1$). Also, $X_2 = Y_1 + \cdots + Y_{X_1}$ because each of the X_1 individuals gives birth to Y individuals and the sum of all these births is X_2; that is,

$$\text{Prob}\{X_2 = j | X_1 = i\} = \text{Prob}\left\{\sum_{k=1}^{i} Y_k = j\right\}.$$

Because X_2 is the sum of X_1 iid random variables, where X_1 is also a random variable, we can apply the results from the previous section. The p.g.f. of X_2 is

$$h_2(t) = h_1(h_1(t)).$$

In generation $n + 1$,

$$X_{n+1} = Y_1 + Y_2 + \cdots + Y_{X_n} = \sum_{i=1}^{X_n} Y_i,$$

X_{n+1} is the sum of X_n iid random variables Y_i. Thus, the **p.g.f. of X_{n+1} is**

$$h_{n+1}(t) = h_n(h_1(t)).$$

However, $h_n(t) = h_{n-1}(h_1(t))$, so that

$$h_{n+1}(t) = h_1(h_1(\cdots(h_1(t))\cdots)),$$

where $h_1(t)$ is the p.g.f. for Y_1, equation (4.3). For simplicity, denote **$h_1(t)$** as $f(t)$. Then h_n is just an n-fold composition of f denoted as

$$f^n(t) = f(f(\cdots(f(t))\cdots)).$$

The derivation of the generating function h_n is based on the fact that $X_0 = 1$. If $X_0 = N$, where N is a positive integer, then $h_0(t) = t^N$ and the process begins with N independent branches (see Figure 4.2). When $X_0 = N$, the process may be considered as N independent branching processes, $X_n = \sum_{i=1}^{N} X_{in}$, where X_{in} are independent and identically distributed random variables. Each of the random variables has a p.g.f. given by $f^n(t)$. Because N is a fixed number, it follows from the previous section that the p.g.f. of X_n is

$$h_n(t) = [f^n(t)]^N, \quad \text{when} \quad X_0 = N.$$

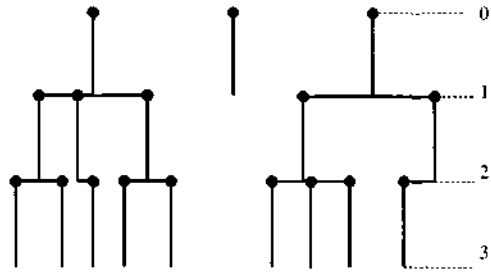

Figure 4.2. A sample path or stochastic realization of a branching process $\{X_n\}_{n=0}^{\infty}$, where $X_0 = 3$, $X_1 = 5$, and $X_2 = 9$.

Example 4.1 Suppose a branching process $\{X_n\}_{n=0}^{\infty}$ with $X_0 = 1$ has an offspring distribution $\{p_k\}$ satisfying $p_0 = 1/4$, $p_1 = 3/4$, and $p_k = 0$, $k = 2, 3, \ldots$. There are either no births or a single birth. Then the p.g.f. for X_1 is

$$f(t) = \frac{1}{4} + \frac{3}{4}t.$$

The p.g.f for X_2 is

$$f^2(t) = f(f(t)) = \frac{1}{4}\left(1 + \frac{3}{4}\right) + \left(\frac{3}{4}\right)^2 t.$$

In general, the p.g.f. for X_n is

$$\begin{aligned} f^n(t) &= \frac{1}{4}\left(1 + \frac{3}{4} + \cdots + \left(\frac{3}{4}\right)^{n-1}\right) + \left(\frac{3}{4}\right)^n t \\ &= 1 - \left(\frac{3}{4}\right)^n + \left(\frac{3}{4}\right)^n t. \end{aligned}$$

If $p_k(n)$ is the probability that the population size is k in **generation n, then**

$$p_0(n) = 1 - \left(\frac{3}{4}\right)^n \quad \text{and} \quad p_1(n) = \left(\frac{3}{4}\right)^n.$$

If $X_0 = N$, then the p.g.f. satisfies

$$\begin{aligned} [f^n(t)]^N &= \binom{N}{0}(p_0(n))^N + \binom{N}{1}(p_0(n))^{N-1}p_1(n)t \\ &\quad + \cdots + \binom{N}{N}(p_1(n))^N t^N. \end{aligned} \tag{4.4}$$

■

We should be careful not to confuse the notation p_k and $p_k(n)$. The notation p_k refers to the probability an individual gives birth to k individuals, and $p_k(n)$ refers to the probability that the total population size is k at generation n. This latter notation is consistent with previous chapters. Note that when $X_0 = 1$, $p_k = p_k(1)$ for $k = 0, 1, 2, \ldots$.

Galton's question (1) can be addressed for this example. After r generations, the probability that all surnames have gone extinct is $[p_0(r)]^N$, the probability that $N-1$ surnames have gone extinct is $N[p_0(r)]^{N-1}p_1(r)$, and so on; so that the probability that no surnames have gone extinct is $[p_1(r)]^N$. It can then be shown that the expected proportion of surnames that have gone extinct by generation r is $p_0(r)$. In Example 4.1, $p_0(r) = 1 - (3/4)^r$. Notice, for this example,

$$\lim_{r\to\infty} p_0(r) = 1.$$

We do not address Galton's question (2) in general but note that in a single branching process, $X_0 = 1$, the probability there are exactly m surnames the same in generation r is $p_m(r)$. In Example 4.1, the probability that there are two or more equal surnames in any generation is zero.

In the next section, the probability of population extinction as $n \to \infty$ is studied for the branching process $\{X_n\}$; that is, $\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = \lim_{n\to\infty} p_0(n)$.

4.4 Probability of Population Extinction

Denote the p.g.f. of X_n by

$$h_n(t) = \sum_{k=0}^{\infty} p_k(n)t^k,$$

where the probability the total population size is k in generation n is given by $p_k(n)$. Denote the probability distribution of the process at time zero by $p(0) = (p_0(0), p_1(0), \ldots)^T$ and the probability distribution of the process at time n as $p(n) = (p_0(n), p_1(n), \ldots)^T$. The probability of total population extinction in the nth generation is $p_0(n) = h_n(0)$. If $X_0 = 1$, then $h_n(0) = f^n(0)$, and if $X_0 = N$, then $h_n(0) = [f^n(0)]^N$.

The following assumptions are made regarding the offspring distribution $\{p_k\}_{k=0}^{\infty}$. Assume

$$0 < p_0 \quad \text{and} \quad 0 < p_0 + p_1 < 1. \tag{4.5}$$

Assumptions (4.5) imply that there is a positive probability of no births occurring and a positive probability of more than one birth. If $p_0 = 0$ or $p_1 = 1$, then in every generation there is at least one birth and there is no chance of extinction, $p_0(n) = 0$. The probability of ultimate extinction is zero in these cases. In the case of a linear p.g.f, $f(t) = p_0 + p_1 t$, $p_0 > 0$, it can be seen from Example 4.1 that the p.g.f. satisfies $f^n(t) = 1 - (p_1)^n + p_1^n t^n$,

where $p_0(n) = 1 - (p_1)^n$. Thus, $\lim_{n\to\infty} p_0(n) = 1$. The assumptions (4.5) exclude these few cases for which the asymptotic results have already been verified.

The p.g.f. for the offspring distribution $\{p_k\}$ is

$$f(t) = \sum_{k=0}^{\infty} p_k t^k. \tag{4.6}$$

Denote the mean number of births as m,

$$m = f'(1) = \lim_{t\to 1^-} f'(t) = \sum_{k=1}^{\infty} k p_k. \tag{4.7}$$

In addition, assume the p.g.f. has the following five properties:

(1) $f(0) = p_0 > 0$ and $f(1) = 1$.

(2) $f(t)$ is continuous for $t \in [0, 1]$.

(3) $f(t)$ is infinitely differentiable for $t \in [0, 1)$.

(4) $f'(t) = \sum_{k=1}^{\infty} k p_k t^{k-1} > 0$ for $t \in (0, 1]$, where $f'(1)$ is defined by (4.7).

(5) $f''(t) = \sum_{k=2}^{\infty} k(k-1) p_k t^{k-2} > 0$ for $t \in (0, 1)$.

The five properties imply that the function f is continuous, strictly increasing, and its first derivative is strictly increasing (concave upward) on $[0, 1]$. The graph of the p.g.f. $y = f(t)$ may intersect $y = t$ in either one or two points on the interval $[0, 1]$. (See Figure 4.3.) Properties (1)–(5) are used to prove two lemmas and the main result concerning ultimate extinction of the branching process $\{X_n\}$.

Lemma 4.1. *Assume the offspring distribution $\{p_k\}$ and the p.g.f. $f(t)$ satisfy inequalities (4.5) and properties (1)–(5). Then $m \leq 1$ if and only if $f'(t) < 1$ for $t \in [0, 1)$.*

Proof. Since f' is strictly increasing on $[0, 1]$ [property (5)] and $m = f'(1) \leq 1$, it follows that $f'(t) < 1$ for $t \in [0, 1)$. The converse is straightforward. □

Lemma 4.2. *Assume the offspring distribution $\{p_k\}$ and the p.g.f. $f(t)$ satisfy inequalities (4.5) and properties (1)–(5). If $m \leq 1$, then $f(t)$ has a unique fixed point at $t = 1$ on the interval $[0, 1]$.*

Proof. To show that the fixed point is unique, note that Lemma 4.1 implies $f'(t) < 1$ for $t \in [0, 1)$. Integration from t to 1 yields $1 - f(t) < 1 - t$ or $t < f(t)$ for $t \in [0, 1)$. Thus, the only fixed point of f on $[0, 1]$ is $t = 1$. □

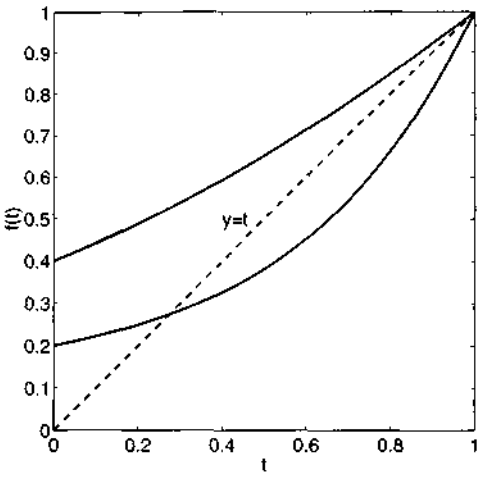

Figure 4.3. Two different probability generating functions $y = f(t)$ intersect $y = t$ in either one or two points on $[0, 1]$.

Theorem 4.1. *Assume the offspring distribution* $\{p_k\}$ *and the p.g.f.* $f(t)$ *satisfy inequalities (4.5) and properties (1)-(5). In addition, assume* $X_0 = 1$. *If* $m \leq 1$, *then*

$$\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = \lim_{n\to\infty} p_0(n) = 1$$

and if $m > 1$, *then there exists* $q < 1$ *such that* $f(q) = q$ ***and***

$$\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = \lim_{n\to\infty} p_0(n) = q.$$

If $m \leq 1$, then Theorem 4.1 states that the probability of ultimate extinction is one. If $m > 1$, then there is a positive probability $1 - q$ that the branching process does not become extinct (e.g., a family name does not become extinct, a mutant gene becomes established, a population does not die out). For a proof with the assumption $p_0 + p_1 < 1$ instead of (4.5), see Schinazi (1999).

The branching process can be divided into three cases: $m > 1$, $m = 1$, and $m < 1$. The case $m > 1$ is referred to as *supercritical*, $m = 1$ is referred to as *critical*, and $m < 1$ is referred to as *subcritical* (Jagers, 1975; Kimmel and Axelrod, 2002).

Proof. First, we show that $\{p_0(n)\}$ is a monotone increasing sequence:

$$p_0 = p_0(1) < p_0(2) < p_0(3) < \cdots < p_0(n) < \cdots \leq 1.$$

By property (4), f is strictly increasing on [0,1], so that $0 < f(0) = p_0 < 1$ implies $f(0) < f(p_0) < f(1) = 1$. But $p_0 = p_0(1)$ and $f(p_0) = f(f(0)) =$

$p_0(2)$ so that $p_0(1) < p_0(2) < 1$. Assume $p_0(n-1) < p_0(n) < 1$. Then again, since f is strictly increasing and $f(p_0(k)) = p_0(k+1)$, it follows $f(p_0(n-1)) < f(p_0(n)) < f(1)$ or $p_0(n) < p_0(n+1) < 1$. The monotonicity of this sequence can also be deduced logically since $p_0(n)$ is the probability of extinction by the time n that includes the probability of extinction at times $1, 2, \ldots, n$.

The sequence $\{p_0(n)\}$ is monotone increasing and bounded above by one. Therefore, it has a limit. Let

$$q = \lim_{n \to \infty} p_0(n).$$

Thus, by the continuity of f,

$$q = \lim_{n \to \infty} f(p_0(n-1)) = f(q).$$

The limit q is a fixed point of f, $f(q) = q$, where $q \leq 1$.

Suppose $m \leq 1$. By Lemma 4.2, the only fixed point of f on $[0,1]$ is one, so that $q = 1$. The graph of f lies above $y = t$. See Figure 4.3.

Suppose $m > 1$. We show that f has only two fixed points on $[0,1]$, q and 1, where $0 < q < 1$. Since f' is strictly increasing and continuous on $[0,1]$, there exists $0 < r < 1$ such that if $r < s < 1$, then $1 < f'(s) < f'(1) = m$. Integration of $f'(t)$ from s to 1 yields $1 - f(s) > 1 - s$ or $s > f(s)$ for $r < s < 1$ [the graph of $y = f(t)$ lies below $y = t$ for $t \in (r,1)$]. See Figure 4.3.

Let $s \in (r,1)$. Consider the function $g(t) = f(t) - t$. Then $g(0) = p_0 > 0$ and $g(s) = f(s) - s < 0$. By the intermediate value theorem, there exists a $q \in (0,s)$ such that $g(q) = 0$ or $f(q) = q$. Now, we show that q is the unique fixed point on (0,1). There can be no fixed point on the interval $(r,1)$ since $f(t) < t$ for $t \in (r,1)$. Suppose there exists another fixed point $u \in (0,1)$. Either $u \in (0,q)$ or $u \in (q,1)$. In either case, $g(q) = 0$, $g(u) = 0$ and $g(1) = 0$. By Rolle's theorem, there exist numbers u_1 and u_2 such that $0 < u < u_1 < q < u_2 < 1$ if $u \in (0,q)$ or $q < u_1 < u < u_2 < 1$ if $u \in (q,1)$ such that $g'(u_1) = 0 = g'(u_2)$. Then $f'(u_1) = 1 = f'(u_2)$. This is a contradiction because f' is strictly increasing on $(0,1)$. Thus, f has only two fixed points on $[0,1]$-namely, q and 1.

Next, we show that $\lim_{n \to \infty} p_0(n) = q < 1$. Suppose

$$\lim_{n \to \infty} p_0(n) = 1.$$

Then, for sufficiently large n, $p_0(n) > r$. But on the interval $(r,1)$, the graph of $f(t)$ lies below the line $y = t$ so that

$$p_0(n) > f(p_0(n)) = p_0(n+1).$$

This contradicts the fact that $\{p_0(n)\}$ is an increasing sequence. Hence, $\lim_{n \to \infty} p_0(n) = q < 1$. □

Although the special case where the p.g.f. is linear, $f(t) = p_0 + p_1 t$, $p_0 > 0$, was verified separately, it satisfies the results of the theorem. In this case, $m = p_1 \leq 1$ so that $\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = 1$.

Example 4.2 Suppose the offspring distribution $\{p_k\}$ satisfies

$$p_0 = 1/5, \quad p_1 = 1/2, \quad p_2 = 3/10,$$

and $p_k = 0$ for $k = 3, 4, \ldots$. Then $m = 1/2 + 2(3/10) = 11/10 > 1$, so that the probability of ultimate extinction is the fixed point of $f(t) = 1/5 + t/2 + 3t^2/10$ on $(0,1)$. The solutions to $f(t) = t$ are $t = 1$ and $t = 2/3$:

$$f(t) - t = \frac{1}{5} - \frac{1}{2}t + \frac{3}{10}t^2 = \frac{1}{10}(3t - 2)(t - 1) = 0.$$

The probability of ultimate extinction is $2/3$. ■

Example 4.3 (Schinazi, 1999) Lotka assumed a geometric distribution to fit the offspring of the American male population. Suppose the number of sons a male has in his lifetime has the following geometric probability distribution:

$$p_0 = 1/2 \quad \text{and} \quad p_k = \left(\frac{3}{5}\right)^{k-1} \frac{1}{5} \quad \text{for} \quad k = 1, 2, \ldots.$$

Note that $\sum_{k=1}^{\infty} p_k = 1/2$ and

$$f(t) = \frac{1}{2} + \frac{1}{5} \sum_{k=1}^{\infty} \left(\frac{3}{5}\right)^{k-1} t^k = \frac{1}{2} + \frac{1}{5} \left(\frac{t}{1 - 3t/5}\right).$$

To find m, note that

$$m = f'(1) = \frac{1/5}{(1 - 3/5)^2} = \frac{5}{4} > 1.$$

The fixed points of $f(t)$ are found by solving

$$\frac{1}{2} + \frac{t}{5 - 3t} = t \quad \text{or} \quad 6t^2 - 11t + 5 = 0.$$

This latter equation factors into $(6t - 5)(t - 1) = 0$, so that $q = 5/6$. A male has a probability of $5/6$ that his line of descent becomes extinct and a probability of $1/6$ that his descendants will continue forever. ■

Theorem 4.1 can be extended to the case $X_0 = N > 1$. Recall that the p.g.f. of X_n in this case is $[f^n(t)]^N$. Thus, the probability of extinction at the nth generation is $[f^n(0)]^N = [p_0(n)]^N$. The result is stated in the following corollary.

Corollary 4.1. *Assume the offspring distribution $\{p_k\}$ and the p.g.f. $f(t)$ satisfy inequalities (4.5) and properties (1)-(5). In addition, assume $X_0 = N$. If $m \leq 1$, then*

$$\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = \lim_{n\to\infty} [p_0(n)]^N = 1.$$

If $m > 1$, then

$$\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = \lim_{n\to\infty} [p_0(n)]^N = q^N < 1.$$

The corollary also holds in the case of a linear p.g.f., $f(t) = p_0 + p_1 t$, $p_0 > 0$, because according to Example 4.1, $[f^n(0)]^N = [1 - p_1^n]^N$. The mean $m = p_1 \leq 1$ and $0 < p_1$, $\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = \lim_{n\to\infty} [p_0(n)]^N = 1$. In the two special cases, $p_0 = 0$ or $p_0 = 1$, it is impossible for ultimate extinction to occur, $\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = 0$.

The zero state of the branching process is positive recurrent. Next, it is shown that the remaining states are transient. In addition, it is shown that either the total population size X_n approaches zero or infinity (see Harris, 1963).

Theorem 4.2. *Assume the offspring distribution $\{p_k\}$ and the p.g.f. $f(t)$ satisfy inequalities (4.5) and properties (1)-(5). In addition, assume $X_0 = 1$. Then the states $1, 2, \ldots,$ are transient. In addition, if the mean $m > 1$, then $\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = q$, where $0 < q < 1$ is the unique fixed point of the p.g.f., $f(q) = q$, and $\text{Prob}\{\lim_{n\to\infty} X_n = \infty\} = 1 - q$.*

Proof. We consider the first return to state k, where $k \in \{1, 2, \ldots\}$. The process begins in state $k = 1$, $X_0 = 1$. Let m be the first time such that $X_m = k$ for $k \neq 1$. If there exists no such time m, then state k is automatically transient. On the other hand, if there exists such a time m, then extend the definition of first return to state k for $k \neq 1$ as follows. Define the first return to state k, $k \neq 1$, at the nth generation as

$$f_{kk}^{(n)} = \text{Prob}\{X_{m+n} = k, X_{m+j} \neq k, j = 1, 2, \ldots, n-1 | X_m = k\},$$

where $f_{kk}^{(0)} = 0$. Then define

$$f_{kk} = \sum_{n=0}^{\infty} f_{kk}^{(n)}.$$

Thus, the first return probability is defined for all states $k = 1, 2, \ldots$. Recall that a state k is transient iff $f_{kk} < 1$.

Let p_{0k} be the probability that beginning in state k, the process is in state 0 in the next generation; that is,

$$p_{0k} = \text{Prob}\{X_{m+1} = 0 | X_m = k\}.$$

Since zero is an absorbing state, $X_n = 0$ for $n \geq m+1$; the process cannot leave the zero state. Therefore, there is a positive probability of at least p_{0k} that the process never returns to state k. Hence,

$$f_{kk} \leq 1 - p_{0k} < 1.$$

Therefore, every state k is transient, where $k \in \{1, 2, \ldots\}$.

For a transient state k, the n step transition probability $p_{kj}^{(n)} \to 0$ as $n \to \infty$, for any state j; that is,

$$\lim_{n \to \infty} \text{Prob}\{X_n = k\} = 0$$

(see Lemma 2.2). Because X_n cannot approach any finite state k as $n \to \infty$, either X_n approaches 0 or X_n approaches infinity. From Theorem 4.1, it follows that $\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = q$. If $m \leq 1$, then $q = 1$ and if $m > 1$, then $0 < q < 1$. □

If $m > 1$, then, as $n \to \infty$, the population size approaches zero with probability q and approaches infinity with probability $1 - q$. Theorem 4.2 also holds in the case $p_0 = 0$. But in this case, $q = 0$; the process approaches infinity with probability one. Next, expressions are derived for the mean and variance of the process.

4.5 Mean and Variance of X_n

Generating functions can be used to find the mean and variance of X_n, the random variable for the total population size in generation n. Recall from Chapter 1 the properties of the p.g.f., $f(x)$, moment generating function (m.g.f.), $M(t) = f(e^t)$, and cumulant generating function (c.g.f.), $K(t) = \ln M(t)$. These functions satisfy

$$f(1) = 1, \quad f'(1) = m = E(X), \quad f''(1) = E(X(X-1)).$$

$$M(0) = 1, \quad M'(0) = m, \quad M''(0) = E(X^2).$$

$$K(0) = 0, \quad K'(0) = m, \quad K''(0) = \sigma^2 = E[(X-m)^2].$$

In general, we shall denote the mean and variance of X_n as m_n and σ_n^2, respectively, and the generating functions associated with X_n as $f^n(x)$, $M_n(t)$, and $K_n(t)$. For the random variable X_1, in the first generation, $m_1 = m$ and $\sigma_1^2 = \sigma^2$; that is,

$$m_1 = m = \sum_{k=1}^{\infty} k p_k \quad \text{and} \quad \sigma_1^2 = \sigma^2 = \sum_{k=1}^{\infty} k^2 p_k - m^2.$$

In addition, the generating functions of X_1 satisfy $f_1 = f$, $M_1 = M$, and $K_1 = K$. The following theorem gives the mean and variance of the process when $X_0 = 1$.

Theorem 4.3. *Assume $X_0 = 1$. The mean of the random variable X_n, the size of the population in generation n, is*

$$m_n = E(X_n) = m^n$$

and the variance is

$$\sigma_n^2 = E[(X_n - m_n)^2] = \begin{cases} \dfrac{m^{n-1}(m^n - 1)}{m - 1}\sigma^2, & m \neq 1 \\ n\sigma^2, & m = 1. \end{cases}$$

Note that when $m = 1$, $m_n = 1$ and $\sigma_n^2 = n\sigma^2 \to \infty$ if $\sigma^2 \neq 0$. Before we prove this theorem we note some properties of the generating functions. Recall that $M_n(t) = f^n(e^t)$. Then

$$\begin{aligned} M_n(t) &= f^n(e^t) = f^{n-1}(f(e^t)) = f^{n-1}(M(t)) \\ &= f^{n-1}(e^{\ln M(t)}) = M_{n-1}(\ln M(t)). \end{aligned}$$

Thus, $M_n(t) = M_{n-1}(K(t))$. By taking natural logarithms of this identity, we obtain the following identity for the c.g.f.:

$$K_n(t) = K_{n-1}(K(t)).$$

Since $f(f^{n-1}(e^t)) = f^n(e^t)$, it can also be shown that

$$K_n(t) = K(K_{n-1}(t)).$$

The first and second derivatives of the first identity yield two relationships that are used to verify Theorem 4.3:

$$K_n'(t) = K_{n-1}'(K(t))K'(t) \tag{4.8}$$

$$K_n''(t) = K_{n-1}''(K(t))[K'(t)]^2 + K_{n-1}'(K(t))K''(t). \tag{4.9}$$

Proof of Theorem 4.3. The proof follows Bailey (1990). First, identity (4.8) is evaluated at $t = 0$,

$$\begin{aligned} K_n'(0) &= K_{n-1}'(K(0))K'(0), \\ m_n &= m_{n-1}m \end{aligned}$$

because $K_n(0) = 0$, $K_n'(0) = m_n$, and $m_1 = m$. The equation $m_n - mm_{n-1} = 0$ is a first-order, homogeneous, constant coefficient, difference equation in m_n. The solution is

$$m_n = m^n.$$

Second, identity (4.9) is evaluated at $t = 0$,

$$\begin{aligned} K_n''(0) &= K_{n-1}''(K(0))[K'(0)]^2 + K_{n-1}'(K(0))K''(0) \\ \sigma_n^2 &= \sigma_{n-1}^2 m^2 + m_{n-1}\sigma^2. \end{aligned}$$

Substituting $m_{n-1} = m^{n-1}$, then $\sigma_n^2 - m^2\sigma_{n-1}^2 = m^{n-1}\sigma^2$ is a first-order, nonhomogeneous, constant coefficient, difference equation in σ_n^2. The general solution to this difference equation is a sum of the general solution to the homogeneous equation and a particular solution. The general solution to the homogeneous equation is cm^{2n}. Assume the particular solution has the form $\sigma_n^2 = km^{n-1}$, $m \neq 1$. Substituting this value into the difference equation yields

$$\begin{aligned} km^{n-1} - m^2km^{n-2} &= m^{n-1}\sigma^2 \\ m^{n-1}[k - km - \sigma^2] &= 0 \\ k(1-m) &= \sigma^2 \\ k &= \frac{\sigma^2}{1-m} \end{aligned}$$

provided that $m \neq 1$. Thus, the general solution to the nonhomogeneous difference equation is

$$\sigma_n^2 = cm^{2n} + \frac{\sigma^2 m^{n-1}}{1-m}, \quad m \neq 1.$$

The constant c is found by setting $\sigma_1^2 = \sigma^2$. Then $\sigma^2 = cm^2 + \sigma^2/(1-m)$ or $c = \sigma^2/[m(m-1)]$. The solution to σ_n^2 is

$$\sigma_n^2 = \frac{m^{n-1}(m^n - 1)}{m-1}\sigma^2, \quad m \neq 1.$$

In the case $m = 1$, the particular solution has the form kn. Substitution of this solution into the difference equation gives

$$kn - k(n-1) = \sigma^2$$

or $k = \sigma^2$. Thus the general solution to the nonhomogeneous difference equation is $\sigma_n^2 = c + n\sigma^2$. Application of $\sigma_1^2 = \sigma^2$ yields $c = 0$. Thus, the solution to the difference equation is

$$\sigma_n^2 = n\sigma^2, \quad m = 1.$$

The proof is complete. □

In the case that $m = 0$, then $p_0 = 1$ and $\sigma^2 = 0$. The population becomes extinct in one generation. In general, in the subcritical case, $m < 1$, the process decays geometrically. In the critical case, $m = 1$, the process is constant, and in the supercritical case, the process increases geometrically.

The conditional expectation in the case $X_0 = 1$ can be expressed in terms of the mean:

$$E(X_{n+1}|X_n) = E\left(\sum_{i=1}^{X_n} Y_i | X_n\right) = E(X_nY_i|X_n) = X_nE(Y_i) = mX_n,$$

because all of the random variables Y_i are identically distributed with mean m. In general, it follows from the Markov property and by induction that

$$E(X_{n+r}|X_n) = m^r X_n \tag{4.10}$$

(see Karlin and Taylor, 1975).

Example 4.4 Consider the p.g.f. in Example 4.2, $f(t) = 1/5+t/2+3t^2/10$. The mean $m = f'(1) = 1.1$ and variance $\sigma^2 = f''(1) + f'(1) - [f'(1)]^2 = 0.6 + 1.1 - (1.1)^2 = 0.49$. Therefore, the mean and variance for X_n satisfy

$$m_n = (1.1)^n \quad \text{and} \quad \sigma_n^2 = \left[(1.1)^{2n-1} - (1.1)^{n-1}\right] 4.9. \qquad \blacksquare$$

The results of the theorems are applied to an example on the spread of a mutant gene (Bailey, 1990).

Example 4.5 Suppose the population size is very large. A new mutant gene appears in N individuals of the population; the remaining individuals in the population are normal; they do not carry the mutant gene. Individuals without the mutant gene and those with the mutant gene reproduce according to a branching process. Suppose the mean number of individuals with a mutant gene that are produced by a mutant individual is m. If $m \leq 1$, then the line of descendants from the individual with a mutant gene will eventually become extinct. Suppose the mean reproductive potential of a normal individual is 1 but the mean of a mutant individual is greater than 1. Then $m > 1$, but suppose it is only slightly larger than 1,

$$m = 1 + \epsilon, \quad \epsilon > 0.$$

Then there is a probability q, $q = f(q)$, that the subpopulation with the mutant gene will become extinct. Note that q is close to 1 since ϵ is small. The value of q can be approximated from the mean and variance without knowing the p.g.f. f. Let $q = e^{\theta}$. Then $e^{\theta} = f(e^{\theta}) = M(\theta)$. Also,

$$\begin{aligned} \theta &= \ln M(\theta) = K(\theta) \\ &= 0 + m\theta + \sigma^2 \frac{\theta^2}{2!} + \cdots \end{aligned}$$

because $K(0) = 0$, $K'(0) = m$, and $K''(0) = \sigma^2$. Now, $q = e^{\theta} \approx 1$, so that $\theta \approx 0$, but $\theta < 0$. Thus, the preceding Maclaurin series can be truncated:

$$\theta \approx m\theta + \sigma^2 \frac{\theta^2}{2}.$$

Solving this equation for θ,

$$\theta \approx \frac{2}{\sigma^2}(1 - m) = -\frac{2}{\sigma^2}\epsilon \quad \text{or} \quad q \approx e^{-\frac{2}{\sigma^2}\epsilon}.$$

N	q^N	$1-q^N$
1	0.9804	0.0196
100	0.1380	0.8620
200	0.0191	0.9809
300	0.0026	0.9974

Table 4.1. Approximations to the probability that a mutant gene becomes extinct, q^N, or becomes established, $1-q^N$, when there are initially N mutant genes in a population with a Poisson offspring distribution and $m = 1.01 = \sigma^2$

For an initial size of N mutants, the chance of extinction is q^N or

$$q^N \approx e^{-\frac{2N}{\sigma^2}\epsilon}.$$

For example, in the case of a Poisson distribution, $m = \sigma^2 = 1+\epsilon$. Suppose $\epsilon = 0.01$. Then

$$q^N \approx e^{-\frac{2N}{1.01}(0.01)} = e^{-N0.01980198\ldots} \approx (0.98039)^N.$$

The probability that the mutant gene becomes established in the population is $1-q^N$. See Table 4.1. ■

4.6 Multitype Branching Processes

In a multitype Galton-Watson process, it is assumed that each individual behaves independently of any other individual. In the Galton-Watson process just described and analyzed, each individual gives birth and is replaced by its progeny. Every individual is of the same "type"–that is, gives birth to new individuals, each with the same probability distribution from generation to generation. In a multitype branching process, each individual may give "birth" to different "types" or "classifications" of individuals in the population. There is an offspring distribution corresponding to each of these different types of individuals. For example, the population may be divided according to age or size and in each generation, individuals may "age" or "grow" to another age or size class. In addition, in each generation, individuals give birth to new individuals in the youngest age or smallest size class. In the next section, the multitype branching process is applied to a population structured according to age, a stochastic model related to the well-known Leslie matrix model (Leslie, 1945). References for multitype branching processes include the books by Harris (1963), Karlin and Taylor (1975), Kimmel and Axelrod (2002), and Mode (1971).

The notation is changed slightly from the previous section. Denote the multitype branching process as $\{X(n)\}_{n=0}^{\infty}$, where $X(n)$ is a vector of random variables, $X(n) = (X_1(n), X_2(n), \ldots, X_k(n))^T$, with k different types of individuals. Here, the subscript i in $X_i(n)$ denotes the ith component of the vector random variable $X(n)$ and n is the time step. In addition, each random variable $X_i(n)$ has k associated random variables, $\{Y_{ji}\}_{j=1}^{k}$.

Each random variable Y_{ji} has an associated offspring distribution describing the probability an individual of type i gives "birth" to an individual of type j, for $j = 1, 2, \ldots, k$. For each $i = 1, 2, \ldots, k$ assume that $(Y_{1i}^l, Y_{2i}^l, \ldots, Y_{ki}^l)$ are iid for $l = 1, 2, \ldots$. Each Y_{ji}^l has the same offspring distribution as Y_{ji} for $l = 1, 2, \ldots$. Let $p_i(s_1, s_2, \ldots, s_k)$ denote the probability an individual of type i gives birth to s_1 individuals of type 1, s_2 individuals of type 2, ..., and s_k individuals of type k; that is,

$$p_i(s_1, s_2, \ldots, s_k) = \text{Prob}\{Y_{1i}^l = s_1, Y_{2i}^l = s_2, \ldots, Y_{ki}^l = s_k\}$$

for $s_j = 0, 1, 2, \ldots$ and $i, j = 1, 2, \ldots, k$. Define the k-dimensional p.g.f.'s $f_i : [0,1]^k \to [0,1]$ as follows:

$$f_i(t_1, t_2, \ldots, t_k) = \sum_{s_k=0}^{\infty} \cdots \sum_{s_2=0}^{\infty} \sum_{s_1=0}^{\infty} p_i(s_1, s_2, \ldots, s_k) t_1^{s_1} t_2^{s_2} \cdots t_k^{s_k},$$

for $i = 1, 2, \ldots, k$.

Let δ_i denote a k-vector with the ith component being one and the remaining components zero, $\delta_i = (\delta_{1i}, \delta_{2i}, \ldots, \delta_{ki})^T$, where δ_{ij} is the Kronecker delta symbol. Then $X(0) = \delta_i$ means there is initially one individual of type i in the population. The p.g.f. for $X_i(0)$ given $X(0) = \delta_i$ is $f_i^0(t_1, t_2, \ldots, t_k) = t_i$. Then the p.g.f. for $X_i(n)$ is denoted $f_i^n(t_1, t_2, \ldots, t_k)$ and defined by

$$\sum_{s_k=0}^{\infty} \cdots \sum_{s_2=0}^{\infty} \sum_{s_1=0}^{\infty} \text{Prob}\{X_1(n) = s_1, \ldots, X_k(n) = s_k | X(0) = \delta_i\} t_1^{s_1} t_2^{s_2} \cdots t_k^{s_k}.$$

For $n = 1$, $f_i^1(t_1, t_2, \ldots, t_k) = f_i(t_1, t_2, \ldots, t_k)$. Let

$$F \equiv F(t_1, \ldots, t_k) = (f_1(t_1, \ldots, t_k), \ldots, f_k(t_1, \ldots, t_k))$$

denote the vector of p.g.f.'s, $F : [0,1]^k \to [0,1]^k$. The function F has a fixed point at $(1, 1, \ldots, 1)$, since for each i, $f_i(1, 1, \ldots, 1) = 1$. Ultimate extinction of the population depends on whether F has another fixed point in $[0,1]^k$. The following theorem on extinction (Theorem 4.4) is an extension of Theorem 4.1, and as in Theorem 4.1, the probability of extinction depends on the value of the mean. The analogue of the mean for a multitype branching process is defined next.

The mean number of births of a j-type of an individual by an i-type individual is defined. Let m_{ji} denote the expected number of "births" of a

type j individual by a type i individual; that is,

$$m_{ji} = E\{X_j(1)|X(0) = \delta_i\} \quad \text{for} \quad i, j = 1, 2, \ldots, k.$$

The means m_{ji} can be defined in terms of the p.g.f.'s,

$$m_{ji} = \left.\frac{\partial f_i(t_1, \ldots, t_k)}{\partial t_j}\right|_{t_1=1,\ldots,t_k=1}.$$

Define the $k \times k$ **expectation matrix,**

$$M = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ m_{21} & m_{22} & \cdots & m_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ m_{k1} & m_{k2} & \cdots & m_{kk} \end{pmatrix}.$$

Matrix M is nonnegative. If matrix M is regular (i.e., some power of M is strictly positive, $M^p > 0$, for some $p > 0$), then M has a simple eigenvalue of maximum modulus (Gantmacher, 1964). Denote this eigenvalue as λ. The main theorem regarding ultimate extinction in a multitype branching process assumes that M is a nonnegative regular matrix. Extinction depends on the magnitude of λ. We state the theorem but do not include a proof. A proof in the two-dimensional case when M is positive can be found in Karlin and Taylor (1975), and for the more general case, see Harris (1963) or Mode (1971).

Theorem 4.4. *Assume each of the components functions f_i of the p.g.f. F, where*

$$F(t_1, \ldots, t_k) = (f_1(t_1, \ldots, t_k), \ldots, f_k(t_1, \ldots, t_k))$$

are nonlinear functions of the variables $t_1, \ldots, t_k$ and the expectation matrix M is regular. If the dominant eigenvalue λ of M satisfies $\lambda \leq 1$, then

$$\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0}|X(0) = \delta_i\} = 1,$$

$i = 1, 2, \ldots, k$. If the dominant eigenvalue of M satisfies $\lambda > 1$, then there exists a vector $q = (q_1, q_2, \ldots, q_k)^T$, $q_i \in [0, 1)$, $i = 1, 2, \ldots, k$, the unique nonnegative solution to $F(t_1, t_2, \ldots, t_k) = (t_1, t_2, \ldots, t_k)$, such that

$$\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0}|X(0) = \delta_i\} = q_i,$$

$i = 1, 2, \ldots, k$.

Theorem 4.4 excludes the case that the p.g.f.'s are linear; that is,

$$f_i(t_1, \ldots, t_k)) \neq p_i(0, 0, \ldots, 0) + p_i(1, 0, \ldots, 0)t_1 + \cdots + p_i(0, 0, \ldots, 1)t_k$$

for all i. In the case of a single branching process, the p.g.f. is linear when $p_0 + p_1 = 1$. This case was studied separately. In Theorems 4.1 and 4.2 and Corollary 4.1, the case of a linear p.g.f. was also excluded. It was assumed that $0 < p_0 + p_1 < 1$; inequalities (4.5) were assumed to hold.

Corollary 4.1 can be extended to the multitype branching case. If $X(0) = (r_1, r_2, \ldots, r_k)^T$, $r_i \geq 0$, $i = 1, 2, \ldots, k$, then

$$\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0} | X(0) = (r_1, r_2, \ldots, r_k)^T\} = q_1^{r_1} q_2^{r_2} \cdots q_k^{r_k}.$$

For the multitype branching process, the zero state is an absorbing state. It can be shown under the hypotheses of Theorem 4.4 that all other states are transient (see Harris, 1963).

Example 4.6 This example shows the importance of the assumptions in Theorem 4.4. Consider a two-dimensional multitype branching process, where $f_1(t_1, t_2) = t_1 t_2$ and $f_2(t_1, t_2) = 1$. The p.g.f. f_2 is linear. The expectation matrix satisfies

$$M = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

Hence, M is not regular. If $X(0) = \delta_1 = (1, 0)^T$, then $X(n) = (1, 1)^T$, for $n \geq 1$, and if $X(0) = \delta_2 = (0, 1)^T$, then $X(n) = (0, 0)^T$, for $n \geq 1$. According to the generating functions, an individual of type 1 gives birth to an individual of type 1 and type 2 with probability 1, but an individual of type 2 gives birth to zero individuals of type 1 or type 2. The chain is reducible. Only states $(0, 0)^T$ and $(1, 1)^T$ can be reached when $X(0) = \delta_i$, $i = 1, 2$. State $(1, 1)^T$ is not transient. ■

Example 4.7 Consider a two-dimensional multitype branching process. Suppose the p.g.f.'s satisfy

$$f_1(t_1, t_2) = \frac{1}{4}(1 + t_1 + t_2^2 + t_1^2 t_2) \quad \text{and} \quad f_2(t_1, t_2) = \frac{1}{4}(1 + t_1 + t_2^2 + t_1 t_2^2).$$

For example, an individual of type 1 gives birth to a single individual of the same type or two individuals of type 2 or two individuals of type 1 and one individual of type 2, each with probability 1/4. The expectation matrix satisfies

$$M = \begin{pmatrix} 3/4 & 1/2 \\ 3/4 & 1 \end{pmatrix}.$$

The expectation matrix is regular with dominant eigenvalue $\lambda = 3/2$ and the p.g.f.'s satisfy the hypothesis of the theorem. Since $\lambda > 1$, there exists $q_1, q_2 \in [0, 1)$ such that $f_1(q_1, q_2) = q_1$ and $f_2(q_1, q_2) = q_2$. The fixed point is $q_1 = \sqrt{2} - 1 \approx 0.4142$ and $q_2 = \sqrt{2} - 1 \approx 0.4142$. If $X(0) = (r_1, r_2)$, then $\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0}\} \approx (0.4142)^{r_1 + r_2}$. ■

The identity for the conditional expectation (4.10) can be extended to multitype branching processes (see, e.g., Karlin and Taylor, 1975 or Harris, 1963). The conditional expectation satisfies

$$E(X(n+1)|X(n)) = MX(n); \tag{4.11}$$

that is, the expectation of $X(n+1)$ given the value of $X(n)$ is the expectation matrix times $X(n)$. In general,

$$E(X(n+r)|X(n)) = M^r X(n).$$

It is important to note that these identities apply only to the conditional expectation (see Exercises 8, 9, and 10).

Example 4.8 Consider the multitype branching process in Example 4.7. If $X(0) = (1,1)^T$, then

$$E(X(1)|X(0)) = \begin{pmatrix} 3/4 & 1/2 \\ 3/4 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5/4 \\ 7/4 \end{pmatrix}$$

and

$$E(X(2)|X(0)) = \begin{pmatrix} 3/4 & 1/2 \\ 3/4 & 1 \end{pmatrix} \begin{pmatrix} 5/4 \\ 7/4 \end{pmatrix} = \begin{pmatrix} 29/16 \\ 43/16 \end{pmatrix}. \quad \blacksquare$$

4.7 An Example: Age-Structured Model

Suppose there are k age classes, $i = 1, 2 \ldots, k$. The first age class, type 1, represents newborns. An individual of age i gives birth to individuals of type 1, then survives, with a given probability, to the next age class becoming an individual of type $i+1$. Age class k is the oldest age class, and individuals in this class do not survive past age k. Assume an individual of type i at time n either survives to become a type $i+1$ individual at time $n+1$ with probability $p_{i+1,i} > 0$ or dies with probability $1 - p_{i+1,i}$, $i = 1, 2, \ldots, k-1$. Probability $p_{k+1,k} = 0$ because age k is the oldest age class. In addition, a type i individual gives birth to r individuals of type 1 at time $n+1$ with probability $b_{i,r}$. The offspring distribution, $\{b_{i,r}\}_{r=0}^{\infty}$, for an individual of time i satisfies

$$b_{i,r} \geq 0, \quad \text{and} \quad \sum_{r=0}^{\infty} b_{i,r} = 1, \quad i = 1, 2, \ldots, k.$$

The mean of this distribution is denoted by

$$b_i = \sum_{r=1}^{\infty} r b_{i,r}.$$

The expectation matrix M can be computed from these probability distributions:

$$M = \begin{pmatrix} b_1 & b_2 & \cdots & b_{k-1} & b_k \\ p_{21} & 0 & \cdots & 0 & 0 \\ 0 & p_{32} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{k,k-1} & 0 \end{pmatrix}. \tag{4.12}$$

The form of matrix M is known as a *Leslie matrix* or a *projection matrix* (Caswell, 2001; Cushing, 1998; Leslie, 1945). In the deterministic Leslie matrix model, the value of each of the age classes at time $n+1$, $X(n+1)$, is found after multiplication by M; that is, $X(n+1) = MX(n)$. In particular, the first age group $x_1(n+1)$ consists of offspring from all of the other age groups; that is,

$$x_1(n+1) = b_1x_1(n) + b_2x_2(n) + \cdots + b_kx_k(n) = \sum_{i=1}^{k} b_ix_i(n).$$

The $i+1$st age group, $i = 1, 2, \ldots, k-1$, $x_{i+1}(n+1)$, consists of individuals from age group i who survived and became age $i+1$:

$$x_{i+1}(n+1) = p_{i+1,i}x_i(n).$$

It is interesting to note that, in the stochastic model, the conditional expectation satisfies a similar identity, equation (4.11).

The expectation matrix M can be determined directly from the p.g.f.'s (see Exercise 11). The p.g.f's f_i, $i = 1, 2, \ldots, k$ satisfy

$$f_i(t_1, t_2, \ldots, t_k) = [p_{i+1,i}t_{i+1} + (1 - p_{i+1,i})] \sum_{r=0}^{\infty} b_{i,r}t_1^r, \quad i = 1, \ldots, k. \tag{4.13}$$

Note that $f_i(1, 1, \ldots, 1) = 1$.

It will be assumed that the expectation matrix M is regular and that the p.g.f.'s are nonlinear. Then Theorem 4.4 can be applied. These assumptions are reasonable for many age-structured models. For example, for matrices M satisfying $p_{i+1,i} > 0$ for $i = 1, \ldots, k-1$ and $b_k > 0$, it can be shown that M is regular if and only if the greatest common divisor of the set of indices i, where $b_i > 0$, is 1, g.c.d.$\{i | b_i > 0\} = 1$ (Sykes, 1969).

Example 4.9 Suppose

$$M_1 = \begin{pmatrix} b_1 & 0 & 0 & b_4 \\ p_{21} & 0 & 0 & 0 \\ 0 & p_{32} & 0 & 0 \\ 0 & 0 & p_{43} & 0 \end{pmatrix} \quad \text{and} \quad M_2 = \begin{pmatrix} 0 & b_2 & 0 & b_4 \\ p_{21} & 0 & 0 & 0 \\ 0 & p_{32} & 0 & 0 \\ 0 & 0 & p_{43} & 0 \end{pmatrix},$$

where the elements $p_{i+1,i} > 0$ and $b_i > 0$. Applying the criteria of Sykes (1969), it can be seen that M_1 is regular (g.c.d.$\{1,4\} = 1$) but that M_2 is not regular (g.c.d.$\{2,4\} = 2$). ■

When Theorem 4.4 applies, the extinction behavior in a multitype branching process is determined from the dominant eigenvalue λ of M and from the p.g.f.'s. If the dominant eigenvalue satisfies $\lambda > 1$, then the limit, $\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0}\}$, depends on the initial distribution and the fixed point of the p.g.f.'s. If $X(0) = (r_1, r_2, \ldots, r_k)^T$, and the fixed point of F is $(q_1, q_2, \ldots, q_k)$, then

$$\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0} | X(0) = (r_1, r_2, \ldots, r_k)^T\} = q_1^{r_1} q_2^{r_2} \cdots q_k^{r_k}.$$

In the particular case $b_{i,0} = 0$, $i = 1, 2, \ldots, r$, every individual has at least one offspring and the fixed point of F is at the origin, $q_i = 0$ for all i (see Exercise 12). In this case the probability of ultimate extinction is zero.

The multitype branching process simplifies to a single branching process when the number of age classes is reduced to one, $k = 1$. In this case, there is one p.g.f. given by

$$f_1(t) = \sum_{r=0}^{\infty} b_{1,r} t^r,$$

where $b_{1,r} = p_r$ is the probability of r births. If the mean number of births $m = f_1'(1) > 1$, then there exists a fixed point $q \in [0, 1)$ of f_1 such that $\lim_{n\to\infty} \text{Prob}\{X_n = 0 | X_0 = 1\} = q$.

Example 4.10 Suppose there are two age classes and the expectation matrix is

$$M = \begin{pmatrix} b_1 & b_2 \\ p_{21} & 0 \end{pmatrix} = \begin{pmatrix} 3/4 & 1 \\ 1/2 & 0 \end{pmatrix}.$$

The characteristic equation of M is

$$\lambda^2 - (3/4)\lambda - 1/2 = 0$$

so that the dominant eigenvalue is $\lambda = (3 + \sqrt{41})/8 \approx \mathbf{1.17539 > 1}$. Suppose the birth probabilities are

$$b_{1,r} = \begin{cases} 1/2, & r = 0 \\ 1/4, & r = 1, 2 \\ 0, & r \neq 0, 1, 2 \end{cases}, \qquad b_{2,r} = \begin{cases} 1/4, & r = 0, 2 \\ 1/2, & r = 1 \\ 0, & r \neq 0, 1, 2 \end{cases}.$$

The mean number of births for each age class is

$$b_1 = 3/4 = \sum_{r=1}^{\infty} r b_{1,r} \quad \text{and} \quad b_2 = 1 = \sum_{r=1}^{\infty} r b_{2,r}$$

$X(0)^T$	Proportion of Extinct Sample Paths Out of 10,000	$\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0} \mid X(0)\}$
(1,0)	0.6318	0.6285
(0,1)	0.6665	0.6631
(1,1)	0.4197	0.4168
(0,2)	0.3892	0.3950
(2,0)	0.4391	0.4396

Table 4.2. Probability of population extinction for Example 4.10 (**Block** and Allen, 2000)

(the values in the first row of M). In addition, the p.g.f.'s for the two age classes satisfy

$$\begin{aligned} f_1(t_1,t_2) &= [(1/2)t_2 + 1/2][1/2 + (1/4)t_1 + (1/4)t_1^2] \\ f_2(t_1,t_2) &= 1/4 + (1/2)t_1 + (1/4)t_1^2. \end{aligned}$$

Since $\lambda > 1$, the preceding system of p.g.f.'s has a unique fixed point on $[0,1)\times[0,1)$. The fixed point (q_1,q_2) satisfies $f_1(q_1,q_2) = q_1$ and $f_2(q_1,q_2) = q_2$ and satisfies

$$(q_1,q_2) \approx (0.6285, 0.6631).$$

For example, if there are initially five individuals of age 1 and three individuals of age 2, then the probability of ultimate extinction of the total population is approximately

$$(0.6285)^5(0.6631)^3 \approx 0.0286. \quad \blacksquare$$

In Block and Allen (2000), sample paths were numerically simulated for the branching process in Example 4.10. Ten thousand sample paths were generated for each set of parameter values. From these 10,000 sample paths, the proportion of paths for which the population size had reached zero by time 39 was calculated and compared with the preceding estimate $(0.6285)^{r_1}(0.6631)^{r_2}$. See Table 4.2.

For additional examples on branching processes in the context of Leslie age-structured or general structured models and models with size-dependent birth or transition probabilities, see Pollard (1966, 1973) and Block and Allen (2000). In addition, please consult Tuljapurkar (1990) for a discussion of stochastic matrix models with applications to structured populations. Tuljapurkar (1990) studies models of the form

$$X(n+1) = M(n+1)X(n),$$

where matrix M is a random matrix depending on time $n+1$. In these models, the population structure depends on environmental variation. Birth and

survival rates depend on the state of the environment (see Exercise 16). We end this chapter with an example of a size-dependent branching process.

The single and multitype branching processes discussed thus far exhibit either exponential growth or decline. Eventually, either the total population size approaches zero or infinity. This is due to the fact that the offspring distribution is constant over time. If the offspring distribution depends on the population size, then the branching process is size dependent. Size-dependent branching processes based on simple discrete time population models have been formulated and analyzed by Högnäs (1997, 2000). The next example illustrates a size-dependent branching process that is based on a population growth model known as the *Ricker model*:

$$x_{n+1} = x_n \exp(r - \gamma x_n), \quad 0 < r, \ 0 < \gamma$$

(May, 1976; Ricker, 1954). The Ricker model has been used frequently in biological applications (see Caswell, 2001). It has very interesting behavior for various values of the parameter r. For example, if $0 < r < 2$, then x_n converges to a stable fixed point r/γ:

$$\lim_{n\to\infty} x_n = \frac{r}{\gamma}.$$

But if $2 < r < 2.526$, then solutions converge to a stable two-cycle; solutions x_n oscillate between two values. For increasing values of r, solutions exhibit what is known as period-doubling behavior (see May, 1976; Elaydi, 2000).

Example 4.11 We describe the size-dependent Ricker branching process formulated by Högnäs (1997). Let X_n be the random variable for the total population size. Let $\{p_k\}_{k=1}^{\infty}$ denote the size-independent offspring distribution, where the probability an individual produces k offspring is p_k, $\sum_{k=1}^{\infty} p_k = 1$. In addition, assume the mean satisfies

$$m = \sum_{k=1}^{\infty} k p_k = e^r > 1.$$

Each individual produces offspring independently of any other individual. For a population of size $x \in \{0, 1, 2, \ldots\}$, let the size-dependent offspring distribution be defined as follows: The probability that an individual produces k offspring is $p_k \exp(-\gamma x)$ for $k = 1, 2, \ldots$ and the probability that an individual produces no offspring is $1 - \exp(-\gamma x)$. Let $\{Y_j\}$ be a set of iid random variables having the aforementioned size-dependent offspring distribution; that is, $\text{Prob}\{Y_j = k\} = p_k \exp(-\gamma x)$ and $\text{Prob}\{Y_j = 0\} = 1 - \exp(-\gamma x)$. Then if $X_n = x$,

$$X_{n+1} = \begin{cases} \sum_{j=1}^{x} Y_j, & x = 1, 2, \ldots, \\ 0, & x = 0. \end{cases}$$

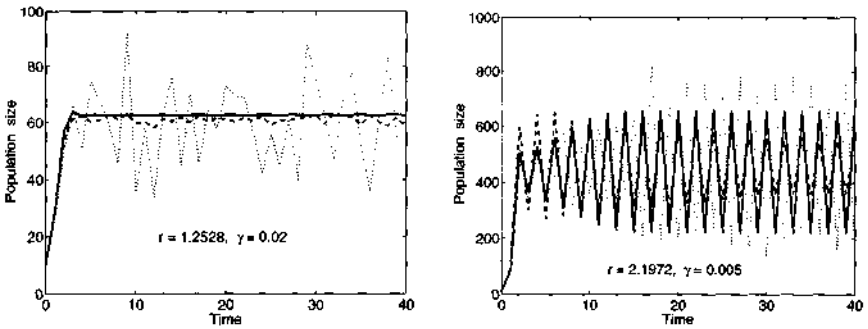

Figure 4.4. Solutions to the deterministic and stochastic Ricker models. The deterministic solution is the solid curve. One sample path for the size-dependent branching Ricker process is the dotted curve and the mean of 100 sample paths is the dashed curve. In the first case, the size-independent offspring distribution satisfies $p_k = 1/6$ for $k = 1, 2, \ldots, 6$, $r = 1.2528 < 2$, and $\gamma = 0.02$. Then the deterministic solution and the stochastic mean converge to $r/\gamma = 62.64$. In the second case, the size-independent offspring distribution satisfies $p_8 = 0.2$, $p_9 = 0.6$, and $p_{10} = 0.2$, $r = 2.1972 > 2$, and $\gamma = 0.005$. In this case, the deterministic solution and the stochastic mean oscillate between two values.

It can be shown that the following conditional expectation of the branching process has the same behavior as the deterministic model:

$$E(X_{n+1}|X_n = x) = x \exp(r - \gamma x).$$

Numerical simulations comparing the deterministic and stochastic Ricker models are graphed in Figure 4.4 for the two cases: $r = 1.2528$ and $r = 2.1972$. ■

4.8 Exercises for Chapter 4

1. Consider Galton's problem in the context of Example 4.1, where $X_0 = N$ adult males each have different surnames. In each generation, a proportion p_0 of the adult males have no male children who reach adult life and p_1 have one such child, $p_0 + p_1 = 1$. Show that the expected proportion of surnames that has gone extinct in generation r is $p_0(r)$. Notice that the proportion that has gone extinct is either $1, (N-1)/N, (N-2)/N, \ldots, 1/N$, or 0. The probabilities for each of these proportions are given by the coefficients in the expansion of $[f^r(t)]^N$, equation (4.4). (*Hint*: Refer to Example 4.1.)

2. Suppose a branching process with $X_0 = 1$ has an offspring distribution satisfying

$$p_k = ab^{k-1}, \ k = 1, 2, \ldots$$

and

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

where $0 < b < a + b < 1$ (Karlin and Taylor, 1975).

(a) Show that the p.g.f satisfies

$$f(t) = \frac{1-(a+b)}{1-b} + \frac{at}{1-bt}.$$

(b) Assume $a > (b-1)^2$. Then find $\lim_{n\to\infty} \text{Prob}\{X_n = 0\}$.

3. The p.g.f. of a branching process with $X_0 = 1$ satisfies $f(t) = at^2 + bt + c$, where a, b, and c are positive and $f(1) = 1$. Assume $f'(1) > 1$. Then show that

$$\lim_{n\to\infty} \text{Prob}\{X_n = 0\} = \frac{c}{a}$$

(Karlin and Taylor, 1975).

4. Suppose the p.g.f. of a branching process satisfies $f(t) = p_0 + p_1 t$, $p_0 > 0$, $p_1 > 0$, and $p_0 + p_1 = 1$.

(a) Show that $f^n(t) = 1 - p_1^n + p_1^n t$. (*Hint*: Refer to Example 4.1.)

(b) Suppose $X_0 = N$ so that the p.g.f. of X_n is $[f^n(t)]^N$. Let T be the random variable for the first time to extinction; that is, the smallest n such that $X_n = 0$ (i.e., the *first* passage time into the state 0). Then $\text{Prob}\{T = 1\} = (1-p_1)^N$. Use the fact that $\text{Prob}\{T \le n\} = [f^n(0)]^N$ to show that $\text{Prob}\{T = n\} = (1-p_1^n)^N - (1-p_1^{n-1})^N$.

5. Suppose a branching process with $X_0 = 1$ has an offspring distribution satisfying

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

(a) Find the p.g.f. $f(t)$; then find the mean and variance of X_1, $m = E(X_1)$, and $\sigma^2 = \text{Var}(X_1)$.

(b) Find the mean and variance of X_n, $m_n = E(X_n)$, and $\sigma_n^2 = \text{Var}(X_n)$.

(c) For $\lambda = 1.5$ and $\lambda = 2$, find $\lim_{n\to\infty} \text{Prob}\{X_n = 0\}$.

6. Suppose $Z_k = \sum_{n=0}^{k} X_n$ for $k = 0, 1, 2, \ldots$ and $Z = \sum_{n=0}^{\infty} X_n$, where $\{X_n\}$ is a branching process with $X_0 = 1$. Suppose the mean m of the offspring distribution satisfies $0 < m < 1$. Show that $E(Z_k) = \sum_{n=0}^{k} m^n$ and $E(Z) = (1-m)^{-1}$ (Taylor and Karlin, 1998).

7. Suppose a branching process with $X_0 = 1$ has an offspring distribution with mean $m > 0$. Let $Z_n = X_n/m^n$. Show that $E(Z_{n+1}|Z_n = k) = k$ (Taylor and Karlin, 1998).

8. Suppose a multitype branching process with two types $X(n) = (X_1(n), X_2(n))^T$ has p.g.f.'s

$$f_1(t_1, t_2) = (1/4)(1 + t_2 + 2t_1^2) \text{ and } f_2(t_1, t_2) = (2/3)(1/2 + t_1^2).$$

   (a) Find the expectation matrix M and show that M is a regular matrix.

   (b) Find the expectation $E\left(X(n)|X(0) = (3,4)^T\right)$ for $n = 1, 2, 3$.

   (c) Find $\lim_{n\to\infty} \text{Prob}\{X(n) = (0,0)^T | X(0) = (3,4)^T\}$.

9. Suppose a multitype branching process with three types $X(n) = (X_1(n), X_2(n), X_3(n))^T$ has p.g.f.'s

$$f_1(t_1, t_2, t_3) = \frac{1}{3} + \frac{1}{3}t_2(t_1 + t_3), \quad f_2(t_1, t_2, t_3) = \frac{1}{2} + \frac{1}{2}t_1 t_2 t_3,$$

   and

$$f_3(t_1, t_2, t_3) = \frac{1}{4}(1 + t_1^2 + t_2^2 + t_3^2).$$

   (a) Find the expectation matrix M and show that M is a regular matrix.

   (b) Find $\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0}|X(0) = (r_1, r_2, r_3)^T\}$.

   (c) If $X(0) = (1,1,1)^T$, find the expectation $E\left(X(n)|X(0)\right)$ for $n = 1, 2, 3$.

10. Suppose a multitype branching process with three types $X(n) = (X_1(n), X_2(n), X_3(n))^T$ has p.g.f.'s

$$f_1(t_1, t_2, t_3) = \frac{1}{2}t_1 t_2 t_3 + \frac{1}{2}t_2^2, \quad f_2(t_1, t_2, t_3) = 1, \text{ and } f_3(t_1, t_2, t_3) = t_3.$$

   (a) Find the expectation matrix M. Is M regular?

   (b) If $X(0) = \delta_1 = (1,0,0)^T$, find the expectation $E(X(n)|X(0))$ for $n = 1, 2, 3$.

   (c) If $X(0) = \delta_2 = (0,1,0)^T$, find the probability distribution for $X(n)$, $n \geq 1$.

   (d) If $X(0) = \delta_3 = (0,0,1)^T$, find the probability distribution for $X(n)$, $n \geq 1$.

11. Use the p.g.f.'s given by (4.13) to verify that the expectation matrix M of the age-structured example satisfies (4.12).

12. Suppose a multitype branching process for an age-structured population model satisfies $b_{i,0} = 0$ and $b_{i,r} = 0$ for $r > 5$ and $i = 1, 2, \ldots, k$. Show that a fixed point of the generating function F is the origin [i.e., $f_i(0, 0, \ldots, 0) = 0$–the probability of population extinction is zero].

13. Suppose the offspring distribution for an age-structured branching process satisfies

$$b_{1,r} = \begin{cases} 1/2, & r = 0 \\ 1/2, & r = 2 \\ 0, & r \neq 0, 2, \end{cases} \qquad b_{2,r} = \begin{cases} 1/6, & r = 0, 2 \\ 2/3, & r = 1 \\ 0, & r \neq 0, 1, 2 \end{cases}.$$

In addition, suppose the probability of surviving from the first to the second age class is $p_{21} = 3/4$.

(a) Find the mean birth rates, b_1 and b_2.

(b) Find the expectation matrix M. Show that M is regular; then find the dominant eigenvalue of M.

(c) Find the p.g.f.'s for the two age classes, $f_1(t_1, t_2)$ and $f_2(t_1, t_2)$. Then find the probability of population extinction given $X(0) = (1, 2)^T$; that is, $\lim_{n\to\infty} \text{Prob}\{X(n) = \mathbf{0} | X(0) = (1, 2)^T\}$.

14. A simple example of a multitype branching process related to cellular dynamics is discussed by Jagers (1975). Cell division results in two identical daughter cells containing the same number of chromosomes as the original cell ($2n$ for a diploid cell). Sometimes a mistake occurs and only one cell is produced having twice the number of chromosomes ($4n$ chromosomes), referred to as *endomitosis*. When this abnormal cell divides again, it will produce two daughter cells with twice the number of chromosomes ($4n$ chromosomes). Endomitosis can occur again for a cell having $4n$ chromosomes to produce a cell with $8n$ chromosomes and, in general, endomitosis occurring in a cell with $2^i n$ chromosomes produces a cell with $2^{i+1} n$ chromosomes. Cells with more than two copies of the genes are known as *polyploid cells*. The incidence of higher ploidies than four is small. Therefore, it is reasonable to consider a cellular model with only two types: diploid cells and polyploid cells (Jagers, 1975). Let p be the probability of endomitosis, $0 < p < 1/2$. The p.g.f.'s for the two types satisfy

$$f_1(t_1, t_2) = (1 - p)t_1^2 + pt_2 \text{ and } f_2(t_1, t_2) = pt_2 + (1 - p)t_2^2.$$

(a) Find the expectation matrix M and the dominant eigenvalue of M. Is M regular?

(b) Find all of the fixed points of (f_1, f_2) on the interval $[0, 1] \times [0, 1]$.

15. (a) Consider the size-dependent Ricker branching process $\{X_n\}_{n=0}^{\infty}$ discussed in Example 4.11. Show that the conditional expectation $E(X_{n+1}|X_n = x)$ satisfies

$$E(X_{n+1}|X_n = x) = x\exp(r - \gamma x).$$

(b) Formulate a size-dependent branching process similar to the one described in Example 4.11 but one based on the following discrete time population model:

$$x_{n+1} = \frac{rx_n}{1 + \gamma x_n}, \quad 0 < r, \ 0 < \gamma.$$

This population model is known as the Beverton-Holt model (Caswell, 2001). Assume the size-independent offspring distribution $\{p_k\}$ satisfies $\sum_{k=1}^{\infty} kp_k = r > 1$. Then show that the model satisfies the following conditional expectation:

$$E(X_{n+1}|X_n = x) = \frac{rx}{1 + \gamma x}.$$

16. Suppose a population has two stages and the birth rates for each stage are constant, $b_1 > 0$ and $b_2 > 0$, but the survival probability from stage 1 to 2 is an environmentally determined, time-dependent random variable p_n, where $0 < \underline{p} \leq p_n \leq \overline{p} \leq 1$, for $n = 1, 2, \ldots$. The stochastic model satisfies

$$X(n+1) = \begin{pmatrix} b_1 & b_2 \\ p_{n+1} & 0 \end{pmatrix} X(n), \quad b_i > 0, \quad i = 1, 2$$

(see Tuljapurkar, 1990). The ratio of stage 2 to stage 1 over time is denoted as $U_n = X_2(n)/X_1(n)$, where $X(n) = (X_1(n), X_2(n))^T$.

(a) Show that

$$U_{n+1} = \frac{p_{n+1}}{b_1 + b_2 U_n}.$$

(b) Use the fact that $\underline{p}/(b_1 + b_2 U_n) \leq U_{n+1} \leq \overline{p}/(b_1 + b_2 U_n)$ to show that there exist constants $0 < c_1 < c_2$ such that $c_1 \leq U_n \leq c_2$ for all time n (i.e., it is possible to obtain bounds on the ratio of the age structure).

4.9 References for Chapter 4

Bailey, N. T. J. 1990. *The Elements of Stochastic Processes with Applications to the Natural Sciences.* John Wiley & Sons, New York.

Block, G. L. and L. J. S. Allen. 2000. Population extinction and quasi-stationary behavior in stochastic density-dependent structured models. *Bull. Math. Biol.* 62: 199–228.

Caswell, H. 2001. *Matrix Population Models: Construction, Analysis and Interpretation.* 2nd ed. Sinauer Assoc. Inc., Sunderland, Mass.

Cushing, J. M. 1998. *An Introduction to Structured Population Dynamics.* CBMS-NSF Regional Conf. Series in Applied Mathematics # 71. SIAM: Philadelphia.

Elaydi, S. N. 2000. *Discrete Chaos.* Chapman & Hall/CRC, Boca Raton.

Gantmacher, F. R. 1964. *The Theory of Matrices.* Vol. II, Chelsea Pub. Co., New York.

Harris, T. E. 1963. *The Theory of Branching Processes.* Prentice Hall, Inc., Englewood Cliffs, N. J.

Högnäs, G. 1997. On the quasi-stationary distribution of a stochastic Ricker model. *Stoch. Proc. Appl.* 70: 243–263.

Högnäs, G. 2000. On some one-dimensional stochastic population models. *Contemp. Math.* 261: 209–220.

Jagers, P. 1975. *Branching Processes with Biological Applications.* Wiley, Chichester.

Karlin, S. and H. Taylor. 1975. *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York.

Kimmel, M. and D. Axelrod. 2002. *Branching Processes in Biology.* Springer-Verlag, New York.

Leslie, P. H. 1945. On the use of matrices in certain population mathematics. *Biometrics* 21: 1–18.

May, R. M. 1976. Simple mathematical models with very complicated dynamics. *Nature.* 261: 459–467.

Mode, C. J. 1971. *Multitype Branching Processes Theory and Applications.* Elsevier, New York.

Pollard, J. H. 1966. On the use of the direct matrix product in analyzing certain stochastic population models. *Biometrika* 53: 397-415.

Pollard, J. H. 1973. *Mathematical Models for the Growth of Human Populations.* Cambridge University Press, Cambridge, U. K.

Ricker, W. E. 1954. Stock and recruitment. *J. Fish. Res. Bd. Can.* 11: 559–623.

Schinazi, R. B. 1999. *Classical and Spatial Stochastic Processes.* Birkhäuser, Boston.

Sykes, Z. M. 1969. On discrete stable population theory. *Biometrics* 25: 285–293.

Taylor, H. M. and S. Karlin. 1998. *An Introduction to Stochastic Modeling,* 3rd ed. Academic Press, New York.

Tuljapurkar, S. 1990. *Population Dynamics in Variable Environments.* Springer-Verlag, Berlin and New York.

In matrix form,

$$P(s)P(t) = P(s+t)$$

for all $s, t \in [0, \infty)$. Notice that these definitions are the continuous analogues of the definitions given for discrete time Markov chains. There are some special cases known as explosive processes when condition (5.1) may not hold for all times, but these special cases can occur only when the state space is infinite. An explosive process is defined, but first the concept of a jump time is defined.

The distinction between discrete and continuous time Markov chains is that in discrete time chains, there is a "jump" to a new state at times $1, 2, \ldots$, but in continuous time chains the "jump" to a new state may occur at any time $t \geq 0$. In a continuous time chain, with beginning state $X(0)$, the process stays in state $X(0)$ for a random amount of time W_1 until it jumps to a new state, $X(W_1)$. Then it stays in state $X(W_1)$ for a random amount of time until it jumps to a new state at time W_2, $X(W_2)$, and so on. In general, W_i is the random variable for the time of the ith jump. Define $W_0 = 0$. The collection of random variables $\{W_i\}$ is referred to as the *jump times* or *waiting times* of the process (Norris, 1999; Taylor and Karlin, 1998). In addition, the random variables $T_i = W_{i+1} - W_i$ are referred to as the *interevent times* or *holding times* or *sojourn times* (Norris, 1999; Taylor and Karlin, 1998). The waiting times W_i and interevent times T_i are illustrated in Figure 5.1.

In an explosive process, the value of the state approaches infinity at a finite time, $\lim_{t \to T^-} X(t) = \infty$ for $T < \infty$. Then $p_{ji}(T) = 0$ for all $i, j = 0, 1, 2, \ldots$, which means $\sum_{j=0}^{\infty} p_{ji}(T) = 0$. Hence, condition (5.1) does not hold. See Norris (1999) or Karlin and Taylor (1975) for further discussion on explosive processes. Most of the well-known birth and death processes are nonexplosive. In particular, all finite, continuous time Markov chains are nonexplosive. Explosive birth processes will be discussed in Chapter 6.

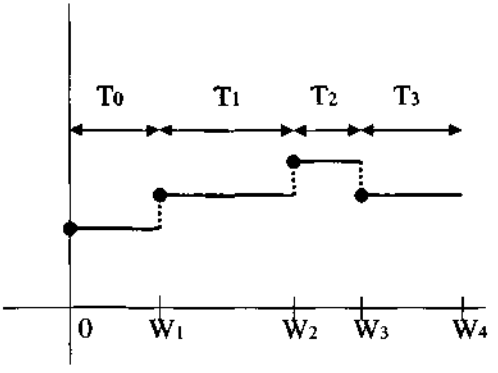

Figure 5.1. One sample path of a continuous time Markov chain, illustrating waiting times and interevent times.

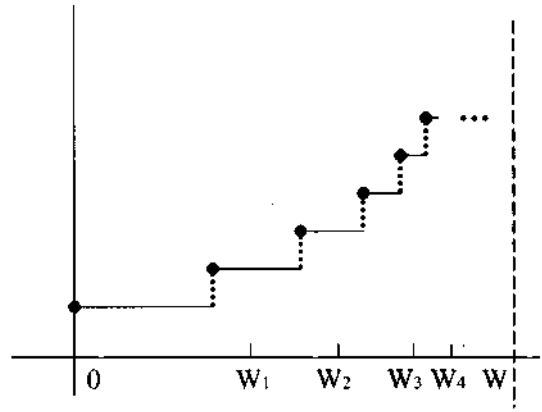

Figure 5.2. One sample path of a continuous time Markov chain that is explosive.

One sample path of an explosive process is graphed in Figure 5.2. The values of the waiting times are approaching a positive constant, $W = \sup\{W_i\}$, and the values of the states are approaching infinity, $\lim_{i\to\infty} X(W_i) = \infty$; the process is explosive. Notice that the sample path in Figure 5.2 is a piecewise constant function that is continuous from the right. However, for ease in sketching sample paths, they will be drawn as connected rectilinear curves (as in Figure 5.3).

As before, our notation differs from the standard notation in that we have defined the transition probability from state $i \to j$ as p_{ji}, rather than the more commonly used notation p_{ij} (e.g., Bailey, 1990; Karlin and Taylor, 1975, 1981; Norris, 1999; Schinazi, 1999; Stewart, 1994; Taylor and Karlin, 1998). This notation is consistent with our notation for discrete time Markov chains. In our notation, the element in the ith row and jth column of $P(t)$ is p_{ij}, which represents the transition $j \to i$.

5.3 The Poisson Process

The *Poisson process* is a continuous time Markov chain $\{X(t)\}$ **defined on** $\{0, 1, 2, \ldots\}$ with the following properties:

(1) For $t = 0$, $X(0) = 0$.

(2) For Δt sufficiently small, the transition probabilities satisfy

$$\begin{aligned} p_{i+1,i}(\Delta t) &= \text{Prob}\{X(t+\Delta t) = i+1 | X(t) = i\} = \lambda \Delta t + o(\Delta t) \\ p_{ii}(\Delta t) &= \text{Prob}\{X(t+\Delta t) = i | X(t) = i\} = 1 - \lambda \Delta t + o(\Delta t) \\ p_{ji}(\Delta t) &= \text{Prob}\{X(t+\Delta t) = j | X(t) = i\} = o(\Delta t), \quad j \geq i+2 \\ p_{ji}(\Delta t) &= 0, \quad j < i, \end{aligned}$$

where the notation $o(\Delta t)$ ("little oh Δt") is the Landau order symbol.

In general, we say that the function $f(\Delta t) = o(\Delta t)$ or $f(\Delta t)$ is $o(\Delta t)$ as $\Delta t \to 0$ if f has the following property:

$$\lim_{\Delta t \to 0} \frac{f(\Delta t)}{\Delta t} = 0.$$

Therefore, in part (2) of the definition of the Poisson process, the functions $p_{i+1,i}(\Delta t) - \lambda \Delta t$, $p_{ii}(\Delta t) - 1 + \lambda\, \Delta t$, and $p_{ji}(\Delta t)$ are $o(\Delta t)$ as $\Delta t \to 0$. In particular,

$$\lim_{\Delta t \to \infty} \frac{p_{i+1,i}(\Delta t) - \lambda \Delta t}{\Delta t} = 0 = \lim_{\Delta t \to \infty} \frac{p_{ii}(\Delta t) - 1 + \lambda \Delta t}{\Delta t}$$

and

$$\lim_{\Delta t \to \infty} \frac{p_{ji}(\Delta t)}{\Delta t} = 0, \quad j \geq i + 2.$$

Frequently, continuous time Markov processes are defined by conditions such as those given in (2). From these "infinitesimal" transition probabilities, other properties of the process can be derived.

In a small time interval Δt, the Poisson process can either stay in the same state or move to the next larger state, $i \to i + 1$; it cannot move to a smaller state. The probability that the Poisson process moves up two or more states is a very small probability and approaches zero when $\Delta t \to 0$. Note that the transition probabilities $p_{ji}(\Delta t)$ are independent of i and j and only depend on the length of the interval Δt. If the intervals $[s, s + \Delta t]$ and $[t, t + \Delta t]$ are nonoverlapping, $s + \Delta t \leq t$, then property (2) and the Markov property imply that the following random variables from the Poisson process, $X(t + \Delta t) - X(t)$ and $X(s + \Delta t) - X(s)$, are independent and have the same probability distributions (i.e., the Poisson process has stationary and independent increments).

The assumptions (1) and (2) are used to derive a system of differential equations satisfied by $p_i(t)$ for $i = 0, 1, 2, \ldots$. The solution to $p_i(t)$ is then shown to be a Poisson probability distribution. Because in the Poisson process $X(0) = 0$, it follows that $p_{i0}(t) = p_i(t)$. Let $p_{00}(t+\Delta t) = p_0(t+\Delta t)$. Then

$$\begin{aligned} p_0(t + \Delta t) &= \text{Prob}\{X(t + \Delta t) = 0\} \\ &= \text{Prob}\{X(t) = 0, X(t + \Delta t) - X(t) = 0\} \\ &= \text{Prob}\{X(t) = 0\}\text{Prob}\{X(t + \Delta t) - X(t) = 0\} \\ &= \text{Prob}\{X(t) = 0\}\text{Prob}\{X(\Delta t) = 0\}, \end{aligned}$$

where we have used the fact that $X(t) - X(0) = X(t)$ and $X(t+\Delta t) - X(t)$ are independent. Therefore,

$$p_0(t + \Delta t) = p_0(t)\left[1 - \lambda \Delta t + o(\Delta t)\right].$$

Subtracting $p_0(t)$ from both sides of the last equation,

$$p_0(t+\Delta t) = p_0(t)\,[1 - \lambda\Delta t + o(\Delta t)]$$

and dividing by Δt,

$$\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + p_0(t)\frac{o(\Delta t)}{\Delta t}.$$

Then taking the limit as $\Delta t \to 0$,

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t). \tag{5.2}$$

Given that $p_0(0) = 1 = \text{Prob}\{X(0) = 0\}$, the solution to this linear first-order differential equation is

$$p_0(t) = e^{-\lambda t}.$$

The differential equations for $i \geq 1$ are derived in a similar manner. Let $p_{i0}(t+\Delta t) = p_i(t+\Delta t)$. Then

$$\begin{aligned} p_i(t+\Delta t) &= \text{Prob}\{X(t+\Delta t) = i\} \\ &= \text{Prob}\{X(t) = i, \Delta X(t) = 0\} \\ &\quad + \text{Prob}\{X(t) = i-1, \Delta X(t) = 1\} \\ &\quad + \sum_{k=2}^{k\leq i} \text{Prob}\{X(t) = i-k, \Delta X(t) = k\}, \end{aligned}$$

where $\Delta X(t) = X(t+\Delta t) - X(t)$. The latter summation is $o(\Delta t)$ since

$$\sum_{k=2}^{k\leq i} \text{Prob}\{X(t) = i-k, \Delta X(t) = k\} = \sum_{k=2}^{k\leq i} p_{i-k}(t)o(\Delta t) = o(\Delta t).$$

Applying the definition of the transition probabilities and the independence of the increments,

$$p_i(t+\Delta t) = p_i(t)[1 - \lambda\Delta t + o(\Delta t)] + p_{i-1}(t)[\lambda\Delta t + o(\Delta t)] + o(\Delta t). \tag{5.3}$$

Note that the equations given in (5.3) can be derived directly from the infinitesimal transition probabilities. If the process is in state i at time $t+\Delta t$, then at the previous time t it was either in state i or $i-1$ [the probability it was in some other state is $o(\Delta t)$]. If the process is in state i at time t, the process stays in state i with probability $1 - \lambda\Delta t + o(\Delta t)$, and if the process is in state $i-1$ at time t, it moves to state i with probability $\lambda\Delta t + o(\Delta t)$.

Subtract $p_i(t)$ from both sides (5.3) and divide by Δt. Then

$$\frac{p_i(t+\Delta t) - p_i(t)}{\Delta t} = -\lambda p_i(t) + \lambda p_{i-1}(t) + \frac{o(\Delta t)}{\Delta t},$$

where all terms with a factor of $o(\Delta t)$ are bounded and can be absorbed in $o(\Delta t)$. Taking the limit as $\Delta t \to 0$, then

$$\frac{dp_i(t)}{dt} = -\lambda p_i(t) + \lambda p_{i-1}(t), \quad i \geq 1. \tag{5.4}$$

The preceding equations represent a system of differential-difference equations, difference equations in the variable i and differential equations in the variable t.

The system of differential-difference equations (5.4) can be solved sequentially. Replace $p_0(t)$ by $e^{-\lambda t}$ and apply the initial conditions $p_i(0) = 0$, $i \geq 1$. The differential equation for $p_1(t)$ is a linear first-order differential equation,

$$\frac{dp_1(t)}{dt} + \lambda p_1(t) = \lambda e^{-\lambda t}, \quad p_1(0) = 0.$$

Multiplying by the factor $e^{\lambda t}$ (known as an *integrating factor*) yields

$$\frac{d\left[e^{\lambda t} p_1(t)\right]}{dt} = \lambda. \tag{5.5}$$

In general, an integrating factor for a linear differential equation of the form $dx/dt + a(t)x = b(t)$ is $e^{\int a(t)dt}$. Integrating both sides of (5.5) from 0 to t and applying the initial condition yields the solution

$$p_1(t) = \lambda t e^{-\lambda t}.$$

Next, the differential equation for $p_2(t)$ is

$$\frac{dp_2(t)}{dt} + \lambda p_2(t) = \lambda^2 t e^{-\lambda t}, \quad p_2(0) = 0.$$

Applying the same technique to $p_2(t)$ as for $p_1(t)$, it follows that

$$\frac{d\left[e^{\lambda t} p_2(t)\right]}{dt} = \lambda^2 t.$$

Integrating both sides and applying the initial condition yields the solution

$$p_2(t) = (\lambda t)^2 \frac{e^{-\lambda t}}{2!}.$$

By induction, it can be shown that

$$p_i(t) = (\lambda t)^i \frac{e^{-\lambda t}}{i!}, \quad i = 0, 1, 2, \ldots.$$

The probability distribution, $\{p_i(t)\}_{i=0}^{\infty}$, represents a Poisson probability distribution with parameter λt. The mean and variance of this Poisson distribution satisfy

$$m(t) = \lambda t = \sigma^2(t).$$

For more general continuous time Markov chains, it can be difficult to solve a system of differential-difference equations in a sequential manner and obtain a general formula for $p_i(t)$. A pattern may not emerge as in the case of the Poisson process. Therefore, other techniques for obtaining information about the probabilities $p_i(t)$ in a continuous time Markov chain will be applied–in particular, techniques that involve solving a partial differential equation for the probability or moment generating function.

The probability $p_0(t) = e^{-\lambda t}$ in the Poisson process can be thought of as a waiting-time probability (i.e., the probability that the first event $0 \to 1$ occurs at a time greater than t). Let W_1 be the random variable for the time until the process reaches state 1, the holding time until the first jump. Then

$$\text{Prob}\{W_1 > t\} = e^{-\lambda t} \quad \text{or} \quad \text{Prob}\{W_1 \leq t\} = 1 - e^{-\lambda t}.$$

This latter expression is the cumulative distribution function for an exponential random variable with parameter λ. Thus, W_1 is an exponential random variable with parameter λ, with c.d.f. $F(t) = 1 - e^{-\lambda t}$ and p.d.f. $f(t) = F'(t) = \lambda e^{-\lambda t}$. In general, it can be shown that it takes an exponential amount of time to move from state i to state $i+1$ (i.e., the random variable for the time between jumps i and $i+1$, $W_{i+1} - W_i$, has an exponential distribution with parameter λ). In fact, sometimes in the definition of the Poisson process it is stated that the interevent times, $T_i = W_{i+1} - W_i$, are independent exponential random variables with parameter λ (Norris, 1999; Schinazi, 1999). Interevent times for more general continuous time Markov chains are discussed in Section 5.9, where it is shown that they are also exponential random variables.

Figure 5.3 is a sample path or realization of a Poisson process when $\lambda = 1$. A general method will be given for generating sample paths of birth and death processes based on this exponential distribution for the interevent time.

5.4 Generator Matrix Q

The transition probabilities p_{ji} are used to derive transition rates q_{ji}. The transition rates form a matrix known as the infinitesimal generator matrix Q. Matrix Q defines a relationship between the rates of change of the transition probabilities. First, we define the elements of the matrix $Q = (q_{ji})$.

Assume the transition probabilities $p_{ji}(t)$ are continuous and differentiable for $t \geq 0$ and at $t = 0$ they satisfy

$$p_{ji}(0) = 0, \quad j \neq i, \quad \text{and} \quad p_{ii}(0) = 1.$$

Define

$$q_{ji} = \lim_{\Delta t \to 0^+} \frac{p_{ji}(\Delta t) - p_{ji}(0)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{p_{ji}(\Delta t)}{\Delta t}, \quad i \neq j. \tag{5.6}$$

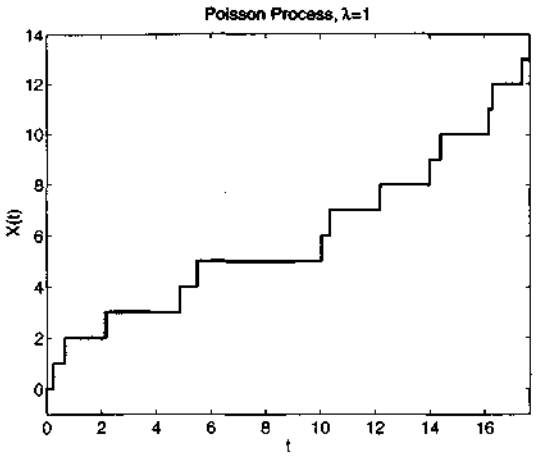

Figure 5.3. Sample path for a Poisson process with $\lambda = 1$.

Notice that $q_{ji} \geq 0$ since $p_{ji}(\Delta t) \geq 0$. In addition, define

$$q_{ii} = \lim_{\Delta t \to 0^+} \frac{p_{ii}(\Delta t) - p_{ii}(0)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{p_{ii}(\Delta t) - 1}{\Delta t}. \tag{5.7}$$

From equation (5.1), $\sum_{j=0}^{\infty} p_{ji}(\Delta t) = 1$, it follows that

$$1 - p_{ii}(\Delta t) = \sum_{j=0, j \neq i}^{\infty} p_{ji}(\Delta t) = \sum_{j=0, j \neq i}^{\infty} [q_{ji}\Delta t + o(\Delta t)].$$

Thus,

$$\begin{aligned} q_{ii} &= \lim_{\Delta t \to 0^+} \frac{-\sum_{j=0, j \neq i}^{\infty} [q_{ji}\Delta t + o(\Delta t)]}{\Delta t} \\ &= -\sum_{j=0, j \neq i}^{\infty} q_{ji}, \end{aligned} \tag{5.8}$$

where it is assumed that $\sum_{j \neq i} o(\Delta t) = o(\Delta t)$. This is certainly true if the summation contains only a finite number of terms (finite state space). If the summation contains an infinite number of terms (state space is infinite), Karlin and Taylor (1981) show that the limit (5.8) does exist on the extended interval, $[-\infty, 0]$, $0 \leq -q_{ii} \leq \infty$. In either case $q_{ii} \leq 0$. If q_{ii} is finite, then $\sum_{j=0}^{\infty} q_{ji} = 0$ and it follows from (5.6) and (5.7) that

$$p_{ji}(\Delta t) = \delta_{ji} + q_{ji}\Delta t + o(\Delta t), \tag{5.9}$$

where δ_{ji} is Kronecker's delta symbol.

Definition 5.2. The matrix of transition rates $Q = (q_{ji})$, where the elements q_{ji} are defined in (5.6) and (5.7), is known as the *infinitesimal generator matrix*,

$$Q = \begin{pmatrix} q_{00} & q_{01} & q_{02} & \cdots \\ q_{10} & q_{11} & q_{12} & \cdots \\ q_{20} & q_{21} & q_{22} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

$$= \begin{pmatrix} -\sum_{i=1}^{\infty} q_{i0} & q_{01} & q_{02} & \cdots \\ q_{10} & -\sum_{i=0, i\neq 1}^{\infty} q_{i1} & q_{12} & \cdots \\ q_{20} & q_{21} & -\sum_{i=0, i\neq 2}^{\infty} q_{i2} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

Sometimes the terms *transition rate matrix* or *infinitesimal matrix* or simply *generator matrix* are used when referring to matrix Q. Matrix Q has the property that each column sum is zero and the ith diagonal element is the negative of the sum of the off-diagonal elements in that column.

The next example shows that the system of differential-difference equations for the Poisson process (5.4) can be expressed in terms of the generator matrix Q (i.e., $dp/dt = Qp$).

Example 5.1 The generator matrix for the Poisson process can be calculated easily:

$$q_{i+1,i} = \lim_{\Delta t \to 0} \frac{p_{i+1,i}(\Delta t)}{\Delta t} = \lim_{t \to \infty} \frac{\lambda \Delta t + o(\Delta t)}{\Delta t} = \lambda.$$

In addition,

$$q_{ii} = -\lambda \quad \text{and} \quad q_{ji} = 0, \quad j \neq i, i+1.$$

The generator matrix satisfies

$$Q = \begin{pmatrix} -\lambda & 0 & 0 & \cdots \\ \lambda & -\lambda & 0 & \cdots \\ 0 & \lambda & -\lambda & \cdots \\ 0 & 0 & \lambda & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

The system of differential-difference equations for the Poisson process, equations (5.2) and (5.4), can be expressed in terms of the generator matrix Q as follows:

$$\frac{dp(t)}{dt} = Qp(t). \qquad \blacksquare$$

The generator matrix Q will be shown to be important for several reasons. In particular, Q is used to define the forward and backward Kolmogorov equations, which express the transition matrix $P(t)$ in terms of a differential equation, $dP/dt = QP$ and $dP/dt = PQ$. In addition, the generator matrix Q is used to define a transition matrix for the embedded Markov chain. We discuss these two concepts in the next two sections.

5.5 Embedded Markov Chain and Classification of States

Every sample path or realization of a continuous time Markov chain remains in a particular state (stays constant) for a random amount of time before making a jump to a new state. Recall that the waiting times are denoted as W_i, $i = 0, 1, 2, \ldots$, and the interevent times as $T_i = W_{i+1} - W_i$, $i = 0, 1, 2, \ldots$. For example, in the Poisson process, the states $0, 1, 2, \ldots$, are visited sequentially with an exponential amount of time between jumps.

Definition 5.3. Let Y_n denote the random variable for the state of a continuous time Markov chain $\{X(t)\}$, $t \in [0, \infty)$ at the nth jump,

$$Y_n = X(W_n), \quad n = 0, 1, 2, \ldots.$$

The set of random variables $\{Y_n\}_{n=0}^{\infty}$ is known as the *embedded Markov chain* or the *jump chain* associated with the continuous time Markov **chain** $\{X(t)\}$, $t \geq 0$.

The embedded Markov chain is a discrete time Markov chain. The embedded Markov chain is useful for classifying states (transient, recurrent, etc.) in the corresponding continuous time Markov chain. We shall define a transition matrix $T = (t_{ji})$ for the embedded Markov chain, $\{Y_n\}_{n=0}^{\infty}$, where $t_{ji} = \text{Prob}\{Y_{n+1} = j | Y_n = i\}$, which will be useful in classifying states. First, we demonstrate how to define the transition matrix for the embedded Markov chain of the Poisson process.

Example 5.2 Consider the Poisson process, where $X(0) = X(W_0) = 0$ and $X(W_n) = n$ for $n = 1, 2, \ldots$. The embedded Markov chain $\{Y_n\}$ satisfies $Y_n = n$, $n = 0, 1, 2, \ldots$. The transition from state n to $n+1$ occurs with probability 1. It is easy to see that the transition matrix for the embedded Markov chain corresponding to the Poisson process satisfies

$$T = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}. \tag{5.10}$$

■

In general, the transition matrix $T = (t_{ji})$ can be defined using the generator matrix Q. First, note that the transition probability t_{ii} is zero, because an assumption inherent in the definition of the embedded Markov chain is that the state must change, unless state i is absorbing. A state i in the continuous time chain is *absorbing* if $q_{ii} = 0$ (i.e., the rate of change is zero since the state does not change). Thus,

$$t_{ii} = \begin{cases} 0, & \text{if } q_{ii} \neq 0 \\ 1, & \text{if } q_{ii} = 0. \end{cases} \tag{5.11}$$

As motivation for the definition of the transition probability t_{ji}, recall that $q_{ji} = \lim_{\Delta t \to 0^+} p_{ji}(\Delta t)/\Delta t$ and $-q_{ii} = \lim_{\Delta t \to 0^+} (1 - p_{ii}(\Delta t))/\Delta t$. Thus,

$$-\frac{q_{ji}}{q_{ii}} = \lim_{\Delta t \to 0^+} \frac{p_{ji}(\Delta t)}{1 - p_{ii}(\Delta t)}.$$

This latter probability is the probability of a transfer from state i to j, given the process does not remain in state i. Hence, we define $t_{ji} = -q_{ji}/q_{ii}$, $q_{ii} \neq 0$. The transition probability t_{ji} for $j \neq i$ satisfies

$$t_{ji} = \begin{cases} \dfrac{q_{ji}}{\sum_{k=0, k\neq i}^{\infty} q_{ki}} = -\dfrac{q_{ji}}{q_{ii}}, & \text{if } q_{ii} \neq 0 \\ 0, & \text{if } q_{ii} = 0. \end{cases} \tag{5.12}$$

Definition 5.4. The matrix $T = (t_{ji})$, where the elements t_{ji} are defined in (5.11) and (5.12), is the *transition matrix of the embedded Markov chain* $\{Y_n\}_{n=0}^{\infty}$. In particular, for $q_{ii} \neq 0$, $i = 0, 1, 2, \ldots$,

$$T = \begin{pmatrix} 0 & -\frac{q_{01}}{q_{11}} & -\frac{q_{02}}{q_{22}} & \cdots \\ -\frac{q_{10}}{q_{00}} & 0 & -\frac{q_{12}}{q_{22}} & \cdots \\ -\frac{q_{20}}{q_{00}} & -\frac{q_{21}}{q_{11}} & 0 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

If any $q_{ii} = 0$, the (i, i) element of T is one and the remaining elements in that column are zero.

Matrix T is a stochastic matrix; the column sums equal one. The transition probabilities are homogeneous (i.e., independent of n). In addition, $T^n = (t_{ji}^{(n)})$, where $t_{ji}^{(n)} = \text{Prob}\{Y_n = j | Y_0 = i\}$. Using the generator matrix Q defined in Example 5.1 for the Poisson process, it can be seen that the transition matrix of the embedded Markov chain has the form given by (5.10).

Example 5.3 Suppose a continuous time, finite Markov chain has a generator matrix given by

$$Q = \begin{pmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{5.13}$$

The transition matrix of the corresponding embedded Markov chain satisfies

$$T = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{5.14}$$

From the embedded Markov chain, we can see that the states communicate in the following manner: $1 \to 2 \to 3 \to 4 \to 1$. ■

The classification schemes for states in continuous time Markov chains are the same as for discrete time Markov chains. The transition probabilities $P(t) = (p_{ji}(t))$ and the transition matrix for the embedded Markov chain $T = (t_{ji})$ are used to define these classification schemes. Definitions for communication class and irreducible in continuous time Markov chains can be defined in a manner similar to those for discrete time Markov chains.

Definition 5.5. State j can be *reached* from state i, $i \to j$, if $p_{ji}(t) > 0$ for some $t \geq 0$. State i *communicates* with state j, $i \leftrightarrow j$, if $i \to j$ and $j \to i$. The set of states that communicate is called a *communication class.* If every state can be reached from every other state, the Markov chain is *irreducible*; otherwise, it is said to be *reducible.* A set of states C is *closed* if it is impossible to reach any state outside of C from a state inside C, $p_{ji}(t) = 0$ for $t \geq 0$ if $i \in C$ and $j \notin C$.

In the case that $p_{ji}(\Delta t)$ equals $\delta_{ji} + q_{ji}\Delta t + o(\Delta t)$, then $p_{ji}(\Delta t) > 0$ iff $q_{ji} > 0$ for $j \neq i$ and Δt sufficiently small. Therefore, $i \leftrightarrow j$ in the continuous time Markov chain iff $i \leftrightarrow j$ in the embedded Markov chain. The generator matrix Q in the continuous Markov chain is irreducible iff the transition matrix T in the embedded Markov chain is irreducible.

Definitions for recurrent and transient states in the continuous time Markov chain can be defined in a manner similar to discrete time Markov chains. Let T_{ii} be the first time the chain is in state i after leaving state i,

$$T_{ii} = \inf\{t > W_1,\ X(t) = i | X(0) = i\}.$$

The random variable T_{ii} is known as the *first return time.* The first return can occur at any time $t > 0$; T_{ii} is a continuous random variable.

Definition 5.6. State i is *recurrent (transient)* in a continuous time Markov chain $\{X(t)\}$, $t \geq 0$, if the first return time is finite (infinite),

$$\text{Prob}\{T_{ii} < \infty | X(0) = i\} = 1 \ (< 1). \tag{5.15}$$

These definitions are similar to the definitions of recurrence and transience in discrete time Markov chains. Recall that state i is said to be *recurrent (transient)* in a discrete time Markov chain $\{Y_n\}$, with $Y_0 = i$, if

$$\sum_{n=0}^{\infty} f_{ii}^{(n)} = 1 \ (< 1),$$

where $f_{ii}^{(n)}$ is the probability that the first return to state i is at step n. The following theorem relates recurrent and transient states in continuous time Markov chains to recurrent and transient states in the corresponding embedded Markov chain. For a proof of this result, please consult Norris (1999) or Schinazi (1999).

Theorem 5.1. *State i in a continuous time Markov chain $\{X(t)\}$, $t \geq 0$, is recurrent (transient) iff state i in the corresponding embedded Markov chain Y_n, $n = 0, 1, 2, \ldots$, is recurrent (transient).*

Recurrence or transience in a continuous time Markov chains can be determined from the properties of the transition matrix T of the embedded Markov chain. For example, a state i in a continuous time Markov chain $\{X(t)\}$, $t \geq 0$, is *recurrent (transient)* iff

$$\sum_{n=0}^{\infty} t_{ii}^{(n)} = \infty \ (< \infty),$$

where $t_{ii}^{(n)}$ is the (i, i) element in the transition matrix of T^n of the embedded Markov chain $\{Y_n\}_{n=0}^{\infty}$. Other properties that determine recurrence and transience in discrete time Markov chains can be applied to continuous time Markov chains. For example, in a finite Markov chain, all states cannot be transient and if the finite Markov chain is irreducible, it is recurrent.

Note that the transition matrix of the embedded Markov chain for the Poisson process (Example 5.2) satisfies $\lim_{n\to\infty} T^n = \mathbf{0}$. For sufficiently large n and all i, $t_{ii}^{(n)} = 0$, which implies $\sum_{n=0}^{\infty} t_{ii}^{(n)} < \infty$. Therefore, every state is transient in the Poisson process. This is an obvious result since each state $X(W_i) = i$ can only advance to state $i + 1$, $X(W_{i+1}) = i + 1$; a return to state i is impossible.

Unfortunately, the concepts of null recurrence and positive recurrence for a continuous time chain cannot be defined in terms of the embedded Markov chain. Positive recurrence depends on the waiting times $\{W_i\}$ so that the embedded Markov chain alone is not sufficient to define positive

recurrence. See Schinazi (1999) for an example of an embedded Markov chain that is null recurrent but the corresponding continuous time Markov chain is positive recurrent.

Recall the definitions of positive recurrence and null recurrence in discrete time Markov chains. State i is *positive recurrent (null recurrent)* in the discrete time Markov chain $\{Y_n\}$ if the mean recurrence time is finite (infinite), $\sum_{n=1}^{\infty} n f_{ii}^{(n)} < \infty \ \ (= \infty)$. Positive and null recurrence in a continuous time Markov chain depends on the expected value of the random variable T_{ii}.

Definition 5.7. State i is *positive recurrent (null recurrent)* in the continuous time Markov chain $\{X(t)\}$, $t \geq 0$, if the mean recurrence time is finite (infinite); that is,

$$\mu_{ii} = E(T_{ii}|X(0) = i) < \infty \ \ (= \infty).$$

This definition is not very useful to show positive or null recurrence. Instead, the next theorem gives a method that can be used to determine μ_{ii}, and hence, positive or null recurrence. There are a number of limit theorems for continuous time Markov chains that give results similar to those for discrete time Markov chains. Recall the basic limit theorem for aperiodic discrete time Markov chains: A discrete time Markov chain $\{Y_n\}_{n=0}^{\infty}$ that is recurrent, irreducible, and aperiodic with transition matrix $T = (t_{ji})$ satisfies

$$\lim_{n\to\infty} t_{ji}^{(n)} = \frac{1}{\mu_{jj}},$$

where μ_{jj} is the mean recurrence time for the discrete time Markov chain. There is no concept of aperiodic and periodic in continuous time Markov chains because the interevent time is random. Therefore, the basic limit theorem for continuous time Markov chains is simpler. See Norris (1999) for a proof of the following result.

Theorem 5.2 (Basic limit theorem for continuous time Markov chains). *If the generator matrix Q of a continuous time, nonexplosive Markov chain $\{X(t)\}$, $t \geq 0$, is irreducible and positive recurrent, then*

$$\lim_{t\to\infty} p_{ij}(t) = -\frac{1}{q_{ii}\mu_{ii}}, \tag{5.16}$$

where μ_{ii} is the mean recurrence time in the continuous time chain $\{X(t)\}$. In particular, if the state space is finite, then the process is nonexplosive and the limit (5.16) exists and is positive if Q is irreducible.

Matrix Q is irreducible iff matrix T is irreducible. The result (5.16) differs slightly from discrete time Markov chains since an additional term $-q_{ii}$ is needed to define the limit. For finite Markov chains, all that is needed to show the existence of a positive limit (5.16) is to show that the

generator matrix Q is irreducible. In addition, since the limit is positive, $0 < \mu_{ii} < \infty$, it follows that irreducible, finite Markov chains are positive recurrent. The following result is reminiscent of the results from discrete time, finite Markov chains.

Corollary 5.1. *A continuous time, finite Markov chain with irreducible generator matrix Q is positive recurrent.*

Example 5.4 Consider Example 5.3. Matrix T given in equation (5.14) and matrix Q given in equation (5.13) are irreducible. All states are positive recurrent. Notice that the embedded Markov chain is periodic with period 4. However, the continuous time Markov chain is not periodic because periodicity is not defined for continuous time Markov chains. ■

In the next section, the forward and backward Kolmogorov differential equations are defined in terms of the generator matrix Q. In addition, the stationary probability distribution is defined for a continuous time Markov chain.

5.6 Kolmogorov Differential Equations

The forward and backward Kolmogorov differential equations are expressions for the rate of change of the transition probabilities. The transition probability $p_{ji}(t + \Delta t)$ can be expanded as follows by applying the Chapman-Kolmogorov equations,

$$p_{ji}(t + \Delta t) = \sum_{k=0}^{\infty} p_{jk}(\Delta t) p_{ki}(t).$$

Given that the generator matrix Q exists, we can apply the identity (5.9),

$$p_{ji}(t + \Delta t) = \sum_{k=0}^{\infty} p_{ki}(t) \left[\delta_{jk} + q_{jk}\Delta t + o(\Delta t)\right].$$

Subtract $p_{ji}(t)$, divide by Δt, and apply the identity $\sum_{k=0}^{\infty} p_{ki}(t) = 1$,

$$\frac{p_{ji}(t + \Delta t) - p_{ji}(t)}{\Delta t} = \sum_{k=0}^{\infty} p_{ki}(t) \left[q_{jk} + \frac{o(\Delta t)}{\Delta t}\right] = \sum_{k=0}^{\infty} p_{ki}(t) q_{jk} + \frac{o(\Delta t)}{\Delta t}.$$

Let $\Delta t \to 0$. Then

$$\frac{dp_{ji}(t)}{dt} = \sum_{k=0}^{\infty} q_{jk} p_{ki}(t), \quad i, j = 0, 1, \ldots. \tag{5.17}$$

Definition 5.8. The system of equations (5.17) represents the *forward Kolmogorov differential equations.* Expressed in matrix form, they are

$$\frac{dP(t)}{dt} = QP(t),$$

where $P(t) = (p_{ji}(t))$ is the matrix of transition probabilities and $Q = (q_{ji})$ is the generator matrix.

In the case that the initial distribution of the process satisfies $X(0) = k$ $(p_i(0) = \delta_{ik})$, then the transition probability $p_{ik}(t)$ is the same as the state probability $p_i(t) = \text{Prob}\{X(t) = i | X(0) = k\}$. Therefore, in this case, the state probabilities satisfy the forward Kolmogorov differential equations,

$$\frac{dp(t)}{dt} = Qp(t), \tag{5.18}$$

where $p(t) = (p_0(t), p_1(t), \ldots)^T$.

The system of differential equations (5.18) can be approximated by a system of difference equations corresponding to a discrete time Markov chain: $p(n+1) = Pp(n)$. This approximation shows the relationship between the Kolmogorov differential equations and the discrete time Markov chain. In particular, if the derivative $dp(t)/dt$ is approximated by the finite difference scheme, $[p(t+\Delta t) - p(t)]/\Delta t$, then the differential equation (5.18) can be expressed as

$$p(t + \Delta t) \approx [Q\Delta t + I]\, p(t),$$

where I is an infinite dimensional identity matrix, $I = \text{diag}(1, 1, \ldots)$. Suppose time is measured in units of Δt and $1 + q_{ii}\Delta t > 0$. Then it can be shown that the matrix $P = Q\Delta t + I$ is a stochastic matrix and

$$p(n+1) \approx Pp(n),$$

where the unit length of time n to $n+1$ is Δt (see Exercise 7).

The backward Kolmogorov differential equations can be derived in a manner similar to the forward Kolmogorov equations. Apply the Chapman-Kolmogorov equations and make the following substitutions:

$$p_{ji}(t + \Delta t) = \sum_{k=0}^{\infty} p_{ki}(\Delta t) p_{jk}(t) = \sum_{k=0}^{\infty} [\delta_{ki} + q_{ki}\ \Delta t + o(\Delta t)]\, p_{jk}(t).$$

Simplifications similar to the derivation of the forward equations and the assumption $\sum_{k=0}^{\infty} p_{jk}(t) < \infty$ yield the following system of differential equations:

$$\frac{dp_{ji}(t)}{dt} = \sum_{k=0}^{\infty} p_{jk}(t) q_{ki}, \quad i, j = 0, 1, \ldots. \tag{5.19}$$

Definition 5.9. The system of equations (5.19) represent the *backward Kolmogorov differential equations.* Expressed in matrix form, they are

$$\frac{dP(t)}{dt} = P(t)Q,$$

where $P(t) = (p_{ji}(t))$ is the matrix of transition probabilities and $Q = (q_{ji})$ is the generator matrix.

These differential equations depend on the existence of the generator matrix Q. For finite-dimensional systems or finite Markov chains, Q always exists. The solution $P(t)$ can be found via the forward or backward equations. In birth and death chains and other applications, the transition matrix $P(t)$ is defined in such a way that the forward and backward Kolmogorov differential equations can be derived.

The Kolmogorov differential equations can be used to define a stationary probability distribution. A constant solution to (5.18) is a stationary probability distribution. A formal definition is given next.

Definition 5.10. Let $\{X(t)\}$, $t \geq 0$, be a continuous time Markov chain with generator matrix Q. Suppose $\pi = (\pi_0, \pi_1, \ldots)^T$ is nonnegative and satisfies

$$Q\pi = 0, \quad \text{and} \quad \sum_{i=0}^{\infty} \pi_i = 1.$$

Then π is called a *stationary probability distribution* of **the continuous time** Markov chain.

A stationary probability distribution π can be defined in terms of the transition matrix $P(t)$ as well. A constant solution π is called a *stationary probability distribution* if

$$P(t)\pi = \pi, \quad \text{for} \quad t \geq 0, \quad \sum_{i=0}^{\infty} \pi_i = 1, \quad \text{and} \quad \pi_i \geq 0$$

for $i = 0, 1, 2 \ldots$. This latter definition can be applied if the transition matrix $P(t)$ is known and the process is nonexplosive.

The two definitions involving $P(t)$ and Q for the stationary probability distribution are equivalent if the transition matrix $P(t)$ is a solution of the forward and backward Kolmogorov differential equations [i.e., $dP(t)/dt = QP(t)$ and $dP(t)/dt = P(t)Q$]. This is always the case for finite Markov chains. The equivalence of these two definitions can be seen as follows. If $Q\pi = 0$, then for a finite Markov chain

$$\left[\frac{dP(t)}{dt}\right]\pi = \frac{d[P(t)\pi]}{dt} = P(t)Q\pi = 0,$$

which implies $P(t)\pi = \text{constant}$ for all t. But $P(0) = I$ implies $P(t)\pi = \pi$. On the other hand, if $P(t)\pi = \pi$ in a finite Markov chain, then

$$0 = \frac{d[P(t)\pi]}{dt} = \left[\frac{dP(t)}{dt}\right]\pi = QP(t)\pi = Q\pi.$$

An explicit solution $P(t)$ cannot be found for many continuous time Markov processes. Definition 5.10 is the one that will be applied most often to find the stationary probability distribution.

5.7 Finite Markov Chains

For some finite Markov chains it is possible to find an explicit solution $P(t)$ to the forward and backward Kolmogorov differential equations. In addition, in the case that the probability distribution $p(t)$ satisfies the forward Kolmogorov equation ($X(0) = k$, $p_i(0) = \delta_{ik}$), it is possible to calculate the probability distribution $p(t)$ directly [i.e., $p(t) = P(t)p(0)$].

Assume the state space of a finite Markov chain is $\{0, 1, 2, \ldots, N\}$. The forward and backward Kolmogorov differential equations have a unique solution. The forward and backward Kolmogorov differential equations satisfy $dP/dt = QP$ and $dP/dt = PQ$, respectively, where $P(0) = I$. The systems are linear, and the unique solution to each of them is given by

$$P(t) = e^{Qt}P(0) = e^{Qt},$$

where e^{Qt} is the matrix exponential,

$$e^{Qt} = I + Qt + Q^2\frac{t^2}{2!} + Q^3\frac{t^3}{3!} + \cdots = \sum_{k=0}^{\infty} Q^k\frac{t^k}{k!}.$$

This result can be verified easily. Let $P(t) = e^{Qt}$. Then differentiation yields

$$\frac{dP(t)}{dt} = \sum_{k=1}^{\infty} Q^k\frac{t^{k-1}}{(k-1)!} = Q\sum_{k=0}^{\infty} Q^k\frac{t^k}{k!} = Qe^{Qt} = QP(t).$$

However, it is also true that

$$\frac{dP(t)}{dt} = \left(\sum_{k=0}^{\infty} Q^k\frac{t^k}{k!}\right)Q = e^{Qt}Q = P(t)Q.$$

The solution e^{Qt} satisfies the forward and backward Kolmogorov differential equations, $dP/dt = QP$ and $dP/dt = PQ$, respectively. Uniqueness follows from the theory of differential equations (Brauer and Nohel, 1969). Some techniques for computing e^{Qt} for finite Markov chains are demonstrated.

Suppose matrix Q is an $n \times n$ diagonalizable matrix with eigenvalues, λ_i, $i = 1, 2, \ldots, n$. Then an expression for Q^k can be obtained by the method presented in Chapter 2, Section 2.8,

$$Q^k = H\Lambda^k H^{-1},$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and the columns of H are the right eigenvectors of H. The expression for e^{Qt} simplifies to

$$e^{Qt} = H \sum_{k=0}^{\infty} \Lambda^k \frac{t^k}{k!} H^{-1} = H\text{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_n t}) H^{-1}. \tag{5.20}$$

Differentiation of $P(t) = e^{Qt}$ can be used to generate information about the derivatives of P evaluated at $t = 0$. Notice that $P'(0) = Q$, $P''(0) = Q^2$ and, in general, $d^k P(t)/dt^k|_{t=0} = Q^k$ or

$$\left.\frac{d^k p_{ji}(t)}{dt^k}\right|_{t=0} = q_{ji}^{(k)}, \tag{5.21}$$

where $q_{ji}^{(k)}$ is the element in the jth row and ith column of $\boldsymbol{Q^k}$. The identity (5.20) shows that the elements of $P(t)$ satisfy

$$p_{ji}(t) = a_1 e^{\lambda_1 t} + a_2 e^{\lambda_2 t} + \cdots + a_n e^{\lambda_n t}.$$

Using the initial conditions (5.21), the coefficients a_k, $k = \mathbf{1, 2, \ldots, n}$ can be determined. Alternately, first computing H, then H^{-1},

$$P(t) = H\text{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_n t}) H^{-1}.$$

There are many other methods for computing the matrix exponential (see e.g., Leonard, 1996; Moler and Van Loan, 1978; Waltman, 1986). One may also use a computer algebra system or numerical methods to compute e^{Qt}. We mention one additional method for computing e^{Qt} which is due to Leonard (1996). This method is similar to one of the methods discussed in Chapter 2, Section 2.8 for computing the power of a matrix.

Suppose Q is an $n \times n$ matrix with characteristic polynomial,

$$det(\lambda I - Q) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0 = 0.$$

This polynomial equation is also a characteristic polynomial of an nth-order scalar differential equation of the form

$$x^{(n)}(t) + a_{n-1}x^{(n-1)}(t) + \cdots a_0 x(t) = 0.$$

To find a formula for e^{Qt} it is necessary to find n linearly independent solutions to this nth order scalar differential equation, $x_1(t)$, $x_2(t), \ldots, x_n(t)$,

with initial conditions

$$\left.\begin{matrix} x_1(0)=1 \\ x_1'(0)=0 \\ \vdots \\ x_1^{(n-1)}(0)=0 \end{matrix}\right\}, \quad \left.\begin{matrix} x_2(0)=0 \\ x_2'(0)=1 \\ \vdots \\ x_2^{(n-1)}(0)=0 \end{matrix}\right\}, \quad \cdots, \quad \left.\begin{matrix} x_n(0)=0 \\ x_n'(0)=0 \\ \vdots \\ x_n^{(n-1)}(0)=1 \end{matrix}\right\}.$$

Then

$$e^{Qt} = x_1(t)I + x_2(t)Q + \cdots + x_n(t)Q^{n-1}, \quad -\infty < t < \infty. \tag{5.22}$$

Some of these techniques are illustrated in the following examples.

Example 5.5 Suppose the generator matrix of a continuous time Markov chain with two states is

$$Q = \begin{pmatrix} -a & b \\ a & -b \end{pmatrix},$$

where $a > 0$ and $b > 0$. The matrix exponential, e^{Qt}, is computed first using the definition. Note that $Q^2 = -(a+b)Q$ and, in general, $Q^n = [-(a+b)]^{n-1}Q$. Then $P(t) = e^{Qt} = I + \sum_{n=1}^{\infty} \frac{(Qt)^n}{n!}$. Applying the identity for Q^n, it follows that

$$P(t) = I - \frac{Q}{a+b}\sum_{n=1}^{\infty}\frac{[-(a+b)t]^n}{n!} = I - \frac{Q}{a+b}[e^{-(a+b)t} - 1]. \tag{5.23}$$

Secondly, we compute the matrix exponential using the identity (5.22). The characteristic polynomial of Q is $\lambda^2 + (a+b)\lambda = 0$. Therefore, the eigenvalues of Q are $\lambda_{1,2} = 0, -(a+b)$. The general solution to this second-order differential equation $x''(t) + (a+b)x'(t) = 0$ is $x(t) = c_1 + c_2 e^{-(a+b)t}$. Applying the initial conditions to find the constants c_1 and c_2, the solutions $x_1(t)$ and $x_2(t)$ are $x_1(t) = 1$ and $x_2(t) = (1-e^{-(a+b)t})/(a+b)$, respectively. Applying the identity (5.22) gives the solution

$$e^{Qt} = x_1(t)I + x_2(t)Q = \frac{1}{a+b}\begin{pmatrix} b+ae^{-(a+b)t} & b-be^{-(a+b)t} \\ a-ae^{-(a+b)t} & a+be^{-(a+b)t} \end{pmatrix}.$$

This latter formula agrees with the solution given in (5.23).

The limit of $P(t)$ exists,

$$\lim_{t\to\infty} P(t) = \begin{pmatrix} \dfrac{b}{a+b} & \dfrac{b}{a+b} \\ \dfrac{a}{a+b} & \dfrac{a}{a+b} \end{pmatrix}$$

and, therefore, for any initial distribution $p(0) = (p_1(0), p_2(0))^T$,

$$\lim_{t\to\infty} P(t)p(0) = \begin{pmatrix} \dfrac{b}{a+b} \\ \dfrac{a}{a+b} \end{pmatrix} = \pi.$$

This limiting distribution π is the unique stationary probability **distribution** satisfying $Q\pi = 0$.

The transition matrix of the embedded Markov chain is

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Matrices T and Q are irreducible. Therefore, according to Theorem 5.2, the limit (5.16) satisfies $b/(a+b) = -1/(q_{11}\mu_{11})$ and $a/(a+b) = -1/(q_{22}\mu_{22})$. Elements $q_{11} = -a$ and $q_{22} = -b$, so that the mean recurrence time for the continuous time Markov chain is

$$\mu_{ii} = \frac{a+b}{ab}, \quad i = 1, 2.$$

It is interesting to note that the mean recurrence time for the corresponding embedded Markov chain is $\mu_{ii} = 2$, $i = 1, 2$. ■

Example 5.6 The forward Kolmogorov differential equations in Example 5.3 satisfy $dP/dt = QP$, where the generator matrix Q and transition matrix T of the embedded Markov chain are

$$Q = \begin{pmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The matrix exponential e^{Qt} can be computed by one of the methods discussed above or a computer algebra system may be used. Matrix $P(t) = e^{Qt}$ is given by

$$\begin{aligned} e^{Qt} &= \frac{1}{4}E + \frac{1}{2}e^{-t}\begin{pmatrix} \cos(t) & -\sin(t) & -\cos(t) & \sin(t) \\ \sin(t) & \cos(t) & -\sin(t) & -\cos(t) \\ -\cos(t) & \sin(t) & \cos(t) & -\sin(t) \\ -\sin(t) & -\cos(t) & \sin(t) & \cos(t) \end{pmatrix} \\ &\quad + \frac{1}{4}e^{-2t}\begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix}, \end{aligned}$$

where E is a 4×4 matrix of ones. Thus,

$$\lim_{t\to\infty} P(t) = \frac{1}{4}E.$$

Matrices T and Q are irreducible. Thus, according to Theorem 5.2, the mean recurrence time of the continuous time chain satisfies $\mu_{ii} = 4$, $i = 1, 2, 3, 4$. This mean recurrence time agrees with that of the embedded Markov chain. There exists a unique stationary probability distribution π satisfying $Q\pi = 0$. The stationary probability distribution $\pi = \frac{1}{4}(1, 1, 1, 1)^T$. In addition, note that $\pi = \lim_{t\to\infty} P(t)p(0)$, where $\pi_i = -1/(q_{ii}\mu_{ii})$ is the limit in Theorem 5.2. ■

These examples have illustrated that the limit in Theorem 5.2 is a stationary probability distribution. This result is true in general for finite Markov chains and is stated in the next theorem.

Theorem 5.3. *Suppose the generator matrix Q of a finite, continuous time Markov chain $\{X(t)\}$, $t \geq 0$, with state space $\{0, 1, 2, \ldots, N\}$ is irreducible. Then the limit in (5.16) is a stationary probability distribution π, where $\pi = (\pi_0, \pi_1, \ldots, \pi_N)^T$ and $\pi_i = -1/(q_{ii}\mu_{ii}) > 0$.*

Proof. Since matrix $P(t)$ is a finite matrix, it is a stochastic matrix for all $t \geq 0$. Hence, $\lim_{t\to\infty} P(t)p(0) = \pi$ has the property that $\sum_{i=0}^{N} \pi_i(i) = 1$. By the Chapman-Kolmogorov equations $P(t)P(s)\pi = P(t+s)\pi$. Hold t fixed and take the limit as $s \to \infty$; then $P(t)L = L$. Since t is arbitrary, π is a stationary probability distribution. □

Example 5.7 The generator matrix Q of a continuous time Markov chain and transition matrix T of the embedded Markov chain are

$$Q = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 2 & 0 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & -3 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Matrices T and Q are reducible because the first state is absorbing. The exponential matrix e^{Qt} is computed using the eigenvalues of e^{Qt}. The eigenvalues of Q are $0, -1, -2$, and -3, $Q = H\text{diag}(0, -1, -2, -3)H^{-1}$ so that

$$e^{Qt} = H\text{diag}(1, e^{-t}, e^{-2t}, e^{-3t})H^{-1}.$$

We compute $p_{44}(t)$ by the method described in this section,

$$p_{44}(t) = a_1 + a_2 e^{-t} + a_3 e^{-2t} + a_4 e^{-3t}.$$

Applying the initial conditions, $p_{44}(0) = 1$, $p'_{44}(0) = q_{44}$, $p''_{44}(0) = q^{(2)}_{44}$, and $p'''_{44}(0) = q^{(3)}_{44}$, yields the four linear equations,

$$\begin{aligned} a_1 + a_2 + a_3 + a_4 &= 1 \\ -a_2 - 2a_3 - 3a_4 &= -3 \\ a_2 + 4a_3 + 9a_4 &= 9 \\ -a_2 - 8a_3 - 27a_4 &= -27. \end{aligned}$$

Figure 5.4. Directed graph of the embedded Markov chain $\{Y_n\}$.

The solution of this linear system is $a_1 = a_2 = a_3 = 0$ and $a_4 = 1$. Thus, $p_{44}(t) = e^{-3t}$. A closed form expression can be obtained for e^{Qt}. Matrix

$$e^{Qt} = \begin{pmatrix} 1 & 1-e^{-t} & 1-2e^{-t}+e^{-2t} & 1-3e^{-t}+3e^{-2t}-e^{-3t} \\ 0 & e^{-t} & 2e^{-t}-2e^{-2t} & 3e^{-t}-6e^{-2t}+3e^{-3t} \\ 0 & 0 & e^{-2t} & 3e^{-2t}-3e^{-3t} \\ 0 & 0 & 0 & e^{-3t} \end{pmatrix}.$$

Let $p(0) = (0,0,0,1)^T$. Then $p(t) = P(t)p(0)$, where $p(t)$ is

$$\left(1-3e^{-t}+3e^{-2t}-e^{-3t}, 3e^{-t}-6e^{-2t}+3e^{-3t}, 3e^{-2t}-3e^{-3t}, e^{-3t}\right)^T.$$

The mean and variance of this distribution satisfy

$$m(t) = \sum_{k=0}^{3} k p_k(t) = 3e^{-t}$$

and

$$\sigma^2(t) = \sum_{k=0}^{3} k^2 p_k(t) - m^2(t) = 3e^{-t}(1-e^{-t}).$$

Although Theorems 5.2 and 5.3 do not apply to this example, note that $\lim_{t\to\infty} p(t) = (1,0,0,0)^T$, which is a stationary probability distribution. The probability of extinction, $p_0(t)$, approaches one,

$$\lim_{t\to\infty} p_0(t) = 1.$$

This result is reasonable because the zero state is absorbing (see Figure 5.4). This example is a special case of a simple death process that will be discussed more fully in Chapter 6. ■

5.8 Generating Function Technique

We present another method for finding the probability distribution associated with the process $\{X(t)\}$. In this method, a partial differential equation satisfied by a generating function is derived, either p.g.f., m.g.f, or c.g.f. Denote the p.g.f. of a continuous time Markov chain $\{X(t)\}$ as

$$P(z,t) = \sum_{i=0}^{\infty} p_i(t) z^i,$$

the m.g.f. as

$$M(\theta,t)=\sum_{i=0}^{\infty}p_i(t)e^{\theta i},$$

and the c.g.f. as $K(\theta,t)=\ln M(\theta,t)$. Notice that the generating functions depend on two continuous variables, z and t or θ and t, where $t\geq 0$, and the domain of z or θ consists of the values where the summation converges [e.g., $|z|<1$ for $P(z,t)$].

Recall that the mean $m(t)$ of the process at time t satisfies

$$m(t)=\left.\frac{\partial P(z,t)}{\partial z}\right|_{z=1}=\sum_{i=0}^{\infty}ip_i(t)$$

In terms of $M(\theta,t)$ and $K(\theta,t)$,

$$m(t)=\left.\frac{\partial M(\theta,t)}{\partial\theta}\right|_{\theta=0}=\left.\frac{\partial K(\theta,t)}{\partial\theta}\right|_{\theta=0}.$$

In addition, the variance $\sigma^2(t)$ satisfies

$$\sigma^2(t)=\left.\frac{\partial^2P(z,t)}{\partial z^2}\right|_{z=1}+\left.\frac{\partial P(z,t)}{\partial z}\right|_{z=1}-\left(\left.\frac{\partial P(z,t)}{\partial z}\right|_{z=1}\right)^2$$

or, in terms of $M(\theta,t)$ and $K(\theta,t)$,

$$\sigma^2(t)=\left.\frac{\partial^2M(\theta,t)}{\partial\theta^2}\right|_{\theta=0}-\left(\left.\frac{\partial M(\theta,t)}{\partial\theta}\right|_{\theta=0}\right)^2=\left.\frac{\partial^2K(\theta,t)}{\partial\theta^2}\right|_{\theta=0}.$$

A partial differential equation for the generating function is derived using the forward Kolmogorov equation. First, the technique is discussed; then an example is given using the Poisson process. When the initial distribution is a fixed value, then the forward Kolmogorov differential equations can be expressed in terms of the state probabilities, $dp/dt=Qp$,

$$\frac{dp_i(t)}{dt}=\sum_{k=0}^{\infty}q_{ik}p_k(t),\quad i=0,1,2,\ldots.$$

If each of these equations is multiplied by z^i and then summed over i,

$$\sum_{i=0}^{\infty}\frac{dp_i(t)}{dt}z^i=\sum_{i=0}^{\infty}\sum_{k=0}^{\infty}q_{ik}p_k(t)z^i.$$

Interchanging the summation and differentiation and the order of the summation (possible for values of t and z, where the summation converges absolutely), yields

$$\frac{\partial P(z,t)}{\partial t}=\sum_{k=0}^{\infty}\left[\sum_{i=0}^{\infty}q_{ik}p_k(t)z^i\right].$$

If for each i, q_{ik} is zero, except for finitely many k, then the right-hand side may be expressed in terms of either $P(z,t)$ or the first-, second-, or higher-order derivatives of P with respect to z.

A partial differential equation for the m.g.f. can be derived in a similar manner. Instead of multiplying by z^i, the forward Kolmogorov differential equation is multiplied by $e^{i\theta}$. Alternately, a differential equation for the m.g.f. can be derived directly from the differential equation for the p.g.f. by making a change of variable. Recall that $M(\theta,t) = P(e^\theta,t)$, so that $z = e^\theta$. In addition, a differential equation for the c.g.f. can be obtained from the one for the m.g.f. by letting $K(\theta,t) = \ln M(\theta,t)$. The generating function technique will be used frequently with birth and death chains and other applications. If the differential equations are linear and first order, they can be solved by the method of characteristics. The last section of this chapter is devoted to a brief review of the method of characteristics. The generating function technique is illustrated in the next example for the Poisson process.

Example 5.8 The forward Kolmogorov differential equations for the Poisson process given in (5.2) and (5.4) are

$$\begin{aligned} \frac{dp_i\ (t)}{dt} &= -\lambda p_i(t) + \lambda p_{i-1}(t), \quad i \geq 1, \\ \frac{dp_0\ (t)}{dt} &= -\lambda p_0(t). \end{aligned}$$

Multiplying by z^i, then summing over i,

$$\sum_{i=0}^{\infty} \frac{dp_i\ (t)}{dt} z^i = -\lambda \sum_{i=0}^{\infty} p_i(t) z^i + \lambda \sum_{i=1}^{\infty} p_{i-1}(t) z^i.$$

Interchanging differentiation and integration yields the differential equation,

$$\frac{\partial P(z,t)}{\partial t} = -\lambda P(z,t) + z\lambda P(z,t) = \lambda(z-1)P(z,t).$$

Because there is no differentiation with respect to z, the variable z can be treated as a constant. The solution to this differential equation is an exponential function in t,

$$P(z,t) = P(z,0)e^{\lambda(z-1)t}.$$

Recall that $X(0) = 0$. Thus, $p_0(0) = \text{Prob}\{X(0) = 0\} = 1$ and $p_i(0) = \text{Prob}\{X(0) = i\} = 0$. Hence, $P(z,0) = 1$ and the p.g.f. satisfies

$$P(z,t) = e^{\lambda t(z-1)}.$$

Replacing z by e^θ yields the m.g.f. $M(\theta,t) = e^{\lambda t(e^\theta - 1)}$, and taking logarithms yields the c.g.f, $K(\theta,t) = \lambda t(e^\theta - 1)$. As expected, the p.g.f., m.g.f., and c.g.f. are the generating functions corresponding to a Poisson distribution with parameter λt. ■

5.9 Interevent Time and Stochastic Realizations

To calculate sample paths of a continuous time Markov chain $\{X(t)\}$, $t \geq 0$, we need to know the distribution for the time between successive events or the *interevent* time. Recall that the random variable for the interevent time is $T_i = W_{i+1} - W_i$, where W_i is the time of the ith jump (see Figure 5.5). In applications, the event may be a birth, death, immigration, or any other event that changes the value of the state variable. Note that $W_{i+1} > W_i$ so that $T_i \in [0, \infty)$; T_i is a continuous random variable. It will be shown that the interevent time T_i is an exponential random variable.

Assume the value of the state at the ith jump is n, $X(W_i) = n$. Let $\alpha(n)\Delta t + o(\Delta t)$ be the probability that the process moves to a state different from n in the time period Δt; that is,

$$\sum_{j=0, j\neq n}^{\infty} p_{jn}(\Delta t) = \alpha(n)\Delta t + o(\Delta t).$$

A change in state could result from a birth, death, or immigration. Then the probability of no change in state is $1 - \alpha(n)\Delta t + o(\Delta t)$; that is,

$$p_{nn}(\Delta t) = 1 - \alpha(n)\Delta t + o(\Delta t).$$

Let $G_i(t)$ be the probability that the process remains in state n for a time of length t, that is, for a time of length $[W_i, W_i + t]$. Then $G_i(t)$ can be expressed in terms of the interevent time T_i,

$$G_i(t) = \text{Prob}\{t + W_i < W_{i+1}\} = \text{Prob}\{T_i > t\}.$$

If state n is not an absorbing state, so that there is a positive probability of moving to another state, then, at $t = 0$, $G_i(0) = \text{Prob}\{T_i > 0\} = 1$. For

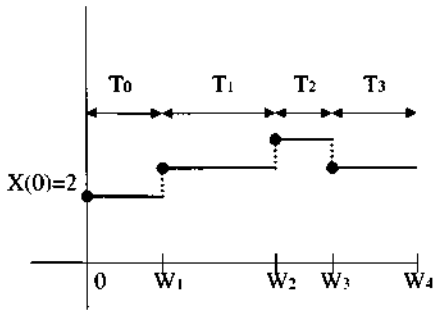

Figure 5.5. A sample path or single realization $X(t)$ of a continuous time Markov chain, $t \in [0, \infty)$ illustrating the jump times $\{W_i\}$ and the interevent times $\{T_i\}$, $X(0) = 2$, $X(W_1) = 3$, $X(W_2) = 4$, $X(W_3) = 3$.

Δt sufficiently small,

$$G_i(t + \Delta t) = G_i(t)p_{nn}(\Delta t) = G_i(t)(1 - \alpha(n)\Delta t + o(\Delta t)). \tag{5.24}$$

Subtract $G_i(t)$ from both sides of the preceding equation and divide by Δt. Then taking the limit as $\Delta t \to 0$, it follows that

$$\frac{dG_i(t)}{dt} = -\alpha(n)G_i(t).$$

The differential equation is first order and homogeneous with initial condition $G_i(0) = 1$. The solution is

$$G_i(t) = \text{Prob}\{T_i > t\} = e^{-\alpha(n)t}.$$

Thus, the probability that $T_i \leq t$ is

$$\text{Prob}\{T_i \leq t\} = 1 - G_i(t) = 1 - e^{-\alpha(n)t} = F_i(t), \quad t \geq 0.$$

The function $F_i(t)$ is the cumulative distribution function for the interevent time T_i, which corresponds to an exponential random variable with parameter $\alpha(n)$. The p.d.f. for T_i is $F_i'(t) = f_i(t) = \alpha(n)e^{-\alpha(n)t}$. Recall that the mean and variance for an exponential random variable with parameter λ satisfy $E(T_i) = 1/\alpha(n)$ and $\text{Var}(T_i) = 1/[\alpha(n)]^2$. These results are summarized in the next theorem.

Theorem 5.4. *Let $\{X(t)\}$, $t \geq 0$, be a continuous time Markov chain such that*

$$\sum_{j=0, j\neq n}^{\infty} p_{jn}(\Delta t) = \alpha(n)\Delta t + o(\Delta t)$$

and

$$p_{nn}(\Delta t) = 1 - \alpha(n)\Delta t + o(\Delta t)$$

for Δt sufficiently small. Then the interevent time, $T_i = W_{i+1} - W_i$ given $X(W_i) = n$, is an exponential random variable with parameter $\alpha(n)$. The c.d.f. for T_i is $F_i(t) = 1 - e^{-\alpha(n)t}$ so that the mean and variance of T_i satisfy

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

For example, in a birth process with birth probability $b_n\Delta t + o(\Delta t)$, given $X(W_i) = n$, there will be a mean waiting time of $E(T_i) = 1/b_n$, until another birth, $X(W_{i+1}) = n + 1$. Suppose there is more than one event, such as a birth or a death and $X(W_i) = n$. If the death probability is $d_n\Delta t + o(\Delta t)$, there will be a mean waiting time of $E(T_i) = 1/(b_n + d_n)$. When an event occurs, it will be a birth with probability $b_n/(b_n + d_n)$ and a death with probability $d_n/(b_n + d_n)$. To ensure that the Markov process defined in this manner is the unique representation of the original

continuous time Markov process requires a more rigorous mathematical justification than presented here; the interested reader is referred to the references (e.g., Karlin and Taylor, 1981; Norris, 1999).

An assumption used in the preceding derivation for $G_i(t)$ is an inherent property of the exponential distribution. It was assumed in (5.24) that $\text{Prob}\{T_i \geq t + \Delta t\} = \text{Prob}\{T_i \geq t\}\text{Prob}\{T_i \geq \Delta t\}$. This property can be written as

$$\text{Prob}\{T_i \geq t + \Delta t | T_i \geq t\} = \text{Prob}\{T_i \geq \Delta t\}.$$

It is a property of the exponential distribution referred to as a *memoryless property.* It is due to this memoryless property that Markov processes have an interevent time that is exponential.

The set of interevent times $\{T_i\}$ have an important probabilistic relationship. The interevent times $\{T_i\}$ are independent if conditioned on the successive states visited by the Markov chain. In particular, given $X(W_i) = Y_i$, then the interevent time T_i is independent of T_{i-1}. $i = 1, 2, \ldots$ (Schinazi, 1999).

Example 5.9 Let $\Delta X(t) = X(t + \Delta t) - X(t)$. A simple birth process is defined. For Δt sufficiently small, the transition probabilities satisfy

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} bi\Delta t + o(\Delta t), & j = 1 \\ 1 - bi\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \geq 2 \\ 0, & j < 0. \end{cases} \end{aligned}$$

Denote the expected time to reach state k, $k \geq 2$, as τ_k. In general, if $X(W_i) = k$, then the interevent time T_i has p.d.f. bke^{-bkt} and the expected time to reach state $k + 1$ from state k is $E(T_i | X(W_i) = k) = 1/(bk)$. If, for example, $X(0) = 1$, then $X(W_1) = 2$, and, in general, $X(W_k) = k + 1$. The expected time to reach state k beginning from state 1 can be computed,

$$\tau_2 = E(T_0) = \frac{1}{b},$$

$$\tau_3 = E(T_0) + E(T_1) = \frac{1}{b} + \frac{1}{2b} = \frac{1}{b}\left[1 + \frac{1}{2}\right],$$

and

$$\tau_k = \sum_{i=0}^{k-2} E(T_i) = \frac{1}{b}\sum_{i=1}^{k-1} \frac{1}{i}. \qquad \blacksquare$$

Next, we show that the random variable T_i can be expressed in terms of the distribution function $F_i(t)$ and a uniform random variable U. This relationship is very useful for computational purposes; that is, when stochastic realizations are numerically simulated.

Theorem 5.5. *Let U be a uniform random variable defined on $[0,1]$ and T be a continuous random variable defined on $[0,\infty)$. Then $T = F^{-1}(U)$, where F is the cumulative distribution of the random variable T.*

Proof. Since $\text{Prob}\{T \le t\} = F(t)$, we want to show that $\text{Prob}\{F^{-1}(U) \le t\} = F(t)$. First note that $F : [0,\infty) \to [0,1)$ is strictly increasing, so that F^{-1} exists. In addition, for $t \in [0,\infty)$,

$$\begin{aligned} \text{Prob}\{F^{-1}(U) \le t\} &= \text{Prob}\{F(F^{-1}(U)) \le F(t)\} \\ &= \text{Prob}\{U \le F(t)\}. \end{aligned}$$

Because U is a uniform random variable, $\text{Prob}(U \le y) = y$ for $y \in [0,1]$. Thus, $\text{Prob}\{U \le F(t)\} = F(t)$. □

In the Poisson process, the only change in state is a birth that occurs with probability $\lambda\Delta t + o(\Delta t)$ in a small interval of time Δt. Because $\alpha(n) = \lambda$, the distribution function for the interevent time is $F_i(t) = \text{Prob}\{T_i \le t\} = 1 - \exp(-\lambda t)$. But because λ is independent of the state of the process, the interevent time is the same for every jump i, $T_i \equiv T$. The interevent time T, expressed in terms of the uniform random variable U, is $T = F^{-1}(U)$. The function $F^{-1}(U)$ is found by solving $F(T) = 1 - \exp(-\lambda T) = U$ for U:

$$T = F^{-1}(U) = -\frac{\ln(1-U)}{\lambda}.$$

However, because U is a uniform random variable on $[0,1]$, so is $1 - U$. It follows that the interevent time can be expressed in terms of a uniform random variable U as follows:

$$T = -\frac{\ln(U)}{\lambda}. \tag{5.25}$$

For more general processes, the formula given in (5.25) for the interevent time depends on the state of the process. In particular, given $X(W_i) = n$, the interevent time T_i satisfies

$$\boxed{T_i = -\frac{\ln(U)}{\alpha(n)},} \tag{5.26}$$

where U is a uniform random variable on [0,1].

The formula given in (5.26) for the interevent time is applied to three simple birth and death processes, known as simple birth, simple death, and simple birth and death processes. In each of these processes, probabilities of births and deaths are linear functions of the population size. These processes will be considered in more detail in Chapter 6. Let $X(t)$ be the random variable for the total population size at time t.

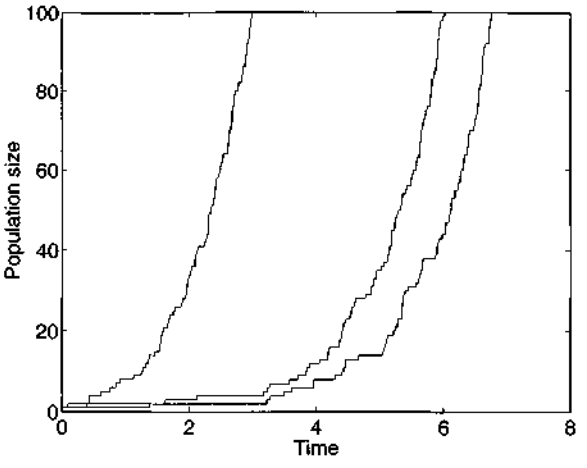

Figure 5.6. Three stochastic realizations of the simple birth process, $X(0) = 1$ and $b = 1$.

Example 5.10 [Simple birth process] Consider the simple birth process defined in Example 5.9. Given $X(W_i) = n$, the probability of a change in the population size is $bn\Delta t + o(\Delta t)$. Thus, $\alpha(n) = bn$. The interevent time T_i satisfies

$$T_i = -\frac{\ln(U)}{bn},$$

where U is a uniform random variable. The next event is a birth $n \to n+1$, $X(Wi+1) = n+1$. The *deterministic* analogue of this simple birth process is the exponential growth model $dn/dt = bn$, $n(0) = N$, whose solution is

$$n(t) = Ne^{bt}. \quad \blacksquare$$

A MATLAB program is given in the Appendix for Chapter 5 that generates three sample paths or realizations of a simple birth process when $b = 1$ and $X(0) = 1$. Three sample paths are graphed in Figure 5.6.

Example 5.11 [Simple death process] In a simple death process, the only event is a death, that is, state $i \to i - 1$. For Δt sufficiently small, the transition probabilities satisfy

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} di\Delta t + o(\Delta t), & j = -1 \\ 1 - di\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \le -2 \\ 0, & j > 0. \end{cases} \end{aligned} \tag{5.27}$$

Given $X(W_i) = n$, then $\alpha(n) = dn$. Therefore, the interevent time T_i is

$$T_i = -\frac{\ln(U)}{dn}.$$

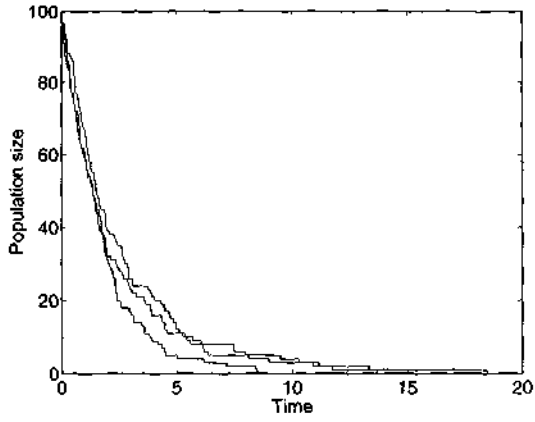

Figure 5.7. Three stochastic realizations of the simple death process, $X(0) = 100$ and $d = 0.5$.

The next event is a death $n \to n-1$, $X(W_{i+1}) = n-1$. The *deterministic* analogue of this simple death process is the differential equation $dn/dt = -dn$, $n(0) = N$, with solution

$$n(t) = Ne^{-dt}.$$

Figure 5.7 graphs three sample paths of this process when $X(0) = 100$ and $d = 0.5$. ■

Example 5.12 [Simple birth and death process] In the simple birth and death process, an event can be a birth or a death, $i \to i+1$ or $i \to i-1$. For Δt sufficiently small, the transition probabilities satisfy

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} di\Delta t + o(\Delta t), & j = -1 \\ bi\Delta t + o(\Delta t), & j = 1 \\ 1 - (b+d)i\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq -1,\ 0,\ 1. \end{cases} \end{aligned}$$

Given $X(W_i) = n$, $\alpha(n) = (b+d)n$. Therefore, the interevent time T_i satisfies

$$T_i = -\frac{\ln(U)}{(b+d)n}.$$

The next event is either a birth or a death; a birth occurs with probability $b/(b+d)$ and a death with probability $d/(b+d)$. The *deterministic* analogue of this simple birth and death process is the differential equation $dn/dt = (b-d)n$, $n(0) = N$, with solution

$$n(t) = Ne^{(b-d)t}.$$

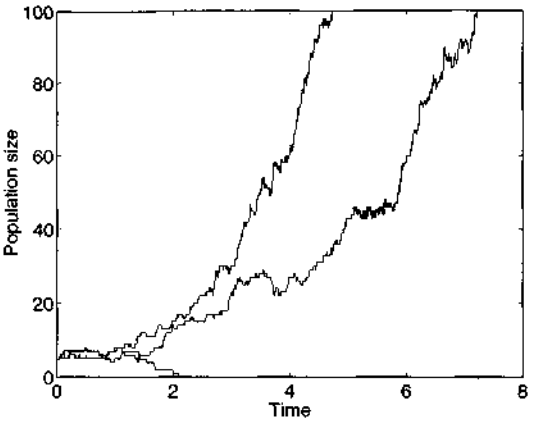

Figure 5.8. Three stochastic realizations of the simple birth and death process, $X(0) = 5$, $b = 1$, and $d = 0.5$.

Figure 5.8 graphs three sample paths of the simple birth and death process when $X(0) = 5$, $b = 1$, and $d = 0.5$. Note that one of the sample paths hits zero before $t = 4$. ■

5.10 Review of Method of Characteristics

For the simple birth, simple death, and simple birth and death processes, the generating functions will be first-order linear partial differential equations. To solve these partial differential equations, the method of characteristics can be applied. For more information on the method of characteristics, please consult a textbook on partial differential equations (Farlow, 1982; John, 1975; Schovanec and Gilliam, 2000). The method of characteristics is illustrated with an example.

Example 5.13 Let $P(z,t)$ satisfy the partial differential equation

$$\frac{\partial P}{\partial t} + (z+1)\frac{\partial P}{\partial z} = 1, \quad P(z,0) = \phi(z). \tag{5.28}$$

The domain of the differential equation corresponds to $t \in [0,\infty)$ and $z \in (-\infty,\infty)$.

In the method of characteristics, it is assumed that the partial differential equation can be expressed as a system of ordinary differential equations along characteristic curves, curves expressed in terms of auxiliary variables s and τ. Assume that $P(z,t) \equiv P(z(s,\tau),t(s,\tau)) \equiv P(s,\tau)$. Along the characteristic curves, the variable $s =$ constant, so that $P(z(s,\tau),t(s,\tau)) =$

$P(z(\tau), t(\tau))$. The characteristic curves are found by solving the following ordinary differential equations:

$$\frac{dt}{d\tau} = 1, \quad \frac{dz}{d\tau} = z + 1, \quad \text{and} \quad \frac{dP}{d\tau} = 1,$$

with initial conditions

$$t(s,0) = 0, \quad z(s,0) = s, \quad \text{and} \quad P(s,0) = \phi(s).$$

The reason this method works is that along characteristic curves, solutions satisfying the ordinary differential equations also satisfy the partial differential equation,

$$\frac{dP}{d\tau} = \frac{\partial P}{\partial z}\frac{dz}{d\tau} + \frac{\partial P}{\partial t}\frac{dt}{d\tau} = (z+1)\frac{\partial P}{\partial z} + \frac{\partial P}{\partial t}.$$

The system of ordinary differential equations is solved and P is expressed in terms of s and τ. Then the variables s and τ are expressed in terms of the original variables z and t.

The solution to the system of ordinary differential equations along the characteristic curves satisfies

$$t(s,\tau) = \tau, \quad z + 1 = (s+1)e^{\tau}, \quad \text{and} \quad P(s,\tau) = \tau + \phi(s).$$

The variables τ and s can be expressed in terms of z and t as follows: $\tau = t$ and $s = (z+1)e^{-t} - 1$. Substituting these values for τ and s into P gives the solution in terms of z and t,

$$P(z,t) = t + \phi((z+1)e^{-t} - 1).$$

For example, if $\phi(z) = z^3$, then

$$P(z,t) = t + \left[(z+1)e^{-t} - 1\right]^3.$$

The solution can be verified by checking that it solves the partial differential equation and the initial condition given in (5.28). ■

5.11 Exercises for Chapter 5

1. Suppose the generator matrix of a continuous time Markov chain is

$$Q = \begin{pmatrix} -a-b & c & e \\ a & -c-d & f \\ b & d & -e-f \end{pmatrix},$$

where the constants a, b, c, d, e, and f are positive. Find the transition matrix T of the embedded Markov chain. Are the states recurrent or transient?

2. Suppose the generator matrix Q of a continuous time Markov chain satisfies

$$Q = \begin{pmatrix} -a & d & 0 & 0 \\ a & -b-d-a & 2d & 0 \\ 0 & b+a & -2b-2d-a & 3d \\ 0 & 0 & 2b+a & -3d \end{pmatrix}.$$

   (a) Suppose $b = d = a > 0$. Find the corresponding transition matrix T for the embedded Markov chain. Are the states recurrent or transient?

   (b) Suppose $a = 0$ and $b = d > 0$. Find the corresponding transition matrix T for the embedded Markov chain. Are the states recurrent or transient?

3. Suppose the generator matrix Q of a continuous time Markov chain satisfies

$$Q = \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}.$$

   (a) Show that $Q^n = (-3)^{n-1}Q$.

   (b) Use (a) to compute $P(t) = e^{Qt} = \sum_{n=0}^{\infty}(Qt)^n/n!$.

   (c) Show that Q is irreducible. Find $\lim_{t\to\infty} P(t)$; then use (5.16) to compute μ_{ii}, for $i = 1, 2$.

   (d) Show that there exists a unique probability stationary distribution π, $Q\pi = 0$.

   (e) Verify that the limit $\pi = \lim_{t\to\infty} P(t)p(0)$ equals the unique stationary probability distribution.

4. Suppose the generator matrix of a continuous time Markov chain satisfies

$$Q = \begin{pmatrix} -2 & 1 & 2 \\ 1 & -1 & 1 \\ 1 & 0 & -3 \end{pmatrix}. \tag{5.29}$$

   (a) Find the eigenvalues λ_1, λ_2, and λ_3 of matrix Q; then express

$$p_{11}(t) = a_1 e^{\lambda_1 t} + a_2 e^{\lambda_2 t} + a_3 e^{\lambda_3 t}.$$

   (b) Find $p_{11}(0)$, $p'_{11}(0) = q_{11}$, and $p''_{11}(0) = q_{11}^{(2)}$; then solve for the coefficients a_1, a_2, and a_3 (see Norris, 1999).

5. Suppose the generator matrix satisfies (5.29).

   (a) Find the corresponding transition matrix T and show that T is irreducible.

(b) Find the matrix $P(t) = e^{Qt}$.

(c) Find $\lim_{t\to\infty} P(t)$. Then apply equation (5.16) to compute μ_{ii} for $i = 1, 2, 3$.

(d) Show that there exists a unique stationary probability distribution π satisfying $Q\pi = 0$.

(e) Verify that the limit $\pi = \lim_{t\to\infty} P(t)p(0)$ equals the unique stationary probability distribution.

6. Suppose the generator matrix of a continuous time Markov chain satisfies

$$Q = \begin{pmatrix} -1 & 4 & 2 \\ 0 & -4 & 1 \\ 1 & 0 & -3 \end{pmatrix}.$$

(a) Find the corresponding transition matrix T of the embedded Markov chain. Is the chain irreducible or reducible?

(b) The $\lim_{t\to\infty} P(t)p(0)$ converges to a unique stationary distribution. Find the stationary distribution.

7. When the initial distribution of the process is a fixed value, then the probability distribution $p(t)$ satisfies the forward Kolmogorov differential equations,

$$\frac{dp(t)}{dt} = Qp(t).$$

(a) Show that the differential equation can be approximated by

$$p(n+1) = Pp(n),$$

where $n = t$, the unit length of time n to $n+1$ is Δt, and $P = Q\Delta t + I$.

(b) Assume the elements q_{ii} are finite and $1 + q_{ii}\Delta t \geq 0$ for $i = 0, 1, 2, \ldots$. Show that P is a stochastic matrix.

8. Suppose $\{X(t)\}$ is a birth process with values in $\{1, 2, \ldots\}$. Let $\Delta X(t) = X(t + \Delta t) - X(t)$. Assume

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} b_i\Delta t + o(\Delta t), & j = 1 \\ 1 - b_i\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \geq 2 \\ 0, & j < 0, \end{cases} \end{aligned}$$

for Δt sufficiently small and $i = 1, 2, \ldots$. Suppose $X(0) = 1$, $b_0 = 0$, $b_1 = 1$, $b_2 = 3$, and $b_3 = 6$. Find the differential equations satisfied by the probabilities $p_1(t)$, $p_2(t)$, and $p_3(t)$. Then solve for $p_1(t)$, $p_2(t)$, and $p_3(t)$.

9. Suppose in a simple death process $X(0) = 100$ and $d > 0$.

   (a) Find the expected time to reach a population size of zero.

   (b) For $d = 0.5$, use part (a) to find the expected time to reach a population size of zero.

10. Suppose in a simple birth process $X(0) = 1$ and $b > 0$.

   (a) Find the expected time to reach a population size of 100.

   (b) For $b = 0.5$, use part (a) to find the expected time to reach a population size of 100.

11. Consider the simple death process described in equations (5.27).

   (a) Derive the differential equations for the probabilities $p_i(t) = \text{Prob}\{X_t = i\}$ in the same manner as for the Poisson process to show that

$$\frac{dp_i}{dt} = d(i+1)p_{i+1}(t) - dip_i(t), \quad i < N.$$

   What is the differential equation satisfied by $p_N(t)$?

   (b) Suppose there are initially N individuals, $p_N(0) = 1$. Use the generating function technique to show that the probability generating function $P(z,t) = \sum_{i=0}^{N} p_i(t)z^i$ satisfies

$$\frac{\partial P}{\partial t} = d(1-z)\frac{\partial P}{\partial z}, \quad P(z,0) = z^N. \tag{5.30}$$

12. Assume the p.g.f. of a process satisfies equation (5.30).

   (a) Apply the method of characteristics to show that the solution of (5.30) is

$$P(z,t) = \left[1 - e^{-dt} + ze^{-dt}\right]^N.$$

   (b) Note that $P(z,t)$ is a p.g.f. of a binomial distribution, $b(n,p)$, where $n = N$, $p = e^{-dt}$, and $q = 1-p$. Find the mean $m(t)$ and the variance $\sigma^2(t)$ of the simple death process.

13. Consider the simple death process described in equations (5.27).

   (a) Use the differential equation satisfied by the p.g.f., equation (5.30), and make a change of variable $z = e^\theta$ to find a differential equation satisfied by the m.g.f., $M(\theta,t) = P(e^\theta,t)$.

   (b) Use the differential equation satisfied by the m.g.f. in part (a) to find a differential equation satisfied by the c.g.f., $K(\theta,t) = \ln M(\theta,t)$.

14. Suppose the m.g.f. $M(\theta, t)$ of a continuous time Markov process satisfies the following first-order partial differential equation:

$$\frac{\partial M}{\partial t} + \frac{e^{-\theta} - 1}{e^{-\theta}} \frac{\partial M}{\partial \theta} = 0,$$

with corresponding initial condition

$$M(\theta, 0) = e^{5\theta}.$$

Apply the method of characteristics to show that the solution $M(\theta, t)$ satisfies

$$M(\theta, t) = [1 + e^t(e^{-\theta} - 1)]^{-5}.$$

15. Write a MATLAB program for a simple death process and graph three sample paths. Assume $d = 0.25$ and $X(0) = 100$.

16. Write a MATLAB program for a simple birth and death process and graph three sample paths. Assume $b = 1$, $d = 1$, and $X(0) = 50$.

5.12 References for Chapter 5

Bailey, N. T. J. 1990. *The Elements of Stochastic Processes with Applications to the Natural Sciences.* John Wiley & Sons, New York.

Brauer, F. and J. A. Nohel. 1969. *The Qualitative Theory of Ordinary Differential Equations An Introduction.* Dover Pub., New York.

Farlow, S. J. 1982. *Partial Differential Equations for Scientists & Engineers.* John Wiley & Sons, New York.

John, F. 1975. *Partial Differential Equations.* 2nd ed. Applied Mathematical Sciences, Vol 1. Springer-Verlag, New York.

Karlin, S. and H. Taylor. 1975. *A First Course in Stochastic Processes.* 2nd ed. Academic Press, New York.

Karlin, S. and H. Taylor. 1981. *A Second Course in Stochastic Processes.* Academic Press, New York.

Leonard, I. E. 1996. The matrix exponential. *SIAM Review* 39: 507–512.

Moler, C. and C. Van Loan. 1978. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review* 20: 801–836.

Norris, J. R. 1999. *Markov Chains.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Schinazi, R. B. 1999. *Classical and Spatial Stochastic Processes.* Birkhäuser, Boston.

Schovanec, L. and D. Gilliam. 2000. Classroom notes for Ode/Pde Class. Available at http://texas.math.ttu.edu/~gilliam/ttu/ttu.htm.

Stewart, W. J. 1994. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, Princeton, N. J.

Taylor, H. M. and S. Karlin. 1998. *An Introduction to Stochastic Modeling.* 3rd ed. Academic Press, New York.

Waltman, P. 1986. *A Second Course in Elementary Differential Equations.* Academic Press, New York.

5.13 Appendix for Chapter 5

5.13.1 MATLAB Program

The following MATLAB program generates three stochastic realizations for the simple birth process.

```
% Matlab program for the simple birth process
clear
set(0,'DefaultAxesFontSize',18); % Increases axes labels.
b=1;
x=linspace(0,100,101);
y=exp(x);
n=linspace(1,100,100); % Defines the population vector.
for j=1:3;
    t(1)=0;
    for i=1:49;
       t(i+1)=t(i)-log(rand)/(b*n(i));
    end % End of i loop.
    s= stairs(t,n); % Draws stairstep graph.
    set(s,'LineWidth',2); % Thickens the line width.
    hold on % Holds current plot.
end % end of j loop
plot(x,y,'k--','LineWidth',2); % Plots the exponential.
axis([0,8,0,100]);
xlabel('Time');
ylabel('Population Size');
hold off  % Erases previous plots before drawing new ones.
```

Note: A statement following % explains the command; these statements are not executable.

Chapter 6

Continuous Time Birth and Death Chains

6.1 Introduction

In this chapter, continuous time birth and death processes are studied. In the next two sections, a general birth and death process is formulated and a necessary and sufficient condition is stated for existence of a unique positive stationary probability distribution. If the birth and death process is non-explosive and a unique positive stationary probability distribution exists, then by applying convergence results from the previous chapter, it follows that the process converges to this stationary probability distribution.

Following the general discussion on birth and death processes, some well-known birth and death processes are discussed: simple birth, simple death, simple birth and death, and simple birth and death with immigration processes. For the simple birth and simple death processes, explicit formulas are derived for the state probabilities, $p_i(t) = \text{Prob}\{X(t) = i\}$. It is shown that the probability distribution for a simple birth process is a negative binomial distribution, and for a simple death process it is a binomial distribution. For the other two processes, simple birth and death and simple birth and death with immigration, explicit formulas are derived for the moment generating functions. In addition, it is shown that the process with immigration has a unique positive stationary probability distribution provided the death rate exceeds the birth rate. An important application of birth and death processes is queueing processes. In queueing theory, the state of the system is the number of individuals in the queue. We discuss a few examples, where arrivals and departures in the system can be modeled as births and deaths.

For many birth and death processes in biology, a positive stationary probability distribution may not exist, because the zero state (extinction

state) is absorbing. For processes with an absorbing state at zero, the probability of extinction $p_0(t)$ and the mean time until population extinction are investigated. General conditions are given that allow determination of the value of the probability of population extinction as $t \to \infty$, $\lim_{t\to\infty} p_0(t)$, and the expected time until population extinction.

A stochastic logistic process is formulated and discussed. The zero state is an absorbing state for the logistic growth model. Using results derived in previous sections, it is shown that $\lim_{t\to\infty} p_0(t) = 1$ and the expected time to extinction is finite. Every state is transient except the zero state. However, if the initial population size is large and the carrying capacity is large, it may take a long time for the process to reach the zero state, and in this case, a quasistationary probability distribution can be defined. For the process conditioned on nonextinction, there exists a stationary distribution, known as the quasistationary probability distribution. Such types of probability distributions were studied in connection with the discrete time logistic process. It is shown that by approximating the probability distribution associated with the process conditioned on nonextinction, an approximation to the quasistationary probability distribution can be easily defined that has the same form as the approximate quasistationary distribution that was derived for the discrete time logistic process.

The last two sections of this chapter present two examples. The first example is an explosive birth process. The second example is a nonhomogeneous birth and death process, a process in which the transition probabilities are not stationary (nonhomogeneous).

Before we discuss the general birth and death process, some notation is introduced that will be used throughout this chapter. Let $\Delta X(t)$ denote the change in state of the stochastic process from t to $t + \Delta t$; that is,

$$\Delta X(t) = X(t + \Delta t) - X(t).$$

When birth and death processes were introduced at the end of the last chapter, the birth rate was denoted as b_i and death rate as d_i when the state of the process was $X(t) = i$. However, in this chapter, we will use notation that has become almost standard in birth and death processes. If the population size is i, then the birth and death rates will be denoted as

$$\lambda_i = \text{birth rate} \quad \text{and} \quad \mu_i = \text{death rate}$$

(see, e.g., Bailey, 1990; Karlin and Taylor, 1975; Norris, 1999; Schinazi, 1999; Taylor and Karlin, 1998).

6.2 General Birth and Death Process

The continuous time birth and death Markov chain $X(t)$ has either a finite or infinite state space $\{0, 1, 2, \ldots, N\}$ or $\{0, 1, \ldots\}$. Assume the infinitesimal

transition probabilities of a general birth and death process satisfy

$$p_{i+j,i}(\Delta t) = \text{Prob}\{\Delta X(t) = j | X(t) = i\} = \begin{cases} \lambda_i \Delta t + o(\Delta t), & j = 1 \\ \mu_i \Delta t + o(\Delta t), & j = -1 \\ 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq -1, 0, 1 \end{cases} \tag{6.1}$$

for Δt sufficiently small, where $\lambda_i \geq 0$, $\mu_i \geq 0$ for $i = 0, 1, 2, \ldots$ and $\mu_0 = 0$. It is often the case that $\lambda_0 = 0$ also, except, for example, when there is immigration. The initial conditions satisfy $p_{ji}(0) = \delta_{ji}$. If the state space is finite, then the initial transition matrix $P(0) = (p_{ji}(0))$ is just the identity matrix, $P(0) = I$. The infinitesimal transition probabilities based on Δt being small, (6.1), are similar to the type of assumptions made in defining the Poisson process and the simple birth, simple death, and simple birth and death processes that were introduced in the last chapter. In a small interval of time Δt, at most one change in state can occur, either a birth or a death. If the population size is i, and a birth occurs, then $i \to i + 1$, but if a death occurs, $i \to i - 1$.

The forward Kolmogorov differential equations for $p_{ji}(t)$ can be derived directly from the assumptions in (6.1). Assume Δt is sufficiently small and consider the transition probability $p_{ji}(t + \Delta t)$. This transition probability can be expressed in terms of the transition probabilities at time t as follows:

$$\begin{aligned} p_{ji}(t + \Delta t) &= p_{j-1,i}(t)[\lambda_{j-1}\Delta t + o(\Delta t)] + p_{j+1,i}(t)[\mu_{j+1}\Delta t + o(\Delta t)] \\ &\quad + p_{ji}(t)[1 - (\lambda_j + \mu_j)\Delta t + o(\Delta t)] + \sum_{k \neq -1,0,1}^{\infty} p_{j+k,i}(t) o(\Delta t) \\ &= p_{j-1,i}(t)\lambda_{j-1}\Delta t + p_{j+1,i}(t)\mu_{j+1}\Delta t \\ &\quad + p_{ji}(t)[1 - (\lambda_j + \mu_j)\Delta t] + o(\Delta t), \end{aligned}$$

which holds for all i and j in the state space with the exception of the endpoints, $j = 0$ and $j = N$. If $j = 0$, then

$$p_{0i}(t + \Delta t) = p_{1i}(t)\mu_1 \Delta t + p_{0i}(t)[1 - \lambda_0 \Delta t] + o(\Delta t).$$

In the case of a finite state space, where $j = N$ is the maximum population size, then

$$p_{Ni}(t + \Delta t) = p_{N-1,i}(t)\lambda_{N-1}\Delta t + p_{Ni}(t)[1 - \mu_N \Delta t] + o(\Delta t).$$

In the finite case, it is assumed that $\lambda_N = 0$ and $p_{kN}(t) = 0$ for $k > N$. In deriving these expressions, note that we determine what has happened prior to time $t + \Delta t$ given that the process will be in state j at time $t + \Delta t$. From this process, we shall see that a forward Kolmogorov equation is derived. Recall in deriving expressions for the probability of absorption a_k (first

step analysis), we determined what will happen in the next time interval given that the process is currently in state k (this process gives a backward equation).

Now the forward Kolmogorov differential equations are derived. Subtract $p_{ji}(t)$, $p_{0i}(t)$, and $p_{Ni}(t)$ from the preceding three equations, respectively. Then, dividing by Δt and taking the limit as $\Delta t \to 0$, the forward Kolmogorov differential equations are obtained for the general birth and death process,

$$\begin{aligned}\frac{dp_{ji}(t)}{dt} &= \lambda_{j-1}p_{j-1,i}(t) - (\lambda_j + \mu_j)p_{ji}(t) + \mu_{j+1}p_{j+1,i}(t)\\ \frac{dp_{0i}(t)}{dt} &= -\lambda_0 p_{0i}(t) + \mu_1 p_{1i}(t)\end{aligned}$$

for $i \geq 0$ and $j \geq 1$. For the finite state space, the differential equation for $p_{Ni}(t)$ satisfies

$$\frac{dp_{Ni}}{dt} = \lambda_{N-1}p_{N-1,i}(t) - p_{Ni}(t)\mu_N.$$

The forward Kolmogorov differential equations satisfy $dP/dt = QP$, where Q is the generator matrix. The generator matrix Q has the following form when the state space is infinite:

$$Q = \begin{pmatrix} -\lambda_0 & \mu_1 & 0 & 0 & \cdots \\ \lambda_0 & -\lambda_1 - \mu_1 & \mu_2 & 0 & \cdots \\ 0 & \lambda_1 & -\lambda_2 - \mu_2 & \mu_3 & \cdots \\ 0 & 0 & \lambda_2 & -\lambda_3 - \mu_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}, \tag{6.2}$$

and when the state space is finite,

$$Q = \begin{pmatrix} -\lambda_0 & \mu_1 & 0 & \cdots & 0 \\ \lambda_0 & -\lambda_1 - \mu_1 & \mu_2 & \cdots & 0 \\ 0 & \lambda_1 & -\lambda_2 - \mu_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu_N \\ 0 & 0 & 0 & \cdots & -\mu_N \end{pmatrix}. \tag{6.3}$$

When the initial distribution $X(0)$ is a fixed value, the state probabilities $p(t) = (p_0(t), p_1(t), \ldots)^T$, $p_i(t) = \text{Prob}\{X(t) = i\}$, satisfy the forward Kolmogorov differential equations, $dp/dt = Qp$. These differential equations can be derived in the same manner as previously:

$$\begin{aligned}p_i(t + \Delta t) &= p_{i-1}(t)\lambda_{i-1}\Delta t + p_{i+1}(t)\mu_{i+1}\Delta t \\ &\quad + p_i(t)[1 - (\lambda_i + \mu_i)\Delta t] + o(\Delta t).\end{aligned}$$

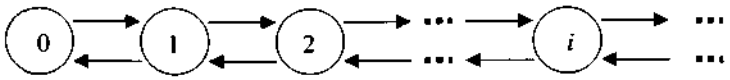

Figure 6.1. The directed graph for the embedded Markov chain of the general birth and death process when $\lambda_0 > 0$ and $\lambda_i + \mu_i > 0$ for $i = 1, 2, \ldots$.

Subtracting $p_i(t)$, dividing by Δt and letting $\Delta t \to 0$ leads to

$$\frac{dp_i}{dt} = \lambda_{i-1}p_{i-1} - (\lambda_i + \mu_i)p_i(t) + \mu_{i+1}p_{i+1}.$$

The transition matrix $T = (t_{ji})$ for the embedded Markov chain $\{Y_n\}$ is easily defined from the generator matrices (6.2) and (6.3). For the generator matrix (6.2), the transition matrix of the embedded Markov chain satisfies

$$T = \begin{pmatrix} 0 & \mu_1/(\lambda_1+\mu_1) & 0 & 0 & \cdots \\ 1 & 0 & \mu_2/(\lambda_2+\mu_2) & 0 & \cdots \\ 0 & \lambda_1/(\lambda_1+\mu_1) & 0 & \mu_3/(\lambda_3+\mu_3) & \cdots \\ 0 & 0 & \lambda_2/(\lambda_2+\mu_2) & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}.$$

For the generator matrix (6.3), the transition matrix of the embedded Markov chain satisfies

$$T = \begin{pmatrix} 0 & \mu_1/(\lambda_1+\mu_1) & 0 & \cdots & 0 \\ 1 & 0 & \mu_2/(\lambda_2+\mu_2) & \cdots & 0 \\ 0 & \lambda_1/(\lambda_1+\mu_1) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

In both cases, it is assumed that $\lambda_i + \mu_i > 0$ for $i = 0, 1, 2, \ldots$. If, for any i, $\lambda_i + \mu_i = 0$, then state i is absorbing.

The embedded Markov chain can be thought of as a generalized random walk model with a reflecting boundary at zero (and at N in the finite case). The probability of moving right (or a birth) is $t_{i+1,i} = \lambda_i/(\lambda_i + \mu_i)$ and the probability of moving left (or a death) is $t_{i-1,i} = \mu_i/(\lambda_i + \mu_i)$. See the directed graph in Figure 6.1. It can be verified easily from the transition matrix T or from the directed graph that the chain is irreducible if and only if $\lambda_i > 0$ and $\mu_{i+1} > 0$ for $i = 0, 1, 2, \ldots$. If any $\lambda_i = 0$, then $t^{(n)}_{i+1,i} = 0$ for all n, and if any $\mu_i = 0$, then $t^{(n)}_{i-1,i} = 0$ for all n.

6.3 Stationary Probability Distribution

In this section, a positive stationary probability distribution is defined for a general birth and death chain. Recall that a stationary probability dis-

tribution $\pi = (\pi_0, \pi_1, \pi_2, \ldots)^T$ of a continuous time Markov chain with generator matrix Q satisfies

$$Q\pi = 0, \quad \sum_{i=0}^{\infty} \pi_i = 1, \quad \text{and} \quad \pi_i \geq 0$$

for $i = 0, 1, 2, \ldots$. The stationary probability distribution also satisfies $P(t)\pi = \pi$ for $t \geq 0$.

For birth and death chains, there is an iterative procedure for computing the stationary probability distribution. The following theorem gives a formula for the stationary probability distribution when the state space is finite or infinite.

Theorem 6.1. *Suppose the continuous time Markov chain $\{X(t)\}$, $t \geq 0$, is a general birth and death chain satisfying (6.1). If the state space is infinite, $\{0, 1, 2, \ldots\}$, a unique positive stationary probability distribution π exists iff*

$$\mu_i > 0 \quad \text{and} \quad \lambda_{i-1} > 0 \quad \text{for} \quad i = 1, 2, \ldots,$$

and

$$\sum_{i=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} < \infty.$$

*The **stationary** probability distribution satisfies*

$$\pi_i = \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} \pi_0, \quad i = 1, 2, \ldots \tag{6.4}$$

and

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i}}. \tag{6.5}$$

If the state space is finite, $\{0, 1, 2, \ldots, N\}$, then a unique positive stationary probability distribution π exists iff

$$\mu_i > 0 \quad \text{and} \quad \lambda_{i-1} > 0 \quad \text{for} \quad i = 1, 2, \ldots, N.$$

The stationary probability distribution is given by (6.4) and (6.5), where the index i and the summation on i extend from 1 to N.

Proof. The stationary probability distribution π satisfies $Q\pi = 0$ or

$$0 = \lambda_{i-1}\pi_{i-1} - (\lambda_i + \mu_i)\pi_i + \mu_{i+1}\pi_{i+1} \quad i = 1, 2, \ldots,$$

$$0 = -\lambda_0 \pi_0 + \mu_1 \pi_1,$$

and $\sum_{i=0}^{\infty} \pi_i = 1$. These equations can be solved recursively. For π_1,

$$\pi_1 = \frac{\lambda_0}{\mu_1} \pi_0.$$

Then π_2 satisfies

$$\begin{aligned}\mu_2\pi_2 &= (\lambda_1+\mu_1)\pi_1-\lambda_0\pi_0\\ &= \left[\frac{(\lambda_1+\mu_1)\lambda_0}{\mu_1}-\lambda_0\right]\pi_0\\ \pi_2 &= \frac{\lambda_0\lambda_1}{\mu_1\mu_2}\pi_0.\end{aligned}$$

The general formula for π_i can be proved by induction. Assume π_j has been defined for $j=1,2,\ldots,i$,

$$\pi_i=\frac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_i}\pi_0.$$

Then

$$\begin{aligned}\mu_{i+1}\pi_{i+1} &= (\lambda_i+\mu_i)\pi_i-\lambda_{i-1}\pi_{i-1}\\ &= \left[\frac{\lambda_0\lambda_2\cdots\lambda_{i-1}(\lambda_i+\mu_i)}{\mu_1\mu_2\cdots\mu_i}-\frac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_{i-1}}\right]\pi_0\\ &= \frac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_{i-1}}\left[\frac{\lambda_i+\mu_i}{\mu_i}-1\right]\pi_0\\ \pi_{i+1} &= \frac{\lambda_0\lambda_1\cdots\lambda_i}{\mu_1\mu_2\cdots\mu_{i+1}}\pi_0.\end{aligned}$$

It follows that π_j is defined for $j=1,2,\ldots$. Applying the additional constraint, $\sum_{i=0}^{\infty}\pi_i=1$ or $\pi_0(1+\sum_{i=1}^{\infty}\pi_i/\pi_0)=1$, it follows that

$$\pi_0=\frac{1}{1+\sum_{i=1}^{\infty}\dfrac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_i}}.$$

A unique positive stationary distribution exists if and only if the following summation is positive and finite:

$$0<\sum_{i=1}^{\infty}\frac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_i}<\infty. \tag{6.6}$$

If the continuous time Markov chain is finite, then the summation in (6.6) is automatically finite, so that a unique positive stationary distribution is given by the preceding formulas for π_i, $i=0,1,2,\ldots,N$, where the summation is from $i=1$ to N. □

Note that if $\lambda_i=0$ for some i and $\mu_i>0$ for $i\geq 1$, a stationary distribution still exists but it is not positive. If $\lambda_0=0$ and $\mu_i>0$ for $i\geq 1$, then $\pi_0=1$ and $\pi_i=0$ for $i\geq 1$.

Example 6.1 Suppose a continuous time birth and death Markov chain satisfies $\lambda_i = b$ and $\mu_i = id$ for $i = 0, 1, 2, \ldots$. Then, applying Theorem 6.1,

$$\frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 u_2 \cdots \mu_i} = \frac{b^i}{d^i i!} = \frac{(b/d)^i}{i!}$$

and

$$1 + \sum_{i=1}^{\infty} \frac{(b/d)^i}{i!} = e^{b/d}.$$

There exists **a unique** stationary probability distribution satisfying $\pi_0 = e^{-b/d}$ and

$$\pi_i = \frac{(b/d)^i}{i!} e^{-b/d}.$$

This stationary probability distribution is a Poisson **probability distribution** with parameter b/d. ■

Example 6.2 Suppose a continuous time birth and death Markov chain satisfies $\mu_i = q > 0$, $i = 1, 2, \ldots$, and $\lambda_i = p > 0$, $i = 0, 1, 2, \ldots$, where $p + q = 1$. The embedded Markov chain is a semi-infinite random walk model with reflecting boundary conditions at zero. The transition matrix for the embedded Markov chain has a directed graph given by Figure 6.1. The value of $q = t_{i-1,i}$ and $p = t_{i+1,i}$. The chain has a unique stationary probability distribution iff

$$\sum_{j=1}^{\infty} \left(\frac{p}{q}\right)^j < \infty$$

iff $p < q$. The **stationary** probability distribution **is a geometric probability** distribution,

$$\pi_0 = \frac{q - p}{q} \quad \text{and} \quad \pi_i = \left(\frac{p}{q}\right)^i \pi_0$$

or

$$\pi_i = \left(1 - \frac{p}{q}\right)\left(\frac{p}{q}\right)^i, \quad i = 0, 1, 2, \ldots.$$ ■

6.4 Simple Birth and Death Processes

In the next four subsections, some classical continuous time birth and death processes are described: simple birth, simple death, simple birth and death, and simple birth and death with immigration.

6.4.1 Simple Birth Process

Let $\{X(t)\}$ for $t \geq 0$ be a continuous time Markov chain, where the random variable $X(t)$ denotes the total population size at time t. Let the initial population size be N, $X(0) = N$, so that $p_i(0) = \delta_{iN}$. In the simple birth process, it is assumed that the only event is a birth. For Δt sufficiently small, the transition probabilities satisfy

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} \lambda i \Delta t + o(\Delta t), & j = 1 \\ 1 - \lambda i \Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \geq 2 \\ 0, & j < 0. \end{cases} \end{aligned}$$

Because there are only births, the population size increases in size by one or stays the same size during a small increment of time Δt; it cannot decrease. This simple birth process is also referred to as a *pure* birth process (Bailey, 1990).

The state probabilities $p_i(t) = \text{Prob}\{X(t) = i\}$ also satisfy the forward Kolmogorov differential equations, $dp/dt = Qp$:

$$\begin{aligned} \frac{dp_i(t)}{dt} &= \lambda(i-1)p_{i-1}(t) - \lambda i p_i(t), \quad i = N+1, \ldots, \\ \frac{dp_i(t)}{dt} &= 0, \quad i = 0, 1, \ldots, N-1, \end{aligned}$$

with initial condition $p_i(0) = \delta_{iN}$. These differential equations can be solved in a sequential fashion as was done for the Poisson process in Chapter 5. For example, note that $p_i(t) = 0$ for $i < N$; then $dp_N/dt = -\lambda N p_N$ so that the solution is $p_N(t) = e^{-\lambda N t}$. However, the generating function technique will be used to find the solutions $p_i(t)$. This technique has wider applicability and is sometimes simpler to apply than solving the differential equations sequentially.

Before applying the generating function technique, we make some observations about the simple birth process. Because the only event is a birth and $X(0) = N$, the state space for this process is $\{N, N+1, \ldots\}$. In addition, it can be shown from the generator matrix and the transition matrix of the embedded Markov chain (with state space $\{N, N+1, \ldots\}$) that all of the states are transient and there is no stationary probability distribution (see the Exercises). It will be shown that the p.g.f. and the m.g.f. for the simple birth process correspond to a negative binomial distribution.

To derive the partial differential equation for the p.g.f., multiply the

differential equations by z^i and sum over i. Then

$$\begin{aligned}\frac{\partial P(z,t)}{\partial t} &= \lambda \sum_{i=N+1}^{\infty} p_{i-1}(i-1)z^i - \lambda \sum_{i=N}^{\infty} p_i i z^i \\ &= \lambda z^2 \sum_{i=N}^{\infty} i p_i z^{i-1} - \lambda z \sum_{i=N}^{\infty} i p_i z^{i-1} \\ &= \lambda z(z-1)\frac{\partial P}{\partial z},\end{aligned}$$

because the terms on the right-hand side are the partial derivatives of P with respect to z. The initial condition is $P(z,0) = z^N$.

The partial differential equation for the m.g.f. can be derived by a change of variable. Let $z = e^\theta$. Then $P(e^\theta, t) = M(\theta, t)$. Because

$$\frac{\partial P}{\partial z} = \frac{\partial M}{\partial \theta}\frac{d\theta}{dz} = \frac{1}{z}\frac{\partial M}{\partial \theta},$$

the m.g.f satisfies the following partial differential equation:

$$\frac{\partial M}{\partial t} = \lambda(e^\theta - 1)\frac{\partial M}{\partial \theta},$$

with corresponding initial condition, $M(\theta, 0) = e^{N\theta}$.

The solutions $P(z,t)$ and $M(\theta,t)$ to the first-order partial differential equations are found by applying the method of characteristics. We demonstrate this method by solving the partial differential equation for $M(\theta,t)$. Rewrite the differential equation for $M(\theta,t)$ as follows:

$$\frac{\partial M}{\partial t} + \lambda(1 - e^\theta)\frac{\partial M}{\partial \theta} = 0.$$

Along characteristic curves, s and τ, $t(s,\tau)$, $\theta(s,\tau)$, and $M(s,\tau)$, and, in addition,

$$\frac{dt}{d\tau} = 1, \quad \frac{d\theta}{d\tau} = \lambda(1 - e^\theta), \quad \text{and} \quad \frac{dM}{d\tau} = 0,$$

with initial conditions

$$t(s,0) = 0, \quad \theta(s,0) = s, \quad \text{and} \quad M(s,0) = e^{Ns}.$$

Separating variables and simplifying leads to

$$\frac{d\theta}{1 - e^\theta} = \lambda d\tau \quad \text{or} \quad \frac{e^{-\theta}d\theta}{e^{-\theta} - 1} = \lambda d\tau.$$

Integrating yields the relations

$$t = \tau, \quad \ln(e^{-\theta} - 1) = -\lambda\tau + c, \quad \text{and} \quad M(s,\tau) = e^{Ns}.$$

Applying the initial condition $\theta(s,0) = s$, the middle expression can be written as

$$e^{-\theta} - 1 = (e^{-s} - 1)e^{-\lambda\tau}.$$

The solution M must be expressed in terms of θ and t. Using the preceding formulas, e^{-s} can be expressed in terms of θ and t, $e^{-s} = 1 - e^{\lambda t}(1 - e^{-\theta})$. Since $e^{Ns} = [e^{-s}]^{-N}$, the m.g.f. for the simple birth process, expressed in terms of θ and t, is

$$M(\theta,t) = [1 - e^{\lambda t}(1 - e^{-\theta})]^{-N}.$$

The p.g.f can be found directly from the m.g.f. by making the change of variable, $\theta = \ln z$,

$$\begin{aligned} P(z,t) &= [1 - e^{\lambda t}(1 - z^{-1})]^{-N} \\ &= z^N e^{-\lambda N t}[ze^{-\lambda t} - (z-1)]^{-N} \\ &= \frac{z^N e^{-\lambda N t}}{[1 - z(1 - e^{-\lambda t})]^N}. \end{aligned}$$

Hence,

$$P(z,t) = \frac{(pz)^N}{[1 - zq]^N}, \tag{6.7}$$

where $p = e^{-\lambda t}$ and $q = 1 - e^{-\lambda t}$. From this latter expression (6.7) for the p.g.f. it can be seen that the p.g.f. corresponds to a negative binomial distribution (see Chapter 1).

The probabilities $p_i(t)$ from a negative binomial distribution with p.g.f. given by (6.7) satisfy

$$p_{i+N}(t) = \binom{N+i-1}{i} p^N q^i, \ i = 0, 1, \ldots.$$

Let $i + N = n$ and replace p and q by $e^{-\lambda t}$ and $1 - e^{-\lambda t}$, respectively,

$$p_n(t) = \binom{n-1}{n-N} e^{-\lambda N t}(1 - e^{-\lambda t})^{n-N}, \ n = N, N+1, \ldots.$$

Notice that $p_N(t) = e^{-\lambda N t}$, which agrees with the solution computed directly from the forward Kolmogorov differential equations, $dp_N/dt = -\lambda N p_N$.

The mean and variance for the simple birth process can be obtained directly from the formulas for mean and variance of a negative binomial distribution. They are

$$m(t) = N/p = Ne^{\lambda t} \ \text{ and } \ \sigma^2(t) = Nq/p^2 = Ne^{2\lambda t}(1 - e^{-\lambda t}).$$

These moments can also be calculated directly from one of the generating functions.

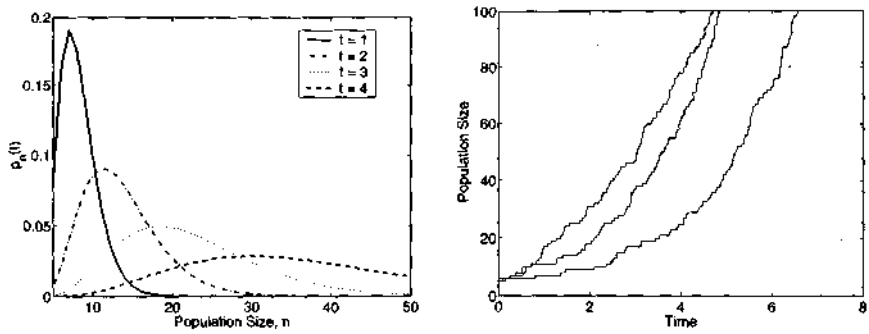

Figure 6.2. Probability distributions for the simple birth process $X(1)$, $X(2)$, $X(3)$, and $X(4)$, and three sample paths when $\lambda = 0.5$ and $X(0) = 5$.

The mean of the simple birth process corresponds to the solution of a deterministic exponential growth model with $X(0) = N$. However, for large t it does not follow the exponential model very closely because the variance increases exponentially with t.

Example 6.3 Suppose $\lambda = 0.5$ and $N = 5$. Then $m(t) = 5e^{0.5t}$, $\sigma^2(t) = 5(e^t - e^{0.5t})$, and

$$p_n(t) = \frac{(n-1)!}{(n-5)!4!} e^{-2.5t}(1 - e^{-0.5t})^{n-5}, \ n = 5, 6, \ldots.$$

The probability distributions for the simple birth process are graphed in Figure 6.2 for times $t = 1, 2, 3, 4$ and, in addition, three sample paths are graphed for $t \in [0, 8]$. ■

6.4.2 Simple Death Process

Let $X(t)$ for $t \geq 0$ be a continuous time Markov chain, where the random variable $X(t)$ denotes the total population size at time t. Let the initial population size be N, $X(0) = N$, $p_i(0) = \delta_{iN}$. In the simple death process, it is assumed that the only event is a death. For Δt sufficiently small, the transition probabilities satisfy

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} \mu i \Delta t + o(\Delta t), & j = -1 \\ 1 - \mu i \Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \leq -2 \\ 0, & j > 0. \end{cases} \end{aligned}$$

In a sufficiently small period of time Δt, either the population size decreases by 1 or stays the same size. Since the process starts in state N, the state space is given by $\{0, 1, 2, \ldots, N\}$.

The transition probabilities lead to the forward Kolmogorov equations, $dP/dt = QP$, and the state probabilities also satisfy this system of differential equations, $dp/dt = Qp$:

$$\begin{aligned} \frac{dp_i(t)}{dt} &= \mu(i+1)p_{i+1}(t) - \mu i p_i(t) \\ \frac{dp_N(t)}{dt} &= -\mu N p_N(t), \end{aligned}$$

for $i = 0, 1, 2, \ldots, N-1$ with initial conditions $p_i(0) = \delta_{iN}$. From the generator matrix Q and the transition matrix of the embedded Markov chain, it can be shown easily that zero is an absorbing state and the unique stationary probability distribution is $\pi = (1, 0, 0, \ldots, 0)$.

The probabilities $p_i(t)$ can be calculated sequentially from the forward Kolmogorov differential equations. For example, it is easy to see that $p_N(t) = e^{-\mu N t}$. However, the generating function method is applied. It is shown that the probability distribution for the simple death process is a binomial distribution.

Multiplying the forward Kolmogorov equations by z^i and summing over i, the partial differential equation for the p.g.f. is

$$\frac{\partial P}{\partial t} = \mu(1-z)\frac{\partial P}{\partial z}, \quad P(z,0) = z^N.$$

Substituting $z = e^\theta$, the partial differential for the m.g.f. is

$$\frac{\partial M}{\partial t} = \mu(e^{-\theta} - 1)\frac{\partial M}{\partial \theta}, \quad M(\theta, 0) = e^{N\theta}.$$

The solutions to these equations can be found by the method of characteristics. They yield the following solutions for the p.g.f. and m.g.f.:

$$P(z,t) = [1 - e^{-\mu t}(1-z)]^N = [1 - e^{-\mu t} + e^{-\mu t}z]^N$$

and

$$M(\theta, t) = [1 - e^{-\mu t}(1 - e^\theta)]^N.$$

Let $p = e^{-\mu t}$ and $q = 1 - e^{-\mu t}$. Then the p.g.f. has the form $P(z,t) = (q + pz)^N$, which is the p.g.f. for a binomial distribution, $b(N,p)$. The probabilities, $p_i(t)$, can now be defined (see Chapter 1):

$$p_i(t) = \binom{N}{i} p^i q^{N-i} = \binom{N}{i} e^{-i\mu t}(1 - e^{-\mu t})^{N-i}$$

for $i = 0, 1, \ldots, N$. The mean and variance of binomial distribution $b(N,p)$ are $m = Np$ and $\sigma^2 = Npq$, but expressed in terms of the original variables they are

$$m(t) = Ne^{-\mu t}, \quad \sigma^2(t) = Ne^{-\mu t}(1 - e^{-\mu t}).$$

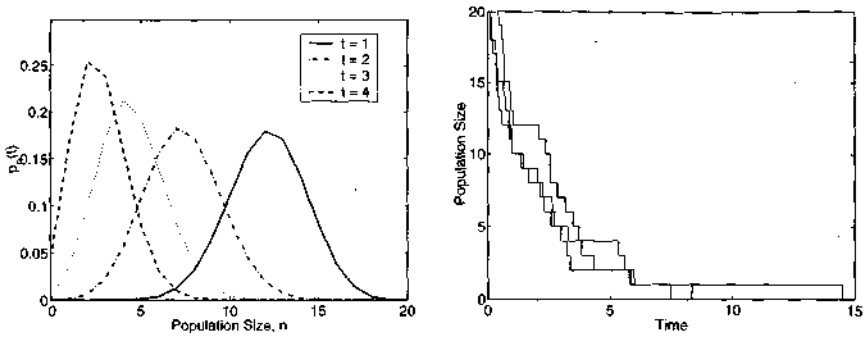

Figure 6.3. Probability distributions for the simple death process $X(1)$, $X(2)$, $X(3)$, and $X(4)$ and three sample paths when $\mu = 0.5$ and $X(0) = 20$.

Notice that the mean of the distribution corresponds to the solution of a deterministic model with exponential decay. The variance decreases exponentially with time. The differences in the variances between the simple birth and simple death models can be seen in the sample paths graphed in Figures 6.2 and 6.3.

Example 6.4 Consider the simple death process with $\mu = 0.5$ and $N = 20$. The mean and variance satisfy

$$m(t) = 20e^{-0.5t} \quad \text{and} \quad \sigma^2(t) = 20(e^{-.5t} - e^{-t}),$$

respectively. Probability distributions for the simple death process are graphed in Figure 6.3 at times $t = 1, 2, 3, 4$. It is evident as time increases that the mean and variance of the distributions decrease. ■

6.4.3 Simple Birth and Death Process

Let $\{X(t)\}$ for $t \geq 0$ be a continuous time Markov chain with $X(t)$ the random variable for the total population size at time t. Let the initial population size be N, $X(0) = N$, $p_i(0) = \delta_{iN}$. In the simple birth and death process, an event can be a birth or a death. For Δt sufficiently small, the transition probabilities satisfy

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} \mu i \Delta t + o(\Delta t), & j = -1 \\ \lambda i \Delta t + o(\Delta t), & j = 1 \\ 1 - (\lambda + \mu) i \Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq -1,\ 0,\ 1. \end{cases} \end{aligned}$$

The state probabilities satisfy the forward **Kolmogorov differential** equations, $dp/dt = Qp$:

$$\begin{aligned}\frac{dp_i(t)}{dt} &= \lambda(i-1)p_{i-1}(t) + \mu(i+1)p_{i+1}(t) - (\lambda+\mu)ip_i(t)\\ \frac{dp_0(t)}{dt} &= \mu p_1(t)\end{aligned}$$

for $i = 1, 2, \ldots$ with initial conditions $p_i(0) = \delta_{iN}$. Since $\lambda_0 = 0$, zero is an absorbing state. It can be shown that $\pi = (1, 0, 0, \ldots)^T$ is the unique stationary probability distribution.

Using the generating function technique, first-order partial differential equations for the m.g.f. and the p.g.f. can be derived. The p.g.f. satisfies

$$\frac{\partial P}{\partial t} = [\mu(1-z) + \lambda z(z-1)]\frac{\partial P}{\partial z}, \quad P(z,0) = z^N$$

and the m.g.f. satisfies

$$\frac{\partial M}{\partial t} = [\mu(e^{-\theta}-1) + \lambda(e^{\theta}-1)]\frac{\partial M}{\partial \theta}, \quad M(\theta,0) = e^{\theta N}.$$

Application of the method of characteristics to the m.g.f. equation leads to

$$\frac{dt}{d\tau} = 1, \quad \frac{d\theta}{\mu(e^{-\theta}-1) + \lambda(e^{\theta}-1)} = -d\tau, \quad \text{and} \quad \frac{dM}{d\tau} = 0,$$

with initial conditions

$$t(s,0) = 0, \quad \theta(s,0) = s, \quad \text{and} \quad M(s,0) = e^{sN}.$$

We omit the mathematical details for solving this equation (please consult the Appendix for Chapter 6). Two cases must be considered $\lambda = \mu$ and $\lambda \neq \mu$. The m.g.f. M is

$$M(\theta,t) = \begin{cases} \left(\dfrac{e^{t(\mu-\lambda)}(\lambda e^{\theta}-\mu) - \mu(e^{\theta}-1)}{e^{t(\mu-\lambda)}(\lambda e^{\theta}-\mu) - \lambda(e^{\theta}-1)}\right)^N, & \text{if } \lambda \neq \mu \\ \left(\dfrac{1-(\lambda t-1)(e^{\theta}-1)}{1-\lambda t(e^{\theta}-1)}\right)^N, & \text{if } \lambda = \mu. \end{cases}$$

Making the change of variable $\theta = \ln z$, the p.g.f. P is

$$P(z,t) = \begin{cases} \left(\dfrac{e^{t(\mu-\lambda)}(\lambda z-\mu) - \mu(z-1)}{e^{t(\mu-\lambda)}(\lambda z-\mu) - \lambda(z-1)}\right)^N, & \text{if } \lambda \neq \mu \\ \left(\dfrac{1-(\lambda t-1)(z-1)}{1-\lambda t(z-1)}\right)^N, & \text{if } \lambda = \mu. \end{cases}$$

Obtaining formulas for the probabilities $p_i(t)$ is not as straightforward as it was for the simple birth and simple death processes because the p.g.f. cannot be associated with a well-known probability distribution. However, recall that

$$P(z,t) = \sum_{i=0}^{\infty} p_i(t) z^i \quad \text{and} \quad p_i(t) = \frac{1}{i!} \frac{\partial^i P}{\partial z^i}\bigg|_{z=0}.$$

A computer algebra system may be helpful in finding the terms in the series expansion. The first term in the expansion of $P(z,t)$ is $p_0(t) = P(0,t)$:

$$p_0(t) = \begin{cases} \left(\dfrac{\mu - \mu e^{(\mu-\lambda)t}}{\lambda - \mu e^{(\mu-\lambda)t}}\right)^N, & \text{if } \lambda \neq \mu \\ \left(\dfrac{\lambda t}{1+\lambda t}\right)^N, & \text{if } \lambda = \mu. \end{cases}$$

The mean and variance of the simple birth and death process can be derived from either the p.g.f. or the m.g.f. For $\lambda \neq \mu$,

$$m(t) = Ne^{(\lambda-\mu)t} \quad \text{and} \quad \sigma^2(t) = N\frac{(\lambda+\mu)}{(\lambda-\mu)} e^{(\lambda-\mu)t}(e^{(\lambda-\mu)t} - 1).$$

The mean corresponds to an exponential growth model if $\lambda > \mu$ and an exponential decay model if $\lambda < \mu$. For the case $\lambda = \mu$, the mean and variance satisfy

$$m(t) = N \quad \text{and} \quad \sigma^2(t) = 2N\lambda t.$$

The probability of extinction, $p_0(t)$, has a simple expression when $t \to \infty$. Taking the limit,

$$p_0(\infty) = \lim_{t\to\infty} p_0(t) = \begin{cases} 1, & \text{if } \lambda \leq \mu \\ \left(\dfrac{\mu}{\lambda}\right)^N, & \text{if } \lambda > \mu \end{cases} \tag{6.8}$$

This latter result is reminiscent of a semi-infinite random walk with an absorbing barrier at $x = 0$–that is, the gambler's ruin problem, where the probability of losing a game is μ and the probability of winning a game is λ. When the probability of losing (death) is greater than or equal to the probability of winning (birth), then, in the long run ($t \to \infty$), the probability of losing all of the initial capital N (probability of absorption) approaches 1. However, if the probability of winning is greater than the probability of losing, then, in the long run, the probability of losing all of the initial capital is $(\mu/\lambda)^N$.

Three sample paths for the simple birth and death process are graphed in Figure 6.4 for parameter values $\lambda = 1 = \mu$ and initial population size $X(0) = 50$.

The mean, variance, and p.g.f. for the simple birth, simple death and simple birth and death processes are summarized in Table 6.1.

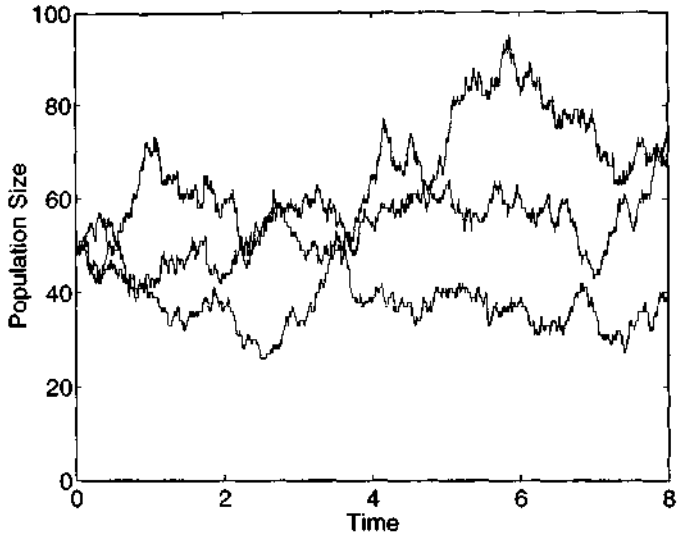

Figure 6.4. Three sample paths for the simple birth and death process when $\lambda = 1 = \mu$ and $X(0) = 50$.

	Simple Birth	Simple Death	Simple Birth and Death
$m(t)$	$Ne^{\lambda t}$	$Ne^{-\mu t}$	$Ne^{(\lambda-\mu)t}$
$\sigma 2(t)$	$Ne^{2\lambda t}(1-e^{-\lambda t})$	$Ne^{-\mu t}(1-e^{-\mu t})$	$N\frac{\lambda+\mu}{\lambda-\mu}\rho(\rho-1)$
$P(z,t)$	$\frac{(pz)^N}{(1-z(1-p))^N}$ Negative binomial $p = e^{-\lambda t}$	$[1-p+pz]^N$ Binomial $b(N,p)$ $p = e^{-\mu t}$	$\left(\frac{\rho^{-1}(\lambda z-\mu)-\mu(z-1)}{\rho^{-1}(\lambda z-\mu)-\lambda(z-1)}\right)^N$

Table 6.1. The mean, variance, and p.g.f. for the simple birth, simple death, and simple birth and death processes, where $X(0) = N$ and $\rho = e^{(\lambda-\mu)t}$, $\lambda \neq \mu$

6.4.4 Simple Birth and Death Process with Immigration

Suppose immigration is included in the simple birth and death process at a constant rate ν. For Δt sufficiently small, the transition probabilities for a simple birth and death process with immigration satisfy

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} \mu i \Delta t + o(\Delta t), & j = -1 \\ (\nu + \lambda i)\Delta t + o(\Delta t), & j = 1 \\ 1 - [\nu + (\lambda + \mu)i]\,\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq -1,\ 0,\ 1. \end{cases} \end{aligned}$$

Due to the immigration, $\lambda_0 = \nu > 0$.

Let $X(0) = N$. Then the state probabilities $p_i(t)$ satisfy the forward Kolmogorov equations, $dp/dt = Qp$:

$$\begin{aligned} \frac{dp_i}{dt} &= [\lambda(i-1) + \nu]p_{i-1} + \mu(i+1)p_{i+1} - [\lambda i + \mu i + \nu]p_i \\ \frac{dp_0}{dt} &= -\nu p_0 + \mu p_1, \end{aligned}$$

for $i = 1, 2, \ldots$ with initial conditions $p_i(0) = \delta_{iN}$. This system of differential equations cannot be solved sequentially. Therefore, we employ the generating function technique.

The m.g.f. $M(\theta, t)$ satisfies

$$\frac{\partial M}{\partial t} = \left[\lambda(e^{\theta} - 1) + \mu(e^{-\theta} - 1)\right] \frac{\partial M}{\partial \theta} + \nu(e^{\theta} - 1)M, \quad M(\theta, 0) = e^{N\theta}.$$

Note that the form of this equation is the same as for the birth and death process if $\nu = 0$. We apply the method of characteristics to solve for $M(\theta, t)$,

$$\frac{dt}{d\tau} = 1, \quad \frac{d\theta}{d\tau} = \lambda(1 - e^{\theta}) + \mu(1 - e^{-\theta}), \quad \text{and} \quad \frac{dM}{d\tau} = \nu(e^{\theta} - 1)M.$$

The initial conditions satisfy

$$t(s, 0) = 0, \quad \theta(s, 0) = s, \quad \text{and} \quad M(s, 0) = e^{N\theta}.$$

Note that the solution for θ is the same as for the simple death process, and M can be solved in terms of θ (separation of variables):

$$\frac{dM}{d\theta} = \frac{\nu(e^{\theta} - 1)M}{\lambda(1 - e^{\theta}) + \mu(1 - e^{-\theta})},$$

which leads to (Bailey, 1990):

$$M(\theta, t) = \frac{(\lambda - \mu)^{\nu/\lambda} \left[\mu(e^{(\lambda-\mu)t} - 1) - e^{\theta}(\mu e^{(\lambda-\mu)t} - \lambda)\right]^N}{\left[(\lambda e^{(\lambda-\mu)t} - \mu) - \lambda(e^{(\lambda-\mu)t} - 1)e^{\theta}\right]^{N+\nu/\lambda}}.$$

The moments of the probability distribution $X(t)$ can be computed from the m.g.f. A computer algebra system was used to calculate the mean $(m(t) = \partial M/\partial\theta|_{\theta=0})$,

$$m(t) = \frac{e^{(\lambda-\mu)t}[N\mu - N\lambda - \nu] + \nu}{\mu - \lambda} \quad \text{if} \quad \lambda \neq \mu. \tag{6.9}$$

For the case $\lambda = \mu$, l'Hôpital's rule is applied. Let $u = \lambda - \mu$. Then

$$m(t) = \lim_{u\to 0} \frac{e^{ut}(Nu + \nu) - \nu}{u} = \nu t + N. \tag{6.10}$$

The mean increases exponentially in t when $\lambda > \mu$ and linearly when $\lambda = \mu$. However, in the case $\lambda < \mu$, the mean approaches a constant:

$$m(\infty) = \frac{\nu}{\mu - \lambda}, \quad \lambda < \mu.$$

Thus, for the case $\lambda < \mu$, we should expect that as time progress the process settles down to a stationary probability distribution with mean $\nu/(\mu - \lambda)$. The stationary probability distribution is independent of the initial distribution and is the limiting distribution.

Theorem 6.1 can be used to determine conditions for the existence of a positive stationary distribution. A stationary probability distribution exists for the birth, death, and immigration process if

$$\sum_{i=1}^{\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_i} = \sum_{i=1}^{\infty} \frac{\nu(\nu+\lambda)\cdots(\nu+(i-1)\lambda)}{i!\mu^i} < \infty.$$

It can be verified easily that the birth and death process with immigration satisfies the preceding condition. There exists a constant $c > 0$ such that $\nu = c\lambda$. Using this identity for ν, the preceding summation simplifies to

$$\sum_{i=1}^{\infty} \frac{c(c+1)\cdots(c+i-1)}{i!} \left(\frac{\lambda}{\mu}\right)^i. \tag{6.11}$$

It is left as an exercise to show that the series converges iff $\lambda < \mu$. Thus, a unique positive stationary distribution exists iff $\lambda < \mu$. In addition, because $\mu_{i+1} > 0$ and $\lambda_i > 0$ for $i = 0, 1, 2, \ldots$, the stationary distribution is the limiting distribution.

Example 6.5 Suppose $c = 1$, $\lambda = 0.5 = \nu$, and $\mu = 1$ so that $\lambda_i = \lambda(c + i) = 0.5(i + 1)$ and $\mu_i = \mu i = i$. To find the stationary probability distribution π, apply Theorem 6.1:

$$\pi_{i+1} = \frac{\lambda_i}{\mu_{i+1}}\pi_i, \quad \pi_i = \frac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_i}\pi_0, \quad \text{and} \quad \sum_{i=0}^{\infty}\pi_i = 1.$$

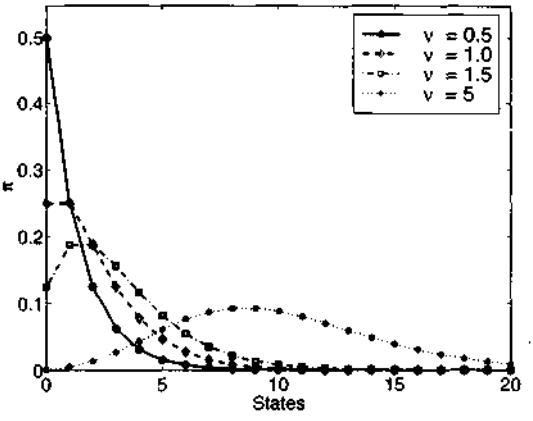

Figure 6.5. Stationary probability distributions for the birth, death and immigration process with $\lambda = 0.5$, $\mu = 1$, and $\nu = 0.5$, 1.0, 1.5, or 5.

Thus,

$$\pi_{i+1} = \frac{0.5(i+1)}{i+1}\pi_i = 0.5\pi_i \quad \text{and} \quad \pi_i = (0.5)^i\pi_0.$$

Also,

$$\sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^i = 2\pi_0.$$

This implies $\pi_0 = 1/2$. The stationary distribution is given by

$$\pi_i = \left(\frac{1}{2}\right)^{i+1}, \quad i = 0, 1, 2, \ldots.$$

The mean of the stationary distribution is $m = \sum_{i=1}^{\infty} i/2^{i+1}$. It can be shown that the mean equals 1. This follows from calculating the derivative of $(1-x)^{-1} = \sum_{i=0}^{\infty} x^i$, which is absolutely convergent for $|x| < 1$. The first and second derivatives satisfy

$$(1-x)^{-2} = \sum_{i=1}^{\infty} ix^{i-1} \quad \text{and} \quad 2(1-x)^{-3} = \sum_{i=2}^{\infty} i(i-1)x^{i-2} \tag{6.12}$$

for $|x| < 1$. Substitution of $x = 1/2$ in the first derivative and multiplying by $(1/2)^2$ yields $m = (1/2)^2(1-1/2)^{-2} = 1$. The second derivative can be used to find the variance, $\sigma^2 = 2$. The value of the mean agrees with the formula $m = \nu/(\mu - \lambda)$, which was derived previously. The stationary probability distribution for this example is graphed in Figure 6.5. ■

When the stationary probability distribution exists, $\lambda < \mu$, it is also the limiting distribution. The p.g.f. for this stationary distribution π can be found by taking the limit of the p.g.f.,

$$P(z,\infty) = \lim_{t\to\infty} P(z,t) = \lim_{t\to\infty} \sum_{i=0}^{\infty} p_i(t) z^i = \sum_{i=0}^{\infty} \pi_i z^n.$$

Let $z = e^\theta$ in $M(\theta, t)$ and let $t \to \infty$. It follows that

$$P(z,\infty) = \left(\frac{\lambda-\mu}{\lambda z-\mu}\right)^{\nu/\lambda} = \left(\frac{1-\dfrac{\lambda}{\mu}}{1-\dfrac{\lambda}{\mu}z}\right)^{\nu/\lambda}.$$

This is a p.g.f. of a negative binomial distribution provided ν/λ is a positive integer. The parameters are $p = 1 - \lambda/\mu$, $q = 1 - p = \lambda/\mu$, and $n = \nu/\lambda$. The mean, variance, and probability distribution of a negative binomial distribution satisfy

$$m = \frac{nq}{p} = \frac{\nu}{\mu-\lambda}, \quad \text{and} \quad \sigma^2 = \frac{nq}{p^2} = \frac{\nu\mu}{(\mu-\lambda)^2},$$

and

$$\pi_i = \binom{x+n-1}{n-1} p^n (1-p)^i, \quad i = 0, 1, 2, \ldots.$$

For the parameter values in Example 6.5, where $\nu = 0.5$ and $\lambda = 0.5$, $n = \nu/\lambda = 1$, the stationary probability distribution is a negative binomial. We can verify that the mean and variance satisfy the formulas given previously.

Figure 6.5 graphs the stationary probability distributions for four different sets of parameter values. In each case, the distributions are negative binomial. The four graphs in Figure 6.5 have parameter values $\lambda = 0.5$, $\mu = 1$ and the immigration rates are $\nu = 0.5, 1, 1.5$, and 5. The mean values for the stationary probability distributions corresponding to each of these values of ν are $m = 1, 2, 3$, and 10, respectively. Four sample paths for each set of parameter values are graphed in Figure 6.6. It can be seen at about time $t = 25$ that the sample paths appear to vary around the respective mean values of the stationary probability distribution.

In practice, we may not be able to find closed form solutions for the stationary distribution. It may be necessary to approximate them. Because $\pi_i \to 0$ as $i \to \infty$, for sufficiently large i, $\pi_i \approx 0$. In Example 6.5, the first 15 terms provide a good estimate of the stationary distribution,

$$\sum_{i=15}^{\infty} \pi_i = \pi_0 (0.5)^{15} \sum_{i=0}^{\infty} (0.5)^i = \pi_0 (0.5)^{14} < (0.5)^{14}.$$

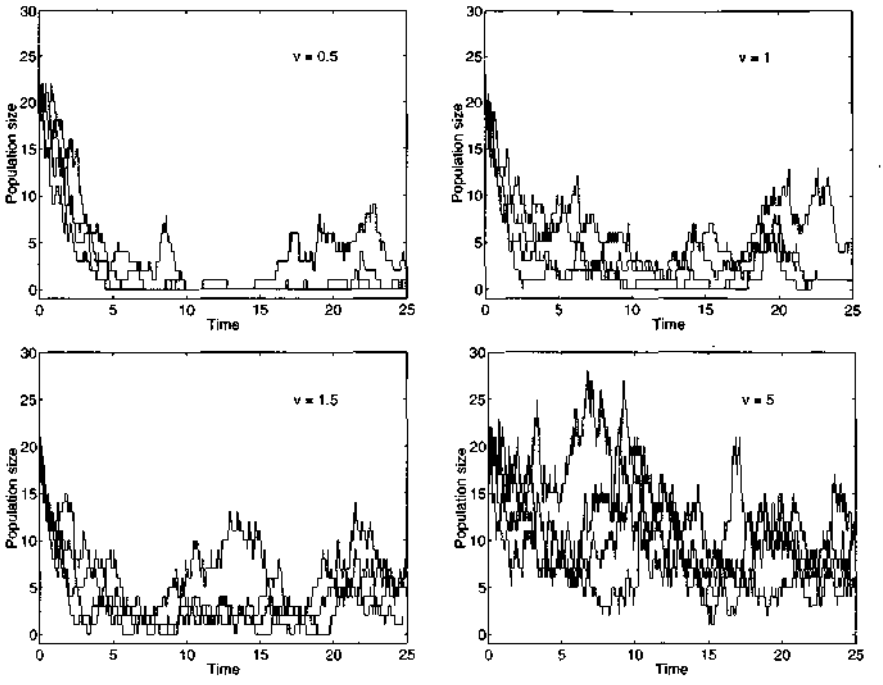

Figure 6.6. Four sample paths corresponding to the birth, death, and immigration process when $X(0) = 20$, $\lambda = 0.5$, $\mu = 1$, and $\nu = 0.5, 1.0, 1.5, 5$; the respective mean values of the stationary distribution are $m = 1, 2, 3, 10$.

Using just the first 15 values of the stationary probability distribution in Example 6.5,

$$\pi_0(1 + 1/2 + 1/2^2 + \cdots + 1/2^{14}) = \pi_0\left(2 - (0.5)^{14}\right) \approx \pi_0(1.99994) = 1,$$

gives a value of $\pi_0 = 1/1.99994$, which is close to its exact value of $1/2$.

The mean, variance, and p.g.f. are summarized for the simple birth and death process with immigration in Table 6.2.

6.5 Queueing Processes

An important application related to birth and death processes is queueing processes, processes that involve waiting in queues or lines. The arrival and departure processes are similar to birth and death processes. The formal study of queueing processes began during the early part of the twentieth century with the work of the Danish engineer A. K. Erlang. The field of queueing theory has expanded tremendously because of the diversity of applications, which include scheduling of patients, traffic regulation, telephone routing, aircraft landing, and restaurant service. Here, we give a very brief

	Simple Birth and Death with Immigration
$m(t)$	$\dfrac{\rho[N(\lambda-\mu)+\nu]-\nu}{\lambda-\mu}$
$\sigma^2(t)$	$N\dfrac{(\lambda^2-\mu^2)\rho[\rho-1]}{(\lambda-\mu)^2}+\nu\dfrac{\mu+\rho(\lambda\rho-\mu-\lambda)}{(\lambda-\mu)^2}$
$P(z,t)$	$\dfrac{(\lambda-\mu)^{\nu/\lambda}\left[\mu(\rho-1)-z(\mu\rho-\lambda)\right]^N}{\left[\lambda\rho-\mu-\lambda(\rho-1)z\right]^{N+\nu/\lambda}}$

Table 6.2. The mean, variance, and p.g.f. for the simple birth and death with immigration process, where $X(0)=N$ and $\rho=e^{(\lambda-\mu)t}$, $\lambda\neq\mu$

Figure 6.7. A simple queueing system.

introduction to some simple queueing processes. For a more thorough but elementary introduction to queueing systems, please consult Bharucha-Reid (1997), Hsu (1997), or Taylor and Karlin (1998). A more in-depth treatment of queueing processes and networks can be found in Kleinrock (1975) and Chao, Miyazawa, and Pinedo (1999).

A queueing process involves three components: (1) arrival process, (2) queue discipline, and (3) service mechanism (see Figure 6.7). The arrival process involves the arrival of customers for service and specifies the sequence of arrival times for the customers. The queue discipline is a rule specifying how customers form a queue and how they behave while waiting (e.g., first-come, first-serve basis). The service mechanism involves how customers are serviced and specifies the sequence of service times. The notation $A/B/s/K$ is used to denote the type of queue. The variables A = arrival process, B = service time distribution, s = number of servers, and K = capacity of the system. If the queue has unlimited capacity, then it is denoted as $A/B/s$. We shall consider a Poisson arrival process so that the interarrival time is exponential (Markov process) and the service-time distribution is exponential (Markov process). In this case, the queue is denoted as $M/M/s$ or $M/M/s/K$.

Consider a queueing system of type $M/M/1$. Assume the arrival process is Poisson with parameter λ (mean arrival or birth rate). The service time is exponentially distributed with parameter μ (mean departure or death rate).

After being serviced, individuals leave the system. Let $X(t)$ = number of individuals in the queue at time t. If $X(t) = 0$, then there are arrivals (births) but no departures (deaths). If λ and μ are constant, then $X(t)$ is a birth and death process as described in Example 6.2 ($\mu = q$ and $\lambda = p$). The probabilities $\text{Prob}\{X(t) = i\} = p_i(t)$ satisfy the forward Kolmogorov equations, $dp/dt = Qp$, where the generator matrix

$$Q = \begin{pmatrix} -\lambda & \mu & 0 & \dots \\ \lambda & -\lambda - \mu & \mu & \dots \\ 0 & \lambda & -\lambda - \mu & \dots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

The transition matrix T of the embedded Markov chain satisfies

$$T = \begin{pmatrix} 0 & \frac{\mu}{\lambda+\mu} & 0 & \dots \\ 1 & 0 & \frac{\mu}{\lambda+\mu} & \dots \\ 0 & \frac{\lambda}{\lambda+\mu} & 0 & \dots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

For this queueing system, there exists a unique stationary probability distribution iff $\lambda < \mu$. The ratio λ/μ is referred to as the *traffic intensity*. The stationary probability distribution is a geometric probability distribution,

$$\pi_i = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i, \quad i = 0, 1, 2, \dots.$$

If $\lambda \geq \mu$, then the queue length will tend to infinity. The mean of the stationary probability distribution represents the average number of customers C in the system (at equilibrium),

$$C = \sum_{i=1}^{\infty} i\pi_i = \left(1 - \frac{\lambda}{\mu}\right)\sum_{i=1}^{\infty} i\left(\frac{\lambda}{\mu}\right)^i.$$

This summation can be simplified by applying the identity (6.12):

$$C = \frac{\lambda/\mu}{1 - \lambda/\mu} = \frac{\lambda}{\mu - \lambda}.$$

The average amount of time W a customer spends in the system (at equilibrium) is the average number of customers divided by the average arrival rate λ:

$$W = \frac{C}{\lambda} = \frac{1}{\mu - \lambda}.$$

Example 6.6 Suppose in an $M/M/1$ queueing system customer arrival rate is 3 per minute, $\lambda = 3$. The goal is to find the average service time μ so that 95% of the queue contains less than 10 customers. First,

$$\begin{aligned} \text{Prob}\{\geq 10 \text{ customers}\} &= \sum_{i=10}^{\infty} \pi_i \\ &= \left(1 - \frac{\lambda}{\mu}\right) \sum_{i=10}^{\infty} \left(\frac{\lambda}{\mu}\right)^i \\ &= \left(\frac{\lambda}{\mu}\right)^{10}. \end{aligned}$$

Then $(\lambda/\mu)^{10} = 0.05$. Substituting the value of $\lambda = 3$ leads to

$$\mu = 4.048 \text{ customers per minute.} \qquad \blacksquare$$

Next, we consider a queueing system of the type $M/M/1/K$. The queue is limited to K customers. Therefore, there are no arrivals after the number of customers has reached K:

$$\lambda_i = \begin{cases} \lambda, & i = 0, 1, \ldots, K-1. \\ 0, & i = K, K+1, \ldots. \end{cases}$$

The mean departure or death rate is μ. For the $M/M/1/K$ queueing system there exists a unique stationary probability distribution given by

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}}, \quad i = 0, 1, \ldots, K \tag{6.13}$$

(see Exercise 14).

For the last example, consider a queueing system of type $M/M/s$. The mean arrival or birth rate is λ. Each of the s servers has an exponential service time with parameter μ. Therefore, the mean departure or death rate is

$$\mu_i = \begin{cases} i\mu, & i = 1, 2, \ldots, s-1, \\ s\mu, & i = s, s+1, \ldots. \end{cases}$$

In this case, there exists a unique stationary probability distribution iff $\lambda < s\mu$. The stationary probability distribution is given by

$$\pi_i = \begin{cases} \dfrac{(\lambda/\mu)^i}{i!}\pi_0, & i = 0, 1, \ldots, s-1, \\ \dfrac{(\lambda/\mu)^i s^{s-i}}{s!}\pi_0, & i = s, s+1, \ldots, \end{cases}$$

where π_0 is determined from $\sum_{i=0}^{\infty} \pi_i = 1$.

6.6 Probability of Population Extinction

In the simple birth, simple death, and simple birth and death processes, $\lambda_0 = 0 = \mu_0$, the set of states does not form an irreducible set and a positive stationary probability distribution does not exist. The zero state is absorbing. For many realistic birth and death processes, the zero state is absorbing. Eventually, the distribution for the total population size is concentrated at zero, $\lim_{t\to\infty} p_0(t) = 1$. The following theorem gives necessary and sufficient conditions for total population extinction as $t \to \infty$.

Theorem 6.2. *Let $\mu_0 = 0 = \lambda_0$ in a general birth and death chain with $X(0) = m \geq 1$.*

(i) Suppose $\mu_i > 0$ and $\lambda_i > 0$ for $i = 1, 2 \ldots$. If

$$\sum_{i=1}^{\infty} \frac{\mu_1 \mu_2 \cdots \mu_i}{\lambda_1 \lambda_2 \cdots \lambda_i} = \infty, \tag{6.14}$$

***then $\lim_{t\to\infty} p_0(t) = 1$** and if*

$$\sum_{i=1}^{\infty} \frac{\mu_1 \mu_2 \cdots \mu_i}{\lambda_1 \lambda_2 \cdots \lambda_i} < \infty, \tag{6.15}$$

then

$$\lim_{t\to\infty} p_0(t) = \frac{\sum_{i=m}^{\infty} \frac{\mu_1 \mu_2 \cdots \mu_i}{\lambda_1 \lambda_2 \cdots \lambda_i}}{1 + \sum_{i=1}^{\infty} \frac{\mu_1 \mu_2 \cdots \mu_i}{\lambda_1 \lambda_2 \cdots \lambda_i}}. \tag{6.16}$$

*(ii) Suppose $\mu_i > 0$ for $i = 1, 2, \ldots$, $\lambda_i > 0$ for $i = 1, 2, \ldots, N-1$ **and** $\lambda_i = 0$ for $i = N, N+1, N+2, \ldots$. Then $\lim_{t\to\infty} p_0(t) = 1$.*

A proof of Theorem 6.2 is given in the Appendix. Case (ii) is stated separately since the ratio in (i) is undefined if $\lambda_i = 0$. However, formally, the sum in case (i) is infinite if $\lambda_i = 0$ for any $i \geq 1$. Case (i) of Theorem 6.2 applies to an infinite Markov chain with state space $\{0, 1, 2, \ldots\}$, and case (ii) applies to a finite Markov chain with state space $\{0, 1, \ldots, N\}$ when $X(0) = m \leq N$. Note that in case (ii), the birth rate, λ_i, is zero for states i greater than or equal to the maximum size N. If $X(0) = m > N$, there is a simple death process occurring until size N is reached. Theorem 6.2(ii) also holds if the death and birth rates are zero, $\mu_i = 0 = \lambda_i$ for $i > N$ and $X(0) = m \leq N$.

Theorem 6.2 is applied to the simple birth and death process in the next example. In a later section, Theorem 6.2 will be applied to the logistic growth model.

Example 6.7 In the simple birth and death process, $\lambda_i = \lambda i$ **and** $\mu_i = \mu i$. Then

$$\sum_{i=1}^{\infty} \frac{\mu_1 \mu_2 \cdots \mu_i}{\lambda_1 \lambda_2 \cdots \lambda_i} = \sum_{i=1}^{\infty} \left(\frac{\mu}{\lambda}\right)^i = \begin{cases} \infty, & \text{if } \mu \geq \lambda, \\ < \infty, & \text{if } \mu < \lambda. \end{cases}$$

According to Theorem 6.2, if $\mu \geq \lambda$, then $\lim_{t\to\infty} p_0(t) = 1$ and if $\mu < \lambda$, then

$$\lim_{t\to\infty} p_0(t) = \left(\frac{\mu}{\lambda}\right)^m.$$

This result was shown in the simple birth and death process, equation (6.8). ■

6.7 Expected Time to Extinction and First Passage Time

In this section, the time until the process reaches a certain stage is studied. Suppose a population size is a and we want to find the time it takes until it reaches a size equal to b. The time spent going from state a to b is referred to as the *first passage time*. Two cases are considered, $a < b$ and $a > b$.

Let $T_{i+1,i}$ be the random variable for the time it takes to go from state i to $i+1$. From the derivation of the interevent time, we know the p.d.f. for the interevent time has an exponential distribution with parameter $\lambda_i + \mu_i$ $(X(0) = i)$:

$$f_i(t) = (\lambda_i + \mu_i)e^{-(\lambda_i+\mu_i)t}.$$

Thus, the expected time to go from state i to either $i+1$ or $i-1$ is the mean of the exponential distribution, $1/(\lambda_i + \mu_i)$. The process jumps to state $i+1$ if there is a birth (probability $\lambda_i/(\lambda_i + \mu_i)$) and jumps to state $i-1$ if there is a death (probability $\mu_i/(\lambda_i + \mu_i)$). Thus, if $a < b$, the time it takes to go from state a to b is

$$T_{b,a} = T_{a+1,a} + T_{a+2,a+1} + \cdots + T_{b,b-1}.$$

Similarly, if $a > b$, the time it takes to go from state a to b is

$$T_{b,a} = T_{a-1,a} + T_{a-2,a-1} + \cdots + T_{b,b+1}.$$

Now, if $a < b$, the expected time to go from state a to b satisfies

$$E(T_{b,a}) = E(T_{a+1,a}) + E(T_{a+2,a+1}) + \cdots + E(T_{b,b-1}).$$

Similarly, an expression can be derived for $a > b$. These are *mean first passage times.*

General expressions for $E(T_{i,i+1})$ and $E(T_{i+1,i})$ are derived for a birth and death process. Suppose the process is in state i. After an exponential amount of time the process jumps from i to $i+1$ with probability $\lambda_i/(\lambda_i+\mu_i)$

and to state $i+1$ with probability $\mu_i/(\lambda_i+\mu_i)$. To find the expected time of going from state i to $i+1$, we must consider that the process may jump to $i-1$; then the expected time it takes to go back to i must be added to this time,

$$\begin{aligned} E(T_{i+1,i}) &= \frac{\lambda_i}{\lambda_i+\mu_i}\left(\frac{1}{\lambda_i+\mu_i}\right) \\ &\quad + \frac{\mu_i}{\lambda_i+\mu_i}\left(\frac{1}{\lambda_i+\mu_i} + E(T_{i,i-1}) + E(T_{i+1,i})\right) \\ &= \frac{1}{\lambda_i+\mu_i} + \frac{\mu_i}{\lambda_i+\mu_i}\left[E(T_{i,i-1}) + E(T_{i+1,i})\right] \end{aligned}$$

(see Schinazi, 1999; Renshaw, 1993). The preceding relation can be simplified to obtain

$$E(T_{i+1,i}) = \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i}E(T_{i,i-1}). \tag{6.17}$$

A similar derivation can be obtained for $E(T_{i-1,i})$:

$$\begin{aligned} E(T_{i-1,i}) &= \frac{\mu_i}{\lambda_i+\mu_i}\left(\frac{1}{\lambda_i+\mu_i}\right) \\ &\quad + \frac{\lambda_i}{\lambda_i+\mu_i}\left(\frac{1}{\lambda_i+\mu_i} + \boldsymbol{E(T_{i,i+1}) + E(T_{i-1,i})}\right). \end{aligned}$$

Simplifying,

$$E(T_{i-1,i}) = \frac{1}{\mu_i} + \frac{\lambda_i}{\mu_i}E(T_{i,i+1}). \tag{6.18}$$

The two identities (6.17) and (6.18) will be applied to the simple birth and simple death processes.

Example 6.8 Consider a simple birth process beginning in state a. Recall that $\lambda_i = \lambda i$ and $\mu_i = 0$. The expected time $E(T_{i+1,i}) = 1/\lambda_i$. Therefore, the expected time to go from state a to state b, $a < b$ is

$$\frac{1}{\lambda}\left(\frac{1}{a} + \frac{1}{a+1} + \cdots + \frac{1}{b-1}\right). \tag{6.19}$$

The summation can be bounded by logarithms as follows:

$$\ln\left(\frac{b}{a}\right) \leq \sum_{i=a}^{b-1}\frac{1}{i} \leq \ln\left(\frac{b-1}{a-1}\right).$$

In particular,

$$\lim_{n\to\infty}\left[\sum_{i=1}^{n}\frac{1}{i} - \ln(n)\right] = \gamma = 0.57721566490\ldots,$$

where the constant γ is known as *Euler's constant.* Thus, for large values of a and b, the expression in (6.19) can be approximated by

$$E(T_{b,a}) \approx \frac{1}{\lambda} \ln\left(\frac{b}{a}\right).$$

We can compare this estimate with the deterministic exponential growth model, $n(t) = ae^{\lambda t}$. The time it takes to go from state a to state b is found by solving the equation $b = ae^{\lambda t}$ for t,

$$t = \frac{1}{\lambda} \ln\left(\frac{b}{a}\right).$$

This estimate agrees with the approximation obtained from the stochastic simple birth process. It is left as an exercise to show that for the simple death process the expected time it takes to go from state a to state b, $a > b$, is $E(T_{a,b}) \approx (1/\mu)\ln(a/b)$. ■

Example 6.9 Consider a general birth and death process, where $\lambda_0 > 0$, $\mu_0 = 0$, and $\lambda_i > 0$ and $\mu_i > 0$ for $i = 1, 2, \ldots$. Then the mean time it takes to go from state 0 to state 3 is calculated. Note that $E(T_{1,0}) = 1/\lambda_0$; then

$$E(T_{2,1}) = \frac{1}{\lambda_1} + \frac{\mu_1}{\lambda_1\lambda_0},$$

$$E(T_{3,2}) = \frac{1}{\lambda_2} + \frac{\mu_2}{\lambda_2} E(T_{2,1}) = \frac{1}{\lambda_2} + \frac{\mu_2}{\lambda_1\lambda_2} + \frac{\mu_1\mu_2}{\lambda_0\lambda_1\lambda_2},$$

so that

$$\begin{aligned} E(T_{3,0}) &= E(T_{1,0}) + E(T_{2,1}) + E(T_{3,2}) \\ &= \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \frac{\mu_1}{\lambda_1\lambda_0} + \frac{\mu_2}{\lambda_2\lambda_1} + \frac{\mu_2\mu_1}{\lambda_2\lambda_1\lambda_0}. \end{aligned} \tag{6.20}$$

■

The expression for $E(T_{3,0})$, equation (6.20), can be extended to $E(T_{b,0})$ (see Taylor and Karlin, 1998). This latter expression is related to a nonexplosive process. If the expected time to extinction approaches infinity, as $b \to \infty$, then the process is nonexplosive; that is, $\lim_{b\to\infty} E(T_{b,0}) = \infty$ (Taylor and Karlin, 1998).

Suppose $\lambda_0 = 0 = \mu_0$ and $\lim_{t\to\infty} p_0(t) = 1$, so that ultimate extinction is certain. In this case, it is possible to derive a formula for $E(T_{0,i})$, the *expected time to extinction* beginning from state i. Notice that $E(T_{0,0}) = 0$. For simplicity, denote $\tau_i = E(T_{0,i})$. For a small interval of time Δt,

$$\tau_i = \lambda_i\Delta t(\tau_{i+1} + \Delta t) + \mu_i\Delta t(\tau_{i-1} + \Delta t) + [1 - (\lambda_i + \mu_i)\Delta t](\tau_i + \Delta t) + o(\Delta t).$$

Simplifying this expression, dividing by Δt, and letting $\Delta t \to 0$ leads to the following difference equation:

$$\tau_i = \frac{1}{\lambda_i + \mu_i} + \frac{\lambda_i}{\lambda_i + \mu_i}\tau_{i+1} + \frac{\mu_i}{\lambda_i + \mu_i}\tau_{i-1}.$$

These equations can be expressed as

$$\mu_i \tau_{i-1} - (\lambda_i + \mu_i)\tau_i + \lambda_i \tau_{i+1} = -\mathbf{1}, \quad i = 1, 2, \ldots. \tag{6.21}$$

If the Markov chain is finite, with states $\{0, 1, 2, \ldots, N\}$, so that $\lambda_N = 0$, then for $i = 1, 2, \ldots, N$, the system of equations given in (6.21) can be expressed in matrix form. The matrix equation satisfies $D\tau = \mathbf{d}$, where D is an $N \times N$ matrix,

$$D = \begin{pmatrix} -\lambda_1 - \mu_1 & \lambda_1 & 0 & \cdots & 0 & 0 \\ \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & \cdots & 0 & 0 \\ 0 & \mu_3 & -\lambda_3 - \mu_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu_N & -\mu_N \end{pmatrix},$$

$\tau = (\tau_1, \tau_2, \ldots, \tau_N)^T$ and $\mathbf{d} = (-\mathbf{1}, -1, \ldots, -1)^T$. Matrix D is irreducibly diagonally dominant and thus, nonsingular. The expected time to extinction τ satisfies

$$\tau = D^{-1}\mathbf{d}.$$

This same system is derived for general birth and death discrete time Markov chains in Chapter 3, Section 3.5.1.

A recursive formula for higher-order moments can be derived. For example, $E(T_{0,i}^r)$, $i = 1, 2, \ldots, N$, the second moment of the extinction time, can be expressed in terms of the $(r-1)$st moments, $E(T_{0,i}^{r-1})$ (Goel and Richter-Dyn, 1972; Norden, 1982; Richter-Dyn and Goel, 1974). If τ^r denotes the vector of rth moments for the extinction time and the Markov chain is finite, then

$$D\tau^r = -r\tau^{r-1}.$$

Goel and Richter-Dyn (1972), Karlin and Taylor (1975), Nisbet and Gurney (1982), and Richter-Dyn and Goel (1974) give an explicit solution for the first moment, τ_i, $i = 1, 2, \ldots$, in a general birth and death chain. This result is stated next; the proof is given in the Appendix for Chapter 6. Also see Chapter 3, Theorem 3.2.

Theorem 6.3. *Suppose* $\{X(t)\}$, $t \geq 0$, *is a continuous time birth and death chain with* $X(0) = m \geq 1$ *satisfying* $\lambda_0 = 0 = \mu_0$ *and* $\lambda_i > 0$ *and* $\mu_i > 0$ *for* $i = 1, 2, \ldots$. *In addition, assume* $\lim_{t\to\infty} p_0(t) = 1$. *The expected*

time until extinction, $\tau_m = E(T_{0,m})$, satisfies

$$\tau_m = \begin{cases} \dfrac{1}{\mu_1} + \displaystyle\sum_{i=2}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i}, & m = 1 \\ \tau_1 + \displaystyle\sum_{s=1}^{m-1} \left[\frac{\mu_1 \cdots \mu_s}{\lambda_1 \cdots \lambda_s} \sum_{i=s+1}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} \right], & m = 2, 3, \ldots. \end{cases} \tag{6.22}$$

The condition $\lim_{t\to\infty} p_0(t) = 1$ is required in Theorem 6.3 (i.e., the condition (6.14) given in Theorem 6.2), since if $\lim_{t\to\infty} p_0(t) < 1$, there is a positive probability that the population size will approach infinity and, thus, $\tau_m = \infty$. Note that there is no positive stationary distribution since $\lambda_0 = 0$. If

$$\sum_{i=2}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} = \infty,$$

then $\tau_m = \infty$. When the Markov chain is finite with maximal population size N, then automatically $\tau_m < \infty$ and the solution for $m = 1, 2, \ldots, N$, $\tau = (\tau_1, \tau_2, \ldots, \tau_N)^T$, can be calculated numerically from $\tau = D^{-1}\mathbf{d}$. However, equation (6.22) applies to the finite case as well. The summation to ∞ is replaced by N,

$$\tau_m = \begin{cases} \dfrac{1}{\mu_1} + \displaystyle\sum_{i=2}^{N} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i}, & m = 1 \\ \tau_1 + \displaystyle\sum_{s=1}^{m-1} \left[\frac{\mu_1 \cdots \mu_s}{\lambda_1 \cdots \lambda_s} \sum_{i=s+1}^{N} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} \right], & m = 2, 3, \ldots, N. \end{cases}$$

The formulas (6.22) and (6.20) show that the expressions $E(T_{0,m})$ and $E(T_{m,0})$ are very different. In the expression (6.22), $\lambda_0 = 0$, the origin is absorbing, and in expression (6.20), $\lambda_0 > 0$, the origin is not absorbing. For example, the expression in (6.22) for the case $m = 3$ is

$$\begin{aligned} E(T_{0,3}) &= \left[\frac{1}{\mu_1} + \frac{\lambda_1}{\mu_1\mu_2} + \frac{\lambda_1\lambda_2}{\mu_1\mu_2\mu_3} + \cdots \right] + \left[\frac{1}{\mu_2} + \frac{\lambda_2}{\mu_2\mu_3} + \frac{\lambda_2\lambda_3}{\mu_2\mu_3\mu_4} + \cdots \right] \\ &\quad + \left[\frac{1}{\mu_3} + \frac{\lambda_3}{\mu_3\mu_4} + \frac{\lambda_3\lambda_4}{\mu_3\mu_4\mu_5} + \cdots \right]. \end{aligned}$$

The reason for this difference is that before reaching 0, the chain may visit states i, where $1 \le i < \infty$, but when moving from 0 to m, the chain may only visit states i, $0 \le i < m$.

Example 6.10 Formula (6.22) is applied to the simple birth and death process, where $\lambda_i = \lambda i$ and $\mu_i = \mu i$. The extinction time $\tau_m < \infty$ if $\lambda < \mu$. Suppose the initial size is $m = 1$. Then

$$\tau_1 = \frac{1}{\lambda} \sum_{i=1}^{\infty} \frac{1}{i} \left(\frac{\lambda}{\mu} \right)^i = \frac{1}{\lambda} \sum_{i=0}^{\infty} \frac{1}{i+1} \left(\frac{\lambda}{\mu} \right)^{i+1}.$$

Let $u = \lambda/\mu$. Then

$$\frac{1}{i+1}u^{i+1} = \int_0^u w^i \, dw.$$

Substituting this expression into the sum and interchanging summation and integration yields

$$\tau_1 = \frac{1}{\lambda}\int_0^u \left(\sum_{i=0}^{\infty} w^i\right) dw = \frac{1}{\lambda}\int_0^u \left(\frac{1}{1-w}\right) dw = -\frac{1}{\lambda}\ln\left(1-\frac{\lambda}{\mu}\right).$$

(Karlin and Taylor, 1975). ■

6.8 Logistic Growth Process

To formulate a stochastic logistic model, recall that the deterministic logistic model satisfies the differential equation

$$\frac{dn}{dt} = rn\left(1-\frac{n}{K}\right),$$

where $n = n(t)$ equals the population size at time t, r is the intrinsic growth rate, and K is the carrying capacity. Solutions $n(t)$ to this differential equation with nonnegative initial conditions approach the carrying capacity K, $\lim_{t\to\infty} n(t) = K$. The derivative, dn/dt, equals the birth rate minus the death rate. Thus, in a stochastic logistic model, the birth and death rates, λ_n and μ_n, should satisfy

$$\lambda_n - \mu_n = rn - \frac{r}{K}n^2,$$

$\lambda_0 = 0 = \mu_0$, and $\lambda_K = \mu_K$. It is reasonable to assume that the birth and death rates are quadratic functions of the population size. The state space could be finite or infinite. In the case of an infinite state space $\{0, 1, \ldots\}$, assume that for each state $i \in \{0, 1, \ldots\}$, the birth and death rates satisfy

$$\lambda_i = b_1 i + b_2 i^2 > 0 \quad \text{and} \quad \mu_i = d_1 i + d_2 i^2 > 0, \tag{6.23}$$

where the coefficients b_j and d_j, $j = 1, 2$ are constants. In the case of a finite state space $\{0, 1, 2 \ldots, N\}$, assume for each state $i \in \{0, 1, 2, \ldots, N\}$,

$$\lambda_i = \begin{cases} b_1 i + b_2 i^2 > 0, & \text{if } i = 1, 2, \ldots, N-1 \\ 0, & \text{if } i = N \end{cases} \tag{6.24}$$

and

$$\mu_i = d_1 i + d_2 i^2 > 0.$$

The initial population size $X(0)$ can be greater than N, but the process acts as a death process until state N is reached and then the process remains in the state space $\{0, 1, 2, \ldots, N\}$. Returning to the differential equation,

$$\frac{dn}{dt} = (b_1 - d_1)n + (b_2 - d_2)n^2 = rn - \frac{r}{K}n^2,$$

so that

$$r = b_1 - d_1 > 0 \quad \text{and} \quad K = \frac{b_1 - d_1}{d_2 - b_2} > 0. \tag{6.25}$$

Notice that there are four constants to be specified in the stochastic logistic model, b_1, b_2, d_1, and d_2, but only two constants in the deterministic model, r and K. There is an infinite number of values that can be chosen for the four constants that give the same values for r and K. Hence, there is an infinite number of stochastic logistic models that correspond to the same deterministic logistic model. Suppose, for example, $b_1 = r$, $b_2 = c$, $d_1 = 0$, and $d_2 = c + r/K$, where c is a constant. Then the relations (6.25) are satisfied for an infinite number of possible choices for the constant c:

$$\lambda_n - \mu_n = (rn + cn^2) - \left(cn^2 + \frac{r}{K}n^2\right) = rn - \frac{r}{K}n^2.$$

If the coefficients b_2 and d_2 are not zero, then the per capita rates of birth and death depend on the population size (i.e., λ_n/n and μ_n/n depend on n). For example, if $b_2 < 0$ (or $d_2 < 0$), the number of births (deaths) decreases as the population size increases, but if the reverse inequality holds, $b_2 > 0$ ($d_2 > 0$), the number of births (deaths) increases as the population size increases. A reasonable assumption for many populations is that $d_2 > 0$, meaning that the death rate is density dependent. Ultimately, the choice of the coefficients b_j and d_j depends on the dynamics of the particular population being modeled.

Example 6.11 Two stochastic logistic models are defined that have the same deterministic logistic model. Define the birth and death rates, λ_i and μ_i, as follows:

(a) $\lambda_i = i$ and $\mu_i = \dfrac{i^2}{10}$, $i = 0, 1, 2, \ldots$

(b) $\lambda_i = \begin{cases} i - \dfrac{i^2}{20}, & i = 0, 1, \ldots, 20 \\ 0, & i > 20 \end{cases}$ **and** $\boldsymbol{\mu_i = \dfrac{i^2}{20}}$, $\boldsymbol{i = 0, 1, 2, \ldots}$

In both cases the deterministic model is

$$\frac{dn}{dt} = n\left(1 - \frac{n}{10}\right),$$

where $r = 1$ and $K = 10$. In the deterministic model, solutions approach the carrying capacity $K = 10$; it is a globally stable equilibrium for positive

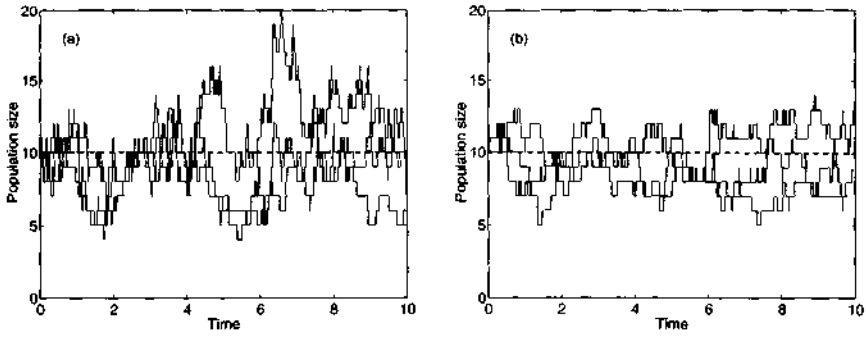

Figure 6.8. Three sample paths of the stochastic logistic model for cases (a) and (b) with $X(0) = 10$.

initial conditions. In the stochastic models, for population sizes greater than $K = 10$, the death rate exceeds the birth rate, and for population sizes less than $K = 10$, the birth rate exceeds the death rate. When $n = 10$, $\lambda_{10} = \mu_{10}$, and the probability of a birth or a death is equal to $1/2$,

$$\frac{\lambda_{10}}{\lambda_{10} + \mu_{10}} = \frac{1}{2} = \frac{\mu_{10}}{\lambda_{10} + \mu_{10}}.$$

Three sample paths for models (a) and (b) are graphed in Figure 6.8. ■

The assumptions (6.23) or (6.24) and application of Theorems 6.2 and 6.3 imply that in the stochastic logistic model, extinction occurs with probability 1, $\lim_{t\to\infty} p_0(t) = 1$, and the expected time to extinction is finite.

Corollary 6.1. *Assume the stochastic logistic model satisfies (6.25) and either (6.23) or (6.24). Then*

$$\lim_{t\to\infty} p_0(t) = 1 \tag{6.26}$$

and the expected time until extinction $\tau_m = E(T_{0,m}) < \infty$.

Proof. Suppose (6.24) is satisfied. Then (6.26) follows directly from Theorem 6.2 part (ii) and $\tau_m < \infty$.

Suppose (6.23) is satisfied. We will show that (6.14) is satisfied. The ratio of two successive terms, a_i/a_{i-1} in the summation of (6.14), equals μ_i/λ_i, where

$$\frac{\mu_i}{\lambda_i} = \frac{d_1 + d_2 i}{b_1 + b_2 i}.$$

Because λ_i and μ_i are positive for all i and from the assumptions in (6.25), it follows that $d_2 > b_2 \geq 0$. Thus,

$$\lim_{i\to\infty} \frac{\mu_i}{\lambda_i} = \begin{cases} \dfrac{d_2}{b_2} > 1, & \text{if } b_2 > 0 \\ \infty, & \text{if } b_2 = 0. \end{cases}$$

In either case, by the ratio test, it follows that (6.14) is satisfied and Theorem 6.2 (i) implies $\lim_{t\to\infty} p_0(t) = 1$. In addition, it is easy to show by the ratio test that $\sum_{i=1}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} < \infty$:

$$\lim_{i\to\infty} \frac{\lambda_i}{\mu_{i+1}} = \begin{cases} \frac{b_2}{d_2} < 1, & \text{if } b_2 > 0 \\ 0, & \text{if } b_2 = 0. \end{cases}$$

Hence, by Theorem 6.3 part (ii), $\tau_m < \infty$. □

The expected time to extinction is calculated **using the techniques** from the previous section in the next two examples.

Example 6.12 Suppose the stochastic logistic **model has the following** birth and death rates:

$$\lambda_i = \begin{cases} i - \frac{i^2}{N}, & i = 0, 1, \ldots, N \\ 0, & i > N \end{cases} \quad \text{and} \quad \mu_i = \frac{i^2}{N}, \quad i = 0, 1, 2, \ldots.$$

The intrinsic growth rate $r = 1$ and the carrying capacity $K = N/2$. The expected time to extinction can be calculated for $X(0) = m$, $m = 1, 2, \ldots, N$ by solving the linear system $D\tau = \mathbf{d}$. The expected time to extinction $\tau = (\tau_1, \ldots, \tau_N)^T$ is graphed in Figure 6.9 in two cases, $N = 10$ and $N = 20$. When $N = 10$, the carrying capacity is $K = 5$, and when $N = 20$, the carrying capacity is $K = 10$. The expected time to extinction for a carrying capacity of $K = 5$ ranges from approximately 234 ($= \tau_1$) to 269 ($= \tau_{10}$), whereas for $K = 10$, the expected time to extinction is on the order of 10^5. ■

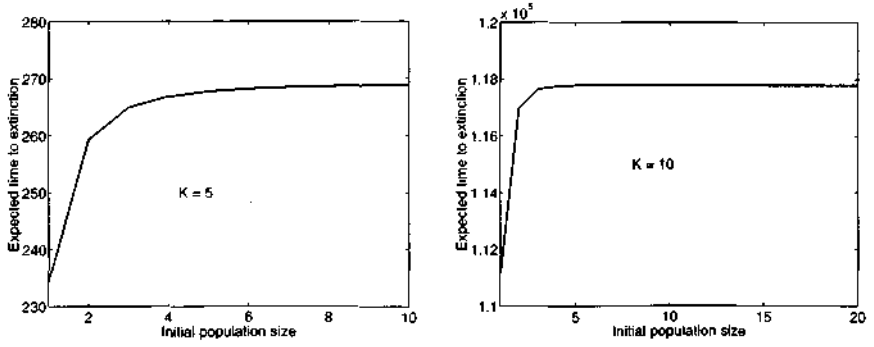

Figure 6.9. Expected time until extinction in the stochastic logistic model with $K = 5$ and $K = 10$.

Example 6.13 Assume the stochastic logistic model satisfies

$$\lambda_i = i \text{ and } \mu_i = \frac{i^2}{10}, \quad i = 0, 1, 2, \ldots.$$

In this case, the carrying capacity is $K = 10$. Formula (6.22) can be used to compute $\tau_1 = E(T_{0,1})$, a lower bound for the expected time to extinction, $\tau_m \geq \tau_1$, $m \geq 1$,

$$\tau_1 = \sum_{i=1}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i}.$$

An estimate for τ_1 is $\tau_1 \approx 2489.35$. ■

The mean and variance can be computed from the forward Kolmogorov differential equations, $dp/dt = Qp$, or from the partial differential equation satisfied by the m.g.f. Applying the generating function technique to the forward Kolmogorov equations, it can be shown that the partial differential equation satisfied by the m.g.f. for the stochastic logistic model is

$$\begin{aligned} \frac{\partial M}{\partial t} &= \left[b_1(e^{\theta} - 1) + d_1(e^{-\theta} - 1)\right] \frac{\partial M}{\partial \theta} \\ &\quad + \left[b_2(e^{\theta} - 1) + d_2(e^{-\theta} - 1)\right] \frac{\partial^2 M}{\partial \theta^2}, \end{aligned} \tag{6.27}$$

where $M(\theta, 0) = e^{N\theta}$ if $X(0) = N$. This differential equation can be used to derive differential equations satisfied by the mean and higher-order moments. Differentiating (6.27) with respect to θ and interchanging the order of the differentiation yields

$$\begin{aligned} \frac{\partial^2 M}{\partial t \, \partial \theta} &= [b_1 e^{\theta} - d_1 e^{-\theta}] \frac{\partial M}{\partial \theta} + [b_1(e^{\theta} - 1) + d_1(e^{-\theta} - 1)] \frac{\partial^2 M}{\partial \theta^2} \\ &\quad + [b_2 e^{\theta} - d_2 e^{-\theta}] \frac{\partial^2 M}{\partial \theta^2} + [b_2(e^{\theta} - 1) + d_2(e^{-\theta} - 1)] \frac{\partial^3 M}{\partial \theta^3}. \end{aligned}$$

Evaluating this differential equation at $\theta = 0$ and using the fact that $\partial^k M / \partial \theta^k$ evaluated at $\theta = 0$ is $E(X^k(t))$ gives the identity

$$\begin{aligned} \frac{dm(t)}{dt} &= [b_1 - d_1] m(t) + [b_2 - d_2] E(X^2(t)) \\ &= r m(t) - \frac{r}{K} E(X^2(t)). \end{aligned} \tag{6.28}$$

The differential equation for $m(t)$ cannot be solved since it depends on $E(X^2(t))$, but $m(t)$ can be compared the to the solution of the deterministic model. In the simple birth and death processes, the mean agrees with the solution of the deterministic model. This is not the case for more complicated models. In particular, for the stochastic logistic model, it will

be shown that the mean of the stochastic process is less than the solution of the deterministic model (Tognetti and Winley, 1980).

The variance $\sigma^2(t) = E(X^2(t)) - m^2(t) > 0$ so that $E(X^2(t)) > m^2(t)$. Because $m(t) \geq 0$, it follows from equation (6.28) that

$$\frac{dm(t)}{dt} < rm(t)\left[1 - \frac{m(t)}{K}\right], \quad t \in [0, \infty).$$

Suppose $n(t)$ is the solution of the logistic differential equation $dn/dt = rn(1 - n/K)$ satisfying $0 < n(0) = m(0)$. By comparing the differential equations for $m(t)$ and $n(t)$, it can be shown that

$$m(t) \leq n(t) \quad \text{for} \quad t \in [0, \infty)$$

(see the comparison theorem in the Appendix for Chapter 6). The solution to the deterministic logistic differential equation is greater than the mean of the stochastic logistic model. This result seems plausible because ultimate extinction is certain in the stochastic logistic model, $\lim_{t\to\infty} p_0(t) = 1$. As $t \to \infty$, the probability distribution becomes concentrated at zero, so that $\lim_{t\to\infty} m(t) = 0$.

A differential equation for $E(X^2(t))$ can be derived in a manner similar to the mean. Use the partial differential equation for the m.g.f. and differentiate twice with respect to θ; then evaluate at $\theta = 0$:

$$\begin{aligned}\frac{dE(X^2(t))}{dt} &= [b_1 + d_1]m(t) + [2(b_1 - d_1) + (b_2 + d_2)]E(X^2(t)) \\ &\quad + 2[b_2 - d_2]E(X^3(t)). \qquad (6.29)\end{aligned}$$

The variance increases if $b_1 + d_1$ and $b_2 + d_2$ are large and positive. This can be seen in the stochastic simulations of the logistic model, Figure 6.8. In (a), $b_1 + d_1 = 1$ and $b_2 + d_2 = 1/10$ and in (b) $b_1 + d_1 = 1$ and $b_2 + d_2 = 0$ for $i \leq 20$ and $1/20$ for $i > 20$. Both models have the same deterministic logistic model, but model (b) has a smaller variance.

6.9 Quasistationary Probability Distribution

In birth and death models, when the origin is absorbing, there is no stationary probability distribution. However, even when $\lim_{t\to\infty} p_0(t) = 1$, prior to extinction, the probability distribution of $X(t)$ can be approximately stationary for a long period of time. This is especially true if the expected time to extinction is very long. This approximate stationary distribution is known as the *quasistationary probability distribution* or *quasiequilibrium probability distribution*.

Denote the probability distribution associated with $X(t)$ conditioned on nonextinction as $q_i(t)$. Then

$$q_i(t) = \frac{p_i(t)}{1 - p_0(t)}, \quad i = 1, 2, \ldots.$$

The quasistationary probabilities satisfy a system of differential equations similar to the forward Kolmogorov differential equations,

$$\begin{aligned}\frac{dq_i}{dt} &= \frac{dp_i/dt}{1-p_0} + \frac{p_i}{(1-p_0)}\frac{dp_0/dt}{(1-p_0)} \\ &= \lambda_{i-1}q_{i-1} - (\lambda_i + \mu_i)q_i + \mu_{i+1}q_{i+1} + q_i[\mu_1 q_1],\end{aligned}$$

where it is assumed that $\lambda_0 = 0 = \mu_0$.

The quasistationary probability distribution can be approximated by making the assumption $\mu_1 = 0$. Then $dq/dt = \tilde{Q}q$, where

$$\tilde{Q} = \begin{pmatrix} -\lambda_1 & \mu_2 & 0 & \cdots \\ \lambda_1 & -\lambda_2 - \mu_2 & \mu_3 & \cdots \\ 0 & \lambda_2 & -\lambda_3 - \mu_3 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

This system will have a unique positive stationary probability distribution given by $\tilde{\pi} = (\tilde{\pi}_1, \tilde{\pi}_2, \ldots)^T$ if the assumptions of Theorem 6.1 are satisfied. The stationary probability distribution must satisfy

$$\tilde{\pi}_i = \frac{\lambda_1\lambda_2\cdots\lambda_{i-1}}{\mu_2\mu_3\cdots\mu_i}\tilde{\pi}_1, \quad \text{and} \quad \sum_{i=1}^{\infty}\tilde{\pi}_i = 1.$$

Therefore, a unique positive stationary probability distribution exists to the system $dq/dt = \tilde{Q}q$ if

$$\sum_{i=2}^{\infty}\frac{\lambda_1\lambda_2\cdots\lambda_{i-1}}{\mu_2\mu_3\cdots\mu_i} < \infty.$$

The solution $\tilde{\pi}$ approximates the quasistationary probability distribution.

Example 6.14 Consider the two stochastic logistic models discussed previously, where the birth and death rates, λ_i and μ_i satisfy

(a) $\lambda_i = i$ and $\mu_i = \dfrac{i^2}{10}$, $i = 0, 1, 2, \ldots$

(b) $\lambda_i = \begin{cases} i - \dfrac{i^2}{20}, & i = 0, 1, \ldots, 20, \\ 0, & i > 20, \end{cases}$ and $\mu_i = \dfrac{i^2}{20}$, $i = 0, 1, 2 \ldots$

In both cases the deterministic model is

$$\frac{dn}{dt} = n\left(1 - \frac{n}{10}\right),$$

so that the intrinsic growth rate is $r = 1$ and the carrying capacity is $K = 10$. The approximate quasistationary probabilities $\tilde{\pi}_n$ are calculated for each model and graphed in Figure 6.10. ■

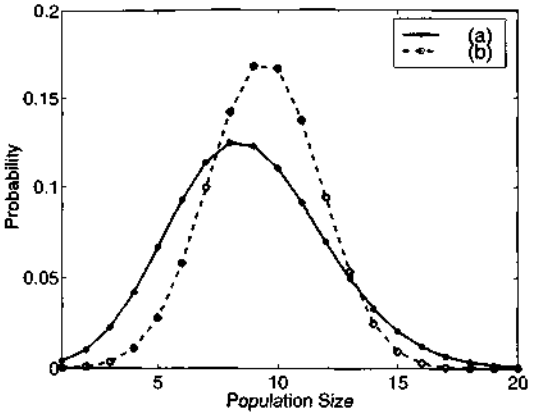

Figure 6.10. Approximate quasistationary probability distribution $\tilde{\pi}$ for cases (a) and (b). The mean and standard deviation for case (a) are $\tilde{m} \approx 8.848$, $\tilde{\sigma} \approx 3.193$ and for case (b), $\tilde{m} \approx 9.435$ and $\tilde{\sigma} \approx 2.309$.

Notice that the means of the quasistationary distributions are less than the equilibrium of the deterministic model, $K = 10$, and that the variance in model (a) is greater than model (b). Recall that it was shown in the last section that the mean $m(t)$ of the stochastic logistic model is always less than the solution of the deterministic model, $m(t) < n(t)$. In the deterministic model, $n(t) \to K$, and in the stochastic model, $m(t) \approx \tilde{m}$, so that it is reasonable to expect $\tilde{m} < K$. In addition, the variance is large when $b_1 + d_1$ and $b_2 + d_2$ are large. These parameter values are larger for case (a) than for case (b), implying that the variance is larger for case (a) than for case (b). This is evident in Figure 6.10.

For a discussion of generalized stochastic logistic models with immigration, see Matis and Kiffe (1999). In their models, a stationary probability distribution exists and the cumulant generating function for this stationary distribution is obtained. In addition, it is shown that the stationary probability distribution is approximately normal. See also Nåsell (2001) for a discussion of the quasistationary probability distribution in stochastic logistic models.

In the next section, an explosive birth process is discussed. Necessary and sufficient conditions are stated for a birth process to be explosive.

6.10 An Explosive Birth Process

We shall expand on the definition of an explosive process. Norris (1999) defines an explosive process in terms of the jump times or waiting times $\{W_i\}_{i=1}^{\infty}$ and the interevent times $\{T_i\}_{i=1}^{\infty}$. Recall that W_i is the time at which the process jumps to a new state and $T_i = W_{i+1} - W_i$. Let

$W = \sup_i\{W_i\}$, $i \in \{1, 2, \ldots\}$, and $T = \sum_{i=1}^{\infty} T_i$. If, for some state i,

$$\text{Prob}\{W < \infty | X(0) = i\} > 0 \quad \text{or} \quad \text{Prob}\{T < \infty | X(0) = i\} > 0, \tag{6.30}$$

then the process is said to be *explosive*; otherwise it is *nonexplosive*. In an explosive process, it follows that given $X(0) = i$, the probability distribution $\{p_i(t)\}_{i=0}^{\infty}$ corresponding to $X(t)$ has the property that there exists a time $t^* < \infty$, where $T = t^*$ such that

$$\sum_{i=0}^{\infty} p_i(t^*) < 1.$$

Therefore, a birth process is not explosive if

$$\sum_{i=0}^{\infty} p_i(t) = 1 \quad \text{for all} \quad t \geq 0.$$

Necessary and sufficient conditions for a birth process to be explosive are given in the next theorem (Feller, 1968; Norris, 1999). Recall that in a birth process, $\lambda_i > 0$ and $\mu_i = 0$ for $i = 1, 2, \ldots$.

Theorem 6.4. *A birth process is explosive or satisfies (6.30) iff*

$$\sum_{i=1}^{\infty} \frac{1}{\lambda_i} < \infty. \tag{6.31}$$

Proof. Suppose $X(0) = m$ or $p_j(0) = \delta_{mj}$. Let $S_k(t) = \sum_{i=0}^{k} p_i(t)$. The forward Kolmogorov differential equations $dp/dt = Qp$ for a birth process satisfy

$$\begin{aligned} \frac{dp_i}{dt} &= \lambda_{i-1} p_{i-1} - \lambda_i p_i \\ \frac{dp_0}{dt} &= -\lambda_0 p_0. \end{aligned}$$

From the forward Kolmogorov differential equations it follows that

$$\frac{dS_k(t)}{dt} = -\lambda_k p_k(t).$$

If $k \geq m$, then $S_k(0) = 1$. Integrating from 0 to t leads to

$$S_k(t) - 1 = -\int_0^t \lambda_k p_k(\tau)\, d\tau, \quad k \geq m.$$

The sequence $\{S_k(t)\}_{k=m}^{\infty}$ is increasing. Therefore, the sequence $\{1 - S_k(t)\}_{k=m}^{\infty}$ is decreasing for $k \geq m$ and is bounded below by zero. The sequence must have a limit. Call this limit $L(t)$:

$$\lim_{k \to \infty} \lambda_k \int_0^t p_k(\tau)\, d\tau = L(t) \geq 0.$$

Also, $\int_0^t p_k(\tau)\,d\tau \geq L(t)/\lambda_k$ for $k \geq m$. It follows from the definition of $S_k(t)$ that for $k \geq m$,

$$\int_0^t S_k(\tau)\,d\tau \geq \sum_{i=m}^{k} \int_0^t p_i(\tau)\,d\tau \geq L(t) \sum_{i=m}^{k} \frac{1}{\lambda_i}. \tag{6.32}$$

Suppose the process is explosive given $X(0) = m$. Then there exists a time t^* such that

$$\lim_{k\to\infty} S_k(t^*) < 1 \quad \text{and} \quad \lim_{k\to\infty} \lambda_k \int_0^{t^*} p_k(\tau)\,d\tau > 0,$$

so that $L(t^*) > 0$. This fact, together with $t^* \geq \int_0^{t^*} S_k(\tau)\,d\tau$ in (6.32), implies that the summation (6.31) is convergent.

To show the converse, concepts presented by Norris (1999) are applied. Suppose the summation (6.31) is convergent and $X(0) = m$. The probability of a birth given the population size is i has an exponential distribution with parameter λ_i. Hence, the expectation

$$E\left(\sum_{i=0}^{k} T_i\right) = \sum_{i=0}^{k} E(T_i) = \sum_{i=m}^{m+k+1} \frac{1}{\lambda_i} < \sum_{i=1}^{\infty} \frac{1}{\lambda_i} < \infty.$$

Because $\sum_{i=0}^{\infty} E(T_i) < \infty$ it follows from the dominated convergence theorem (Rudin, 1987) that the limit of the left side as $k \to \infty$ is also finite, $E(T|X(0) = m) = E(\sum_{i=0}^{\infty} T_i|X(0) = m) < \infty$. Because m is arbitrary and the expectation of T is finite,

$$\text{Prob}\,\{T = \infty | X(0) = m\} = 0.$$

Thus, for any initial state m, Prob$\{T < \infty|X(0) = m\} = 1$; the process is explosive. □

The following example describes a birth process that is explosive.

Example 6.15 Suppose a birth process satisfies $\lambda_i = bi^k > 0$, $i = 1, 2, \ldots$, where $k > 1$. Note that

$$\sum_{i=1}^{\infty} \frac{1}{bi^k} = \frac{1}{b} \sum_{i=1}^{\infty} \frac{1}{i^k}$$

is a multiple of a convergent p-series with $p = k > 1$. According to Theorem 6.4, the birth process is explosive. A deterministic analogue of this model is the differential equation

$$\frac{dn}{dt} = bn^k.$$

Integration of n with initial condition $n(0) = N$ leads to the solution

$$n(t) = \left[N^{1-k} - (k-1)bt\right]^{-1/(k-1)}.$$

As $t \to N^{1-k}/[b(k-1)]$, then $n(t) \to \infty$. Hence, the deterministic solution approaches infinity or "explodes" at a finite time. ■

6.11 Nonhomogeneous Birth and Death Process

In all of the birth and death processes discussed thus far the transition probabilities have been homogeneous with respect to time; that is, time independent. Suppose the birth and death rate parameters λ_i and μ_i satisfy

$$\lambda_i \equiv \lambda(i,t) \quad \text{and} \quad \mu_i \equiv \mu(i,t).$$

Then the transition probabilities are nonhomogeneous. The forward Kolmogorov equation is

$$\frac{dp_i(t)}{dt} = \lambda(i-1,t)p_{i-1}(t) + \mu(i+1,t)p_{i+1}(t) - (\lambda(i,t) + \mu(i,t))p_i(t).$$

Multiply by z^i and sum from $i = 0$ to ∞ to obtain a partial differential equation for the probability generating function $P(z,t)$:

$$\begin{aligned}\frac{\partial P}{\partial t} &= \sum_{i=0}^{\infty} \lambda(i-1,t)p_{i-1}(t)z^i + \sum_{i=0}^{\infty} \mu(i+1,t)p_{i+1}(t)z^i \\ &\quad - \sum_{i=0}^{\infty} [\lambda(i,t) + \mu(i,t)]p_i(t)z^i.\end{aligned}$$

The right-hand side depends on the form of $\lambda(i,t)$ and $\mu(i,t)$. One example is discussed next.

Example 6.16 Suppose $\lambda(i,t) = \lambda(t)i$ and $\mu(i,t) = \mu(t)i$ (Bailey, 1990). Then the partial differential equation for the probability generating function satisfies

$$\frac{\partial P}{\partial t} = \left[\lambda(t)(z^2 - z) + \mu(t)(1-z)\right]\frac{\partial P}{\partial z}, \quad P(z,0) = z^N.$$

Along characteristic curves, τ and s,

$$\frac{dt}{d\tau} = 0, \quad \frac{dz}{d\tau} = (1-z)[\lambda(t)z - \mu(t)], \quad \text{and} \quad \frac{dP}{d\tau} = 0$$

with initial conditions

$$t(s,0) = 0, \quad z(s,0) = s, \quad \text{and} \quad P(s,0) = s^N.$$

The differential equation for z can be solved by letting $t = \tau$ and making the change of variable $1 - z = 1/y$ so that $y^2 = 1/(1-z)^2$ and $dz/d\tau = (1/y^2)\,dy/d\tau$. The differential equation for z can be transformed into a linear differential equation in y:

$$\frac{dy}{d\tau} = (\lambda(\tau) - \mu(\tau))y - \lambda(\tau).$$

Use of an integrating factor,

$$e^{\int_0^\tau [\mu(\alpha)-\lambda(\alpha)]\,d\alpha} = e^{\rho(\tau)},$$

leads to

$$ye^{\rho(\tau)} - y(s,0) = -\int_0^\tau \lambda(\beta)e^{\rho(\beta)}\,d\beta.$$

Now, $y(s,0) = 1/(1-s)$ and $y = 1/(1-z)$, so that

$$\frac{1}{s-1} = \frac{e^{\rho(\tau)}}{z-1} - \int_0^\tau \lambda(\beta)e^{\rho(\beta)}\,d\beta.$$

Solving for s,

$$s = 1 + \frac{1}{e^{\rho(\tau)}z - 1 - \int_0^\tau \lambda(\beta)e^{\rho(\beta)}\,d\beta}.$$

The solution to the p.g.f. is

$$P(z,t) = \left[1 + \frac{1}{\frac{e^{\rho(t)}}{z-1} - \int_0^t \lambda(\tau)e^{\rho(\tau)}\,d\tau}\right]^N,$$

where $\rho(t) = \int_0^t [\mu(\tau) - \lambda(\tau)]\,d\tau$ and τ represents a dummy variable.

An expression for $p_0(t)$, the probability of extinction, can be found by evaluating $P(z,t)$ at $z = 0$:

$$p_0(t) = \left[1 - \frac{1}{e^{\rho(t)} + \int_0^t \lambda(\tau)e^{\rho(\tau)}\,d\tau}\right]^N.$$

Because

$$\begin{aligned} e^{\rho(t)} + \int_0^t \lambda(\tau)e^{\rho(\tau)}\,d\tau &= e^{\rho(t)} - \int_0^t \rho(\tau)e^{\rho(\tau)}d\tau + \int_0^t \mu(\tau)e^{\rho(\tau)}\,d\tau \\ &= e^{\rho(t)} - e^{\rho(\tau)}\Big|_0^t + \int_0^t \mu(\tau)e^{\rho(\tau)}\,d\tau \\ &= 1 + \int_0^t \mu(\tau)e^{\rho(\tau)}\,d\tau, \end{aligned}$$

the probability of extinction simplifies to

$$p_0(t) = \left[\frac{\int_0^t \mu(\tau)e^{\rho(\tau)}d\tau}{1+\int_0^t \mu(\tau)e^{\rho(\tau)}\,d\tau}\right]^N.$$

Note that if $\lim_{t\to\infty}\int_0^t \mu(\tau)e^{\rho(\tau)}\,d\tau = \infty$, then $\lim_{t\to\infty} p_0(t) = 1$. **But**

$$\begin{aligned}\int_0^t \mu(\tau)e^{\rho(\tau)}\,d\tau &= \int_0^t \rho(\tau)e^{\rho(\tau)}\,d\tau + \int_0^t \lambda(\tau)e^{\rho(\tau)}\,d\tau \\ &\geq e^{\rho(t)} - 1.\end{aligned}$$

Thus, if $\lim_{t\to\infty}\rho(t) = \infty$, it follows that $\lim_{t\to\infty} p_0(t) = 1$.

The analogous deterministic model satisfies the differential equation,

$$\frac{dn}{dt} = [\lambda(t) - \mu(t)]n, \quad n(0) > 0.$$

The solution to this differential equation is

$$n(t) = n(0)e^{-\int_0^t[\mu(\tau)-\lambda(\tau)]\,d\tau} = n(0)e^{-\rho(t)}.$$

For the deterministic model population extinction occurs, $\lim_{t\to\infty} n(t) = 0$, if and only if $\lim_{t\to\infty}\rho(t) = \infty$. This result differs from the stochastic model in that population extinction is still possible in the stochastic model even if the limit of $\rho(t)$ is not infinite. ■

6.12 Exercises for Chapter 6

1. Suppose a continuous time Markov chain, $\{X(t)\}$, $t \geq 0$, is defined by the following infinitesimal transition probabilities:

$$\begin{aligned}p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} \mu i\Delta t + o(\Delta t), & j = -1 \\ \lambda i\Delta t + o(\Delta t), & j = 1 \\ \nu\Delta t + o(\Delta t), & j = 2 \\ 1 - (\nu + \lambda i + \mu i)\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq -1, 0, 1, 2. \end{cases}\end{aligned}$$

The initial distribution satisfies $\text{Prob}\{X(0) = N\} = 1$.

(a) Find the differential equations satisfied by each of the state probabilities

$$p_i(t) = \text{Prob}\{X(t) = i\} \quad \text{for} \quad i = 0, 1, 2, \ldots.$$

(b) Find the generator matrix Q.

2. Let $X(t)$ be the random variable for the total population size for a birth-death-emigration process and $p_i(t) = \text{Prob}\{X(t) = i\}$. The infinitesimal transition probabilities satisfy:

$$\begin{aligned} p_{i+j,i}(\Delta t) &= \text{Prob}\{\Delta X(t) = j | X(t) = i\} \\ &= \begin{cases} (\mu i + \nu)\Delta t + o(\Delta t), & j = -1 \\ \lambda i^2 \Delta t + o(\Delta t), & j = 1 \\ 1 - (\lambda i^2 + \mu i + \nu)\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq -1, 0, 1. \end{cases} \end{aligned}$$

The initial probability distribution satisfies $\text{Prob}\{X(0) = N\} = 1$.

(a) Show that the differential equations satisfied by $p_i(t)$ for $i = 1, 2, \ldots$ are

$$\frac{dp_i}{dt} = p_{i-1}[\lambda(i-1)^2] + p_{i+1}[\mu(i+1) + \nu] - p_i[\lambda i^2 + \mu i + \nu].$$

(b) Assume $\lambda = 0$. In addition, assume $\nu = 0$ when $i = 0$ (i.e., there is no emigration when the population size is zero). Notice that the state space consists of $\{0, 1, \ldots, N\}$, since it is a death-emigration process. Write the first-order partial differential equation satisfied by the probability generating function, $P(z, t) = \sum_{i=0}^{N} p_i(t) z^i$.

3. Suppose a general birth and death process satisfies $\lambda_i = b(i+1)$ and $\mu_i = di$ for $i = 0, 1, 2, \ldots, b, d > 0$.

(a) Determine conditions on b and d such that a unique positive stationary probability distribution exists.

(b) Find the unique stationary probability distribution.

4. Consider the simple birth process.

(a) Find the generator matrix Q and the transition matrix T for the embedded Markov chain based on the state space $\{N, N+1, \ldots\}$.

(b) Show that all states are transient and there does not exist a stationary probability distribution.

5. Consider the simple death process.

(a) Find the generator matrix Q and the transition matrix T for the embedded Markov chain based on the state space $\{0, 1, \ldots, N\}$.

(b) Show that zero is an absorbing state, the remaining states are transient, and the unique stationary probability distribution is $\pi = (1, 0, 0, \ldots, 0)^T$.

6. Consider the simple birth and death process with immigration such that $\nu = 1 = \lambda$. Find the stationary probability distribution π when $\mu > 1$.

7. Consider the simple birth and death process with immigration.

   (a) Use the m.g.f. $M(\theta, t)$ to find an expression for $\sigma^2(t)$.

   (b) Find $\lim_{t\to\infty} \sigma^2(t)$ when $\lambda < \mu$. Does $\sigma^2(\infty)$ agree with the variance of the stationary distribution in Example 6.5?

8. Suppose a general birth and death process has birth and death rates given by

$$\lambda_i = b_0 + b_1 i + b_2 i^2, \quad \text{and} \quad \mu_i = d_1 i + d_2 i^2, \quad \text{for} \quad i = 0, 1, 2, \ldots.$$

   (a) Find the forward Kolmogorov equations; then use the generating function technique to find the differential equations satisfied by the p.g.f. and m.g.f. In particular, show that

$$\begin{aligned} \frac{\partial M}{\partial t} &= (e^{\theta} - 1)\left[b_0 + b_1 \frac{\partial}{\partial \theta} + b_2 \frac{\partial^2}{\partial \theta^2}\right] M \\ &\quad + (e^{-\theta} - 1)\left[d_1 \frac{\partial}{\partial \theta} + d_2 \frac{\partial^2}{\partial \theta^2}\right] M. \end{aligned}$$

   (b) Compare the partial differential equation for M with that for the stochastic logistic model and find the form for the partial differential equation for $M(\theta, t)$ in the more general case where $\lambda_i = \sum_{k=0}^{n} b_k i^k$ and $\mu_i = \sum_{k=1}^{n} d_k i^k$.

9. Consider a death and immigration process, where $\lambda_i = \nu$ and $\mu_i = \mu i$.

   (a) Show that the m.g.f. satisfies

$$\frac{\partial M}{\partial t} = \mu(e^{-\theta} - 1)\frac{\partial M}{\partial \theta} + \nu(e^{\theta} - 1)M.$$

   (b) Use the method of characteristics to solve for $M(\theta, t)$. Show that

$$M(\theta, t) = [1 + (e^{\theta} - 1)e^{-\mu t}]^N \exp\left(\frac{\nu}{\mu}\left[1 + e^{-\mu t(e^{\theta}-1)}\right]\right).$$

   (*Hint*: Express the characteristic equation for M in terms of θ; solve $dM/d\theta$.)

   (c) Use the expression for $M(\theta, t)$ to find the mean, $m(t)$, and variance, $\sigma^2(t)$, of the process.

10. For the simple death process, show that the expected time to go from state a to state b can be approximated if a and b are large. Show that

$$E(T_{b,a}) \approx \frac{1}{\mu} \ln\left(\frac{a}{b}\right).$$

11. For the birth and death process with immigration, show that the series (6.11) converges iff $\lambda < \mu$ iff there exists a unique stationary probability distribution.

12. Consider the simple birth and death process with immigration.

    (a) Assume $\lambda = 0.5$ and $\nu = 1 = \mu$. Show that the stationary probability distribution satisfies

$$\pi_i = \frac{i+1}{2^{i+2}}, \quad i = 0, 1, 2, \ldots.$$

    Find the mean of this distribution. This probability distribution is graphed in Figure 6.5.

    (b) Assume $\lambda = 0.5$, $\mu = 1$ and $\nu = 1.5$. Show that the stationary probability distribution satisfies

$$\pi_i = \frac{(i+2)(i+1)}{2^{i+4}}, \quad i = 0, 1, 2, \ldots.$$

    Find the mean. This probability distribution is graphed in Figure 6.5.

13. In the simple birth and death process with immigration, the corresponding deterministic model satisfies

$$\frac{dn}{dt} = (\lambda - \mu)n + \nu, \quad \nu(0) = N.$$

    Show that the solution to the deterministic model $n(t)$ agrees with the mean of the stochastic model, equation (6.9), when $\lambda \neq \mu$ and equation (6.10), when $\lambda = \mu$.

14. In a queueing system of type $M/M/1/K$, derive the stationary probability distribution given in equation (6.13):

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}}, \quad i = 0, 1, \ldots, K.$$

    Then find the average number of customers in the system, $C = \sum_{i=1}^{K} i\pi_i$.

15. In a queueing system of type $M/M/\infty$, there is an infinite number of servers. The arrival and departure rates satisfy

$$\lambda_i = \lambda \text{ and } \mu_i = i\mu, \quad i = 0, 1, 2, \ldots.$$

(a) Find the stationary probability distribution.

(b) What is the average number of customers C in the system?

(c) What is the average amount of time W each customer spends in the system?

16. Assume the random variable $X(t)$ represents the total population size for the stochastic logistic model, where the birth and death rates satisfy either

(i) $\lambda_i = \begin{cases} i - \dfrac{i^2}{100}, & i = 0, 1, \ldots, 100, \\ 0, & i > 100 \end{cases}$ and $\mu_i = \dfrac{i^2}{100}$, $i = 1, 2, \ldots$

(ii) $\lambda_i = i$ and $\mu_i = \frac{i^2}{50}$, $i = 0, 1, 2, \ldots$

Note that in the deterministic model, $r = 1$ and $K = 50$. Assume $X(0) = 50$.

(a) Write a computer program to simulate sample paths for the stochastic logistic models in (i) and (ii) up to time $t = 10$. Graph three sample paths for (i) and (ii).

(b) Calculate the mean and variance of 1000 sample paths at $t = 10$, $m(10)$, and $\sigma^2(10)$, for (i) and (ii).

(c) Calculate the quasistationary distributions $\{\tilde{\pi}_i\}_{i=1}^{\infty}$ for the logistic models (i) and (ii) and graph them.

(d) Calculate the mean and variance of the quasistationary distributions in (i) and (ii). How do they compare?

17. Consider a general birth and death process with

$$\lambda_i = b_1 i + b_2 i + b_3 i^2 \text{ and } \mu_i = d_1 + d_2 i + d_3 i^2$$

for $i = 0, 1, 2, \ldots$, where $d_3 \neq 0$ and $d_3 > b_3$. Assume $\lambda_i, \mu_i > 0$ for $i = 1, 2, \ldots$.

(a) Show that $d_3 > 0$ and $\lim_{t \to \infty} p_0(t) = 1$.

(b) Show that the expected time until population extinction is finite.

18. Consider the stochastic logistic model with m.g.f. $M(\theta, t)$ satisfying equation (6.27).

(a) Find the differential equation satisfied by the c.g.f., $K(\theta, t)$.

(b) Use the differential equation satisfied by $K(\theta, t)$ in part (a) to find a differential equation satisfied by the mean $m(t)$ and variance $\sigma^2(t)$.

19. The partial differential equation satisfied by the m.g.f. for the simple birth process satisfies

$$\frac{\partial M}{\partial t} = \lambda(e^{\theta} - 1)\frac{\partial M}{\partial \theta}, \quad M(\theta, 0) = e^{N\theta}.$$

(a) Differentiate the partial differential equation with respect to θ and evaluate at $\theta = 0$ to find the differential equation satisfied by the mean, $m(t)$. Solve for $m(t)$.

(b) Differentiate the partial differential equation twice with respect to θ and evaluate at $\theta = 0$ to find the differential equation satisfied by $E(X^2(t))$. Solve for $E(X^2(t))$.

(c) Find the differential equation satisfied by $\sigma^2(t)$. Then solve for $\sigma^2(t)$.

20. The following birth and death process is known as a *Prendiville process* (named after B. J. Prendiville) (Iosifescu and Tăutu, 1973). The birth and death rates are

$$\begin{aligned} \lambda_n &= \alpha\left(\frac{n_2}{n} - 1\right), \quad 0 < n_1 \le n \le n_2 \\ \mu_n &= \beta\left(1 - \frac{n_1}{n}\right), \quad 0 < n_1 \le n \le n_2, \end{aligned}$$

where α and β are positive constants. Outside of the interval $[n_1, n_2]$, the birth and death rates are zero, that is, $\lambda_n = 0 = \mu_n$ for $n < n_1$ or $n > n_2$. The p.g.f. $P(z, t)$ for this process satisfies the following partial differential equation:

$$\frac{\partial P}{\partial t} = (1 - z)(\alpha z + \beta)\frac{\partial P}{\partial z} + (z - 1)\left(\alpha n_2 + \frac{\beta n_1}{z}\right)P$$

with initial condition $P(z, 0) = z^{n_0}$ (Iosifescu and Tăutu, 1973).

(a) Show that the solution of this first-order partial differential equation is

$$P(z, t) = \frac{z^{n_1}\left[\alpha(1 - \rho(t))z + \alpha\rho(t) + \beta\right]^{n_2 - n_0}}{(\alpha + \beta)^{n_2 - n_1}\left[(\alpha + \beta\rho(t))z + \beta(1 - \rho(t))\right]^{n_1 - n_0}},$$

where $\rho(t) = e^{-(\alpha+\beta)t}$.

(b) Find $\lim_{t\to\infty} P(z, t)$ and denote this limit as $P(z, \infty)$. This limit is the p.g.f. for the stable stationary probability distribution. Use $P(z, \infty)$ to find the mean of this stationary distribution.

21. The deterministic model of the Prendiville process discussed in Exercise 20 is given by the differential equation

$$\frac{dn}{dt} = \lambda_n - \mu_n = \frac{\alpha n_2 + \beta n_1}{n} - (\alpha + \beta).$$

(a) Show that the equilibrium solution $\bar{n}$ (where $dn/dt = 0$) to this differential equation is

$$\bar{n} = \frac{\alpha n_2 + \beta n_1}{\alpha + \beta}.$$

Then show that $dn/dt > 0$ for $0 < n < \bar{n}$ and $dn/dt < 0$ for $n > \bar{n}$. Conclude that $\lim_{t\to\infty} n(t) = \bar{n}$.

(b) Compare the behavior of this deterministic model to the stochastic Prendiville process discussed in Exercise 20.

22. Consider the birth process with immigration, where $\lambda_i = b_0 + b_1 i^k$, $b_0 > 0$, $b_1 > 0$, for $i = 1, 2, \ldots$. Show that the process is not explosive if $k = 1$ and is explosive if $k = 2, 3, \ldots$.

23. Suppose a nonhomogeneous birth and death process satisfies $\lambda(i,t) = \lambda(t)i$ and $\mu(i,t) = \mu(t)i$ with $\mu(t) = t^2$ and $\lambda(t) = 2t^2$. Find $\lim_{t\to\infty} p_0(t)$.

24. Consider a nonhomogeneous birth and death process, where $\mu(t)$ and $\lambda(t)$ are linear functions of t,

$$\mu(t) = \alpha t \quad \text{and} \quad \lambda(t) = \beta t.$$

(a) Show that if $\alpha > \beta > 0$, then $\lim_{t\to\infty} \rho(t) = \infty$, and if $\alpha = \beta > 0$, $\lim_{t\to\infty} \int_0^t \mu(\tau) e^{\rho(\tau)}\, d\tau = \lim_{t\to\infty} \int_0^t \alpha\tau\, d\tau = \infty$. Find $\lim_{t\to\infty} p_0(t)$.

(b) Show that if $0 < \alpha < \beta$ (births exceed deaths), then

$$\lim_{t\to\infty} \int_0^t \mu(\tau) e^{\rho(\tau)}\, d\tau = \frac{\alpha}{(\alpha - \beta)^2}.$$

Find $\lim_{t\to\infty} p_0(t)$.

25. A birth and death process with immigration was developed by Alonso and McKane (2002) based on a spatially implicit patch model. Let $p(t)$ be the fraction of patches occupied by a particular species. Empty patches can be colonized from occupied patches or via migration from the mainland. The deterministic model is expressed in terms of the following differential equation:

$$\frac{dp}{dt} = (m + cp)(1 - p) - ep, \quad 0 < p(0) < 1. \tag{6.33}$$

The positive constants m, c, and e are the rates of immigration from the mainland, colonization, and extinction, respectively. Model (6.33) was first discussed by Hanski (1999) in the context of metapopulation models. In the original model, studied by Levins (1969, 1970), there was no migration ($m = 0$).

(a) Show that model (6.33) has a unique positive equilibrium given by

$$E_1 = \frac{c - m - e + \sqrt{(c - m - e)^2 + 4cm}}{2c}$$

(when $dp/dt = 0$) and $\lim_{t\to\infty} p(t) = E_1$.

(b) In the stochastic formulation by Alonso and McKane (2002), it is assumed that $X(t)$ is the random variable for the number of patches occupied at time t. The maximal number of patches occupied is M, $X(t) \in \{0, 1, 2, \ldots, M\}$. The birth rate of new patches occupied is

$$\lambda_i = ci\left(1 - \frac{i}{M}\right) + m(M - i)$$

for $i = 0, 1, 2, \ldots, M$, where the first term in the preceding expression represents a colonization event and the second term represents an immigration event from the mainland. The death rate satisfies

$$\mu_i = ei, \quad i = 0, 1, 2, \ldots, M,$$

where death means population extinction on one of the patches. To relate this model to equation (6.33), note that if $p = i/M$, then

$$\frac{dp}{dt} = \frac{(\lambda_i - \mu_i)}{M}.$$

For the birth, death, and immigration process, find a formula for the stationary probability distribution.

(c) Compare the equilibrium value E_1 to the stationary probability distribution when $M = 40$, $c = 2$, $e = 1$, and $m = 0.1$.

6.13 References for Chapter 6

Alonso, D. and A. McKane. 2002. Extinction dynamics in mainland-island metapopulations: an N-patch stochastic model. *Bull. Math. Biol.* 64: 913–958.

Bailey, N. T. J. 1990. *The Elements of Stochastic Processes with Applications to the Natural Sciences.* John Wiley & Sons, New York.

Bharucha-Reid, A. T. 1997. *Elements of the Theory of Markov Processes and their Applications.* Dover Pub., New York.

Chao, X, M. Miyazawa, and M. Pinedo. 1999. *Queueing Networks.* John Wiley & Sons, Chichester, New York.

Corduneanu, C. 1977. *Principles of Differential and Integral Equations.* Chelsea Pub. Co., The Bronx, New York.

Feller, W. 1968. *An Introduction to probability Theory and Its Applications.* Vol. 1. 3rd ed. John Wiley & Sons, New York.

Goel, N. S. and N. Richter-Dyn. 1974. *Stochastic Models in Biology.* Academic Press, New York.

Hanski, I. 1999. *Metapopulation Ecology.* Oxford University Press, New York.

Hsu, H. P. 1997. *Schaum's Outline of Theory and Problems of Probability, Random Variables, and Random Processes.* McGraw-Hill, New York.

Iosifescu, M. and P. Tăutu. 1973. *Stochastic Processes and Applications in Biology and Medicine II. Models.* Springer-Verlag, Berlin, Heidelberg, New York.

Karlin, S. and H. Taylor. 1975. *A First Course in Stochastic Processes,* 2nd ed. Academic Press, New York.

Kleinrock, L. 1975. *Queueing Systems, Vol. 1, Theory.* John Wiley & Sons, New York.

Levins, R. 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull. Entomol. Soc. Am.* 15: 227–240.

Levins, R. 1970. Extinction. *Lecture Notes Math.* 2: 75–107.

Matis, J. H. and T. R. Kiffe. 1999. Effects of immigration on some stochastic logistic models: a cumulant truncation analysis. *Theor. Pop. Biol.* 56: 139–161.

Nåsell, I. 2001. Extinction and quasi-stationarity in the Verhulst logistic model. *J. Theor. Biol.* 211: 11–27.

Nisbet, R. M. and W. S. C. Gurney. 1982. *Modelling Fluctuating Populations.* John Wiley & Sons, Chichester and New York.

Norden, R. H. 1982. On the distribution of the time to extinction in the stochastic logistic population model. *Adv. Appl. Prob.* 14: 687–708.

Norris, J. R. 1999. *Markov Chains.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Ortega, J. M. 1987. *Matrix Theory A Second Course.* Plenum Press, New York.

Renshaw, E. 1993. *Modelling Biological Populations in Space and Time.* Cambridge University Press, Cambridge.

Richter-Dyn, N. and N. S. Goel. 1972. On the extinction of colonizing species. *Theor. Pop. Biol.* 3: 406-433.

Rudin, W. 1987. *Real and Complex Analysis.* 3rd ed. McGraw-Hill, New York.

Schinazi, R. B. 1999. *Classical and Spatial Stochastic Processes.* Birkhäuser, Boston.

Taylor, H. M. and S. Karlin. 1998. *An Introduction to Stochastic Modeling.* 3rd ed. Academic Press, New York.

Tognetti, K. and G. Winley. 1980. Stochastic growth models with logistic mean population. *J. Theor. Biol.* 82: 167–169.

6.14 Appendix for Chapter 6

6.14.1 Generating Functions for the Simple Birth and Death Process

The m.g.f. and the p.g.f. for the simple birth and death process are found by solving the following first-order partial differential equations. The p.g.f. satisfies

$$\frac{\partial P}{\partial t} = [\mu(1-z) + \lambda z(z-1)]\frac{\partial P}{\partial z}, \quad P(z,0) = z^N$$

and the m.g.f. satisfies

$$\frac{\partial M}{\partial t} = [\mu(e^{-\theta} - 1) + \lambda(e^{\theta} - 1)]\frac{\partial M}{\partial \theta}, \quad M(\theta, 0) = e^{\theta N}.$$

Application of the method of characteristics to the m.g.f. equation leads to

$$\frac{dt}{d\tau} = 1, \quad \frac{d\theta}{\mu(e^{-\theta} - 1) + \lambda(e^{\theta} - 1)} = -d\tau, \quad \text{and} \quad \frac{dM}{d\tau} = 0,$$

with initial conditions

$$t(s,0) = 0, \quad \theta(s,0) = s, \quad \text{and} \quad M(s,0) = e^{sN}.$$

The solutions satisfy

$$t = \tau \quad \text{and} \quad M(s,\tau) = e^{sN}.$$

Integrating the differential equation in θ (integration is made easier by a change of variable $x = e^{\theta}$),

$$\tau = \begin{cases} \dfrac{1}{\mu-\lambda}\left(\ln\left(\dfrac{e^{\theta}-1}{\lambda e^{\theta}-\mu}\right)+\ln(\tau_1)\right), & \text{if } \lambda \neq \mu \\ \dfrac{1}{\lambda(e^{\theta}-1)}+\tau_2, & \text{if } \lambda = \mu, \end{cases}$$

where τ_1 and τ_2 are constants. The two cases, $\lambda = \mu$ and $\lambda \neq \mu$, must be solved separately. The initial condition for θ is used to solve for the constants,

$$\tau = \begin{cases} \dfrac{1}{\mu-\lambda}\left[\ln\dfrac{(e^{\theta}-1)(\lambda e^{s}-\mu)}{(\lambda e^{\theta}-\mu)(e^{s}-1)}\right], & \text{if } \lambda \neq \mu \\ \dfrac{1}{\lambda(e^{\theta}-1)}-\dfrac{1}{\lambda(e^{s}-1)}, & \text{if } \lambda = \mu. \end{cases}$$

Because $M(s,\tau) = [e^s]^N$, these relations are solved for e^s,

$$e^s = \begin{cases} \dfrac{e^{\tau(\mu-\lambda)}(\lambda e^{\theta}-\mu)-\mu(e^{\theta}-1)}{e^{\tau(\mu-\lambda)}(\lambda e^{\theta}-\mu)-\lambda(e^{\theta}-1)}, & \text{if } \lambda \neq \mu \\ \dfrac{1-(\lambda\tau-1)(e^{\theta}-1)}{1-\lambda\tau(e^{\theta}-1)}, & \text{if } \lambda = \mu. \end{cases}$$

Now the m.g.f. M can be expressed in terms of θ and t,

$$M(\theta,t) = \begin{cases} \left(\dfrac{e^{t(\mu-\lambda)}(\lambda e^{\theta}-\mu)-\mu(e^{\theta}-1)}{e^{t(\mu-\lambda)}(\lambda e^{\theta}-\mu)-\lambda(e^{\theta}-1)}\right)^N, & \text{if } \lambda \neq \mu \\ \left(\dfrac{1-(\lambda t-1)(e^{\theta}-1)}{1-\lambda t(e^{\theta}-1)}\right)^N, & \text{if } \lambda = \mu. \end{cases}$$

Making the change of variable $\theta = \ln z$, the p.g.f. P is

$$P(z,t) = \begin{cases} \left(\dfrac{e^{t(\mu-\lambda)}(\lambda z-\mu)-\mu(z-1)}{e^{t(\mu-\lambda)}(\lambda z-\mu)-\lambda(z-1)}\right)^N, & \text{if } \lambda \neq \mu \\ \left(\dfrac{1-(\lambda t-1)(z-1)}{1-\lambda t(z-1)}\right)^N, & \text{if } \lambda = \mu. \end{cases}$$

6.14.2 Proofs of Theorems 6.2 and 6.3

Proof of Theorem 6.2. Note that in a general birth and death chain

$$\frac{dp_0(t)}{dt} = \mu_1 p_1(t).$$

Since $0 \leq p_i(t) \leq 1$, it follows that $p_0(t)$ is an increasing function that is bounded above. Thus, $\lim_{t\to\infty} p_0(t)$ exists.

For part (i) of the theorem, let E_i be the probability of extinction given the population size is i (see Karlin and Taylor, 1975; Renshaw, 1993). The ratio $\lambda_i/(\mu_i + \lambda_i)$ is the probability of a birth and the ratio $\mu_i/(\mu_i + \lambda_i)$ is the probability of a death given that an event has occurred (transition probabilities in the embedded Markov chain). Then

$$\begin{aligned} E_i &= \text{Prob}\{\text{first event is a birth}\}E_{i+1} + \text{Prob}\{\text{first event is a death}\}E_{i-1} \\ &= \frac{\lambda_i}{\mu_i + \lambda_i} E_{i+1} + \frac{\mu_i}{\mu_i + \lambda_i} E_{i-1}. \end{aligned}$$

Because $\lambda_0 = 0 = \mu_0$, $E_0 = 1$. Rewriting the preceding expression,

$$\begin{aligned} E_{i+1} &= \frac{\mu_i + \lambda_i}{\lambda_i}\left[E_i - \frac{\mu_i}{\mu_i + \lambda_i} E_{i-1}\right] \\ &= \left(1 + \frac{\mu_i}{\lambda_i}\right) E_i - \frac{\mu_i}{\lambda_i} E_{i-1}. \end{aligned}$$

For $i = 1$ and 2,

$$\begin{aligned} E_2 &= \left(1 + \frac{\mu_1}{\lambda_1}\right) E_1 - \frac{\mu_1}{\lambda_1} \\ &= E_1 + (E_1 - 1)\frac{\mu_1}{\lambda_1}. \\ E_3 &= \left(1 + \frac{\mu_2}{\lambda_2}\right) E_2 - \frac{\mu_2}{\lambda_2} E_1 \\ &= \left(1 + \frac{\mu_2}{\lambda_2}\right)\left(E_1 + (E_1 - 1)\frac{\mu_1}{\lambda_1}\right) - \frac{\mu_2}{\lambda_2} E_1 \\ &= E_1 + (E_1 - 1)\left(\frac{\mu_1}{\lambda_1} + \frac{\mu_1\mu_2}{\lambda_1\lambda_2}\right). \end{aligned}$$

By induction it follows that

$$E_n = E_1 + (E_1 - 1)\left(\sum_{i=1}^{n-1} \frac{\mu_1 \cdots \mu_i}{\lambda_1 \cdots \lambda_i}\right) \tag{6.34}$$

for all $n = 2, 3, \ldots$. Also, $0 \leq E_n \leq 1$ for $n = 1, 2, \ldots$. Let $n \to \infty$. If the factor multiplied by $(E_1 - 1)$ approaches infinity, then $E_1 = 1$. But

if $E_1 = 1$, then $E_n = E_1$, for $n > 1$, so that $E_n = 1$, the probability of extinction is 1, given the population size is n. Hence, if (6.14) holds, then

$$\lim_{t\to\infty} p_0(t) = 1.$$

Suppose (6.15) holds so that a solution $0 < E_1 < 1$ of (6.34) exists. Then E_n is a decreasing function of n. According to Karlin and Taylor (1975) and Nisbet and Gurney (1982), $\lim_{n\to\infty} E_n = 0$. Let $n \to \infty$ in (6.34). Then

$$E_1 = \frac{\sum_{i=1}^{\infty} \frac{\mu_1\mu_2\cdots\mu_i}{\lambda_1\lambda_2\cdots\lambda_i}}{1 + \sum_{i=1}^{\infty} \frac{\mu_1\mu_2\cdots\mu_i}{\lambda_1\lambda_2\cdots\lambda_i}}.$$

Substitution of E_1 into (6.34) with $n = m$ yields (6.16).

For part (ii) of the theorem, if the initial population size $m > N$, there will be only deaths until the population size reaches N. The states $\{N+1, N+2, \ldots\}$ are transient. Once the population size reaches N, the population size will remain less than or equal to N for all time; $\{0, 1, 2, \ldots, N\}$ is a closed set. Thus, to find the probability of extinction, we need only consider population sizes $\{0, 1, \ldots, N\}$. Assume $p_n(t) = 0$ for $n > N$ and $t > T$, T sufficiently large. The system of differential equations satisfied by $p(t) = (p_0(t), \ldots, p_N(t))^T$ for $t > T$ is

$$\begin{aligned}
\frac{dp_0(t)}{dt} &= \mu_1 p_0(t) \\
\frac{dp_n(t)}{dt} &= \lambda_{n-1}p_{n-1}(t) - (\lambda_n + \mu_n)p_n(t) + \boldsymbol{\mu_{n+1}p_{n+1}(t)} \\
\frac{dp_N(t)}{dt} &= \lambda_{N-1}p_{N-1}(t) - \mu_N p_N(t),
\end{aligned}$$

for $n = 1, 2, \ldots, N-1$, where $0 \le p(T) \le 1$ and $\sum_{i=0}^{N} p_0(T) = 1$. Note that for this system of differential equations, $d[\sum_{i=0}^{N} p_i(t)]/dt = 0$, so that $\sum_{i=0}^{N} p_i(t) =$ constant. Because $\sum_{i=0}^{N} p_0(T) = 1$, it follows that $\sum_{i=0}^{N} p_i(t) = 1$. The forward Kolmogorov differential equations satisfy $dp/dt = Qp$, where matrix Q is

$$Q = \begin{pmatrix}
0 & \mu_1 & 0 & \cdots & 0 & 0 \\
0 & -\lambda_1 - \mu_1 & \mu_2 & \cdots & 0 & 0 \\
0 & \lambda_1 & -\lambda_2 - \mu_2 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \lambda_{N-1} & -\mu_N
\end{pmatrix}.$$

The solution is $p(t) = e^{Qt}p(0)$. The limit as $t \to \infty$ depends on the eigenvalues and eigenvectors of Q. Matrix Q has one zero eigenvalue with corresponding eigenvector $e_1 = (1, 0, 0, \ldots, 0)^T$. Matrix Q also has N other

eigenvalues that have negative real part. To show this latter assertion, we can apply Gershgorin's circle theorem and irreducible diagonal dominance to the submatrix of Q formed by deleting the first row and column (Ortega, 1987). Thus, the solution $p(t)$ satisfies $\lim_{t\to\infty} p(t) = c_0 e_1$. Because $\sum_{i=0}^{N} p_i(t) = 1$, it follows that $c_0 = 1$ and, hence,

$$\lim_{t\to\infty} p_0(t) = 1. \qquad \square$$

Proof of Theorem 6.3. Let $z_i = \tau_i - \tau_{i+1} \le 0$ and subtract τ_i from both sides of equation (6.21). Then

$$\begin{aligned} 0 &= \frac{1}{\lambda_i + \mu_i} + \frac{\lambda_i}{\lambda_i + \mu_i}(\tau_{i+1} - \tau_i) + \frac{\mu_i}{\lambda_i + \mu_i}(\tau_{i-1} - \tau_i) \\ \tau_i - \tau_{i+1} &= \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i}(\tau_{i-1} - \tau_i) \\ z_i &= \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i} z_{i-1}. \end{aligned}$$

By induction, it follows that

$$\begin{aligned} z_m &= \frac{1}{\lambda_m} + \frac{\mu_m}{\lambda_{m-1}\lambda_m} + \cdots \frac{\mu_2 \cdots \mu_m}{\lambda_1 \lambda_2 \cdots \lambda_m} + \frac{\mu_1 \cdots \mu_m}{\lambda_1 \cdots \lambda_m} z_0 \\ &= \frac{\mu_1 \cdots \mu_m}{\lambda_1 \cdots \lambda_m}\left[\frac{1}{\mu_1} + \frac{\lambda_1}{\mu_1 \mu_2} + \cdots + \frac{\lambda_1 \cdots \lambda_{m-1}}{\mu_1 \cdots \mu_m} + z_0\right] \\ &= \frac{\mu_1 \cdots \mu_m}{\lambda_1 \cdots \lambda_m}\left[\sum_{i=1}^{m} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} - \tau_1\right], \end{aligned}$$

since $z_0 = \tau_0 - \tau_1 = -\tau_1$. Then

$$\frac{\lambda_1 \cdots \lambda_m}{\mu_1 \cdots \mu_m} z_m = \frac{1}{\mu_1} + \sum_{i=2}^{m} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} - \tau_1. \tag{6.35}$$

Suppose

$$\sum_{i=2}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} = \infty.$$

Then $\tau_1 = \infty$. **But** since $\{\tau_i\}_{i=1}^{\infty}$ is a nondecreasing sequence, it follows that $\tau_m = \infty$.

Suppose

$$\sum_{i=2}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} < \infty.$$

Then, for large m, deaths are much greater than births, so that $z_m \to \tau_m - \tau_{m+1} \approx 1/\mu_{m+1}$ as $m \to \infty$, which is the mean time for a death to

occur when the population size is $m+1$ (see Nisbet and Gurney, 1982). Then let $m \to \infty$ so that the left side of (6.35) is

$$\frac{\lambda_1 \cdots \lambda_m}{\mu_1 \cdots \mu_m} z_m \to \frac{\lambda_1 \cdots \lambda_m}{\mu_1 \cdots \mu_m \mu_{m+1}} \to 0$$

(Karlin and Taylor, 1975; Nisbet and Gurney, 1982). Since the expression on the left of (6.35) approaches zero,

$$\tau_1 = \frac{1}{\mu_1} + \sum_{i=2}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i}$$

and

$$z_m = -\frac{\mu_1 \cdots \mu_m}{\lambda_1 \cdots \lambda_m} \left[\sum_{i=m+1}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} \right].$$

Now,

$$\begin{aligned} \tau_m - \tau_1 &= -\sum_{s=1}^{m-1} z_s \\ &= \sum_{s=1}^{m-1} \left[\frac{\mu_1 \cdots \mu_s}{\lambda_1 \cdots \lambda_s} \sum_{i=s+1}^{\infty} \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} \right]. \end{aligned}$$

The value of τ_m is given by (6.22). □

6.14.3 Comparison Theorem

Theorem 6.5 (Comparison Theorem). *Suppose that $m(t)$ and $n(t)$ are continuous with continuous derivatives on $[0, \infty)$ and $0 < m(0) = n(0)$. If*

$$\frac{dm(t)}{dt} < f(m(t)) \quad \text{and} \quad \frac{dn(t)}{dt} = f(n(t)), \quad \text{for} \quad t \in [0, \infty),$$

then

$$m(t) \leq n(t) \quad \text{for} \quad t \in [0, \infty).$$

Proof. At $t = 0$, $d[m(t) - n(t)]/dt < 0$. By continuity of the derivatives, it follows that $d[m(t) - n(t)]/dt \leq 0$ on some interval $[0, T)$, $T > 0$, so that $m(t) \leq n(t)$ for $t \in [0, T)$. We show $T = \infty$. Suppose the interval $[0, T)$ is the largest possible such that $m(t) \leq n(t)$ for $t \in [0, T)$. Thus, $m(T) = n(T)$ and $m(t) > n(t)$ for some interval $(T, T + \epsilon_1)$, $\epsilon_1 > 0$. However, at $t = T$, $f(m(T)) = f(n(T))$, so that $d[m(t) - n(t)]/dt < 0$. Therefore, $m(t) < n(t)$ for some interval $(T, T + \epsilon_2)$, $0 < \epsilon_2 < \epsilon_1$, a contradiction. □

Consult Corduneanu (1977) for a discussion of comparison results for scalar differential equations.

Chapter 7

Epidemic, Competition, Predation and Population Genetics Processes

7.1 Introduction

In this chapter, continuous time Markov chain models for a variety of biological processes are discussed–in particular, epidemic, competition, predation, and population genetics processes. These models are based on deterministic models, which can be expressed as systems of ordinary differential equations. First the deterministic model will be introduced and discussed. Then the stochastic model will be formulated and discussed.

We begin by defining a continuous time branching process, an extension of a discrete time branching process discussed in Chapter 4. There are many recent biological applications of branching processes, especially in the fields of cellular and molecular biology (see, for example, Kimmel and Axelrod, 2002). Branching processes are introduced in Section 7.2, and several biological examples are presented.

In Section 7.3, we discuss SI and SIS epidemic models. These models involve two random variables, but since the total population size is assumed constant, they can be reduced to a single random variable and the techniques from Chapter 6 can be applied. The SI epidemic model is a birth process, and the SIS epidemic model is a birth and death process. Stochastic models for many biological applications are multivariate processes. Some notation and terminology associated with multivariate processes are reviewed. Then, in Sections 7.5, 7.6, 7.7 and 7.8, multivariate processes for epidemic, competition, predation, and other population processes are presented.

7.2 Continuous Time Branching Processes

The study of continuous time branching processes is an active field of research and has many biological applications related to population, cellular, molecular, and gene dynamics (Jagers, 1975; Kimmel and Axelrod, 2002). Continuous time branching processes are an extension of discrete time branching processes discussed in Chapter 4. In the Galton-Watson process discussed in Chapter 4, an individual's lifetime is a fixed length of time, which for convenience is denoted as one unit of time or one generation, $\tau = 1$. At the end of that time interval, the individual is replaced by his progeny (if we are speaking of male heirs). However, in the continuous time process, an individual's lifetime is not fixed but may have an arbitrary distribution. This process is known as an age-dependent branching process. In the case of an exponentially distributed lifetime, the branching process is *Markovian* and, if not, it is a general *age-dependent* process known as a *Bellman-Harris* branching process (Bharucha-Reid, 1997; Harris, 1963; Kimmel and Axelrod, 2002). Here, we shall briefly discuss these two types of continuous time branching processes.

First, consider a Markov branching process, where the age of an individual has an exponential distribution with parameter λ. Let τ be the random variable for the lifetime of an individual. Then the cumulative distribution function for τ is

$$G(t) = \text{Prob}\{\tau \leq t\} = 1 - e^{-\lambda t}$$

and the p.d.f. is $g(t) = G'(t) = \lambda e^{-\lambda t}$. Assume a single individual is born at time $t = 0$ and lives for a period of time τ. Prior to death, the individual produces a random number of progeny according to the p.g.f. $f(z)$, where

$$f(z) = \sum_{k=0}^{\infty} p_k z^k.$$

Each of the progeny behaves independently, lives for an exponential period of time, and produces a random number of progeny, following the same distributions as the original individual. The process continues with subsequent generations behaving in the same manner. See Figure 7.1.

Let $X(t)$ be the total population size at time t; then $X(t)$ for $t \geq 0$ is a continuous time Markov process. Assume at $t = 0$, $X(0) = 1$. Let $P(z,t)$ be the p.g.f. for $X(t)$. We shall derive a differential equation satisfied by P. Let Δt be sufficiently small. Then, as in discrete time branching processes, the p.g.f. $P(z, t + \Delta t)$ is a composition of p.g.f.'s; that is, it can be shown that

$$P(z, t + \Delta t) = P(P(z,t), \Delta t). \tag{7.1}$$

(Kimmel and Axelrod, 2002). At time $t = 0$, $P(z,0) = z$. For a small

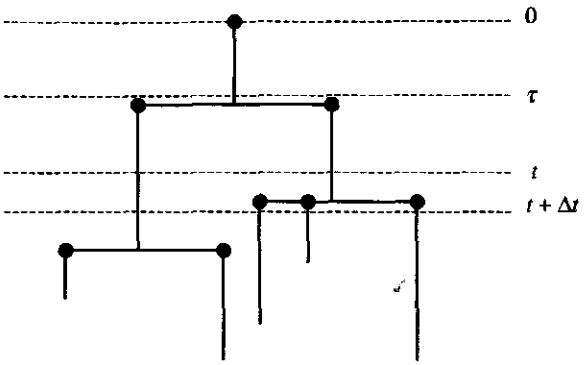

Figure 7.1. A sample path of a continuous time branching process. At time τ, an individual gives birth to two individuals and these individuals give birth to two and three individuals, respectively.

period of time Δt,

$$\begin{aligned} P(z,\Delta t) &= zProb\{\tau > \Delta t\} + f(z)\text{Prob}\{\tau \le \Delta t\} + \boldsymbol{o(\Delta t)} \\ &= ze^{-\lambda\Delta t} + f(z)(1-e^{-\lambda\Delta t}) + o(\Delta t). \end{aligned} \tag{7.2}$$

Subtracting $P(z,t)$ from (7.1) and applying the identity in **(7.2),**

$$\begin{aligned} P(z,t+\Delta t) - P(z,t) &= P(P(z,t),\Delta t) - P(z,t) \\ &= P(z,t)e^{-\lambda\Delta t} + f(P(z,t))(1-e^{-\lambda\Delta t}) \\ &\quad -P(z,t) + o(\Delta t) \\ &= [-P(z,t) + f(P(z,t))](1-e^{-\lambda\Delta t}) + o(\Delta t). \end{aligned}$$

Dividing by Δt and letting $\Delta t \to 0$, the following differential equation is obtained:

$$\frac{\partial P(z,t)}{\partial t} = -\lambda[P(z,t) - f(P(z,t))]. \tag{7.3}$$

We can treat the ordinary derivative as a partial derivative because there is no explicit dependence on z in the right-hand side of (7.3). Equation (7.3) with initial condition $P(z,0) = z$ has a unique solution provided $\lim_{z\to 1^-} P(z,t) = 1$ for all times [i.e., the process does not explode] (Kimmel and Axelrod, 2002).

As was the case for discrete time branching processes, the mean number of births determines the asymptotic behavior. Let $m = f'(1)$. The asymptotic behavior depends on whether $m \le 1$ or $m > 1$. Theorem 4.1 in Chapter 4 can be extended to the continuous time process when the process is nonexplosive. This is stated in the next theorem. For a proof of this result, please see Harris (1963).

Theorem 7.1. *Suppose $X(t)$, $t \geq 0$, is a nonexplosive, continuous time, Markov branching process with $X(0) = 1$. Assume $f(z)$ is the p.g.f. of the birth process, where $m = f'(1)$ and $P(z,t)$ is the p.g.f. of $X(t)$. If $m \leq 1$, then*

$$\lim_{t\to\infty} \text{Prob}\{X(t) = 0\} = \lim_{t\to\infty} p_0(t) = 1$$

and if $\mathbf{m > 1}$, there exists a q satisfying $f(q) = q$ such that

$$\lim_{t\to\infty} \text{Prob}\{X(t) = 0\} = \lim_{t\to\infty} p_0(t) = q < 1.$$

The next example expresses the simple birth and death process as a continuous time Markov branching process.

Example 7.1 Suppose the cumulative distribution for the lifetime distribution is $G(t) = 1 - e^{-(\lambda+\mu)t}$ and the p.g.f. for the birth process is

$$f(z) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} z^2,$$

where $\lambda > 0$ is the birth rate and $\mu > 0$ is the death rate. Either an individual dies or survives and gives birth. The lifetime distribution is exponential with parameter $\lambda + \mu$. The differential equation satisfied by $P(z,t)$ takes the form

$$\frac{dP(z,t)}{dt} = -(\lambda+\mu)P(z,t) + (\lambda+\mu)\left[\frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} P^2(z,t)\right].$$

We have used the notation for an ordinary derivative with respect to t rather than a partial derivative because the problem can be treated as an ordinary differential equation. Simplifying,

$$\frac{dP}{dt} = \mu - (\lambda+\mu)P + \lambda P^2,$$

with initial condition $P(z,0) = z$. This differential equation can be solved by separation of variables, yielding the solution

$$P(z,t) = \begin{cases} \dfrac{e^{t(\mu-\lambda)}(\lambda z - \mu) - \mu(z-1)}{e^{t(\mu-\lambda)}(\lambda z - \mu) - \lambda(z-1)}, & \text{if } \lambda \neq \mu \\ \dfrac{\lambda t(z-1) - z}{\lambda t(z-1) - 1}, & \text{if } \lambda = \mu. \end{cases} \tag{7.4}$$

See Chapter 6, Section 6.4.3. It follows that

$$\lim_{t\to\infty} p_0(t) = \begin{cases} \dfrac{\mu}{\lambda}, & \text{if } \lambda > \mu \\ 1, & \text{if } \lambda \leq \mu. \end{cases}$$

■

In addition, a multitype branching process can be defined. Assume there are k types of individuals, $X_i(t)$ is the random variable for the number of individuals of type i, $i = 1, \ldots, k$, and $X(t) = (X_1(t), \ldots, X_k(t))^T$. Let $f_i(z_1, \ldots, z_k)$ denote the birth p.g.f. and $P_i(z_1, \ldots, z_k, t)$ denote the p.g.f. for an individual of type i, $i = 1, \ldots, k$. Assume the cumulative lifetime distribution for an individual is exponential with parameter λ, $G(t) = 1 - e^{-\lambda t}$. It can be shown that each of the P_i satisfy a differential equation of the following form:

$$\frac{dP_i}{dt} = -\lambda P_i + \lambda f_i(P_1, \ldots, P_k), \quad i = 1, \ldots, k.$$

A multitype process involving the development of drug resistance in cancer cells is discussed in the next example (Kimmel and Axelrod, 2002).

Example 7.2 Assume there are two types of cells. Type 1 cells are sensitive to a particular drug and type 2 cells are resistant to that drug. Assume a population of cancer cells begins with a single type 1 cell. The time it takes for a cell to divide is a random variable τ with an exponential distribution having parameter λ. At each division of a type 1 sensitive cell, with probability p, one of the two daughter cells mutates and becomes resistant to the drug (type 2). A type 2 resistant cell produces two daughter cells that are both resistant (type 2) (i.e., mutations are irreversible). Then the birth p.g.f.'s satisfy

$$f_1(z_1, z_2) = (1-p)z_1^2 + pz_1z_2, \quad \text{and} \quad f_2(z_1, z_2) = z_2^2.$$

The differential equations satisfied by $X(t) = (X_1(t), X_2(t))^T$ are

$$\begin{aligned} \frac{dP_1}{dt} &= -\lambda P_1 + \lambda[(1-p)P_1^2 + pP_1P_2] \\ \frac{dP_2}{dt} &= -\lambda P_2 + \lambda P_2^2. \end{aligned}$$

The initial conditions are $P_1(z_1, z_2, 0) = z_1$ and $P_2(z_1, z_2, 0) = z_2$. The differential equation for P_2 can be solved by separation of variables (this is a logistic-type differential equation). Then the solution P_2 can be substituted into the differential equation for P_1 to find the solution for P_1. It can be shown that the solutions satisfy

$$P_1(z_1, z_2, t) = \frac{z_1 e^{-\lambda t}[z_2 e^{-\lambda t} + 1 - z_2]^{-p}}{1 + z_1([e^{-\lambda t}z_2 + 1 - z_2]^{1-p} - 1)z_2^{-1}} \tag{7.5}$$

$$P_2(z_1, z_2, t) = \frac{z_2}{z_2 + (1 - z_2)e^{\lambda t}} \tag{7.6}$$

(Kimmel and Axelrod, 2002). When the tumor is first identified, it is important to find out what proportion of the cells are resistant. Of course,

it is hoped that there are no resistant cells. The probability there are no resistant cells at time t can be computed as follows:

$$\lim_{z_1 \to 1} \lim_{z_2 \to 0} P_1(z_1, z_2, t) = \frac{1}{1 - p + pe^{\lambda t}}.$$

Eventually, as $t \to \infty$, all cells will be resistant. Therefore, it is important to discover the tumor very early, that is, when t is small. ■

The general age-dependent branching process is known as a *Bellman-Harris process*, named after Richard Bellman and Theodore Harris, the first investigators of this type of process (Harris, 1963). In the Bellman-Harris process, the cumulative lifetime distribution, $G(t) = \text{Prob}\{\tau \leq t\}$, has a general distribution. Age-dependent branching processes belong to a general class of stochastic processes known as *regenerative processes* (see, e.g., Beichelt and Fatti, 2002; Ross, 1983).

An integral equation can be derived for the p.g.f. $P(z, t)$. Note that $P(z, t)$ satisfies

$$P(z,t) = \begin{cases} z, & t < \tau \\ f(P(z, t-\tau)), & t \geq \tau. \end{cases}$$

Then the integral equation for $P(z, t)$ is

$$P(z,t) = z(1 - G(t)) + \int_0^t f(P(z, t-u))\, dG(u). \tag{7.7}$$

In the special case where the lifetime distribution is exponential (Markov branching process), it can be shown that the integral equation (7.7) reduces to the differential equation (7.3). Let $G(t) = 1 - e^{-\lambda t}$. Then

$$P(z,t) = ze^{-\lambda t} + \int_0^t f(P(z, t-u))\lambda e^{-\lambda u}\, du.$$

Multiplying the preceding equation by $e^{\lambda t}$ and then making a change of variable $v = t - u$ in the integral leads to

$$e^{\lambda t} P(z,t) = z + \lambda \int_0^t f(p(z,v))e^{\lambda v}\, dv.$$

Differentiating the last equation with respect to t leads to

$$e^{\lambda t}\left[\lambda P(z,t) + \frac{dP}{dt}\right] = \lambda f(P(z,t))e^{\lambda t}.$$

Finally, solving for dP/dt leads to the differential equation (7.3).

We end this section by presenting an example of a multitype Bellman-Harris process applied to brain cell differentiation. Since there are two types of cells in this process, the integral equation (7.7) is extended to a system of two integral equations.

Example 7.3 This example is based on a paper by Yakovlev et al. (1998). Brain cell development begins with precursor cells. A precursor cell divides and proliferates into daughter cells or transforms into another type of cell that does not divide and proliferate. Brain cell differentiation can be described simply by two types of cells in the central nervous system, the precursor cell, known as the progenitor cell (type 1 cell), which may be transformed into oligodendrocyte cells (type 2 cells). The progenitor cell is a stem cell, whereas an oligodendrocyte cell is a cell responsible for producing a fatty protein known as myelin, which insulates nerve cell axons. Myelinated axons are able to transmit nerve signals faster than unmyelinated ones. For example, in multiple sclerosis, a disease of the central nervous system characterized by neurological dysfunction, oligodendrocyte cells are often destroyed.

In the model of Yakovlev et al. (1998), it is assumed that the process begins with a single type 1 cell. The type 1 cell divides and produces two daughter cells of the same type with probability p or transforms into a single type 2 cell with probability $1-p$. A type 2 cell neither divides nor proliferates. Let $G(t)$ denote the cumulative lifetime distribution for a type 1 cell. Based on these assumptions, the birth p.g.f.'s $f_1(z_1, z_2)$ and $f_2(z_1, z_2)$ can be expressed as follows:

$$f_1(z_1, z_2) = pz_1^2 + (1-p)z_2 \quad \text{and} \quad f_2(z_1, z_2) = z_2.$$

The p.g.f.'s for the number of type 1 and type 2 cells, $X_1(t)$ and $\boldsymbol{X_2(t)}$, satisfy the following integral equations:

$$\begin{aligned} P_1(z_1, z_2, t) &= z_1(1-G(t)) + p\int_0^t P_1^2(z_1, z_2, t-u)\, dG(u) \\ &\quad + (1-p)\int_0^t P_2(z_1, z_2, t-u)\, dG(u) \\ P_2(z_1, z_2, t) &= z_2. \end{aligned}$$

This process can be studied further by assuming $G(t)$ has a gamma distribution (Yakovlev et al., 1998). ■

Extensive biological applications and more thorough discussions about continuous time branching processes can be found in the books by Harris (1963), Jagers (1975), Kimmel and Axelrod (2002), and Mode (1971).

7.3 SI and SIS Epidemic Processes

First, the dynamics of the deterministic SI and SIS epidemic model are reviewed. Recall that $S(t)$ = number of susceptible individuals, $I(t)$ = number of individuals infected, and N =the total population size, where $N = S(t) + I(t)$ is constant. It is assumed that infected individuals are

infectious (no latent period). Therefore, an infected individual will also be referred to as infective. The SI epidemic model has been applied to diseases such as influenza or the common cold, where generally no one is immune and over the course of the epidemic everyone eventually becomes infected. The SIS epidemic model has been applied to sexually transmitted diseases, where there is recovery but no immunity; individuals can become infected immediately following recovery (see discussion of the SIS epidemic model in Chapter 3).

The differential equations for the SI epidemic model satisfy

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI \\ \frac{dI}{dt} &= \frac{\beta}{N}SI, \end{aligned}$$

$S(0) + I(0) = N$. The parameter β = contact rate, the number of contacts that result in an infection of a susceptible by one infectious individual. The SIS epidemic model has an additional parameter for recovery, γ. Sometimes a birth and death rate is included, b. Since the population size is constant the birth rate equals the death rate. The differential equations for the SIS epidemic model are

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI + (\gamma + b)I \\ \frac{dI}{dt} &= \frac{\beta}{N}SI - (\gamma + b)I, \end{aligned}$$

$S(0) + I(0) = N$. It is easy to see that adding the differential equations gives $d(S + I)/dt = 0$ so that $S(t) + I(t) = N$.

For the SI epidemic model, substitution of $S = N - I$ into the differential equation for S yields a logistic differential equation for I. This differential equation can be solved directly via separation of variables; the solution $I(t)$ satisfies

$$I(t) = \frac{NI(0)}{I(0) + (N - I(0))e^{-\beta t}}. \tag{7.8}$$

Eventually, everyone becomes infected, $\lim_{t\to\infty} I(t) = N$.

The SIS epidemic model has an endemic equilibrium given by

$$S^* = \frac{\gamma + b}{\beta}N, \quad I^* = \frac{N(\beta - \gamma - b)}{\beta}.$$

The endemic equilibrium is positive if the *basic reproduction number*,

$$\mathcal{R}_0 = \frac{\beta}{b + \gamma}, \tag{7.9}$$

satisfies $\mathcal{R}_0 > 1$. If $\mathcal{R}_0 > 1$, then solutions approach the endemic equilibrium and if $\mathcal{R}_0 \leq 1$ solutions approach the disease-free state.

	N			
β	10	100	1000	10000
1	4.605	9.210	13.816	18.421
10	0.461	0.921	1.382	1.842
100	0.046	0.092	0.138	0.184

Table 7.1. Approximate duration, T, until the entire population is infected in the deterministic SI epidemic model for various population sizes N and contact rates β when $I(0) = 1$

The contact rate β may be a function of population size. If, for example, $\beta = cN$, then the expression $cNSI/N = cSI$ in the SI or SIS epidemic models is referred to as *mass action incidence rate*, and when β does not depend on N, the expression is referred to as the *standard incidence rate*. In models where the population size is constant, this distinction makes little difference because β is constant in either case. However, the form of β can have a significant impact on the population dynamics when the population size is not constant.

For both models, the infected population size satisfies $0 \leq I(t) \leq N$. In the stochastic models, it will be shown that $I = N$ is the unique absorbing state for the SI epidemic model and $I = 0$ is the unique absorbing state for the SIS epidemic model. The expected duration until absorption is calculated. For comparison purposes, the duration of time until the infected population size reaches N is approximated in the deterministic model. The time T until $I(T) = N$ is actually infinite in the deterministic model because N is approached asymptotically. However, to obtain a realistic estimate of the time to absorption, solve for T in the identity, $I(T) = N - 1$. Using the formula (7.8) for $I(t)$ and solving the following equation for T,

$$N - 1 \approx \frac{NI(0)}{I(0) + (N - I(0))e^{-\beta T}}$$

yields

$$T \approx \frac{\ln[(N - I(0))(N - 1)]}{\beta}.$$

Table 7.1 gives the approximate duration until the entire population is infected for various values of N and β when the initial population size is one, $I(0) = 1$.

7.3.1 Stochastic SI Epidemic Model

The stochastic SI epidemic model is a birth process. Let $I(t)$ denote the random variable for the number of individuals infected at time t. The state

space for $I(t)$ is $\{0, 1, 2 \ldots, N\}$. The transition probabilities satisfy

$$\text{Prob}\{\Delta I(t) = j | I(t) = i\} = \begin{cases} \frac{\beta}{N} i(N-i)\Delta t + o(\Delta t), & j = 1 \\ 1 - \frac{\beta}{N} i(N-i)\Delta t + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq 0, 1. \end{cases}$$

The forward Kolmogorov equations, $dp/dt = Qp$, have generator matrix

$$Q = \begin{pmatrix} -\beta(N-1)/N & 0 & \cdots & 0 & 0 \\ \beta(N-1)/N & -\beta 2(N-2)/N & \cdots & 0 & 0 \\ 0 & \beta 2(N-2)/N & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\beta(N-1)/N & 0 \\ 0 & 0 & \cdots & \beta(N-1)/N & 0 \end{pmatrix}.$$

It is easy to see that the transition matrix of the embedded Markov chain satisfies

$$T = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{pmatrix}.$$

State N is absorbing. The unique stationary probability distribution is $\pi = (0, 0, \ldots, 0, 1)^T$. The stationary probability distribution is the limiting distribution, $\lim_{t\to\infty} p(t) = \lim_{t\to\infty} e^{Qt} p(0) = (0, 0, \ldots, 1)^T$.

The time until absorption beginning from state 1 satisfies

$$T_{N,1} = \sum_{i=1}^{N-1} T_{i+1,i}.$$

The expected duration until absorption or until the entire population is infected is given by

$$E(T_{N,1}) = \sum_{i=1}^{N} E(T_{i+1,i}).$$

The time between events $i \to i+1$ is exponentially distributed with parameter $\beta i(N-i)/N$. Therefore, the mean time between events i and $i+1$ is $E(T_{i+1,i}) = N/(\beta i(N-i))$. The expected duration until the entire population is infected satisfies

$$E(T_{N,1}) = \sum_{i=1}^{N-1} \frac{N}{\beta i(N-i)} = \frac{1}{\beta} \sum_{i=1}^{N-1} \left[\frac{1}{i} + \frac{1}{N-i} \right] = \frac{2}{\beta} \sum_{i=1}^{N-1} \frac{1}{i}.$$

The variance of this distribution also has a simple expression:

$$\mathrm{Var}(T_{N,1}) = \sum_{i=1}^{N-1} \mathrm{Var}(T_{i+1,i}),$$

since the random variables for the time $T_{i+1,i}$ are independent (the states at each jump are known). Recall that the variance of an exponentially distributed random variable with parameter λ is $1/\lambda^2$; in this case, $\lambda = \beta i(N-i)/N$. It follows that

$$\begin{aligned}
\mathrm{Var}(T_{N,1}) &= \frac{1}{\beta^2} \sum_{i=1}^{N-1} \frac{N^2}{[i(N-i)]^2} \\
&= \frac{1}{\beta^2} \sum_{i=1}^{N-1} \left[\frac{2/N}{i} + \frac{1}{i^2} + \frac{2/N}{N-i} + \frac{1}{(N-i)^2} \right] \\
&= \frac{1}{\beta^2} \sum_{i=1}^{N-1} \left[\frac{4}{N}\frac{1}{i} + \frac{2}{i^2} \right].
\end{aligned}$$

For large N, the mean and variance can be approximated using the following two identities:

$$\lim_{N\to\infty} \left[\sum_{i=1}^{N} \frac{1}{i} - \ln(N) \right] = \gamma \quad \text{and} \quad \lim_{N\to\infty} \sum_{i=1}^{N} \frac{1}{i^2} = \frac{\pi^2}{6},$$

where γ is Euler's constant, $\gamma \approx 0.5772156649\ldots$. Thus, for large N, approximations for the mean and variance satisfy

$$\begin{aligned}
E(T_{N,1}) &\approx \frac{2}{\beta}[\gamma + \ln(N)] \approx \frac{2}{\beta}\ln(N-1) \\
\mathrm{Var}(T_{N,1}) &\approx \frac{1}{\beta^2}\left[\frac{4}{N}(\gamma + \ln(N-1)) + \frac{\pi^2}{3} \right] \\
&\approx \frac{\pi^2}{3\beta^2} \approx \frac{3.290}{\beta^2}.
\end{aligned}$$

For a large population size, the variance does not depend on N. Notice that the approximate mean duration equals the approximate duration derived from the deterministic model when $I(0) = 1$. Table 7.2 compares the approximate mean duration to the exact mean duration and also gives the exact variance for the distribution for the time until the entire population is infected. For the parameter $\beta = 1$, $\mathrm{Var}(T_{N,1}) \approx 3.290$. We can see that the approximations are closer to the exact values as N increases.

The time units for an epidemic depend on the particular disease and the population being modeled. For humans, the time units are on the order of of days, weeks, or months. For example, if $\beta = 1$ successful contact/day,

Mean and Variance	N			
	10	100	1000	10000
Approx. $E(T_{N,1})$	4.605	9.210	13.816	18.421
Exact $E(T_{N,1})$	5.658	10.355	14.969	19.579
Exact $Var(T_{N,1})$	4.211	3.477	3.318	3.294

Table 7.2. Approximate and exact mean durations $E(T_{N,1})$ and the exact variance $\mathrm{Var}(T_{N,1})$ of the distribution for the time until absorption in the SI stochastic epidemic model, $I(0) = 1$ and $\beta = 1$

then if $I(0) = 1$ and $N = 1000$, the approximate duration until all 1000 individuals are infected would be, on the average, 15 days. We must realize that the SI epidemic model is oversimplified. We have not included, for example, a latent period where individuals have been exposed to the disease but are not yet infectious nor the possibility of recovery or immunity.

7.3.2 Stochastic SIS Epidemic Model

In this section, we assume that individuals can recover from the disease but do not develop immunity. They immediately become susceptible again. The SIS epidemic model discussed here is the continuous analogue of the discrete time epidemic model discussed in Chapter 3. Let the transition probabilities satisfy

$$\begin{aligned}\mathrm{Prob}\{\Delta I(t) &= j | I(t) = i\} \\ &= \begin{cases} \dfrac{\beta}{N} i(N-i)\,\Delta t + o(\Delta t), & j = 1 \\ (b+\gamma) i\,\Delta t + o(\Delta t), & j = -1 \\ 1 - \left[\dfrac{\beta}{N} i(N-i) + (b+\gamma)\right]\Delta t & \\ \quad + o(\Delta t), & j = 0 \\ o(\Delta t), & j \neq -1, 0, 1, \end{cases}\end{aligned}$$

where $i \in \{0, 1, \ldots, N\}$. The SIS epidemic model is a birth and death process with

$$\lambda_i = \max\left\{0, \frac{\beta}{N} i(N-i)\right\} \quad \text{and} \quad \mu_i = (\gamma + b)i,$$

for $i = 0, 1, \ldots, N$. There is a single absorbing state at zero, $\lim_{t\to\infty} p_0(t) = 1$.

If $\mathcal{R}_0 > 1$, then the SIS epidemic model is a special case of the logistic model considered in Chapter 6. In this case, $\lambda_n = b_1 n + b_2 n^2$ and $\mu_n = d_1 n + d_2 n^2$, where $b_1 = \beta$, $b_2 = -\beta/N$, $d_1 = b + \gamma$ and $d_2 = 0$. In order that

the requirement $\lambda_n - \mu_n = rn(1 - n/K)$ in the logistic model be satisfied,

$$r = \beta - (b + \gamma) > 0 \quad \text{and} \quad -r/K = \beta/N < 0.$$

The expression $r > 0$ is equivalent to $\mathcal{R}_0 > 1$. A generator matrix Q and transition matrix T can be defined easily.

When $\mathcal{R}_0$ and N are large, the time until absorption can be very long. The expected duration until extinction satisfies $D\tau = \mathbf{d}$, where $\tau = (\tau_1, \ldots, \tau_N)^T$ and τ_i is the expected duration given $I(0) = i$. Matrix D and vector $\mathbf{d}$ are defined in Chapter 6 and are given here for reference purposes, $\mathbf{d} = (-1, -1, \ldots, -1)^T$ and

$$D = \begin{pmatrix} -\lambda_1 - \mu_1 & \lambda_1 & 0 & \cdots & 0 & 0 \\ \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & \cdots & 0 & 0 \\ 0 & \mu_3 & -\lambda_3 - \mu_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu_N & -\mu_N \end{pmatrix}.$$

Matrix D is nonsingular (irreducibly diagonally dominant) and, hence, $\tau = D^{-1}\mathbf{d}$. See also Theorem 6.3 for an explicit expression for τ.

Prior to extinction, the probability distribution for the infected population has a quasistationary distribution. Recall that the approximate quasistationary distribution, $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_N)^T$, satisfies

$$\tilde{\pi}_{i+1} = \frac{\lambda_i}{\mu_{i+1}} \tilde{\pi}_i, \quad i = 1, 2, \ldots, N - 1.$$

Example 7.4 Let $\beta = 2$, $N = 100$, and $\gamma + b = 1$. Then $\mathcal{R}_0 = 2$. Graphs of τ and $\tilde{\pi}$ are given in Figure 7.2. The expected time until population extinction when $N = 100$ is on the order of 10^8. The mean and standard deviation of the approximate quasistationary probability distribution $\tilde{\pi}$ are approximately 48.9 and 7.2, respectively. Notice that the mean of $\tilde{\pi}$ is close to the value of the deterministic endemic equilibrium, $I^* = N(1 - 1/\mathcal{R}_0) = 50$. In addition, the distribution of $\tilde{\pi}$ is approximately normal. ■

7.4 Multivariate Processes

In all of the continuous time Markov chains discussed thus far, the processes have been univariate, concerned only with a single random variable $X(t)$. In this section, the notation for multivariate processes will be introduced, processes for which there are two or more dependent random variables. (See also Chapter 1.)

For competition and predation processes, where there is more than one population, models need to be formulated with several random variables,

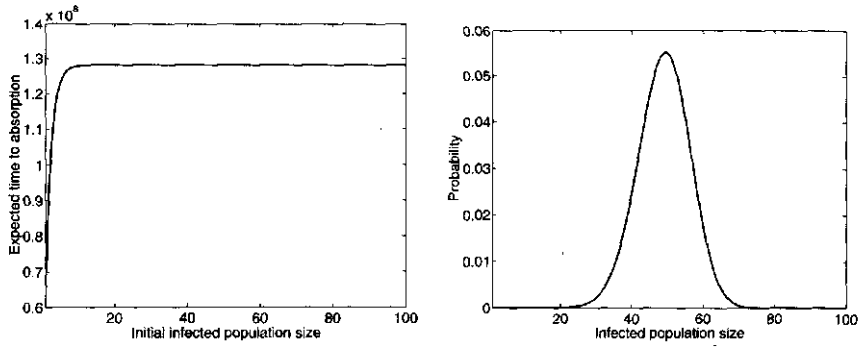

Figure 7.2. Expected duration until absorption at $I = 0$ and the quasistationary distribution of the stochastic SIS model, $\beta = 2$, $b + \gamma = 1$, and $N = 100$.

one random variable corresponding to each of the populations. For simplicity, consider a bivariate process. Let $(X(t), Y(t))$ for $t \geq 0$ denote a continuous time, bivariate Markov process, where $X(t)$ and $Y(t) \in \{0, 1, 2, \ldots\}$. The *joint probability mass function* (or *joint p.d.f.*) is

$$p_{(m,n)}(t) = \text{Prob}\{X(t) = m, Y(t) = n\}.$$

Be careful not to confuse this notation with the notation used in previous sections for the transition from state n to state m in a univariate process. The transition probability for the bivariate process is denoted as

$$p_{(m,n),(i,j)}(\Delta t) = \text{Prob}\{X(t + \Delta t) = m, Y(t + \Delta t) = n | X(t) = i, Y(t) = j\}.$$

The process is assumed to be homogeneous in time (i.e., the transition probabilities only depend on the length of time between transitions and do not depend on the time at which they occur).

The probability generating function satisfies

$$P(w, z, t) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} p_{(m,n)}(t) w^m z^n$$

and the moment and cumulant generating functions satisfy $M(\theta, \phi, t) = P(e^\theta, e^\phi, t)$ and $K(\theta, \phi, t) = \ln M(\theta, \phi, t)$. The marginal probability distributions of $X(t)$ and $Y(t)$ are

$$\sum_{n=0}^{\infty} p_{(m,n)}(t) \quad \text{and} \quad \sum_{m=0}^{\infty} p_{(m,n)}(t),$$

respectively. Their means and variances are

$$m_X(t) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} m p_{(m,n)}(t), \quad m_Y(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} n p_{(m,n)}(t),$$

$$\sigma_X^2(t) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} m^2 p_{(m,n)}(t) - m_X^2(t),$$

and

$$\sigma_Y^2(t) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} n^2 p_{(m,n)}(t) - m_Y^2(t).$$

The means and higher order moments can be obtained from the generating functions; for example,

$$\left.\frac{\partial M}{\partial \theta}\right|_{\theta=0=\phi} = m_X(t), \quad \left.\frac{\partial M}{\partial \phi}\right|_{\theta=0=\phi} = m_Y(t),$$

$$\left.\frac{\partial^2 M}{\partial \theta^2}\right|_{\theta=0=\phi} = E(X^2(t)), \text{ and } \left.\frac{\partial^2 M}{\partial \phi^2}\right|_{\theta=0=\phi} = E(Y^2(t)).$$

Forward Kolmogorov differential equations can be derived from the transition probabilities in a manner similar to the univariate process. Let S be a finite subset of $Z \times Z$, $\Delta X(t) = X(t+\Delta t) - X(t)$ and $\Delta Y(t) = Y(t+\Delta t) - Y(t)$, where $Z = \{0, \pm 1, \pm 2, \ldots\}$ is the set of integers. Then

$$\begin{aligned}\text{Prob}\{\Delta X(t) &= k, \Delta Y(t) = l | (X(t), Y(t))\} \\ &= \begin{cases} h_{kl}(X(t), Y(t))\,\Delta t + o(\Delta t), & (k,l) \in S \\ 1 - \sum\limits_{(i,j)\in S} h_{ij}(X(t), Y(t))\,\Delta t + o(\Delta t), & (k,l) \notin S. \end{cases}\end{aligned}$$

For $(X(t), Y(t)) = (i, j)$, the preceding transition probability is denoted as $p_{(m,n),(i,j)}(\Delta t)$, where $m = i + k$ and $n = j + l$. Thus, the joint p.d.f. satisfies

$$\begin{aligned}p_{(m,n)}(t+\Delta t) &= \sum_{(j,k)\in S} p_{(m-j,n-k)}(t) h_{jk}(m-j, n-k)\,\Delta t \\ &\quad + p_{(m,n)}(t)\left[1 - \sum_{(j,k)\in S} h_{jk}(m,n)\,\Delta t\right] + o(\Delta t).\end{aligned}$$

Subtract $p_{(m,n)}(t)$ and divide by Δt to obtain the forward Kolmogorov differential equation for the bivariate process,

$$\frac{dp_{(m,n)}}{dt} = -p_{(m,n)} \sum_{(j,k)\in S} h_{jk}(m,n) + \sum_{(j,k)\in S} p_{(m-j,n-k)} h_{jk}(m-j, n-k).$$

Differential equations for the probability or moment generating functions can be derived using the generating function technique.

Example 7.5 Suppose $S = \{(1,0),(0,1),(-1,0),(0,-1)\}$; that is, for Δt small, the bivariate process can increase or decrease by one in each of its components. In addition, let $h_{10}(X,Y) = \lambda_1 X$, $h_{01}(X,Y) = \lambda_2 Y$, $h_{-1,0}(X,Y) = \mu_1 X$, and $h_{0,-1}(X,Y) = \mu_2 Y$ (Bailey, 1990). Then the forward Kolmogorov differential equation has the form

$$\begin{aligned}\frac{dp_{(m,n)}}{dt} &= -p_{(m,n)}\left[\lambda_1 m + \lambda_2 n + \mu_1 m + \mu_2 n\right] \\ &\quad + p_{(m-1,n)}\lambda_1(m-1) + p_{(m,n-1)}\lambda_2(n-1) \\ &\quad + p_{(m+1,n)}\mu_1(m+1) + p_{(m,n+1)}\mu_2(n+1). \qquad (7.10)\end{aligned}$$

In addition, the moment generating function has the form

$$\begin{aligned}\frac{\partial M}{\partial t} &= \left[\lambda_1(e^{\theta}-1) + \mu_1(e^{-\theta}-1)\right]\frac{\partial M}{\partial \theta} \\ &\quad + \left[\lambda_2(e^{\phi}-1) + \mu_2(e^{-\phi}-1)\right]\frac{\partial M}{\partial \phi}, \qquad (7.11)\end{aligned}$$

where $M(\theta,\phi,0) = e^{N_1\theta + N_2\phi}$, $X(0) = N_1$, and $Y(0) = N_2$. Using the differential equation for the m.g.f., it can be shown that the mean of $X(t)$ and $Y(t)$ satisfy

$$\frac{dm_X(t)}{dt} = (\lambda_1 - \mu_1)m_X(t), \quad \text{and} \quad \frac{dm_Y(t)}{dt} = (\lambda_2 - \mu_2)m_Y(t). \qquad (7.12)$$

■

Numerical simulations of multivariate Markov chain processes can be performed in a manner similar to univariate processes. The interevent time is exponential. Suppose the process is in state (i,j) at time t; then, assuming the process can jump to at most a finite number of states, the time until the next event has an exponential distribution with parameter $\sum_{(k,l)\in S} h_{k,l}(i,j)$. For example, the bivariate process in Example 7.5 has an exponential interevent time distribution with parameter $(\lambda_1 + \mu_1)i + (\lambda_2 + \mu_2)j$ when the process is in state (i,j).

7.5 SIR Epidemic Process

First, the deterministic Susceptible-Infected-Removed (SIR) model is reviewed. In the SIR epidemic model, individuals recover and develop permanent immunity. The class R represents the individuals that are permanently immune. Such types of models have been applied to childhood diseases such as measles, mumps, and chickenpox (see, e.g., Allen and Thrasher, 1998; Allen, Jones, and Martin, 1991; Anderson and May, 1992; Hethcote, 2000 and references therein). The differential equations for the SIR epidemic

model are given by

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI \\ \frac{dI}{dt} &= \frac{\beta}{N}SI - \gamma I \\ \frac{dR}{dt} &= \gamma I, \end{aligned}$$

where $S(0), I(0), R(0) \geq 0$, and $S(0)+I(0)+R(0) = N$. It can be shown that $\lim_{t\to\infty} I(t) = 0$. Also, $\lim_{t\to\infty} S(t) = S(\infty)$ and $\lim_{t\to\infty} R(t) = R(\infty)$ are finite but depend on initial conditions. The *effective reproduction number or replacement number* is defined as

$$\mathcal{R} = \frac{S(0)}{N}\frac{\beta}{\gamma}$$

(Anderson and May, 1992; Hethcote, 2000). Hethcote (2000) defines the replacement number, $\mathcal{R}$, as the average number of secondary infections produced by a typical infective during the entire period of infectiousness. Recall that $\mathcal{R}_0$ is the average number of secondary infections that occur when one infective is introduced into a completely susceptible population (Hethcote, 2000). Notice that $\mathcal{R} = \mathcal{R}_0 S(0)/N$. If $\mathcal{R} \leq 1$, then in the SIR model there is no epidemic; solutions $I(t)$ decrease monotonically to zero. But if $\mathcal{R} > 1$; then $I(t)$ increases first before decreasing to zero; an epidemic occurs.

Although the epidemic eventually ends, the severity of the epidemic can be determined by the total number of cases or the *final size* of the epidemic. If we assume $R(0) = 0$, then $R(\infty)$ represents the total number of cases or final size. Suppose $I(0) = 1$; then the number of cases equals the initial case plus all others infected during the course of the epidemic. The value of $R(\infty)$ can be obtained from the differential equations dI/dt and dS/dt,

$$\frac{dI}{dS} = \frac{(\beta/N)SI - \gamma I}{-(\beta/N)SI} = -1 + \frac{N\gamma}{\beta S}.$$

Separating variables, integrating, and applying the initial conditions,

$$S(0) = N - 1 \quad \text{and} \quad I(0) = 1,$$

the following solution is obtained:

$$I(t) + S(t) = \frac{N\gamma}{\beta}\ln S(t) + N - \frac{N\gamma}{\beta}\ln(N-1).$$

Letting $t \to \infty$ and using the fact that $I(\infty) = 0$ yields

$$S(\infty) = \frac{N\gamma}{\beta}\ln\left(\frac{S(\infty)}{N-1}\right) + N. \tag{7.13}$$

	N		
β	20	100	1000
0.5	1.87	1.97	2.00
1	5.74	13.52	44.07
2	16.26	80.02	797.15
5	19.87	99.31	993.03
10	20.00	100.00	999.95

Table 7.3. The final size of an SIR epidemic, $R(\infty)$, **when** $\gamma = 1$, $S(0) = N - 1$, and $I(0) = 1$

Equation (7.13) gives an implicit solution for $S(\infty)$, which can be used to find the value of $R(\infty) = N - S(\infty)$. The next example uses this formula to calculate the final size $R(\infty)$ for particular parameter values.

Example 7.6 Let $N = 100$, $\beta = 2$, and $\gamma = 1$. Also, $S(0) = 99$, $I(0) = 1$, and $R(0) = 0$. Then $S(\infty) = 50 \ln(S(\infty)/99) + 100$. The solution $S(\infty) \approx 19.98$ so that the final size satisfies $R(\infty) \approx 80.02$. Table 7.3 gives the final size for the SIR epidemic model for various parameter values. ■

As the replacement number $\mathcal{R}$ increases, the number of cases increases until the entire population is infected. When $\mathcal{R} \leq 1$ ($\beta \leq 1$ in Table 7.3), there is no epidemic, so the total number of cases is relatively small. The data in Table 7.3 will be compared to the stochastic SIR model.

7.5.1 Stochastic SIR Epidemic Model

Let $S(t)$, $I(t)$, and $R(t)$ denote random variables for the number of susceptible, infected, and immune individuals, respectively, where $S(t) + I(t) + R(t) = N$. There is no latent period so that infected individuals are also infectious. Only two of the random variables are independent. Assume the transition probabilities satisfy

$$\begin{aligned}
\text{Prob}\{\Delta S(t) &= i, \Delta I(t) = j | (S(t), I(t))\} \\
&= \begin{cases}
\dfrac{\beta}{N} S(t) I(t)\, \Delta t + o(\Delta t), & (i,j) = (-1,1) \\
\gamma I(t)\, \Delta t + o(\Delta t), & (i,j) = (0,-1) \\
1 - \left[\dfrac{\beta}{N} S(t) I(t) + \gamma I(t)\right] \Delta t & \\
\quad + o(\Delta t), & (i,j) = (0,0) \\
o(\Delta t), & \text{otherwise.}
\end{cases}
\end{aligned}$$

For example, when $\Delta I(t) = -1$, then $\Delta R(t) = 1$.

Assume the initial distribution is $(S(0), I(0)) = (s_0, i_0)$, where $s_0 + i_0 = N$, $s_0 \geq 0$ and $i_0 > 0$. Let $p_{(i,j)}(t) = \text{Prob}\{S(t) = i, I(t) = j\}$; then the

state probabilities satisfy the forward Kolmogorov equations:

$$\frac{dp_{(i,j)}(t)}{dt} = \frac{\beta}{N}(i+1)(j-1)p_{(i+1,j-1)}(t) + \gamma(j+1)p_{(i,j+1)}(t) - \left[\frac{\beta}{N}ij + \gamma j\right] p_{(i,j)}(t),$$

where $i = 0, 1, 2, \ldots, N$, $j = 0, 1, 2, \ldots, N - i$, and $i + j \leq N$. If (i, j) lies outside of this range, the probabilities are assumed to be zero. For example, for $j = 0$,

$$\frac{dp_{(i,0)}}{dt} = \gamma p_{(i,1)}, \quad \text{and} \quad \frac{dp_{(N,0)}}{dt} = 0$$

for $i = 0, 1, 2, \ldots, N - 1$. The $N + 1$ states $(i, 0)$, where $i = 0, 1, 2, \ldots, N$, represent a set of closed states. There are no transitions out of any one of these states.

It was found in the deterministic SIR epidemic model that the occurrence of an epidemic depends on the replacement number

$$\mathcal{R} = \frac{S(0)}{N}\frac{\beta}{\gamma}.$$

When $S(0) = N - j \approx N$ and $I(0) = j$ is small, the replacement number $\mathcal{R} \approx \beta/\gamma = \mathcal{R}_0$. The model can be related to the simple birth and death process. Death of an infected individual corresponds to recovery, $\mu = \gamma$, and birth of an infected individual corresponds to a new infection, $\lambda \approx \beta$. At the beginning of an epidemic, when $S(0) = N - j \approx N$ and $I(0) \approx j$ is small, then the probability that the epidemic ends quickly or that there is no epidemic can be approximated by a simple birth and death process,

$$\text{probability no epidemic} = \begin{cases} 1, & \mathcal{R}_0 \leq 1 \\ \left(\dfrac{1}{\mathcal{R}_0}\right)^j, & \mathcal{R}_0 > 1. \end{cases}$$

[See also Daley and Gani (1999) and Whittle (1955).] If, for example, $N = 100$, $\mathcal{R}_0 = 2$, and $I(0) = 2$, an epidemic occurs with probability $3/4 = 1 - (1/\mathcal{R}_0)^2$ and no epidemic with probability $1/4 = (1/\mathcal{R}_0)^2$.

Even though the process is bivariate, an expression for the generator matrix Q and transition matrix T corresponding to the embedded Markov chain can be obtained. The form of matrices Q and T depends on how the states are ordered. There are $(N + 1)(N + 2)/2$ pairs of states in the SIR epidemic process. Order these pairs of states as follows:

$$(N,0), (N-1,0), \ldots, (0,0), (N-1,1), (N-2,1), \ldots, (0,1), (N-2,2), (N-3,2), \ldots, (0,2), \ldots, (0,N). \tag{7.14}$$

Then $p(t) = (p_{(N,0)}, \ldots, p_{(0,N)})^T$ and the generator matrix Q of $dp/dt = Qp$ depend on this particular ordering of the states. Matrix Q is a $(N+1)(N+$

$2)/2 \times (N+1)(N+2)/2$ matrix with the elements in the first $N+1$ columns zero because there are no transitions out of any states with zero infectives; they are absorbing states. This, in turn, means that the transition matrix T corresponding to the embedded Markov chain, $T = (t_{kl})$, satisfies $t_{ll} = 1$ for $l = 1, 2, \ldots, N+1$, the first $N+1$ states corresponding to $(i, 0)$, $i = 0, 1, 2, \ldots, N+1$. The transition matrix T of the embedded Markov chain is useful in calculating the distribution for the final size of the epidemic.

7.5.2 Final Size of the Epidemic

Explicit formulas for calculating the elements of the transition matrix T of the embedded Markov chain are given. In the embedded Markov chain, there are transitions from state (i, j) to either state $(i+1, j-1)$ representing a susceptible that becomes infected or to $(i, j-1)$ representing a recovery of an infected individual. The probability of recovery is

$$p_i = \frac{\gamma j}{\gamma j + (\beta/N)ij} = \frac{\gamma}{\gamma + (\beta/N)i}. \tag{7.15}$$

The probability that a susceptible becomes infected is

$$1 - p_i = \frac{(\beta/N)ij}{\gamma j + (\beta/N)ij} = \frac{(\beta/N)i}{\gamma + (\beta/N)i}.$$

In addition, it can be seen that the embedded Markov chain satisfies

$$p_{(i,j)} = \begin{cases} p_i p_{(i,j+1)}, & j = 0, 1 \\ p_i p_{(i,j+1)} + (1 - p_{i+1}) p_{(i+1,j-1)}, & j = 2, \ldots, N, \end{cases}$$

with the restriction that $0 \le i + j \le N$; otherwise the probabilities are **zero** (Daley and Gani, 1999).

Example 7.7 A stochastic SIR epidemic model with population size $N = 4$ has 15 states. The states and the directed digraph corresponding to the embedded Markov chain are graphed in Figure 7.3. Group the states into five sets corresponding to the ordering in (7.14):

I: $(4,0)$, $(3,0)$, $(2,0)$, $(1,0)$, $(0,0)$
II: $(3,1)$, $(2,1)$, $(1,1)$, $(0,1)$
III: $(2,2)$, $(1,2)$, $(0,2)$
IV: $(1,3)$, $(0,3)$
V: $(0,4)$.

The **transition matrix** of the embedded chain has **the following block form:**

$$T = \begin{pmatrix} I & A_1 & 0 & 0 & 0 \\ 0 & 0 & A_2 & 0 & 0 \\ 0 & B_1 & 0 & A_3 & 0 \\ 0 & 0 & B_2 & 0 & A_4 \\ 0 & 0 & 0 & B_3 & 0 \end{pmatrix},$$

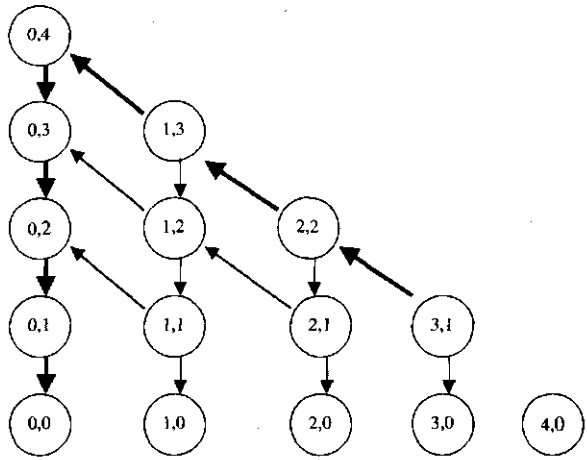

Figure 7.3. Directed graph of the embedded Markov chain of the SIR epidemic model with $N = 4$. The maximum path length beginning from state $(3,1)$ is indicated by the thick arrows.

corresponding to the grouping into the five sets. The state probability vector can also be divided into the five sets $p = (p_{(i,j)}) = (p_{\mathrm{I}}, p_{\mathrm{II}}, p_{\mathrm{III}}, p_{\mathrm{IV}}, p_{\mathrm{V}})^T$, where, for example, $p_I = (p_{(4,0)}, p_{(3,0)}, p_{(2,0)}, p_{(1,0)}, p_{(0,0)})^T$. Each of the block matrices in T has different dimensions and represents different transitions between these sets. Matrix I is a 5×5 identity matrix, which means this set is absorbing. Matrix A_j represents recovery, transitions from j infected individuals to $j-1$ infected individuals, $j = 1, 2, 3, 4$, and matrix B_j represents infection, transitions from j infected individuals to $j+1$ infected individuals. For example, matrices A_1 and B_1 have the following forms:

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ p_3 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 \\ 0 & 0 & p_1 & 0 \\ 0 & 0 & 0 & p_0 \end{pmatrix}$$

and

$$B_1 = \begin{pmatrix} 1-p_3 & 0 & 0 & 0 \\ 0 & 1-p_2 & 0 & 0 \\ 0 & 0 & 1-p_1 & 0 \end{pmatrix}.$$

If the initial state is $(3,1)$, then the maximal number of transitions until absorption is 7. If we follow the path of the thick arrows represented in the digraph, it is easy to see that there are 7 transitions. In general, for any population of size N, beginning with one infective or $p_{(N-1,1)}(0) = 1$, the maximal number of transitions until absorption is $2N - 1$. ■

Suppose, initially, the number of infected individuals is 1 and there are no immune individuals, $S(0) = N - 1$, $I(0) = 1$, and $R(0) = 0$. The probabilities associated with the final size of the epidemic $\{p_j^f\}$, $j = 1, 2, \ldots, N$, can be determined by computing the absorption probabilities. Because the states with no infected individuals are absorbing and $I(0) = 1$, $\lim_{t\to\infty} \sum_{i=0}^{N-1} p_{(i,0)}(t) = 1$. Thus, the probabilities for the final size distribution satisfy

$$\lim_{t\to\infty} p_{(i,0)}(t) = p_{N-i}^f$$

for $i = 0, 1, 2, \ldots, N-1$. If there are i susceptible individuals when the number of infected individuals has reached zero, the final size of the epidemic is $N - i$. Beginning with one infected individual, the maximal number of time steps until absorption is $2N - 1$. Hence, the absorption probabilities can be found using the transition matrix T of the embedded Markov chain (Daley and Gani, 1999). In particular,

$$\lim_{t\to\infty} p(t) = p(2N-1) = T^{2N-1} p(0). \tag{7.16}$$

Example 7.8 Let $I(0) = 1$ and $S(0) = N-1$. The distribution for the final size of the epidemic is computed for $N = 20$ and $N = 100$ for $\mathcal{R}_0 = 0.5$, 2, and 5, where $\gamma = 1$ using formula (7.16). A MATLAB program for computing the final size is given in the Appendix for Chapter 7. The largest probabilities are confined to the tails of the distribution, either near 1 or N (see Figure 7.4). The largest part of the distribution is near 0 when $\mathcal{R}_0$ is less than 1 and near N when $\mathcal{R}_0$ is greater than 1 and sufficiently large. This distribution agrees with the conclusion derived from the preceding approximation; that is, when $\mathcal{R}_0 \leq 1$, there are no epidemics, so the final size of the epidemic should be small. When $\mathcal{R}_0 > 1$, the probability no

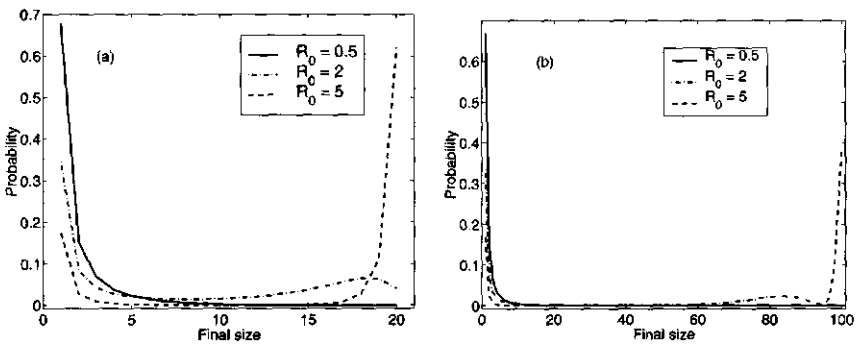

Figure 7.4. Probability distribution for the final size of a stochastic SIR epidemic model when $I(0) = 1$, $S(0) = N - 1$, $\gamma = 1$, and $\beta = 0.5$, 2, and 5 ($\mathcal{R}_0 = 0.5$, 2, and 5). In (a), $N = 20$ and in (b), $N = 100$.

	N	
β	20	100
0.5	1.76	1.93
1	3.34	6.10
2	8.12	38.34
5	15.66	79.28
10	17.98	89.98

Table 7.4. The mean of the final size of a stochastic SIR epidemic with $\gamma = 1$, $S(0) = N - 1$, and $I(0) = 1$. Compare with Table 7.3.

epidemic occurs is approximately $1/\mathcal{R}_0$, so approximately $1 - 1/\mathcal{R}_0$ of the epidemics should be of large size. ■

The mean values of the final size distributions are given in Table 7.4. Notice that the mean values for the stochastic SIR epidemic model do not agree with the final size calculated for the deterministic model in Example 7.6, Table 7.3, especially for $\mathcal{R}_0 > 1$. This difference is due to the fact that there is a positive probability of no epidemic ($1/\mathcal{R}_0$) in the stochastic model, but, in the deterministic model, solutions always approach an endemic equilibrium.

7.5.3 Expected Duration of an SIR Epidemic

The expected duration of an SIR epidemic can be calculated in a manner similar to the SI epidemic. Let $\tau_{(i,j)}$ be the expected duration of an SIR epidemic given that there are i susceptible individuals and j infectives. Notice that $\tau_{(i,0)} = 0$. It can be seen that $\tau_{(i,j)}$ satisfies

$$\tau_{(i,j)} = \eta_{ij} + p_i \tau_{(i,j-1)} + (1 - p_i)\tau_{(i-1,j+1)},$$

where $\eta_{ij} = 1/[\gamma j + (\beta/N)ij]$ is the mean interevent time given the state of the process is (i, j), $i = 0, 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N - i$. Reordering terms,

$$[\gamma j + (\beta/N)ij][p_i \tau_{(i,j-1)} - \tau_{(i,j)} + (1 - p_i)\tau_{(i-1,j+1)}] = -1.$$

The system of equations is linear, $D\tau = \mathbf{d}$, where matrix D is a $(N+1)(N+2)/2 \times (N+1)(N+2)/2$ nonsingular matrix. (The form of D depends on the specific ordering of the states.) The solution for the expected duration satisfies $\tau = D^{-1}\mathbf{d}$.

The last example illustrates several sample paths of the stochastic SIR epidemic model, the approximate duration of an epidemic and the expectation and standard deviation of the duration for different parameter values and initial conditions. A MATLAB program that generated the three sample paths in Figure 7.5 is given in the Appendix for Chapter 7.

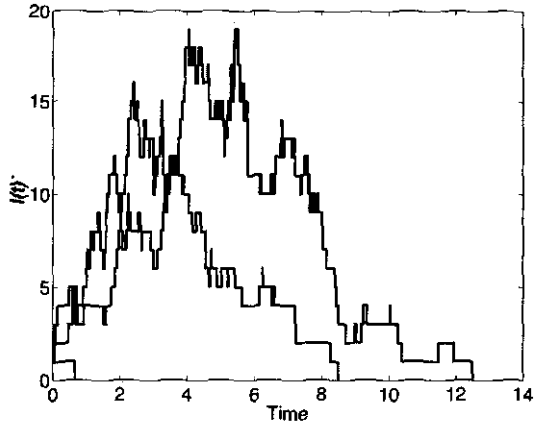

Figure 7.5. Three sample paths of a stochastic SIR epidemic model with $N = 100$, $\beta = 2$, $\gamma = 1$, $S(0) = 99$, and $I(0) = 1$ are graphed; $\mathcal{R}_0 = 2$.

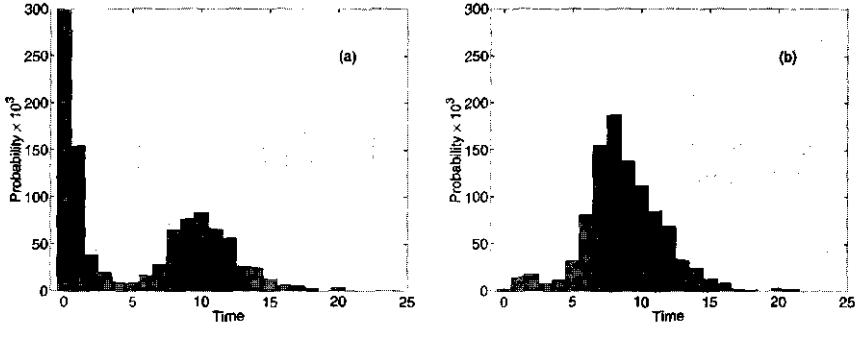

Figure 7.6. The probability distribution for the duration of an SIR epidemic, $N = 100$, $\beta = 2$, and $\gamma = 1$ (estimated from 1000 sample paths). In (a), $I(0) = 1$ and $S(0) = 99$ and in (b), $I(0) = 5$ and $S(0) = 95$.

Example 7.9 Let $N = 100$, $\beta = 2$, $\gamma = 1$, $S(0) = 99$, and $I(0) = 1$. Three sample paths are graphed in Figure 7.5. The durations of the three sample paths are 0.65, 8.49, and 12.49; two of the sample paths represent large epidemics. Since the probability of no epidemic is approximately $(1/\mathcal{R}_0) = 1/2$, approximately half of the sample paths will be large epidemics. In addition, the distribution for the duration of the epidemic is estimated from 1000 sample paths for two cases $I(0) = 1$ and $S(0) = 99$, and $I(0) = 5$ and $S(0) = 95$ in Figure 7.6. Finally, the expected duration of an SIR epidemic and the corresponding standard deviation are graphed as a function of the initial number of infectives, $I(0) = i$ and $S(0) = 100 - i$, $i = 1, 2, \ldots, 100$ in Figure 7.7 ■

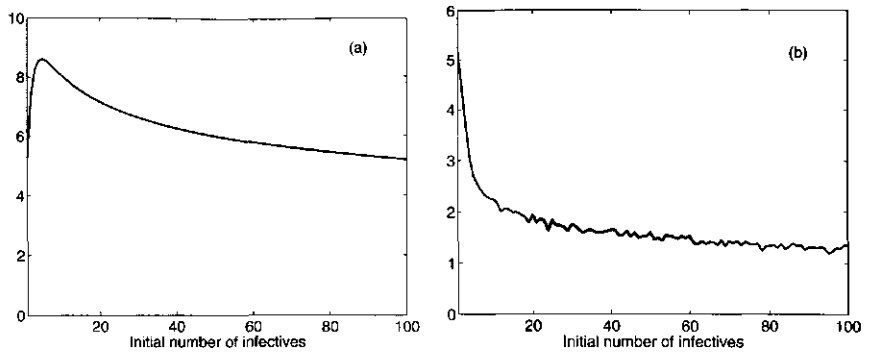

Figure 7.7. The mean and standard deviation of the distribution for the duration of an SIR epidemic as a function of the initial number of infected individuals, $I(0) = i$ and $S(0) = N - i$, $i = 1, 2, \ldots, 100$, $N = 100$, when $\beta = 2$, and $\gamma = 1$, (a) mean and (b) standard deviation (estimated from 1000 sample paths).

The expected duration increases to a maximum, then decreases as the initial number of infected individuals decreases. The standard deviation is a decreasing function of the initial number of infectives. This behavior is not unusual given the bimodal distribution of the duration and the fact that for a large number of infected individuals, the number of cases will not increase very much before the epidemic ends. But note that the mean duration of the SIR epidemic in Figure 7.7 is much less than that of the SIS epidemic. The SIR epidemic always ends quickly as compared to the SIS epidemic, where a quasistationary distribution may be established. See Example 7.4 and Figure 7.2.

Numerous other multivariate epidemic processes can be studied. An interesting application to schistosomiasis is a paper by Chan and Isham (1998). For the spread of an epidemic among a population comprising a large number of small households, please consult the work of Ball (1999) and Ball et al. (1997). For discussions about other deterministic and stochastic epidemic models, please consult Anderson and May (1992), Bailey (1975), Brauer and Castillo-Chávez (2001), Daley and Gani (1999), Gabriel, Lefèvre, and Picard (1990), Goel and Richter-Dyn (1974), Hethcote (2000) and the citations in these works.

7.6 Competition Processes

One of the most well-known competition model is the Lotka-Volterra competition model for two species. The deterministic Lotka-Volterra competition model is reviewed first; then an analogous stochastic model is formulated. In Lotka-Volterra competition, two species compete either directly or indirectly for the same resource; an increase in the density of one species

results in a decrease in the other species that is proportional to the product of both species. The deterministic model has the following form:

$$\frac{dx_1}{dt} = x_1(a_{10} - a_{11}x_1 - a_{12}x_2) \tag{7.17}$$

$$\frac{dx_2}{dt} = x_2(a_{20} - a_{21}x_1 - a_{22}x_2), \tag{7.18}$$

where $x_i(0) > 0$, $a_{ij} > 0$ for $i = 1, 2$ and $j = 0, 1, 2$. The coefficients a_{i0} are intrinsic growth rates for species i, a_{ii} are the intraspecific competition coefficients, and a_{ij}, $i \neq j$ are the interspecific competition coefficients, $i, j = 1, 2$, the effect species j has on species i. If the interspecific competition coefficients are zero, $a_{ij} = 0$, $i \neq j$, then the models are just the logistic growth equation. In this case, each species grows to its carrying capacity,

$$\lim_{t\to\infty} x_i(t) = \frac{a_{i0}}{a_{ii}}, \quad i = 1, 2.$$

Competition between the species, $a_{ij} > 0$, $i \neq j$, changes the dynamics.

The isoclines ($dx_i/dt = 0$) are straight lines, and the asymptotic behavior depends on how the isoclines cross. There are four different cases:

I. If $\frac{a_{20}}{a_{22}} \leq \frac{a_{10}}{a_{12}}$ and $\frac{a_{20}}{a_{21}} \leq \frac{a_{10}}{a_{11}}$, then $\lim_{t\to\infty}(x_1(t), x_2(t)) = (0, a_{10}/a_{11})$.

II. If $\frac{a_{20}}{a_{22}} \geq \frac{a_{10}}{a_{12}}$ and $\frac{a_{20}}{a_{21}} \geq \frac{a_{10}}{a_{11}}$, then $\lim_{t\to\infty}(x_1(t), x_2(t)) = (a_{20}/a_{22}, 0)$.

III. If $\frac{a_{20}}{a_{22}} > \frac{a_{10}}{a_{12}}$ and $\frac{a_{20}}{a_{21}} < \frac{a_{10}}{a_{11}}$, then $\lim_{t\to\infty}(x_1(t), x_2(t)) = (0, a_{10}/a_{11})$ or $\lim_{t\to\infty}(x_1(t), x_2(t)) = (a_{20}/a_{22}, 0)$.

IV. If $\frac{a_{20}}{a_{22}} < \frac{a_{10}}{a_{12}}$ and $\frac{a_{20}}{a_{21}} > \frac{a_{10}}{a_{11}}$, then $\lim_{t\to\infty}(x_1(t), x_2(t)) = (x_1^*, x_2^*)$.

In these cases, x_1^* and x_2^* represent positive solutions of the following linear equations (isoclines):

$$a_{10} = a_{11}x_1 + a_{12}x_2$$

$$a_{20} = a_{21}x_1 + a_{22}x_2.$$

At least one of the inequalities in cases I and II must be a strict inequality otherwise there exists an infinite number of equilibria and the asymptotic behavior depends on initial conditions. Generally, survival of both species (case IV) requires that the interspecific competition coefficients, a_{ij}, $i \neq j$, be less than intraspecific competition coefficients, a_{ii}. For more information about other types of competition models, please consult Waltman (1983) or Smith and Waltman (1995).

7.6.1 Stochastic Competition Model

Let $X_1(t)$ and $X_2(t)$ be random variables for the population size of two competing species, $X_1, X_2 \in \{0, 1, 2, \ldots\}$ and $t \in [0, \infty)$. Let $p_{(i,j)}(t) = \text{Prob}\{X_1(t) = i, X_2(t) = j\}$. The competition model is a birth and death process for two species in which births and deaths depend on the population sizes of one or both of the species. As in the case of logistic growth, there is a multitude of stochastic models corresponding to the one deterministic model.

Suppose for two competing species, the stochastic birth rates are denoted $\lambda_i(X_1, X_2)$ and death rates $\mu_i(X_1, X_2)$. A general competition model assumes the birth and death rates satisfy

$$\lambda_i(X_1, X_2) = \max\{0, X_i(b_{i0} + b_{i1}X_1 + b_{i2}X_2)\}$$

and

$$\mu_i(X_1, X_2) = \max\{0, X_i(d_{i0} + d_{i1}X_1 + d_{i2}X_2)\},$$

where

$$b_{i0} - d_{i0} = a_{i0}, \quad b_{i1} - d_{i1} = -a_{i1}, \quad \text{and} \quad b_{i2} - d_{i2} = -a_{i2},$$

for $i = 1, 2$. The max in the definitions of λ_i and μ_i are to ensure that the expressions are nonnegative, and the assumptions on the coefficients are to ensure that the deterministic model is of the form (7.17) and (7.18), $dx_i/dt = \lambda_i(x_1, x_2) - \mu_i(x_1, x_2)$, $i = 1, 2$. For example, one form for the birth and death rates is

$$\lambda_i(X_1, X_2) = a_{i0}X_i \quad \text{and} \quad \mu_i(X_1, X_2) = X_i(a_{i1}X_1 + a_{i2}X_2). \tag{7.19}$$

The distributions resulting from various birth and death rate assumptions can differ markedly. Here, we only consider case (7.19), where per capita birth rates are constant and per capita death rates depend linearly on the density of both species. Another example is discussed in the Exercises.

Assume the transition probabilities satisfy

$$\begin{aligned}
&\text{Prob}\{\Delta X_1(t) = i, \Delta X_2(t) = j | (X_1(t), X_2(t))\} \\
&= \begin{cases}
a_{10}X_1(t)\Delta t + o(\Delta t), & (i,j) = (1,0) \\
a_{20}X_2(t)\Delta t + o(\Delta t), & (i,j) = (0,1) \\
X_1(t)[a_{11}X_1(t) + a_{12}X_2(t)]\Delta t + o(\Delta t), & (i,j) = (-1,0) \\
X_2(t)[a_{21}X_1(t) + a_{22}X_2(t)]\Delta t + o(\Delta t), & (i,j) = (0,-1) \\
1 - X_1(t)[a_{11}X_1(t) + a_{12}X_2(t)]\Delta t & \\
\quad - X_2(t)[a_{21}X_1(t) + a_{22}X_2(t)]\Delta t + o(\Delta t), & (i,j) = (0,0) \\
o(\Delta t), & \text{otherwise.}
\end{cases}
\end{aligned}$$

The forward Kolmogorov equations satisfy

$$\begin{aligned}
\frac{dp_{(i,j)}}{dt} &= \lambda_1(i-1,j)p_{(i-1,j)} + \lambda_2(i,j-1)p_{(i,j-1)} \\
&\quad + \mu_1(i+1,j)p_{(i+1,j)} + \mu_2(i,j+1)p_{(i,j+1)} \\
&\quad - [\lambda_1(i,j) + \lambda_2(i,j) + \mu_1(i,j) + \mu_2(i,j)]\, p_{(i,j)}.
\end{aligned}$$

Using the generating function technique, the partial differential equation for the m.g.f. can be obtained,

$$\begin{aligned}\frac{\partial M}{\partial t} &= a_{10}(e^{\theta}-1)\frac{\partial M}{\partial \theta}+a_{20}(e^{\phi}-1)\frac{\partial M}{\partial \phi} \\ &\quad +(e^{-\theta}-1)\left[a_{11}\frac{\partial^2 M}{\partial \theta^2}+a_{12}\frac{\partial^2 M}{\partial \theta\,\partial \phi}\right] \\ &\quad +(e^{-\phi}-1)\left[a_{21}\frac{\partial^2 M}{\partial \theta\,\partial \phi}+a_{22}\frac{\partial^2 M}{\partial \phi^2}\right], \qquad (7.20)\end{aligned}$$

where $M(\theta,\phi,0)=e^{(N_1\theta+N_2\phi)}$, $X_1(0)=N_1$ and $X_2(0)=N_2$. Using the differential equation for the m.g.f., differential equations for the means of X_1 and X_2 can be derived. For simplicity, the notation of Bailey (1990) is used for the higher order moments of the distribution,

$$m_{kl}(t)=E(X_1^k(t)X_2^l(t)).$$

The means satisfy

$$\frac{dm_{10}(t)}{dt} = a_{10}m_{10}(t)-a_{11}m_{20}(t)-a_{12}m_{11}(t) \qquad (7.21)$$

$$\frac{dm_{01}(t)}{dt} = a_{20}m_{01}(t)-a_{21}m_{02}(t)-a_{22}m_{11}(t). \qquad (7.22)$$

The two differential equations for the means depend on five unknown variables, $m_{ij}(t)$, and cannot be solved explicitly. However, note that the form of these equations is similar to the deterministic differential equations. Sometimes specific assumptions about $E(X^k(t)X_2^l(t))$ are made to approximate the higher-order moments of the distribution [moment closure; see, e.g., Chan and Isham (1998)].

Example 7.10 Let $a_{10}=2$, $a_{20}=1.5$, $a_{11}=0.03$, $a_{12}=0.02$, $a_{21}=0.01$, and $a_{22}=0.04$. Case IV holds; a positive equilibrium exists and is stable $(x_1^*,x_2^*)=(50,25)$. A sample path is graphed in Figure 7.8 when the initial sizes are the equilibrium values, $X_1(0)=50$ and $X_2(0)=25$. The MATLAB program that generated Figure 7.8 is given in the Appendix for Chapter 7.

At $t=5$, the means and variances for each of the populations are calculated from 1000 sample paths,

$$m_{X_1}(5)=m_{10}(5)=49.9, \quad m_{X_2}(5)=m_{01}(5)=23.2,$$

$$\sigma_{X_1}(5)=9.4, \quad \text{and} \quad \sigma_{X_2}(5)=6.8.$$

The means are close to their equilibrium values. ■

It is possible, in principle, to determine the expected duration until total population extinction or absorption if both population sizes are finite.

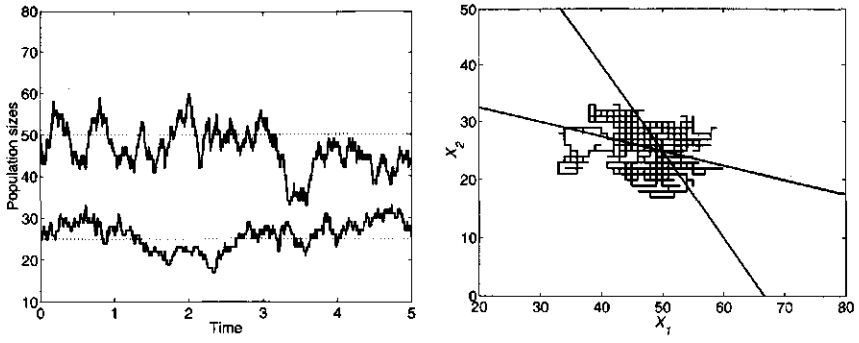

Figure 7.8. A sample path of the Lotka-Volterra competition model graphed as a function of time and in the phase plane with birth and death rates given by (7.19), $a_{10} = 2$, $a_{20} = 1.5$, $a_{11} = 0.03$, $a_{12} = 0.02$, $a_{21} = 0.01$, $a_{22} = 0.04$, $X_1(0) = 50$, and $X_2(0) = 25$. The dotted lines indicate the equilibrium values.

Suppose the random variables for two competing populations have state space $X_1(t) \in \{0, 1, \ldots, N_1\}$ and $X_2 \in \{0, 1, \ldots, N_2\}$. If the expected duration until extinction from state (i, j) is denoted as $\tau_{(i,j)}$, then $\tau = (\tau_{(i,j)})$ is a vector of length $(N_1+1)(N_2+1)$ with a particular order specified. The vector τ is the unique nonnegative solution to a linear system $D\tau = \mathbf{d}$, where matrix D is nonsingular (see Allen, 1999). For example,

$$\begin{aligned}\tau_{(i,j)} &= \eta_{ij} + p_{(i+1,j),(i,j)}\tau_{(i+1,j)} + p_{(i-1,j),(i,j)}\tau_{(i-1,j)} \\ &\quad + p_{(i,j+1),(i,j)}\tau_{(i,j+1)} + p_{(i,j-1),(i,j)}\tau_{(i,j-1)},\end{aligned}$$

where η_{ij} is the mean interevent time given that the state is (i, j),

$$\eta_{ij} = \frac{1}{\lambda_1(i,j) + \lambda_2(i,j) + \mu_1(i,j) + \mu_2(i,j)},$$

$p_{(k,l),(i,j)}$ is the transition probability from state (i, j) to state (k, l) calculated from the embedded Markov chain, and

$$\sum_{\{\Delta i, \Delta j\} \in \{1,-1\}} \left[p_{(i+\Delta i,j),(i,j)} + p_{(i,j+\Delta j),(i,j)}\right] = 1.$$

Matrix D is a sparse, banded matrix. Efficient numerical methods can be used to solve the linear system.

7.7 Predator-Prey Processes

The Lotka-Volterra predator-prey model has the form

$$\frac{dx}{dt} = x(a_{10} - a_{12}y) \tag{7.23}$$

$$\frac{dy}{dt} = y(a_{21}x - a_{20}), \tag{7.24}$$

where $a_{ij} > 0$. The equilibrium $(a_{20}/a_{21}, a_{10}/a_{12})$ is neutrally stable; that is, for any initial condition there exists a unique periodic solution $(x(t), y(t))$ encircling the equilibrium. If a density-dependent factor is added to the prey equation,

$$\frac{dx}{dt} = x(a_{10} - a_{11}x - a_{12}y),$$

then solutions converge to predator extinction or to a positive equilibrium:

I. If $\dfrac{a_{20}}{a_{21}} \geq \dfrac{a_{10}}{a_{11}}$, then $\lim_{t\to\infty}(x(t), y(t)) = (a_{10}/a_{11}, 0)$.

II. If $\dfrac{a_{20}}{a_{21}} < \dfrac{a_{10}}{a_{11}}$, then $\lim_{t\to\infty}(x(t), y(t)) = (a_{20}/a_{21}, [a_{10} - a_{11}a_{20}/a_{21}]/a_{12})$.

Numerous other formulation for predator-prey models exist in the literature (Edelstein-Keshet, 1988; Hassell, 1978; Murray, 1993). The functional response of the predator (prey eaten per predator per unit of time), corresponding to the terms $a_{21}x$, generally has a saturation effect. For example, some well-known forms for the functional response include

$$\text{Ivlev, } a(1 - \exp(-bx)),$$

$$\text{Holling Type II or Michaelis Menten, } \frac{ax}{x+d},$$

and

$$\text{ratio dependent, } \frac{ax}{y+bx}$$

(see, for example, May, 1976; Hassell, 1978; Kuang and Beretta, 1998). Here, we have only considered the dynamics of the simple Lotka-Volterra predator-prey system.

7.7.1 Stochastic Predator-Prey Model

Let $X(t)$ and $Y(t)$ denote random variables for the size of the prey and predator populations respectively, in a stochastic Lotka-Volterra model. Assume the transition probabilities satisfy

$$\text{Prob}\{\Delta X(t) = i, \Delta Y(t) = j | (X(t), Y(t))\} = \begin{cases} a_{10}X(t)\Delta t + o(\Delta t), & (i,j) = (1,0) \\ a_{21}X(t)Y(t)\Delta t + o(\Delta t), & (i,j) = (0,1) \\ a_{12}X(t)Y(t)\Delta t + o(\Delta t), & (i,j) = (-1,0) \\ a_{20}Y(t)\Delta t + o(\Delta t), & (i,j) = (0,-1) \\ 1 - X(t)[a_{10} + a_{12}Y(t)]\Delta t & \\ \quad + Y(t)[a_{20} + a_{21}X(t)]\Delta t + \boldsymbol{o(\Delta t)}, & (i,j) = (0,0) \\ o(\Delta t), & \text{otherwise.} \end{cases}$$

It is straightforward to write the forward Kolmogorov differential equations. Let $p_{(i,j)}(t) = \text{Prob}\{X(t) = i, Y(t) = j\}$. Then

$$\begin{aligned}\frac{dp_{(i,j)}}{dt} &= a_{10}(i-1)p_{(i-1,j)} + a_{21}i(j-1)p_{(i,j-1)} \\ &\quad + a_{12}(i+1)jp_{(i+1,j)} + a_{20}(j+1)p_{(i,j+1)} \\ &\quad - [a_{10}i + a_{21}ij + a_{12}ij + a_{20}j]p_{(i,j)}.\end{aligned}$$

From these equations, differential equations for higher-order moments or the moment generating function can be obtained. The next example illustrates the dynamics for the stochastic model.

Example 7.11 Let $a_{10} = 1$, $a_{20} = 1$, $a_{12} = 0.02$, $a_{21} = 0.01$ in the simple predator-prey model. For the initial conditions, $X(0) = 120$ and $Y(0) = 40$, graphs of the deterministic and stochastic models are compared in Figure 7.9. It can be seen that the stochastic model jumps between cycles. For this single realization, extinction did not occur. But extinction will occur as time is increased and is even more likely if the equilibrium values are smaller. ■

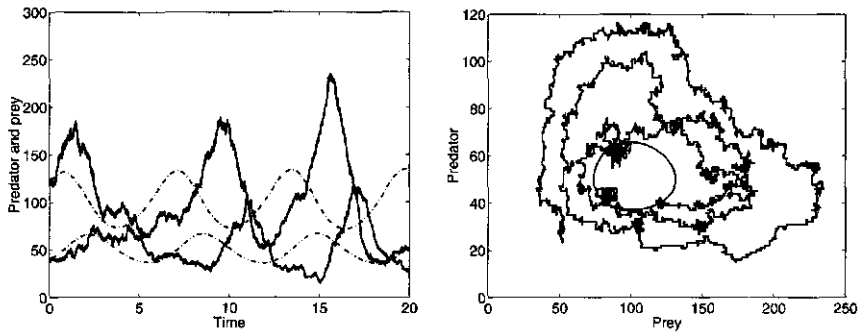

Figure 7.9. A sample path of the Lotka-Volterra predator-prey model is graphed with the solution to the deterministic model. Solutions are graphed over time and in the phase plane. The parameter values and initial conditions satisfy $a_{10} = 1$, $a_{20} = 1$, $a_{12} = 0.02$, $a_{21} = 0.01$, $X(0) = 120$, and $Y(0) = 40$. Solutions with the smaller amplitude represent the predator.

Continuous time Markov chain models for predator and prey and competing species can be formulated in terms of queueing networks. Each species is represented by a node in the network. Arrivals and departures at each node can be births, deaths, and migration. A queueing network model for one predator and two prey is described by Chao, Miyazawa, and Pinedo (1999).

7.8 Other Population Processes

There are numerous other population and epidemic models that can be formulated and analyzed. For example, there are population models with multiple competitors, predators or prey, models with mutualistic interactions, population models with spatial spread, and epidemic models with classes for latent individuals or individuals with maternal antibody protection. Please consult some of the references for the many variations on models for competition, predation, mutualism, and epidemics (e.g., Anderson and May, 1992; Brauer and Castillo-Chávez, 2001; Edelstein-Keshet, 1988; Goel and Richter-Dyn, 1974; Hallam and Levin, 1986; Hethcote, 2000; Kot, 2001; Murray, 1993, 2002, 2003; Renshaw, 1993). Three examples will be discussed here that differ from the models in the previous sections. The first example is an SEIR epidemic model; the second, a spatial predator-prey model; and the third, a population genetics model.

7.8.1 SEIR Epidemic Model

Many variations in the basic deterministic SIS and SIR epidemic models are presented by Hethcote (2000) and Anderson and May (1992). The model presented here was developed by Anderson and May (1986, 1992) to model measles epidemics in different countries. It is an a SEIR model with births and deaths. There is an additional class of exposed or latent individuals, E, individuals that are not yet infectious. The differential equations for the deterministic model are as follows:

$$\begin{aligned}
\frac{dS}{dt} &= b(N-S) - \beta SI \\
\frac{dE}{dt} &= \beta SI - \sigma E - bE \\
\frac{dI}{dt} &= \sigma E - bI - \gamma I \\
\frac{dR}{dt} &= \gamma I - bR,
\end{aligned}$$

where $S(t) + E(t) + I(t) + R(t) = N$. In this model, a mass action rate of incidence is assumed, βSI, rather than the standard incidence $\beta SI/N$. The birth rate b is the same as the death rate. The new parameter σ is the rate of becoming infectious or $1/\sigma$ is the average length of the latent period.

The basic reproductive number for this model is

$$\mathcal{R}_0 = \left(\frac{\beta N}{\gamma + b}\right)\left(\frac{\sigma}{\sigma + b}\right).$$

The disease-free state, $S = N$ and $E = I = R = 0$, is locally asymptotically stable if $\mathcal{R}_0 < 1$. Anderson and May (1986, 1992) were interested in

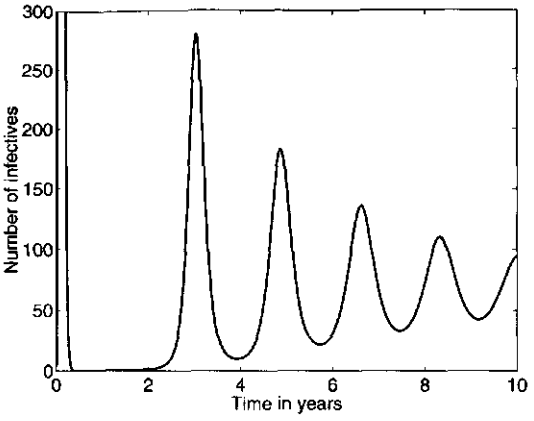

Figure 7.10. SEIR epidemic model with immigration of infectives; the system exhibits oscillations before convergence to an endemic equilibrium. Initial conditions satisfy $S(0) = 249,995$, $E(0) = 0$, $I(0) = 5$, and $R(0) = 0$.

recurrent epidemics and added an immigration term Λ to the differential equation for the number of infected individuals:

$$\frac{dI}{dt} = \sigma E - bI - \gamma I + \Lambda.$$

With immigration, the disease-free state is not possible and the population size is not constant, $N(t) = S(t) + E(t) + I(t) + R(t)$. The dynamics of the SEIR immigration model are illustrated in Figure 7.10 for parameter values corresponding to measles in Iceland (Anderson and May, 1992):

$$\begin{aligned} \beta &= \frac{0.008}{\text{year}}, \\ \frac{1}{\gamma} &= 7 \text{ days}, \\ \frac{1}{\sigma} &= 7 \text{ days}, \\ \Lambda &= \frac{7}{\text{year}}, \\ \text{and } \frac{1}{b} &= 70 \text{ years}. \end{aligned} \tag{7.25}$$

The initial size is $N = 250,000$. The maximum of the first wave of the epidemic is not shown in Figure 7.10, but the first wave includes more than 90,000 infectives.

Let $S(t)$, $E(t)$, $I(t)$, and $R(t)$ denote random variables for the number of susceptible, latent, infectious, and immune individuals. The transition

probabilities in the stochastic model satisfy (Anderson and May, 1992)

$$\text{Prob}\{\Delta S(t) = i, \Delta E(t) = j, \Delta I(t) = k, \Delta R(t) = l | (S(t), E(t), I(t), R(t))\}$$

$$= \begin{cases} bN(t)\Delta t + o(\Delta t), & (i,j,k,l) = (1,0,0,0) \\ bS(t)\Delta t + o(\Delta t), & (i,j,k,l) = (-1,0,0,0) \\ bE(t)\Delta t + o(\Delta t), & (i,j,k,l) = (0,-1,0,0) \\ bI(t)\Delta t + o(\Delta t), & (i,j,k,l) = (0,0,-1,0) \\ bR(t)\Delta t + o(\Delta t), & (i,j,k,l) = (0,0,0,-1) \\ \beta S(t)I(t)\Delta t + o(\Delta t), & (i,j,k,l) = (-1,1,0,0) \\ \sigma E(t)\Delta t + o(\Delta t), & (i,j,k,l) = (0,-1,1,0) \\ \gamma I(t)\Delta t + o(\Delta t), & (i,j,k,l) = (0,0,-1,1) \\ \Lambda\Delta t + o(\Delta t), & (i,j,k,l) = (0,0,1,0) \\ 1 - [\beta S(t)I(t) + \sigma E(t) + \gamma I(t) + \Lambda]\Delta t & \\ \quad - 2b[N(t)]\Delta t + o(\Delta t), & (i,j,k,l) = (0,0,0,0) \\ o(\Delta t), & \text{otherwise.} \end{cases}$$

See Anderson and May (1992) for some stochastic simulations of this model corresponding to population sizes of $N = 230,000$ and $N = 100,000$. In both cases there are recurrent epidemics, but when the population size is smaller, they are less frequent. Consult Bailey (1975), Daley and Gani (1999), Gabriel et al. (1990), Ludwig and Cooke (1975) and references therein for additional examples of stochastic epidemic models.

7.8.2 Spatial Predator-Prey Model

A stochastic spatial predator-prey model was formulated and studied by Renshaw (1993). The prey and predator move among a discrete set of spatial locations or patches, $i = 1, 2, \ldots, n$. For each spatial location there is a random variable for the sizes of the prey and predator, $X_i(t)$ and $Y_i(t)$, $i = 1, 2, \ldots, n$. In a small period of time Δt, the prey moves from location i to j with probability $u_{ji}X_i(t)\Delta t + o(\Delta t)$ and the predator moves from location i to j with probability $v_{ji}Y_i(t)\Delta t + o(\Delta t)$. The prey and predator dynamics within each spatial location or patch follow the simple Lotka-Volterra model, where there is cyclic behavior. The model mimics some of the biological experiments performed by Huffaker in 1958 on mites. These experiments involved a predatory mite, *Typhlodromus occcidentalis*, and another mite that served as the prey, *Eotetranychus sexmaculatus*. Oranges served as food for the prey, and the mites could move from one orange to another (also consult Maynard Smith, 1974).

In the model of Renshaw (1993) for two patches $i = 1, 2$, the transition probabilities satisfy

$$\text{Prob}\{\Delta X_i(t) = k_i, \Delta Y_i(t) = l, i = 1, 2 | (X_1(t), Y_1(t), X_2(t), Y_2(t))\}$$

$$= \begin{cases} a_{10}X_1(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (1, 0, 0, 0) \\ a_{21}X_1(t)Y_1(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (0, 1, 0, 0) \\ a_{12}X_1(t)Y_1(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (-1, 0, 0, 0) \\ a_{20}Y_1(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (0, -1, 0, 0) \\ a_{10}X_2(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (0, 0, 1, 0) \\ a_{21}X_2(t)Y_2(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (0, 0, 0, 1) \\ a_{12}X_2(t)Y_2(t)\,\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (0, 0, -1, 0) \\ a_{20}Y_2(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (0, 0, 0, -1) \\ u_{21}X_1(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (-1, 0, 1, 0) \\ u_{12}X_2(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (1, 0, -1, 0) \\ v_{21}Y_1(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (0, -1, 0, 1) \\ v_{12}Y_2(t)\Delta t + o(\Delta t), & (k_1, l_1, k_2, l_2) = (0, 1, 0, -1) \end{cases}$$

In addition, the probability of no change in state is

$$1 - \sum_{i=1}^{2} X_i(t)[a_{10} + a_{12}Y_i(t)]\Delta t - \sum_{i=1}^{2} Y_i(t)[a_{20} + a_{21}X_i(t)]\Delta t$$
$$- \sum_{j=1, j\neq i}^{2} \sum_{i=1}^{2} [u_{ji}X_i(t) + v_{ji}Y_i(t)]\Delta t + o(\Delta t).$$

The corresponding deterministic model with n patches has the following form:

$$\frac{dx_i}{dt} = x_i(a_{10} - a_{12}y_i) + \sum_{j=1, j\neq i}^{n} (u_{ij}x_j - u_{ji}x_i)$$

$$\frac{dy_i}{dt} = y_i(-a_{20} + a_{21}x_i) + \sum_{j=1, j\neq i}^{n} (v_{ij}y_j - v_{ji}y_i)$$

for $i = 1, 2, \ldots, n$.

Renshaw (1993) simulated the dynamics with two different types of movement patterns, with $u = u_{ji}$ and $v = v_{ji}$, an equilibrium at $(25, 15)$, and $X_1(0) = 10$, $Y_1(0) = 5$, $X_i(0) = 0 = Y_i(0)$, $i = 2, \ldots, n$. When there was no migration, $u = 0 = v$, extinction of the prey occurred in the range of 5 to 6 time units. With some spatial movement, the time until extinction could be significantly increased and sustained cycles could be maintained. Renshaw (1993) found that sustained cycles can be maintained if $m \geq 10$;

predator movement is greater than prey movement, $v > u$; predator movement is sufficiently large to prevent prey explosion; and, finally, u and v are not so large so that the spatial structure becomes synchronized. Predator movement greater than prey movement is in contrast to the results of Huffaker. This may be an artifact of the simple predator-prey assumption: In the absence of predation, the prey grows exponentially. In the Exercises, the model of Renshaw is modified to include density dependence in the prey growth rate.

This spatial predator-prey model can be put in the more general context of spatial population models known as *stepping-stone models* (Renshaw, 1993). A stepping-stone model for a single population with two patches was applied to the spread of Africanized honey bees by Matis, Kiffe, and Otis (1994). Generalizations of this stepping-stone model were studied by Matis, Zheng, and Kiffe (1995) (see also Matis and Kiffe, 2000). Durrett (1995, 1999) discusses some interesting examples on stochastic spatial models with examples from genetics, epidemiology, and ecology. In these models space is represented by a grid of sites.

7.8.3 Population Genetics Model

Suppose the population is diploid; each individual has two copies of every gene. In the simplest case, suppose a trait is determined by a single gene at a particular locus or site. In addition, suppose there are only two different alleles for this gene, denoted as A and a. Therefore, the genotype or the pair that actually occurs in an individual can be one of three different forms, either

$$AA, \quad Aa, \quad \text{or} \quad aa.$$

Let $x_{AA}(t)$, $x_{Aa}(t)$, and $x_{aa}(t)$ denote the sizes of the population corresponding to the number of individuals with these particular genotypes at time t and let $N(t)$ be the total population size,

$$N(t) = x_{AA}(t) + x_{Aa}(t) + x_{aa}(t).$$

A deterministic model is formulated for changes in the size of each of these three genotypes under the assumption of random mating, no selection and no migration. Suppose b is the per capita population birth rate and d is the per capita population death rate. In random mating, a single gene from each parent forms the new gene pair in the next generation.

A model for the change in the genotypic population sizes is

$$\frac{dx_i(t)}{dt} = bf_i - dx_i, \quad i \in \{AA, Aa, aa\}, \tag{7.26}$$

where under the assumption of random mating,

$$f_{AA}(x_{AA}, x_{Aa}) = \frac{(x_{AA} + x_{Aa}/2)^2}{N} \tag{7.27}$$

$$f_{Aa}(x_{AA}, x_{Aa}, x_{aa}) = \frac{2(x_{AA} + x_{Aa}/2)(x_{aa} + x_{Aa}/2)}{N} \quad (7.28)$$

$$f_{aa}(x_{aa}, x_{Aa}) = \frac{(x_{aa} + x_{Aa}/2)^2}{N}. \quad (7.29)$$

For example, the probability two A genes form an AA **genotype** is

$$\left(\frac{(x_{AA} + x_{Aa}/2)}{N}\right)\left(\frac{(x_{AA} + x_{Aa}/2)}{N}\right).$$

If the population birth rate is bN, the birth rate for genotype AA is

$$bN\left(\frac{(x_{AA} + x_{Aa}/2)}{N}\right)\left(\frac{(x_{AA} + x_{Aa}/2)}{N}\right) = bf_{AA}.$$

Since $f_{AA} + f_{Aa} + f_{aa} = N$, it follows that

$$\frac{dN}{dt} = (b - d)N.$$

For this model, it can be shown that the size of the allele populations changes at the rate $(b - d)$ but that the proportion or frequency of the alleles remains constant over time and equals the initial frequency. Let the sizes of the allele populations be denoted as $z_A = x_{AA} + x_{Aa}/2$ and $z_a = x_{aa} + x_{Aa}/2$, and the allele frequencies be denoted by $p_A = z_A/N$ and $p_a = z_a/N$, where $p_A + p_a = 1$. Thus,

$$\frac{dz_i}{dt} = (b - d)z_i, \quad p_A(t) = p_A(0), \quad \text{and} \quad p_a(t) = p_a(0). \quad (7.30)$$

The relationships given in (7.30) mean that the population is at a *Hardy-Weinberg equilibrium*, an equilibrium where the allele frequencies remain the same over time.

The next example illustrates the dynamics of the deterministic population genetics model when there is not a Hardy-Weinberg equilibrium; selection acts on the death rate of genotype AA.

Example 7.12 When the equations (7.26), (7.27), (7.28), and (7.29) are satisfied, the allele frequencies are at a Hardy-Weinberg equilibrium. The equilibrium is determined by the initial values. Let $x_{AA}(0) = 500$, $x_{Aa}(0) = 100$, and $x_{aa}(0) = 200$ and in addition, $b = 0.1 = d$. Then

$$p_A = 0.6875, \quad p_a = 0.3125,$$

and in the limit, each genotype satisfies

$$\lim_{t\to\infty}(x_{AA}(t), x_{Aa}(t), x_{aa}(t)) = \left(\frac{z_A^2(0)}{N(0)}, \frac{2z_a(0)z_A(0)}{N(0)}, \frac{z_a^2(0)}{N(0)}\right) \quad \textbf{(7.31)}$$

$$= (378.125, 343.75, 78.125).$$

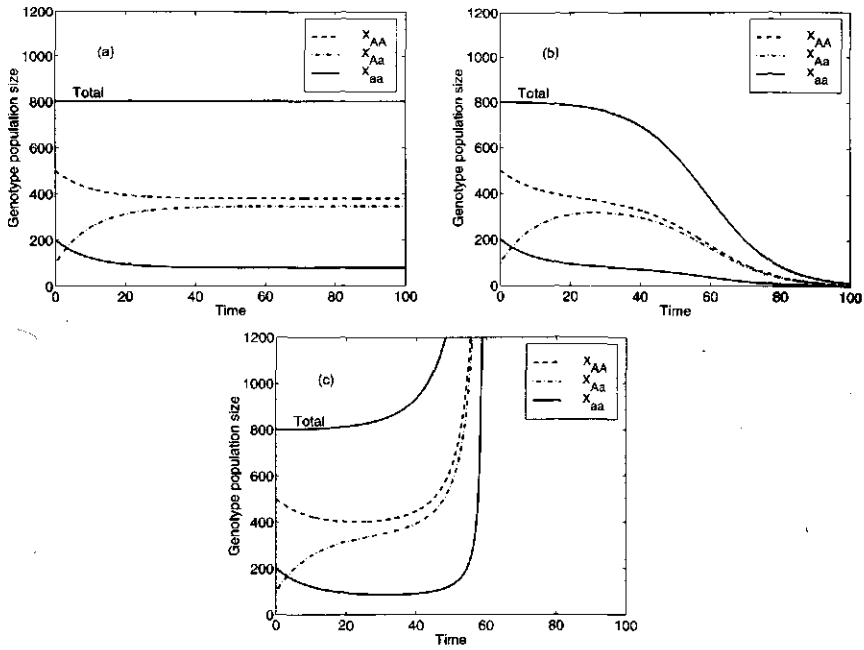

Figure 7.11. Deterministic population genetics model with either no selection or AA selection. The initial conditions satisfy $x_{AA}(0) = 500$, $x_{Aa}(0) = 100$, and $x_{aa}(0) = 200$. In (a), $b = 0.1 = d$, and the genotype population sizes approach a Hardy-Weinberg equilibrium. In (b), $d_{AA} = 0.1005$; the population size approaches zero and the proportion of genotype AA decreases. In (c), $d_{AA} = 0.0995$; the population size increases exponentially and the proportion of genotype AA increases.

The dynamics in this case are illustrated in Figure 7.11(a). Now, suppose there is selection for or against genotype AA. When the death rate of genotype AA, $d_{AA} > d$, genotype AA has a selective disadvantage but if $d_{AA} < d$, then genotype AA has a selective advantage. In Figure 7.11(b), $d_{AA} = 0.1005 > d$, and in Figure 7.11(c), $d_{AA} = 0.0995 < d$. It can be seen that the population size approaches zero when the death rate of genotype AA is increased, and the population size increases exponentially if the death rate is decreased. Although not evident from Figures 7.11(b) and (c), the proportion of genotype AA approaches zero in (b), and the proportion of genotype AA approaches one in (c). ■

Next we formulate a stochastic population genetics model. Let $X_{AA}(t)$, $X_{Aa}(t)$, and $X_{aa}(t)$ denote random variables for the genotypic population sizes and $N(t)$ denote the random variable for the total population size. Let the probabilities be denoted $p_{(i,j,k)}(t) = \text{Prob}\{X_{AA}(t) = i, X_{Aa}(t) = j,$

$X_{aa}(t) = k\}$, $i, j, k \in \{0, 1, 2, \ldots\}$, and

$$p_i^N(t) = \text{Prob}\{N(t) = i\}, \quad i \in \{0, 1, 2, \ldots\}.$$

The transition probabilities for this multivariate stochastic process satisfy

$$\text{Prob}\{\Delta X_{AA}(t) = i, \Delta X_{Aa}(t) = j, \Delta X_{aa}(t) = k | (X_{AA}(t), X_{Aa}(t), X_{aa}(t))\}$$

$$= \begin{cases} bf_{AA}(X_{AA}, X_{Aa})\Delta t + o(\Delta t), & (i,j,k) = (1,0,0) \\ bf_{Aa}(X_{AA}, X_{Aa}, X_{aa})\Delta t + o(\Delta t), & (i,j,k) = (0,1,0) \\ bf_{aa}(X_{aa}, X_{Aa})\Delta t + o(\Delta t), & (i,j,k) = (0,0,1) \\ dX_{AA}\Delta t + o(\Delta t), & (i,j,k) = (-1,0,0) \\ dX_{Aa}\Delta t + o(\Delta t), & (i,j,k) = (0,-1,0) \\ dX_{aa}\Delta t + o(\Delta t), & (i,j,k) = (0,0,-1) \\ 1 - (b+d)N(t)\Delta t + \boldsymbol{o(\Delta t)}, & (i,j,k) = (0,0,0) \\ o(\Delta t), & \text{otherwise.} \end{cases}$$

The stochastic process for the total population size, $\{N(t)\}$, $t \geq 0$, is a simple birth and death process provided $b = \lambda =$ constant and $d = \mu =$ constant. In this case, the transition probabilities satisfy

$$\text{Prob}\{\Delta N(t) = i | N(t)\} = \begin{cases} bN(t)\Delta t + o(\Delta t), & i = 1 \\ dN(t)\Delta t + o(\Delta t), & i = -1 \\ 1 - [b+d]N(t)\Delta t + \boldsymbol{o(\Delta t)}, & i = 0 \\ o(\Delta t), & \text{otherwise.} \end{cases}$$

The forward Kolmogorov differential equation has the form

$$\frac{dp_i^N}{dt} = b(i-1)p_{i-1}^N + d(i+1)p_{i+1}^N - [b+d]ip_i^N$$

and the moment generating function $M(z,t)$ satisfies the partial differential equation

$$\frac{\partial M}{\partial t} = [d(e^{-\theta} - 1) + b(e^{\theta} - 1)]\frac{\partial M}{\partial \theta}, \quad M(\theta, 0) = e^{\theta n_0},$$

where $N(0) = n_0$. The dynamics of the total population size is well understood in this case. For example, when $b = d =$ constant and the population size is large, the mean values of the random variables in the multivariate process, $(X_{AA}(t), X_{Aa}(t), X_{aa}(t))$, $t \geq 0$, are close to their equilibrium values. For small population sizes, extinction may occur rapidly, or fixation at one of the equilibria $Z_a = 0$ or $Z_A = 0$. Consult Nagylaki (1992) for more information on theoretical population genetics.

7.9 Exercises for Chapter 7

1. Consider the simple birth and death process discussed in Example 7.1.

   (a) Show that the solutions (7.4) satisfy the differential equation, $dP/dt = \mu - (\lambda + \mu)P + \lambda P^2$.

   (b) Find $m = f'(1)$. When $m > 1$, use the p.g.f. $f(z)$ to compute $\lim_{t\to\infty} p_0(t)$. Compare this limit with the one obtained for the simple birth and death process in Chapter 6, Section 6.4.3.

2. Consider the development of drug resistance in cancer cells discussed in Example 7.2.

   (a) Show that the solutions (7.5) and (7.6) satisfy the differential equations for P_1 and P_2.

   (b) Find the mean number of sensitive (type 1) and resistant (type 2) cells produced by type 1 cells; that is, $m_{11}(t) = \partial P_1(1,1,t)/\partial z_1$ and $m_{21}(t) = \partial P_1(1,1,t)/\partial z_2$ (Kimmel and Axelrod, 2002).

3. Calculate the quasistationary probability distribution and the expected duration of an epidemic for an SIS epidemic model when $\beta = 1.5$, $b + \gamma = 1$ and $N = 100$. Sketch their graphs and compare them to Figure 7.2.

4. For the bivariate birth and death process in Example 7.5, show that the m.g.f. satisfies (7.11). Then use the forward Kolmogorov differential equation (7.10) to show that the mean and variance of the process in Example 7.5 satisfy (7.12).

5. Consider the transition matrix T corresponding to the embedded Markov chain of the SIR epidemic model in Example 7.7.

   (a) Identify the remaining submatrices A_2, A_3, A_4, B_2, and B_3 of the transition matrix T.

   (b) Let $N = 4$, $\gamma = 1$, and $\alpha = 2$. Then find $T^{2N-1} = T^7$ and show that the final size distribution when $I(0) = 1$ is $p_I = (0, 0.4, 0.15, 0.1556, 0.2944)^T$. What is the final size distribution when $I(0) = 2$?

6. For the Lotka-Volterra competition model given in (7.17) and (7.18), assume the stochastic birth rates and death rates satisfy

$$\lambda_i(X_1, X_2) = \max\left\{0, X_i\left(a_{i0} - \frac{a_{ii}}{2}X_i\right)\right\}, \quad i = 1, 2.$$

$$\mu_i(X_1, X_2) = X_i\left(\frac{a_{ii}}{2}X_i + a_{ij}X_j\right), \quad i, j = 1, 2, \ i \neq j.$$

(a) Write the forward Kolmogorov differential equation for $p_{ij}(t) = \text{Prob}\{X_1(t) = i, X_2(t) = j\}$ and use it to find the differential equations satisfied by the means, $m_{10}(t) = E(X_1(t))$ and $m_{01}(t) = E(X_2(t))$.

(b) For the same parameter values and initial conditions as in Example 7.10, $a_{10} = 2$, $a_{20} = 1.5$, $a_{11} = 0.03$, $a_{12} = 0.02$, $a_{21} = 0.01$, $a_{22} = 0.04$, $X_1(0) = 50$ and $X_2(0) = 25$. Graph one sample path in the phase plane. Then generate 1000 sample paths and find the mean and variance at $t = 5$ and compare your answers with those given for Example 7.10.

7. The deterministic predator-prey system,

$$\begin{aligned} \frac{dx}{dt} &= x\left[r\left(1-\frac{x}{K}\right) - \frac{ay}{x+d}\right] \\ \frac{dy}{dt} &= yb\left(1-\frac{cy}{x}\right), \end{aligned}$$

was shown by Murray (1993) to have a limit cycle for some parameter values.

(a) For the parameter values $r = 1$, $K = 100$, $a = 1$, $d = 20$, $b = 0.02$, and $c = 1$, show that a positive equilibrium exists and that it is unstable. It can be shown that a limit cycle exists.

(b) Assume for the prey x that the birth rate is rx and death rate is $x[r/K + ay/(x+d)]$ and for the predator y that the birth rate is by and the death rate is bcy^2/x. Write the transition probabilities for the stochastic model and the forward Kolmogorov differential equations for $p_{(i,j)}(t) = \text{Prob}\{X(t) = i, Y(t) = j\}$.

(c) For the parameter values in (a) and assumptions in (b), write a computer program to generate a sample path for the stochastic predator-prey system with initial conditions $X(0) = 20$ and $Y(0) = 30$ and graph the sample path.

(d) Due to the oscillations in the predator-prey model, it is very likely that the predator or prey become extinct. What happens in the model if the predator y becomes extinct? Notice that there is a singularity in the differential equations when $x = 0$. When the prey becomes extinct, it is assumed that the predator also becomes extinct. An illustration of prey extinction is given in Figure 7.12.

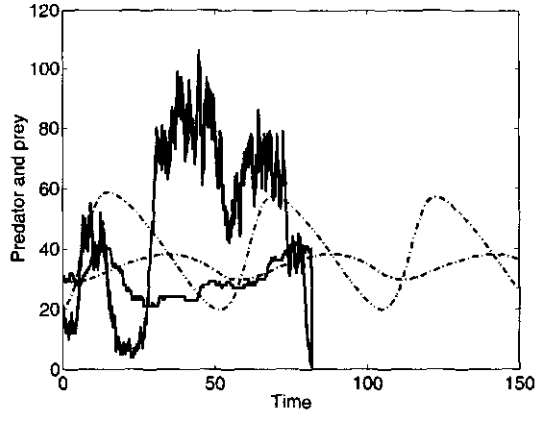

Figure 7.12. A sample path of the stochastic predator-prey model is graphed against the solution to the deterministic model. The parameter values and initial conditions satisfy $r = 1$, $K = 100$, $a = 1$, $d = 20$, $b = 0.02$, $c = 1$, $X(0) = 20$, and $Y(0) = 30$. Solutions with the smaller amplitude are the predator.

8. For the three-species competition model,

$$\frac{dx_i}{dt} = x_i(a_{i0} - a_{i1}x_1 - a_{i2}x_2 - a_{i3}x_3), \quad i = 1, 2, 3,$$

define three random variables $X_i(t)$, $i = 1, 2, 3$ that correspond to the sizes of the three competing species. Define possible birth and death rates, $\lambda_i(X_1, X_2, X_3)$ and $\mu_i(X_1, X_2, X_3)$. Then define transition probabilities corresponding to a stochastic model and express the corresponding forward Kolmogorov differential equations for $p_{(i,j,k)}(t) = \text{Prob}\{X_1(t) = i, X_2(t) = j, X_3(t) = k\}$. Assume the initial distribution satisfies $X_1(0) = a$, $X_2(0) = b$, and $X_3(0) = c$, where a, b and c are positive constants.

9. Consider the SEIR epidemic model with no immigration, $\Lambda = 0$.

   (a) Show that the disease-free equilibrium $S = N$, $E = I = R = 0$ is locally asymptotically stable if $\mathcal{R}_0 < 1$.

   (b) Convert the parameters given by Anderson and May (1992) in (7.25) to units of $(\text{year})^{-1}$. Then find the minimum value of N such that the disease-free equilibrium is no longer locally asymptotically stable. Anderson and May used the stochastic model with immigration to illustrate the importance of a sufficiently large population size to sustain an epidemic.

10. Consider Renshaw's spatial predator-prey model with density dependence in the prey population and dispersal between two patches. Assume the deterministic model satisfies

$$\begin{aligned} \frac{dx_i}{dt} &= x_i(a_{10} - a_{11}x_i - a_{12}y_i) + u(x_j - x_i), \\ \frac{dy_i}{dt} &= y_i(-a_{20} + a_{21}x_i) + v(y_j - y_i), \end{aligned}$$

where $i, j = 1, 2$ and $j \neq i$. There are $n = 2$ patches with $u_{ij} = u$ and $v_{ij} = v$.

Let $X_i(t)$ and $Y_i(t)$ denote the random variables for the corresponding continuous time, stochastic process and

$$[a_{11}X_i(t)X_i(t) + a_{12}X_i(t)Y_i(t)]\Delta t + o(\Delta t)$$

be the transition probability for death of a prey, $\Delta X_i(t) = -1$. Suppose the parameter values are $a_{10} = 1$, $a_{11} = 0.02$, $a_{12} = 0.1$, $a_{20} = 0.15$, and $a_{21} = 0.01$ and initial conditions are $X_1(0) = 15$, $Y_1(0) = 7$, and $X_2(0) = 0 = Y_2(0) = 0$.

(a) Show that $x_i = 15$ and $y_i = 7$, $i = 1, 2$ is an equilibrium of the deterministic model (i.e., $dx_i/dt = 0 = dy_i/dt$).

(b) Let $u = 0 = v$ so that there is essentially one patch; no movement. Write a computer program to simulate 50 sample paths, and record the time when either (i) the prey population size or predator population size equals zero (extinction of one species) or (ii) time has reached $t = 200$ and neither population size is zero (coexistence). In what proportion of the sample paths are the prey extinct? predators extinct? both coexist?

(c) For two patches, consider three cases: (i) $u = 0.001 = v$ (prey and predator movement rates are equal), (ii) $u = 0.001$ and $v = 0.01$ (predator movement > prey movement), and (iii) $u = 0.01$ and $v = 0.001$ (prey movement > predator movement). Write a computer program to simulate 50 sample paths for each of these sets of parameter values and record the time when either the total prey population size or total predator population size equals zero (extinction of one species in both patches) or time has reached $t = 200$ and neither of the population sizes are zero (coexistence). In what proportion of the sample paths are the prey extinct? predators extinct? both coexist?

(d) Repeat part (c) with three patches, $X_1(0) = 15$, $Y_1(0) = 7$, $X_j(0) = 0$, and $Y_j(0) = 0$, $j = 2, 3$. Compare the results of (b) and (c), (b) and (d), and (c) and (d). When is extinction most likely to occur? in one, two, or three patches? When is

coexistence most likely to occur? when prey movement is greater than predator movement or when predator movement is greater than prey movement?

11. Consider the deterministic population genetics model (7.26), (7.27), (7.28), and (7.29).

    (a) Show that the relationships given for the allele population sizes and frequencies in (7.30) are satisfied.

    (b) If $b = d$, show that the genotypic population sizes $x_i(t)$, $i \in \{AA, Aa, aa\}$ have the limit specified by (7.31).

    (c) For the corresponding stochastic population genetics model, find the forward Kolmogorov differential equations for

    $$p_{(i,j,k)} = \text{Prob}\{X_{AA}(t) = i, X_{Aa}(t) = j, X_{AA}(t) = k\}.$$

    Assume the initial distribution satisfies $X_{AA}(0) = a$, $X_{Aa}(0) = b$, $X_{aa}(0) = c$, where a, b, and c are positive constants.

    (d) Write a computer program to simulate sample paths corresponding to the Hardy-Weinberg equilibrium in Example 7.12 with $b = 0.1 = d$, $X_{AA}(0) = 500$, $X_{Aa}(0) = 100$, and $X_{aa}(0) = 200$. Plot X_{AA}, X_{Aa}, X_{aa}, and N. Simulate 100 sample paths. Then find the mean and standard deviation for each of the random variables at $t = 40$ and compare the mean values with the solutions to the deterministic model.

12. Suppose two competing populations disperse between two patches. The deterministic model satisfies

    $$\begin{aligned} \frac{dx_i}{dt} &= x_i(a_{10} - a_{11}x_i - a_{12}y_i) + u(x_j - x_i), \\ \frac{dy_i}{dt} &= y_i(a_{20} - a_{21}x_i - a_{22}y_i) + v(y_j - y_i), \end{aligned}$$

    where $i, j = 1, 2$ and $j \neq i$. Select parameter values a_{ij} such that there exists a positive stable equilibrium $(\bar{x}, \bar{y})$, $5 \leq \bar{x} \leq 20$ and $5 \leq \bar{y} \leq 20$, when $u = 0 = v$. Let $X_i(t)$ and $Y_i(t)$, $i = 1, 2$ denote the random variables for the stochastic process with initial conditions $X_1(0) = \bar{x}$, $Y_1(0) = \bar{y}$ and $X_2(0) = 0 = Y_2(0)$.

    (a) Develop a continuous time Markov chain model based on the preceding deterministic model.

    (b) Let $u = 0 = v$ so that there is essentially one patch; no movement. Write a computer program to simulate 50 sample paths for a continuous time Markov chain model. Fix a time $T \geq 200$. Then record either (i) the time when either species 1 or species

2 equals zero (species extinction) or (ii) time has reached T and neither species is zero (coexistence). In what proportion of the sample paths are species 1 extinct? species 2 extinct? both coexist?

(c) For two patches, consider two cases: (i) $u = v$ and (ii) $u < v$. Write a computer program to simulate 50 sample paths for each of these sets of parameter values, and record the time when either species 1 or species 2 equals zero (species extinction in both patches) or time has reached T and neither species is zero (coexistence). In what proportion of the sample paths is species 1 extinct? species 2 extinct? both coexist? Discuss your results and compare them to part (b).

13. Search the literature and formulate a stochastic model based on a predator-prey, competition, or a population genetics process that differs from the ones discussed in this Chapter. In particular,

(a) Define the transition probabilities.

(b) Find the forward Kolmogorov differential equations satisfied by this process.

(c) Select some parameter values. Write a computer program for this process. Compute at least 100 sample paths. Analyze and discuss your results.

7.10 References for Chapter 7

Allen, E. J. 1999. Stochastic differential equations and persistence time for two interacting populations. *Dyn. Cont., Discrete and Impulsive Systems* 5: 271–281.

Allen, L. J. S., M. A. Jones, and C. F. Martin. 1991. A discrete-time model with vaccination for a measles epidemic. *Math. Biosci.* 105: 111–131.

Allen, L. J. S. and D. Thrasher. 1998. The effects of vaccination in an age-dependent model for varicella and herpes zoster. *IEEE Trans. Aut. Control* 43: 779–789.

Anderson, R. M. and R. May. 1986. The invasion, persistence and spread of infectious diseases within animal and plant communities. *Phil. Trans. Royal Soc.* B314: 533–570.

Anderson, R. M. and R. M. May. 1992. *Infectious Diseases of Humans Dynamics and Control.* Oxford University Press, Oxford.