A FIRST COURSE IN STOCHASTIC PROCESSES

SECOND EDITION

SAMUEL KARLIN

STANFORD UNIVERSITY
AND
THE WEIZMANN INSTITUTE OF SCIENCE

HOWARD M. TAYLOR

CORNELL UNIVERSITY

ACADEMIC PRESS New York San Francisco Londo

A Subsidiary of Harcourt Brace Jovanovich, Publishers

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1

Library of Congress Cataloging in Publication Data

Karlin, Samuel, (date)
A first course in stochastic processes. Second edition.

Includes bibliographical references.
1. Stochastic processes. I. Taylor, Howard M., joint author. II. Title.
QA274.K37 1974 519'.2 74-5705
ISBN 0-12-398552-8

PRINTED IN THE UNITED STATES OF AMERICA

CONTENTS

Chapter 3

THE BASIC LIMIT THEOREM OF MARKOV CHAINS AND APPLICATIONS

Chapter 4

CLASSICAL EXAMPLES OF CONTINUOUS TIME MARKOV CHAINS

Chapter 5

RENEWAL PROCESSES

Chapter 6

MARTINGALES

Chapter 7

BROWNIAN MOTION

Chapter 8

BRANCHING PROCESSES

Chapter 9

STATIONARY PROCESSES

Appendix

REVIEW OF MATRIX ANALYSIS

PREFACE

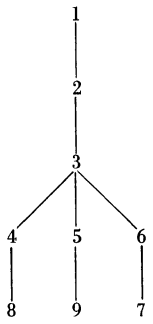
The purposes, level, and style of this new edition conform to the tenets set forth in the original preface. We continue with our tack of developing simultaneously theory and applications, intertwined so that they refurbish and elucidate each other.

We have made three main kinds of changes. First, we have enlarged on the topics treated in the first edition. Second, we have added many exercises and problems at the end of each chapter. Third, and most important, we have supplied, in new chapters, broad introductory discussions of several classes of stochastic processes not dealt with in the first edition, notably martingales, renewal and fluctuation phenomena associated with random sums, stationary stochastic processes, and diffusion theory.

Martingale concepts and methodology have provided a far-reaching apparatus vital to the analysis of all kinds of functionals of stochastic processes. In particular, martingale constructions serve decisively in the investigation of stochastic models of diffusion type. Renewal phenomena are almost equally important in the engineering and managerial sciences especially with reference to examples in reliability, queueing, and inventory systems. We discuss renewal theory systematically in an extended chapter. Another new chapter explores the theory of stationary processes and its applications to certain classes of engineering and econometric problems. Still other new chapters develop the structure and use of

diffusion processes for describing certain biological and physical systems and fluctuation properties of sums of independent random variables useful in the analyses of queueing systems and other facets of operations research.

The logical dependence of chapters is shown by the diagram below. Section 1 of Chapter 1 can be reviewed without worrying about details. Only Sections 5 and 7 of Chapter 7 depend on Chapter 6. Only Section 9 of Chapter 9 depends on Chapter 5.



An easy one-semester course adapted to the junior–senior level could consist of Chapter 1, Sections 2 and 3 preceded by a cursory review of Section 1, Chapter 2 in its entirety, Chapter 3 excluding Sections 5 and/or 6, and Chapter 4 excluding Sections 3, 7, and 8. The content of the last part of the course is left to the discretion of the lecturer. An option of material from the early sections of any or all of Chapters 5–9 would be suitable.

The problems at the end of each chapter are divided into two groups. The first, more or less elementary; the second, more difficult and subtle.

The scope of the book is quite extensive, and on this account, it has been divided into two volumes. We view the first volume as embracing the main categories of stochastic processes underlying the theory and most relevant for applications. In *A Second Course* we introduce additional topics and applications and delve more deeply into some of the issues of *A First Course*. We have organized the edition to attract a wide spectrum of readers including theorists and practitioners of stochastic analysis pertaining to the mathematical, engineering, physical, biological, social, and managerial sciences.

The second volume of this work, *A Second Course in Stochastic Processes*, will include the following chapters: (10) Algebraic Methods in Markov Chains; (11) Ratio Theorems of Transition Probabilities and Applications; (12) Sums of Independent Random Variables as a Markov Chain; (13)

Order Statistics, Poisson Processes, and Applications; (14) Continuous Time Markov Chains; (15) Diffusion Processes; (16) Compounding Stochastic Processes; (17) Fluctuation Theory of Partial Sums of Independent Identically Distributed Random Variables; (18) Queueing Processes.

As noted in the first preface, we have drawn freely on the thriving literature of applied and theoretical stochastic processes. A few representative references are included at the end of each chapter; these may be profitably consulted for more advanced material.

We express our gratitude to the Weizmann Institute of Science, Stanford University, and Cornell University for providing a rich intellectual environment, and facilities indispensable for the writing of this text. The first author is grateful for the continuing grant support provided by the Office of Naval Research that permitted an unencumbered concentration on a number of the concepts and drafts of this book. We are also happy to acknowledge our indebtedness to many colleagues who have offered a variety of constructive criticisms. Among others, these include Professors P. Brockwell of La Trobe, J. Kingman of Oxford, D. Iglehart and S. Ghurye of Stanford, and K. Itô and S. Stidham, Jr. of Cornell. We also thank our students M. Nedzela and C. Macken for their assistance in checking the problems and help in reading proofs.

SAMUEL KARLIN
HOWARD M. TAYLOR

PREFACE TO FIRST EDITION

Stochastic processes concern sequences of events governed by probabilistic laws. Many applications of stochastic processes occur in physics, engineering, biology, medicine, psychology, and other disciplines, as well as in other branches of mathematical analysis. The purpose of this book is to provide an introduction to the many specialized treatises on stochastic processes. Specifically, I have endeavored to achieve three objectives: (1) to present a systematic introductory account of several principal areas in stochastic processes, (2) to attract and interest students of pure mathematics in the rich diversity of applications of stochastic processes, and (3) to make the student who is more concerned with application aware of the relevance and importance of the mathematical subleties underlying stochastic processes.

The examples in this book are drawn mainly from biology and engineering but there is an emphasis on stochastic structures that are of mathematical interest or of importance in more than one discipline. A number of concepts and problems that are currently prominent in probability research are discussed and illustrated.

Since it is not possible to discuss all aspects of this field in an elementary text, some important topics have been omitted, notably stationary stochastic processes and martingales. Nor is the book intended in any sense as an authoritative work in the areas it does cover. On the contrary, its primary aim is simply to bridge the gap between an elementary

probability course and the many excellent advanced works on stochastic processes.

Readers of this book are assumed to be familiar with the elementary theory of probability as presented in the first half of Feller's classic *Introduction to Probability Theory and Its Applications*. In Section 1, Chapter 1 of my book the necessary background material is presented and the terminology and notation of the book established. Discussions in small print can be skipped on first reading. Exercises are provided at the close of each chapter to help illuminate and expand on the theory.

This book can serve for either a one-semester or a two-semester course, depending on the extent of coverage desired.

In writing this book, I have drawn on the vast literature on stochastic processes. Each chapter ends with citations of books that may profitably be consulted for further information, including in many cases bibliographical listings.

I am grateful to Stanford University and to the U.S. Office of Naval Research for providing facilities, intellectual stimulation, and financial support for the writing of this text. Among my academic colleagues I am grateful to Professor K. L. Chung and Professor J. McGregor of Stanford for their constant encouragement and helpful comments; to Professor J. Lamperti of Dartmouth, Professor J. Kiefer of Cornell, and Professor P. Ney of Wisconsin for offering a variety of constructive criticisms; to Dr. A. Feinstein for his detailed checking of substantial sections of the manuscript, and to my students P. Milch, B. Singer, M. Feldman, and B. Krishnamoorthi for their helpful suggestions and their assistance in organizing the exercises. Finally, I am indebted to Gail Lemmond and Rosemarie Stampfel for their superb technical typing and all-around administrative care.

SAMUEL KARLIN

A FIRST COURSE IN STOCHASTIC PROCESSES

SECOND EDITION

Chapter 1

ELEMENTS OF STOCHASTIC PROCESSES

The first part of this chapter summarizes the necessary background material and establishes the terminology and notation of the book. It is suggested that the reader not dwell here assidously, but rather quickly. It can be reviewed further if the need should arise later.

Section 2 introduces the celebrated Brownian motion and Poisson processes, and Section 3 surveys some of the broad types of stochastic processes that are the main concern of the remainder of the book.

The last section, included for completeness, discusses some technical considerations in the general theory. The section should be skipped on a first reading.

1: *Review of Basic Terminology and Properties of Random Variables and Distribution Functions*

The present section contains a brief review of the basic elementary notions and terminology of probability theory. The contents of this section will be used freely throughout the book without further reference. We urge the student to tackle the problems at the close of the chapter; they provide practice and help to illuminate the concepts. For more detailed treatments of these topics, the student may consult any good standard text for a first course in probability theory (see references at close of this chapter).

The following concepts will be assumed familiar to the reader:

(1) A real random variable X.
(2) The distribution function F of X [defined by $F(\lambda) = \Pr\{X \leq \lambda\}$] and its elementary properties.
(3) An event pertaining to the random variable X, and the probability thereof.
(4) $E\{X\}$, the expectation of X, and the higher moments $E\{X^n\}$.
(5) The law of total probabilities and Bayes rule for computing probabilities of events.

The abbreviation r.v. will be used for "real random variables." A r.v.

X is called *discrete* if there is a finite or denumerable set of distinct values $\lambda_1, \lambda_2, \ldots$ such that $a_i \equiv \Pr\{X = \lambda_i\} > 0$, $i = 1, 2, 3, \ldots$, and $\sum_i a_i = 1$. If $\Pr\{X = \lambda\} = 0$ for every value of λ, the r.v. X is called ***continuous***. If there is a nonnegative function $p(t)$, defined for $-\infty < t < \infty$ such that the distribution function F of the r.v. X is given by

$$F(\lambda) = \int_{-\infty}^{\lambda} p(t)\, dt,$$

then p is said to be the probability density of X. If X has a probability density, then it is necessarily continuous; however, examples are known of continuous r.v.'s which do not possess probability densities.

If X is a discrete r.v., then its mth moment is given by

$$E[X^m] = \sum_i \lambda_i^m \Pr\{x = \lambda_i\}$$

(where the λ_i are as earlier), if the series converges absolutely.

If X is a continuous r.v. with probability density $p(\cdot)$, its mth moment is given by

$$E[X^m] = \int_{-\infty}^{\infty} x^m p(x)\, dx,$$

provided the integral converges absolutely.

The first moment of X, commonly called the *mean*, is denoted by m_X or μ_X. The mth central moment of X is defined as the mth moment of the r.v. $X - m_X$ if m_X exists. The first central moment is evidently zero; the second central moment is called the *variance* (σ_X^2) *of* X. The *median* of a r.v. X is any value v with the property that $\Pr\{X \geq v\} \geq \frac{1}{2}$ and $\Pr\{X \leq v\} \geq \frac{1}{2}$.

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

$$E[g(X)] = \sum_{i=1}^{\infty} g(x_i) \Pr\{X = x_i\} \tag{1.1}$$

provided the sum converges absolutely. If X is continuous and has the probability density function p_X then the expectation of $g(X)$ is computed from

$$E[g(X)] = \int g(x) p_X(x)\, dx. \tag{1.2}$$

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

$$E[g(X)] = \int g(x)\, dF_X(x) \tag{1.3}$$

where F_X is the distribution function of the random variable X. Technically speaking, the integral in (1.3) is called a Lebesgue-Stieltjes integral. We do not require knowledge of such integrals in this text but interpret (1.3) to signify (1.1) when X is a discrete random variable and to represent (1.2) when X possesses a probability density function p_X.

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

$$\begin{aligned} E[Y] &= \sum y_i \Pr\{Y = y_i\} \\ &= \sum g(x_i) \Pr\{X = x_i\} \end{aligned}$$

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

$$\begin{aligned} E[Y] &= \int y \, dF_Y(y) \\ &= \int g(x) \, dF_X(x). \end{aligned} \tag{1.4}$$

If X is a discrete random variable then so is $Y = g(X)$. It may be, however, that X is a continuous random variable while Y is discrete (the student should provide an example). Even so, one may compute $E[Y]$ from either form in (1.4) with the same result.

A. JOINT DISTRIBUTION FUNCTIONS

Given a pair (X, Y) of r.v.'s, their joint distribution function is the function F_{XY} of two real variables given by

$$F(\lambda_1, \lambda_2) = F_{XY}(\lambda_1, \lambda_2) = \Pr\{X \leq \lambda_1, Y \leq \lambda_2\}.$$

(The subscripts X, Y will usually be omitted unless there is possible ambiguity.)

The function $F(\lambda, +\infty) \equiv \lim_{\lambda_2 \to \infty} F(\lambda, \lambda_2)$ is a probability distribution function, called the *marginal distribution function* of X. Similarly, the function $F(+\infty, \lambda)$ is called the marginal distribution of Y. If it happens that $F(\lambda_1, +\infty) \cdot F(+\infty, \lambda_2) = F(\lambda_1, \lambda_2)$ for every choice of λ_1, λ_2, then the r.v.'s X and Y are said to be *independent*. A joint distribution function F_{XY} is said to possess a (joint) probability density if there exists a function $p_{XY}(s, t)$ of two real variables such that

$$F_{XY}(\lambda_1, \lambda_2) = \int_{-\infty}^{\lambda_2} \int_{-\infty}^{\lambda_1} p_{XY}(s, t) \, ds \, dt$$

for all λ_1, λ_2. If X and Y are independent, then $p_{XY}(s, t)$ is necessarily of

the form $p_X(s)p_Y(t)$, where p_X and p_Y are the probability densities of the marginal distribution of X and Y, respectively.

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

$$\begin{aligned} F(\lambda_1, \ldots, \lambda_n) &= F_{X_1,\ldots,X_n}(\lambda_1, \ldots, \lambda_n) \\ &= \Pr\{X_1 \leq \lambda_1, \ldots, X_n \leq \lambda_n\}. \end{aligned}$$

The distribution function

$$F_{X_{i_1},\ldots,X_{i_k}}(\lambda_{i_1}, \ldots, \lambda_{i_k}) = \lim_{\lambda_i \to \infty,\, i \neq i_1,\ldots,\, i_k} F(\lambda_1, \ldots, \lambda_n)$$

is called the marginal distribution of the random variables $X_{i_1}, \ldots, X_{i_k}$.

If $F(\lambda_1, \ldots, \lambda_n) = F_{X_1}(\lambda_1) \cdot \ldots \cdot F_{X_n}(\lambda_n)$ for all values of $\lambda_1, \lambda_2, \ldots, \lambda_n$, the random variables $X_1, \ldots, X_n$ are said to be independent.

A joint distribution function $F(\lambda_1, \ldots, \lambda_n)$ is said to have a probability density if there exists a nonnegative function $p(t_1, \ldots, t_n)$ of n variables such that

$$F(\lambda_1, \ldots, \lambda_n) = \int_{-\infty}^{\lambda_n} \cdots \int_{-\infty}^{\lambda_1} p(t_1, \ldots, t_n)\, dt_1 \cdots dt_n$$

for all real $\lambda_1, \ldots, \lambda_n$.

If X and Y are jointly distributed random variables having means m_X and m_Y, respectively, their covariance (σ_{XY}) is the product moment

$$\sigma_{XY} = E[(X - m_X)(Y - m_Y)].$$

If X_1 and X_2 are independent random variables having the distribution functions F_1 and F_2, respectively, then the distribution function F of the sum $X = X_1 + X_2$ is the *convolution* of F_1 and F_2:

$$\begin{aligned} F(x) &= \int F_1(x - y)\, dF_2(y) \\ &= \int F_2(x - y)\, dF_1(y). \end{aligned}$$

Specializing to the situation where X_1 and X_2 have the probability densities p_1 and p_2, the density function p of the sum $X = X_1 + X_2$ is the convolution of the densities p_1 and p_2:

$$\begin{aligned} p(x) &= \int p_1(x - y)p_2(y)\, dy \\ &= \int p_2(x - y)p_1(y)\, dy. \end{aligned}$$

B. CONDITIONAL DISTRIBUTIONS AND CONDITIONAL EXPECTATIONS

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

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

and is left undefined, or assigned an arbitrary value, when $\Pr\{B\} = 0$. Let X and Y be random variables which can attain only countably many different values, say 1, 2, ... The *conditional distribution function* $F_{X|Y}(\cdot\,|\,y)$ of X given $Y = y$ is defined by

$$F_{X|Y}(x|y) = \frac{\Pr\{X \leq x,\ Y = y\}}{\Pr\{Y = y\}}, \quad \text{if} \quad \Pr\{Y = y\} > 0,$$

and any arbitrary discrete distribution function whenever $\Pr\{Y = y\} = 0$. This last prescription is consistent with the subsequent calculations invoked on conditional distribution functions.

Suppose X and Y are jointly distributed continuous random variables having the joint probability density function $p_{XY}(x, y)$. Then the conditional distribution of X given $Y = y$ is given by

$$F_{X|Y}(x|y) = \frac{\int_{\xi \leq x} p_{XY}(\xi, y)\, d\xi}{p_Y(y)}$$

wherever $p_Y(y) > 0$, and with an arbitrary specification where $p_Y(y) = 0$.

Note that $F_{X|Y}$ satisfies

(C.P.1) $F_{X|Y}(x|y)$ is a probability distribution function in x for each fixed y;

(C.P.2) $F_{X|Y}(x|y)$ is a function of y for each fixed x; and

(C.P.3) For any values x, y

$$\Pr\{X \leq x,\ Y \leq y\} = \int_{\eta \leq y} F_{X|Y}(x|\eta)\, dF_Y(\eta)$$

where $F_Y(\eta) = \Pr\{Y \leq \eta\}$ is the marginal distribution of Y. As noted earlier, in this book, the reader need only deal with the integral in (C.P.3) for the discrete and continuous versions. To wit, when Y is a continuous random variable having the probability density function $p_Y(y)$ the integral in (C.P.3) is computed as

$$\Pr\{X \leq x,\ Y \leq y\} = \int_{\eta \leq y} F_{X|Y}(x|\eta) p_Y(\eta)\, d\eta.$$

And when Y is discrete the formula is

$$\Pr\{X \leq x,\ Y \leq y\} = \sum_{\eta \leq y} F_{X|Y}(x|\eta)\ \Pr\{Y = \eta\}.$$

These three properties capture the essential features of conditional distributions. In fact, from (C.P.3) we obtain

$$\begin{aligned}\Pr\{X \leq x,\ Y = y\} &= \Pr\{X \leq x,\ Y \leq y\} - \Pr\{X \leq x,\ Y < y\} \\ &= \sum_{\eta \leq y} F_{X|Y}(x|\eta)\ \Pr\{Y = \eta\} - \sum_{\eta < y} F_{X|Y}(x|\eta)\ \Pr\{Y = \eta\} \\ &= F_{X|Y}(x|y)\ \Pr\{Y = y\}\end{aligned}$$

which then implies the definition $F_{X|Y}(x|y) = \Pr\{X \leq x,\ Y = y\}/\Pr\{Y = y\}$, at least where $\Pr\{Y = y\} > 0$.

In advanced work,* (C.P.1-3) is taken as the basis for the definition of conditional distributions. It can be established that such conditional distributions exist for arbitrary real random variables X and Y, and even for real random vectors $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$.

The application of (C.P.3) in the case $y = \infty$ produces the *law of total probability*

$$\begin{aligned}\Pr\{X \leq x\} &= \Pr\{X \leq x,\ Y \leq \infty\} \\ &= \int_{-\infty}^{+\infty} F_{X|Y}(x|y)\, dF_Y(y),\end{aligned}$$

which is one of the most fundamental formulas of probability analysis. When Y is discrete this relation becomes

$$\Pr\{X \leq x\} = \sum_{y} \Pr\{X \leq x | Y = y\}\ \Pr\{Y = y\}$$

and where Y has the probability density function $p_Y(y)$ we have

$$\Pr\{X \leq x\} = \int_{-\infty}^{+\infty} \Pr\{X \leq x | Y = y\} p_Y(y)\, dy.$$

When X and Y are jointly distributed continuous random variables,

* For more explication, including rigorous and intuitive discussions on conditional expectations the reader can consult Section 7, Chapter 6. These concepts play a fundamental role in the modern development of martingale theory.

we may define the *conditional density function* $p_{X|Y}(x|y)$ of X given $Y=y$ by

$$p_{X|Y}(x|y) = \frac{d}{dx} F_{X|Y}(x|y)$$

$$= \frac{p_{XY}(x,y)}{p_Y(y)}$$

at values y for which $p_Y(y) > 0$, and as a fixed arbitrary probability density function when $p_Y(y) = 0$.

Let g be a function for which the expectation of $g(X)$ is finite. The conditional expectation of $g(X)$ given $Y=y$ can be expressed in the form

$$E[g(X)|\,Y=y] = \int_x g(x)\, dF_{X|Y}(x|y).$$

When X and Y are jointly continuous random variables, $E[g(X)|\,Y=y]$ may be computed from

$$E[g(X)|\,Y=y] = \int g(x) p_{X|Y}(x|y)\, dx$$

$$= \frac{\int g(x) p_{XY}(x,y)\, dx}{p_Y(y)}, \quad \text{if} \quad p_Y(y) > 0, \tag{1.5}$$

and if X and Y are jointly distributed discrete random variables, taking the possible values $x_1, x_2, \ldots$, then the detailed formula reduces to

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

$$= \frac{\sum_{i=1}^{\infty} g(x_i)\ \Pr\{X=x_i,\ Y=y\}}{\Pr\{Y=y\}}, \quad \text{if} \quad \Pr\{Y=y\} > 0.$$

In parallel with (C.P.1–3) we see that the conditional expectation of $g(X)$ given $Y=y$ satisfies

(C.E.1) $E[g(X)|\,Y=y]$ is a function of y for each function g for which $E[|g(X)|] < \infty$; and

(C.E.2) For any bounded function h we have

$$E[g(X)h(Y)] = \int E[g(X)|\,Y=y]h(y)\, dF_Y(y)$$

where F_Y is the marginal distribution function for Y.

Let us validate the latter formula in the continuous case.

We will stipulate that the set of values y for which $p_Y(y) > 0$ is an interval (a, b) where $-\infty \le a < b \le +\infty$. We first insert the appropriate

probability density functions, and then substitute (1.5). This gives

$$\begin{aligned}\int E[g(X)|Y=y]h(y)\,dF_Y(y) &= \int_a^b E[g(X)|Y=y]h(y)p_Y(y)\,dy \\ &= \int_a^b \left(\int_{-\infty}^{+\infty} g(x)p_{X|Y}(x|y)\,dx\right)h(y)p_Y(y)\,dy \\ &= \int_a^b \left(\int_{-\infty}^{+\infty} g(x)\frac{p_{XY}(x,y)}{p_Y(y)}\,dx\right)h(y)p_Y(y)\,dy \\ &= \int_a^b \int_{-\infty}^{+\infty} g(x)h(y)p_{XY}(x,y)\,dx\,dy \\ &= E[g(X)h(Y)].\end{aligned}$$

In the last step we have used that $p_{XY}(x, y) > 0$ only when $a < y < b$.

The special case in (C.E.2) with $h(y) \equiv 1$ produces the formula expressing the *law of total probability* for expectations,

$$E[g(X)] = \int E[g(X)|Y=y]\,dF_Y(y),$$

which, when Y is discrete, becomes

$$E[g(X)] = \sum_{i=1}^{\infty} E[g(X)|Y=y_i]\Pr\{Y=y_i\}$$

and, when Y has a probability density function p_Y, becomes

$$E[g(X)] = \int E[g(X)|Y=y]p_Y(y)\,dy.$$

Since the conditional expectation of $g(X)$ given $Y=y$ is the expectation with respect to the conditional distribution $F_{X|Y}$, conditional expectations behave in many ways like ordinary expectations. In particular, if a_1 and a_2 are fixed numbers and g_1 and g_2 are given functions for which $E[|g_i(X)|] < \infty$, $i = 1, 2$, then

$$\begin{aligned}&E[a_1g_1(X) + a_2g_2(X)|Y=y] \\ &\qquad = a_1E[g_1(X)|Y=y] + a_2E[g_2(X)|Y=y].\end{aligned}$$

According to (C.E.1), $E[g(X)|Y=y]$ is a function of the real variable y. If we evaluate this function at the random variable Y, we obtain a random variable which we denote by $E[g(X)|Y]$. The basic property (C.E.2) then is stated for any bounded function h of y,

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

When $h(y) = 1$ for all y, we get the law of total probability in the form

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

The following list summarizes these and other properties of conditional expectations. Here, with or without affixes, X and Y are random variables, c is a real number, g is a function for which $E[|g(X)|] < \infty$, f is a bounded function and h is a function of two variables for which $E[|h(X, Y)|] < \infty$.

$$E[a_1 g(X_1) + a_2 g(X_2)|\,Y] = a_1 E[g(X_1)|\,Y] + a_2 E[g(X_2)|\,Y], \tag{1.6}$$

$$g \geq 0 \quad \text{implies} \quad E[g(X)|\,Y] \geq 0, \tag{1.7}$$

$$E[h(X, Y)|\,Y = y] = E[h(X, y)|\,Y = y], \tag{1.8}$$

$$E[g(X)|\,Y] = E[g(X)] \quad \text{if } X \text{ and } Y \text{ are independent}, \tag{1.9}$$

$$E[g(X)f(Y)|\,Y] = f(Y)E[g(X)|\,Y], \tag{1.10}$$

and

$$E[g(X)f(Y)] = E\{E[g(X)|\,Y]f(Y)\}. \tag{1.11}$$

As consequences of (1.6), (1.10) and (1.11), with either $g \equiv 1$ or $f \equiv 1$, we obtain,

$$E[c|\,Y] = c, \tag{1.12}$$

$$E[f(Y)|\,Y] = f(Y), \tag{1.13}$$

and

$$E[g(X)] = E\{E[g(X)|\,Y]\}. \tag{1.14}$$

C. INFINITE FAMILIES OF RANDOM VARIABLES

In dealing with an infinite family of random variables, a direct generalization of the preceding definitions involves substantial difficulties. We need to adopt a slightly modified approach.

Given a denumerably infinite family $X_1, X_2, \ldots$ of r.v.'s, their statistical properties are regarded as defined by prescribing, for each integer $n \geq 1$ and every set $i_1, \ldots, i_n$ of n distinct positive integers, the joint distribution function $F_{X_{i_1}, \ldots, X_{i_n}}$ of the random variables $X_{i_1}, \ldots, X_{i_n}$. Of course, some consistency requirements must be imposed upon the infinite family $F_{X_{i_1}, \ldots, X_{i_n}}$, namely, that

$$F_{X_{i_1}, \ldots, X_{i_{j-1}}, X_{i_{j+1}}, \ldots, X_{i_n}}(\lambda_1, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_n) = \lim_{\lambda_j \to \infty} F_{X_{i_1}, \ldots, X_{i_n}}(\lambda_1, \ldots, \lambda_{j-1}, \lambda_j, \lambda_{j+1}, \ldots, \lambda_n)$$

and that the distribution function obtained from

$$F_{X_{i_1}, X_{i_2}, \ldots, X_{i_n}}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

by interchanging two of the indices i_ν and i_μ and the corresponding variables λ_ν and λ_μ should be invariant. This simply means that the manner of labeling the random variables $X_1, X_2, \ldots$ is not relevant.

The joint distributions $\{F_{X_{i_1}, \ldots, X_{i_n}}\}$ are called the *finite-dimensional distributions* associated with $\{X_n\}_{n=1}^{\infty}$. In principle, all important probabilistic quantities of the variables $\{X_n\}_{n=1}^{\infty}$ can be computed in terms of the finite-dimensional distributions.

D. CHARACTERISTIC FUNCTIONS

An important function associated with the distribution function F of a r.v. X is its characteristic function $\phi(t)$ (abbreviated c.f.), where t is a real variable $-\infty < t < \infty$. We write it suggestively in the form

$$\begin{aligned} \phi(t) &= \int_{-\infty}^{\infty} e^{it\lambda}\, dF(\lambda), \quad i = \sqrt{-1} \\ &= E[e^{itX}]. \end{aligned} \tag{1.15}$$

Again the reader should interpret (1.15) symbolically. If F has a probability density function p, the characteristic function becomes

$$\phi(t) = \int_{-\infty}^{\infty} e^{it\lambda} p(\lambda)\, d\lambda.$$

When F is a distribution of a discrete r.v. X with possible values $\{\lambda_k\}_{k=0}^{\infty}$ and $\Pr\{X = \lambda_k\} = a_k$ $(k = 0, 1, \ldots)$, then (1.1) reduces to the series expression

$$\phi(t) = \sum_{k=0}^{\infty} e^{it\lambda_k} a_k.$$

Much of the importance of characteristic functions derives from the following three results:

(a) The relation between distribution functions and characteristic functions is one-to-one. Thus, knowing the characteristic function is synonomous to knowing the distribution function. The equation which expresses the distribution function in terms of its characteristic function is known as Levy's inversion formula; as we do not need it, we refer the reader to one of the references for a discussion of this matter.

(b) If $X_1, \ldots, X_n$ are independent r.v.'s, the characteristic function of their sum is the product of their characteristic functions. This simple

result makes characteristic functions extremely expeditious for dealing with problems involving sums of independent random variables.

(c) When they are finite, the moments of a random variable may be determined by differentiating the characteristic function. The explicit relation is

$$E[X^k] = \frac{1}{i^k}\phi^{(k)}(0)$$

where $i = \sqrt{-1}$ and $\phi^{(k)}(t) = d^k\phi(t)/dt^k$ is the kth derivative of the c.f. $\phi(t)$.

The one-to-one correspondence between distribution functions and their characteristic functions is also preserved by various limiting processes. In fact, if F, F_1, F_2, ... are distribution functions such that $\lim_{n\to\infty} F_n(\lambda) = F(\lambda)$ for every λ at which F is continuous and $\phi_n(t)$ is the c.f. of F_n, then

$$\phi_n(t) = \int_{-\infty}^{\infty} e^{it\lambda}\, dF_n(\lambda) \to \phi(t) = \int_{-\infty}^{\infty} e^{it\lambda}\, dF(\lambda)$$

uniformly in every finite interval. Conversely, if ϕ_1, ϕ_2, ... are the characteristic functions of distribution functions F_1, F_2, ... and $\lim_{n\to\infty} \phi_n(t) = \phi(t)$ for every t, and $\phi(t)$ is continuous at $t = 0$, then $\phi(t)$ is the c.f. of a distribution function F and $\lim_{n\to\infty} F_n(\lambda) = F(\lambda)$ for every λ at which F is continuous. This result is known as Levy's convergence criterion.

E. GENERATING FUNCTIONS AND LAPLACE TRANSFORMS

For random variables whose only possible values are the nonnegative integers, a function related to the characteristic function is the generating function, defined by

$$\begin{aligned} g(s) &= \sum_{k=0}^{\infty} p_k s^k \\ &= E[s^X], \end{aligned}$$

where

$$p_k = \Pr\{X = k\}.$$

Since by hypothesis $p_k \geq 0$ and $\sum_{k=0}^{\infty} p_k = 1$, $g(s)$ is defined at least for

$|s| \leq 1$ (s is a complex variable) and is infinitely differentiable for $|s| < 1$. The generating function of a nonnegative integer valued random variable X is related to the characteristic function ϕ of X formally through a change of variable $s = e^{it}$:

$$\begin{aligned}\phi(t) &= E[e^{itX}] \\ &= E[(e^{it})^X] \\ &= g(e^{it}).\end{aligned}$$

Thus generating functions inherit the three basic properties of characteristic functions;

(a) A generating function determines the distribution function uniquely;

(b) The generating function of a sum of independent nonnegative integer valued random variables is the product of their generating functions; and

(c) The moments may be obtained through successive differentiation. The factorial moments are given by

$$E[X(X-1)\cdot\ldots\cdot(X-k)] = g^{(k+1)}(1),$$

where $g^{(k)}(s) = d^k g(s)/ds^k$ is the *kth* derivative of g. Hence

$$E[X] = g^{(1)}(1)$$

and

$$E[X^2] = g^{(2)}(1) + g^{(1)}(1).$$

We give an example of the use of generating functions in working with sums of independent random variables. Let $N, X_1, X_2, \ldots$ be independent nonnegative integer valued random variables and suppose we wish to determine the generating function $g_R(s)$ of the sum $R = X_1 + \cdots + X_N$, a sum of random variables with a random number of terms.

Let $g_N(s)$ be the generating function of N and suppose the X_i have the same distribution function with common generating function $g(s)$. Then, using (1.14) and (1.9),

$$\begin{aligned}g_R(s) &= E[s^R] \\ &= E[s^{X_1+\cdots+X_N}] \\ &= E\{E[s^{X_1+\cdots+X_N}|N]\} \\ &= \sum_{n=0}^{\infty} E[s^{X_1+\cdots+X_n}|N=n]\Pr\{N=n\} \\ &= \sum_{n=0}^{\infty} E[s^{X_1+\cdots+X_n}]\Pr\{N=n\}\end{aligned}$$

(since N and X_i are independent)

$$\begin{aligned}&= \sum_{n=0}^{\infty} g(s)^n \Pr\{N=n\} \\ &= E[g(s)^N] \\ &= g_N[g(s)].\end{aligned}$$

To sum up:

$$g_R(s) = g_N[g(s)].$$

Using the chain rule for differentiation we calculate

$$g'_R(s) = g'_N[g(s)] \cdot g'(s)$$

and setting $s = 1$ we can infer

$$E[R] = E[N] \cdot E[X].$$

In a similar fashion we calculate the variance of R, σ_R^2, given by

$$\sigma_R^2 = E[X]^2 \cdot \sigma_N^2 + E[N] \cdot \sigma_X^2,$$

where σ_N^2 and σ_X^2 are the variances of N and X, respectively. (See Elementary Problem 4.)

The following slight extension is also available. Let $X_1, X_2, \ldots$ be arbitrary independent identically distributed random variables (i.e., not necessarily integer valued), and let N be as above. Then

$$\phi_R(t) = g_N(\phi(t)),$$

where ϕ_R and g_N are the characteristic function and generating function of $R = X_1 + \cdots + X_N$ and N, respectively, and ϕ is the common characteristic function of the X_i.

When considering nonnegative r.v.'s it is more natural to replace the characteristic function by the Laplace transform of the distribution function. If the distribution F_X has a density p_X, the Laplace transform is defined as

$$\psi_X(s) = \int_0^\infty e^{-sx} p_X(x)\, dx.$$

This integral exists for a complex variable s, where $s = \sigma + it$, σ and t real, $\sigma \geq 0$. When s is purely imaginary, $s = it$, $\psi_X(s)$ reduces to the characteristic function $\phi_X(-t)$. For a discrete nonnegative r.v. the Laplace transform is defined as

$$\psi_X(s) = \sum_{n=0}^{\infty} e^{-s\lambda_n} \Pr\{X = \lambda_n\}.$$

As in the case of characteristic functions, if $X_1, X_2, \ldots, X_n$ are nonnegative independent r.v.'s then

$$\psi_{X_1 + \cdots + X_n}(s) = \prod_{k=1}^{n} \psi_{X_k}(s).$$

In the case of general distribution functions we write

$$\psi_X(s) = \int_0^\infty e^{-s\xi}\, dF_X(\xi)$$

for the Laplace transform.

As in the case of c.f.'s the Laplace transform uniquely determines the distribution function.

F. EXAMPLES OF DISTRIBUTION FUNCTIONS

Some elementary properties of several distribution functions are given in Tables I and II.

Two multivariate distributions of fundamental importance are:

(a) *Multivariate Normal*

Let σ_1, σ_2, m_1, m_2 and ρ be real constants subject to $\sigma_i > 0$, $i = 1, 2$ and $0 < \rho < 1$. Let

$$Q(x_1, x_2) = \frac{1}{1-\rho^2}\left\{\left(\frac{x_1 - m_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1 - m_1}{\sigma_1}\right)\left(\frac{x_2 - m_2}{\sigma_2}\right) + \left(\frac{x_2 - m_2}{\sigma_2}\right)^2\right\}.$$

If X_1 and X_2 are r.v.'s for which

$$\Pr\{X_1 \leq a,\ X_2 \leq b\} = \int_{-\infty}^{a}\int_{-\infty}^{b} \frac{1}{2\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}Q(x_1, x_2)\right\} dx_1\, dx_2$$

then X_1 and X_2 are said to have a *joint normal distribution*. It can be verified that $E[X_i] = m_i$ for $i = 1, 2$ and that the variance of X_i is σ_i^2. The covariance is given by

$$E[(X_1 - m_1)(X_2 - m_2)] = \rho\sigma_1\sigma_2$$

and ρ (a dimensionless variable) is called the *correlation coefficient*. The joint characteristic function is

$$\begin{aligned}\phi_{X_1, X_2}(t_1, t_2) &= E[e^{i(t_1X_1 + t_2X_2)}] \\ &= \exp\{i(t_1m_1 + t_2m_2) - \tfrac{1}{2}(t_1^2\sigma_1^2 + 2\rho t_1\sigma_1 t_2\sigma_2 + t_2^2\sigma_2^2)\}.\end{aligned}$$

If X_1 and X_2 have a joint normal distribution then the conditional distribution of X_2 given $X_1 = x_1$ is also normal with probability density function

$$p_{X_2|X_1}(x_2|x_1) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_2 - m}{\sigma}\right)^2\right], \quad -\infty < x_2 < \infty,$$

where

$$m = m_2 + \frac{\sigma_2}{\sigma_1}\rho(x_1 - m_1)$$

and

$$\sigma = \sigma_2\sqrt{1-\rho^2}.$$

Some Frequently Encountered Continuous Probability Distributions

Continuous distribution function	*Density*, $p(x)$	*Range of parameters*	*Characteristic Function* $\phi(t)$	*Mean*	*Variance*
Normal	$\frac{1}{\sqrt{2\pi}\sigma}\exp-\frac{(x-m)^2}{2\sigma^2}$ for $-\infty<x<\infty$	m real $\sigma>0$	$\exp\left[-\frac{\sigma^2t^2}{2}+imt\right]$	m	σ^2
Exponential	$\lambda e^{-\lambda x}$ for $x>0$	$\lambda>0$	$\frac{\lambda}{\lambda-it}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$\frac{\lambda}{\Gamma(\alpha)}(\lambda x)^{\alpha-1}e^{-\lambda x}$ for $x>0$	$\lambda>0$ $\alpha>0$	$\frac{\lambda^\alpha}{(\lambda-it)^\alpha}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Uniform	$\frac{1}{b-a}$ for $a<x<b$	$a<b$	$\frac{e^{iub}-e^{iua}}{iu(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Beta with parameters p, q	$\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}x^{p-1}(1-x)^{q-1}$ for $0<x<1$	$p>0$ $q>0$	$\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^1 e^{itx}x^{p-1}(1-x)^{q-1}\,dx$	$\frac{p}{p+q}$	$\frac{qp}{(p+q)^2(p+q+1)}$

Note: The gamma distribution with $\alpha=1$ is the exponential distribution where the parameter λ occurs as a scale factor. The beta distribution of parameters $p=q=1$ is the uniform distribution on (0, 1) sometimes abbreviated "uniform (0, 1)".

TABLE II

Some Frequently Encountered Discrete Probability Distributions

Discrete distribution function	*Probability Mass Function*	*Possible values of parameters*	*Generating function*	*Mean*	*Variance*
Poisson with parameter $\lambda > 0$	$\frac{e^{-\lambda}\lambda^n}{n!}$ for $n = 0, 1, 2, \ldots$	$\lambda > 0$	$e^{-\lambda+\lambda s}$	λ	λ
Binomial	$\binom{N}{n} p^n q^{N-n}$ for $n = 0, 1, \ldots, N$	$N = 1, 2, \ldots$ $0 < p < 1$ $q = 1 - p$	$(1 - p + ps)^N$	Np	Npq
Negative binomial (Pascal)	$\binom{\alpha + n - 1}{n} p^\alpha q^n$ for $n = 0, 1, 2, \ldots$	$\alpha > 0$ $0 < p < 1$	$\left(\frac{p}{1 - qs}\right)^\alpha$	$\frac{\alpha q}{p}$	$\frac{\alpha q}{p^2}$
Geometric	$p(1 - p)^n$ for $n = 0, 1, 2, \ldots$	$0 < p < 1$	$\frac{p}{1 - qs}$	$\frac{q}{p}$	$\frac{q}{p^2}$

Let $\|a_{ij}\|$ be an $n \times n$ symmetric positive definite matrix, let $\|b_{ij}\|$ be the inverse matrix of $\|a_{ij}\|$ and let $B = \det\|b_{ij}\|$ be the determinant of $\|b_{ij}\|$. Let m_i, $i = 1, \ldots, n$ be any real constants. The random variables $X_1, \ldots, X_n$ are said to have a joint normal distribution if they possess a probability density function of the form

$$p(x_1, \ldots, x_n) = \frac{\sqrt{B}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2} Q(x_1, \ldots, x_n)\right\}, \qquad -\infty < x_i < \infty$$

where

$$Q(x_1, \ldots, x_n) = \sum_{i,j} (x_i - m_i) b_{ij} (x_j - m_j).$$

The joint characteristic function is

$$\phi(t_1, \ldots, t_n) = E[\exp\{i \sum_{i=1}^{n} t_i X_i\}]$$

$$= \exp\left\{i \sum_{i=1}^{n} t_i m_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} t_i a_{ij} t_j\right\}.$$

From this one can compute

$$E[X_i] = m_i \quad \text{for} \quad i = 1, \ldots, n$$

and

$$E[(X_i - m_i)(X_j - m_j)] = a_{ij},$$

which justifies the name *covariance matrix* for the matrix $\|a_{ij}\|$.

From the nature of the characteristic function it is easily checked that $X_1, \ldots, X_n$ have a joint normal distribution if and only if $Y = a_1 X_1 + \cdots + a_n X_n$ has a normal distribution for every choice of real numbers $a_1, \ldots, a_n$.

(b) *The Multinomial Distribution*

This is a discrete joint distribution of r variables in which only non-negative integer values $0, \ldots, n$ are possible. It is defined by

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

where $p_i > 0$, $i = 1, \ldots, r$, and $\sum_{i=1}^{r} p_i = 1$.

The joint generating function is given by

$$g(s_1, \ldots, s_r) = E[s_1^{X_1} \cdots s_r^{X_r}]$$

$$= (p_1 s_1 + \cdots + p_r s_r)^n.$$

G. LIMIT THEOREMS

A sequence $\{a_n\}$ of real numbers is said to converge to a real number a, written $\lim_{n\to\infty} a_n = a$, if for every positive ε there exists a number $N(\varepsilon)$ such that $|a_n - a| < \varepsilon$ for all $n > N(\varepsilon)$. There are several ways to generalize this concept to random variables. Let Z, Z_1, Z_2, ... be jointly distributed random variables.

(a) *Convergence with probability one*

We say Z_n converges to Z with probability one if

$$\Pr\{\lim_{n\to\infty} Z_n = Z\} = 1.$$

In words, $\lim_{n\to\infty} z_n = z$ for a set of outcomes $Z = z$, $Z_1 = z_1$, $Z_2 = z_2$, ... having total probability one.

(b) *Convergence in probability*

We say Z_n converges to Z in probability if for every positive ε

$$\lim_{n\to\infty} \Pr\{|Z_n - Z| > \varepsilon\} = 0,$$

or conversely, if for every positive ε

$$\lim_{n\to\infty} \Pr\{|Z_n - Z| \leq \varepsilon\} = 1.$$

In words, by taking n sufficiently large, one can achieve arbitrarily high probability that Z_n is arbitrarily close to Z.

(c) *Convergence in quadratic mean*

We say Z_n converges to Z in quadratic mean if

$$\lim_{n\to\infty} E[|Z_n - Z|^2] = 0.$$

In words, by making n sufficiently large, one can ensure that Z_n is arbitrarily close to Z in the sense of mean square difference.

(d) *Convergence in distribution* (= *Convergence in law*)

Let $F(t) = \Pr\{Z \leq t\}$ and $F_k(t) = \Pr\{Z_k \leq t\}$, $k = 1, 2, \ldots$. We say Z_n converges in distribution to Z (or F_n converges in distribution to F) if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

for all t at which F is continuous.

It can be proved that if Z_n converges to Z with probability one, then Z_n converges to Z in probability, and that this in turn implies that Z_n converges to Z in distribution. Thus convergence in distribution is the

weakest form of convergence. In fact it can be shown that every family $\{F_\alpha\}$ of distribution functions contains a sequence $\{F_{\alpha_n}\}$ that converges to a function F at all t at which F is continuous (*the Helly-Bray Lemma*) but F may not be a proper distribution in that $F(\infty)$ may be less than one.

Many of the basic results of probability theory are in the form of limit theorems and we will mention a few here. (We do not state these results under the weakest possible hypotheses.)

Let $X_1, X_2, \ldots$ be independent identically distributed random variables with finite mean m. Let $S_n = X_1 + \cdots + X_n$ and let $\bar{X}_n = S_n/n$ be the sample mean.

Law of Large Numbers (Weak). $\bar{X}_n$ converges in probability to m. That is, for any positive ε

$$\lim \Pr\{|\bar{X}_n - m| > \varepsilon\} = 0.$$

Law of Large Numbers (Strong). $\bar{X}_n$ converges to m with probability one. That is

$$\Pr\{\lim_{n\to\infty} \bar{X}_n = m\} = 1.$$

Central Limit Theorem. Suppose each X_k has the finite variance σ^2. Let

$$Z_n = \frac{S_n - nm}{\sigma\sqrt{n}}$$

$$= \frac{1}{\sigma}(\bar{X}_n - m)\sqrt{n},$$

and let Z be a normally distributed random variable having mean zero and variance one. Then Z_n converges in distribution to Z. That is, for all real a,

$$\lim_{n\to\infty} \Pr\{Z_n \leq a\} = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du.$$

Borel-Cantelli Lemma. Let $A_1, A_2, \ldots$ be an infinite sequence of independent events. Then the event $\{A_i, \text{i.o.}\}$ (where i.o. stands for infinitely often), which is the occurrence of an infinite number of the A_i, is given by

$$\bar{A}_\infty = \{A_i, \text{i.o.}\} = \bigcap_{j=1}^{\infty} \bigcup_{i=j}^{\infty} A_i.$$

The Borel–Cantelli lemma states that the probability of $\bar{A}_\infty$ is zero or one, according to whether $\sum_{i=1}^{\infty} \Pr\{A_i\} < \infty$ or $\sum_{i=1}^{\infty} \Pr\{A_i\} = \infty$.

H. INEQUALITIES

There are a number of inequalities that play an important role in the analytic study of stochastic processes. We mention two here.

Chebyshev's Inequality. Let Z be a nonnegative random variable. Then for any positive number c

$$\Pr\{Z > c\} \leq \frac{1}{c} E[Z]. \tag{1.16}$$

Proof. Since Z is nonnegative,

$$E[Z] = \int_0^\infty z\, dF(z) \geq \int_c^\infty z\, dF(z)$$

$$\geq c \cdot \int_c^\infty dF(z) = c \Pr\{Z > c\},$$

which gives the inequality. If X is a random variable with mean μ and variance σ^2 and we apply (1.16) with $Z = (X - \mu)^2$ we obtain,

$$\Pr\{Z > \varepsilon^2\} = \Pr\{|X - \mu| > \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}.$$

The Schwarz Inequality. Let X and Y be jointly distributed random variables having finite second moments. Then

$$(E[XY])^2 \leq E[X^2]\, E[Y^2].$$

Proof. For all real λ

$$0 \leq E[(X + \lambda Y)^2] = E[X^2] + 2\lambda E[XY] + \lambda^2 E[Y^2].$$

Considered as a quadratic function of λ, there is, then, at most one real root. Equivalently, the discriminant of the quadratic expression is nonpositive. That is

$$4(E[XY])^2 \leq 4E[X^2]E[Y^2]$$

which completes the proof.

2: Two Simple Examples of Stochastic Processes

The developments in this book are intended to serve as an introduction to various aspects of stochastic processes. The theory of stochastic processes is concerned with the investigation of the structure of families of

random variables X_t, where t is a parameter running over a suitable index set T. Sometimes, when no ambiguity can arise we write $X(t)$ instead of X_t.

A *realization* or *sample function* of a stochastic process $\{X_t, t \in T\}$ is an assignment, to each $t \in T$, of a possible value of X_t. The index set t may correspond to discrete units of time $T = \{0, 1, 2, 3, \ldots\}$ and $\{X_t\}$ could then represent the outcomes at successive trials like the result of tossing a coin, the successive reactions of a subject to a learning experiment, or successive observations of some characteristic of a population, etc.

The values of the X_t may be one-dimensional, two-dimensional, or n-dimensional, or even more general. In the case where X_n is the outcome of the nth toss of a die, its possible values are contained in the set $\{1, 2, 3, 4, 5, 6\}$ and a typical realization of the process would be 5, 1, 3, 2, 2, 4, 1, 6, 3, 6, ... This is shown schematically in Fig. 1, where the ordinate for $t = n$ is the value of X_n. In this example, the random

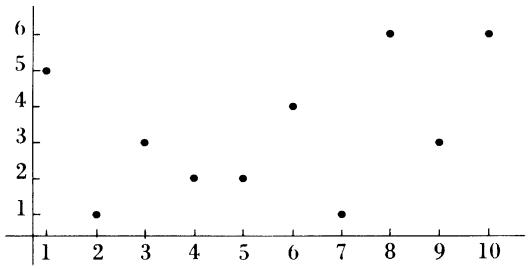


FIG. 1

variables X_n are mutually independent but generally the random variables X_n are dependent.

Stochastic processes for which $T = [0, \infty)$ are particularly important in applications. Here t can usually be interpreted as time.

We will content ourselves, for the moment, with a very brief discussion of some of the concepts of stochastic processes and two examples thereof; a summary of various types of stochastic processes is presented at the end of the chapter, while the examples themselves will be treated in greater detail in succeeding chapters.

Example 1. A very important example is the celebrated *Brownian motion process.* This process has the following characteristics:

(a) Suppose $t_0 < t_1 < \cdots < t_n$; then the increments $X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}$ are mutually independent r.v.'s. (A process with this property is said to be a process with independent increments, and

expresses the fact that the changes of X_t over nonoverlapping time periods are independent r.v.'s.)

(b) The probability distribution of $X_{t_2} - X_{t_1}$, $t_2 > t_1$, depends only on $t_2 - t_1$ (and not, for example, on t_1).

(c) $\Pr[X_t - X_s \leq x] = [2\pi B(t-s)]^{-1/2} \int_{-\infty}^{x} \exp[-u^2/2B(t-s)]\, du$, $t > s$ (B is a positive constant).

Assume for each path that $X_0 = 0$. Note that $EX_t = 0$, $\sigma^2(X_t) = Bt$, where B is a fixed positive constant. It can be proved that, if $0 < t_1 < t_2 < \cdots < t_n < t$, then the conditional probability distribution of X_t, where the values of $X_{t_1}, \ldots, X_{t_n}$ are known, is given by (see Chapter 7)

$$\Pr\{X_t \leq x | X_{t_1} = x_1, \ldots, X_{t_n} = x_n\} = [2\pi B(t-t_n)]^{-1/2} \int_{-\infty}^{x-x_n} \exp[-u^2/2B(t-t_n)]\, du.$$

The history of this process began with the observation by R. Brown in 1827 that small particles immersed in a liquid exhibit ceaseless irregular motions. In 1905 Einstein explained this motion by postulating that the particles under observation are subject to perpetual collision with the molecules of the surrounding medium. The analytical results derived by Einstein were later experimentally verified and extended by various physicists and mathematicians.

Let X_t denote the displacement (from its starting point, along some fixed axis) at time t of a Brownian particle. The displacement $X_t - X_s$ over the time interval (s, t) can be regarded as the sum of a large number of small displacements. The central limit theorem is essentially applicable and it seems reasonable to assert that $X_t - X_s$ is normally distributed. Similarly it seems reasonable to assume that the distribution of $X_t - X_s$ and that of $X_{t+h} - X_{s+h}$ are the same, for any $h > 0$, if we suppose the medium to be in equilibrium. Finally, it is intuitively clear that the displacement $X_t - X_s$ should depend only on the length $t - s$ and not on the time we begin observation.

The Brownian motion process (also called the Wiener process) has proved to be fundamental in the study of numerous other types of stochastic processes. In Chapter 7 we will discuss more fully an example of the one-dimensional Brownian motion process.

Example 2. Another basic example of a continuous time ($T = [0, \infty)$) stochastic process is the Poisson process. The sample function X_t counts the number of times a specified event occurs during the time period from 0 to t. Thus, each possible X_t is represented as a nondecreasing step function.

Figure 2 corresponds to a situation where the event occurred first at time t_1, then at time t_2, at time t_3, at time t_4, etc.; obviously the total

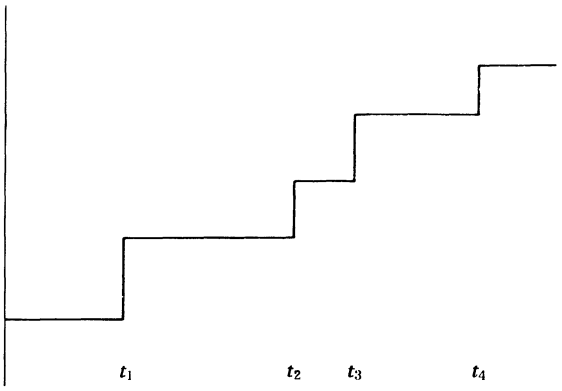


FIG. 2

number of occurrences of the event increases only in unit jumps, and $X_0 = 0$. Concrete examples of such processes are the number of x-rays emitted by a substance undergoing radioactive decay; the number of telephone calls originating in a given locality; the occurrence of accidents at a certain intersection; the occurrence of errors in a page of typing; breakdowns of a machine; and the arrival of customers for service. The justification for viewing these examples as Poisson processes is based on the concept of the law of rare events. We have a situation of many Bernoulli trials with small probability of success where the expected number of successes is constant. Under these conditions it is a familiar theorem that the actual number of events occurring follows a Poisson law. In the case of radioactive decay the Poisson approximation is excellent if the peroid of observation is very short with respect to the half-life of the radioactive substance.

We postulate that the numbers of events happening in two disjoint intervals of time are independent [see (a) above]. Analogously to (b), we also assume that the random variable $X_{t_0+t} - X_{t_0}$ depends only on t and not on t_0 or on the value of X_{t_0}. We set down the following further postulates, which are consistent with the intuitive descriptions given above:

I. The probability of at least one event happening in a time period of duration h is

$$p(h) = ah + o(h), \qquad h \to 0, \qquad a > 0$$

[$g(t) = o(t)$, $t \to 0$ is the usual symbolic way of writing the relation $\lim_{t\to 0} g(t)/t = 0$].

II. The probability of two or more events happening in time h is $o(h)$.

Postulate II is tantamount to excluding the possibility of the simultaneous occurrence of two or more events. In the concrete illustrations cited above, this requirement is usually satisfied.

Let $P_m(t)$ denote the probability that exactly m events occur in time t, i.e.,

$$P_m(t) = \Pr\{X_t = m\}, \qquad m = 0, 1, 2, \ldots .$$

The requirement II can be stated in the form

$$\sum_{m=2}^{\infty} P_m(h) = o(h),$$

and clearly

$$p(h) = P_1(h) + P_2(h) + \ldots .$$

Because of the assumption of independence,

$$P_0(t+h) = P_0(t)P_0(h) = P_0(t)(1 - p(h)),$$

and therefore

$$\frac{P_0(t+h) - P_0(t)}{h} = -P_0(t)\frac{p(h)}{h}.$$

But on the basis of Postulate I we know that $p(h)/h \to a$. Therefore, the probability $P_0(t)$ that the event has not happened during $(0, t)$ satisfies the differential equation

$$P_0'(t) = -aP_0(t),$$

whose well-known solution is $P_0(t) = ce^{-at}$. The constant c is determined by the initial condition $P_0(0) = 1$, which implies $c = 1$. Thus, $P_0(t) = e^{-at}$. We will now calculate $P_m(t)$ for every m. It is easy to see that

$$P_m(t+h) = P_m(t)P_0(h) + P_{m-1}(t)P_1(h) + \sum_{i=2}^{m} P_{m-i}(t)P_i(h). \tag{2.1}$$

By definition $P_0(h) = 1 - p(h)$. The requirement II implies that

$$\begin{aligned} &P_1(h) = p(h) + o(h) \qquad \text{and} \\ &\sum_{i=2}^{m} P_{m-i}(t)P_i(h) \leq \sum_{i=2}^{m} P_i(h) = o(h), \end{aligned} \tag{2.2}$$

since obviously $P_k(t) \leq 1$. Therefore, with the aid of (2.2) we rearrange (2.1) into the form

$$\begin{aligned} P_m(t+h) - P_m(t) &= P_m(t)[P_0(h) - 1] + P_{m-1}(t)P_1(h) + \sum_{i=2}^{m} P_{m-i}(t)P_i(h) \\ &= -P_m(t)p(h) + P_{m-1}(t)P_1(h) + \sum_{i=2}^{m} P_{m-i}(t)P_i(h) \\ &= -aP_m(t)h + aP_{m-1}(t)h + o(h). \end{aligned}$$

Therefore

$$\frac{P_m(t+h) - P_m(t)}{h} \to -aP_m(t) + aP_{m-1}(t) \qquad \text{as} \quad h \to 0$$

and, formally, we get

$$P_m'(t) = -aP_m(t) + aP_{m-1}(t), \qquad m = 1, 2, \ldots, \tag{2.3}$$

subject to the initial conditions

$$P_m(0) = 0, \qquad m = 1, 2, \ldots.$$

In order to solve (2.3), we introduce the functions

$$Q_m(t) = P_m(t)e^{at}, \qquad m = 0, 1, 2, \ldots.$$

Substituting the above in (2.3) gives

$$Q_m'(t) = aQ_{m-1}(t), \qquad m = 1, 2, \ldots, \tag{2.4}$$

where $Q_0(t) \equiv 1$ and the initial conditions are $Q_m(0) = 0$, $m = 1, 2, \ldots$. Solving (2.4) recursively we obtain

$$\begin{aligned} &Q_1'(t) = a \quad \text{or} \quad Q_1(t) = at + c \quad \text{so} \quad Q_1(t) = at \\ &Q_2(t) = \frac{a^2t^2}{2} + c \quad \text{so} \quad Q_2(t) = \frac{a^2t^2}{2!} \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ &\qquad\qquad\qquad\qquad Q_m(t) = \frac{a^mt^m}{m!} \end{aligned}$$

Therefore

$$P_m(t) = \frac{a^mt^m}{m!}e^{-at}.$$

In other words, for each t, X_t follows a Poisson distribution with parameter at. In particular, the mean number of occurrences in time t is at.

Often the Poisson process arises in a form where the time parameter is replaced by a suitable spatial parameter. The following formal example illustrates this vein of ideas. Consider an array of points distributed in a space E (E is a Euclidean space of dimension $d \geq 1$). Let N_R denote the number of points (finite or infinite) contained in the region R of E. We postulate that N_R is a random variable. The collection $\{N_R\}$ of random variables, where R varies over all possible subsets of E, is said to be a homogeneous Poisson process if the following assumptions are fulfilled:

(i) The numbers of points in nonoverlapping regions are independent random variables.

(ii) For any region R of finite volume, N_R is Poisson distributed with mean $\lambda V(R)$, where $V(R)$ is the volume of R. The parameter λ is fixed and measures in a sense the intensity component of the distribution, which is independent of the size or shape. Spatial Poisson processes arise in considering the distribution of stars or galaxies in space, the spatial distribution of plants and animals, of bacteria on a slide, etc. These ideas and concepts will be further studied in Chapter 16.

3: *Classification of General Stochastic Processes*

The main elements distinguishing stochastic processes are in the nature of the *state space*, the *index parameter* T, and the dependence relations among the random variables X_t.

STATE SPACE S

This is the space in which the possible values of each X_t lie. In the case that $S = (0, 1, 2, \ldots)$, we refer to the process at hand as integer valued, or alternately as a discrete state process. If $S =$ the real line $(-\infty, \infty)$, then we call X_t a real-valued stochastic process. If S is Euclidean k space then X_t is said to be a k-vector process.

As in the case of a single random variable, the choice of state space is not uniquely specified by the physical situation being described, although usually one particular choice stands out as most appropriate.

INDEX PARAMETER T

If $T = (0, 1, \ldots)$ then we shall always say that X_t is a discrete time stochastic process. Often when T is discrete we shall write X_n instead of X_t. If $T = [0, \infty)$, then X_t is called a continuous time process.

We have already cited examples where the index set T is not one dimensional (spatial Poisson processes). Another example is that of waves in oceans. We may regard the latitude and longitude coordinates as the value of t, and X_t is then the height of the wave at the location t.

CLASSICAL TYPES OF STOCHASTIC PROCESSES

We now describe some of the classical types of stochastic processes characterized by different dependence relationships among X_t. In the examples, we take $T=[0, \infty)$ unless we state the contrary explicitly. For simplicity of exposition, we assume that the random variables X_t are real valued.

(a) *Process with Stationary Independent Increments*

If the random variables

$$X_{t_2}-X_{t_1}, X_{t_3}-X_{t_2}, \ldots, X_{t_n}-X_{t_{n-1}}$$

are independent for all choices of $t_1, \cdots, t_n$ satisfying

$$t_1<t_2<\cdots<t_n,$$

then we say that X_t is a process with *independent increments*. If the index set contains a smallest index t_0, it is also assumed that

$$X_{t_0}, X_{t_1}-X_{t_0}, X_{t_2}-X_{t_1}, \ldots, X_{t_n}-X_{t_{n-1}}$$

are independent. If the index set is discrete, that is, $T=(0, 1, \ldots)$, then a process with independent increments reduces to a sequence of independent random variables $Z_0=X_0$, $Z_i=X_i-X_{i-1}$ $(i=1, 2, 3, \ldots)$, in the sense that knowing the individual distributions of Z_0, Z_1, ... enables one to determine (as should be fairly clear to the reader) the joint distribution of any finite set of the X_i. In fact,

$$X_i=Z_0+Z_1+\cdots+Z_i, \qquad i=0, 1, 2, \ldots.$$

If the distribution of the increments $X(t_1+h)-X(t_1)$ depends only on the length h of the interval and not on the time t_1 the process is said to have *stationary increments*. For a process with stationary increments the distribution of $X(t_1+h)-X(t_1)$ is the same as the distribution of $X(t_2+h)-X(t_2)$, no matter what the values of t_1, t_2 and h.

If a process $\{X_t, t\in T\}$, where $T=[0, \infty)$ or $T=(0, 1, 2, \ldots)$ has stationary independent increments and has a finite mean, then it is elementary to show that $E[X_t]=m_0+m_1t$ where $m_0=E[X_0]$ and

$m_1 = E[X_1] - m_0$. A similar assertion holds for the variance:

$$\sigma_{X_t}^2 = \sigma_0^2 + \sigma_1^2 t$$

where

$$\sigma_0^2 = E[(X_0 - m_0)^2]$$

and

$$\sigma_1^2 = E[(X_1 - m_1)^2] - \sigma_0^2 .$$

We will indicate the proof in the case of the mean. Let $f(t) = E[X_t] - E[X_0]$. Then for any t and s

$$\begin{aligned} f(t+s) &= E[X_{t+s} - X_0] \\ &= E[X_{t+s} - X_s + X_s - X_0] \\ &= E[X_{t+s} - X_s] + E[X_s - X_0] \\ &= E[X_t - X_0] + E[X_s - X_0] \end{aligned}$$

(using the property of stationary increments)

$$= f(t) + f(s).$$

The only solution, subject to mild regularity conditions, to the functional equation $f(t+s) = f(t) + f(s)$ is $f(t) = f(1) \cdot t$. We indicate the proof of the above statement assuming $f(t)$ differentiable, although much less would suffice. Differentiation with respect to t and independently in s verifies that

$$f'(t+s) = f'(t) = f'(s).$$

Therefore for $s = 1$, we find $f'(t) = \text{constant} = f'(1) = c$. Integration of this elementary differential equation yields $f(t) = ct + d$. But, $f(0) = 2f(0)$ implies $f(0) = 0$ and therefore $d = 0$ is necessary. The expression $f(t) = f(1)t$ for the case at hand is

$$E[X_t] - m_0 = (E[X_1] - m_0) \cdot t$$

or

$$E[X_t] = m_0 + m_1 t$$

as desired.

Both the Brownian motion process and the Poisson process have stationary independent increments.

(b) *Martingales*

Let $\{X_t\}$ be a real-valued stochastic process with discrete or continuous parameter set. We say that $\{X_t\}$ is a *martingale* if, $E[|X_t|] < \infty$ for all t,

and if for any $t_1 < t_2 < \cdots < t_{n+1}$, $E(X_{t_{n+1}} | X_{t_1} = a_1, \ldots, X_{t_n} = a_n) = a_n$ for all values of $a_1, \cdots, a_n$. Martingales may be considered as appropriate models for fair games, in the sense that X_t signifies the amount of money that a player has at time t. The martingale property states, then, that the average amount a player will have at time t_{n+1}, given that he has amount a_n at time t_n, is equal to a_n regardless of what his past fortune has been. The reader can readily verify that the process $X_n = Z_1 + \cdots + Z_n$, $n = 1, 2, \ldots$, is a discrete time martingale if the Z_i are independent and have means zero. Similarly, if X_t, $0 \leq t < \infty$ has independent increments whose means are zero, then $\{X_t\}$ is a continuous time martingale (see Elementary Problem 6).

Martingales are the subject matter of Chapter 6.

(c) *Markov Processes*

Roughly speaking, a *Markov process* is a process with the property that, given the value of X_t, the values of X_s, $s > t$, do not depend on the values of X_u, $u < t$; that is, the probability of any particular future behavior of the process, when its present state is known exactly, is not altered by additional knowledge concerning its past behavior. We should make it clear, however, that if our knowledge of the present state of the process is imprecise, then the probability of some future behavior will in general be altered by additional information relating to the past behavior of the system. In formal terms a process is said to be Markov if

$$\begin{aligned} \Pr\{a < X_t \leq b | X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n\} \\ = \Pr\{a < X_t \leq b | X_{t_n} = x_n\} \end{aligned} \tag{3.1}$$

whenever $t_1 < t_2 < \cdots < t_n < t$.

Let A be an interval of the real line. The function

$$P(x, s; t, A) = \Pr\{X_t \in A | X_s = x\}, \qquad t > s, \tag{3.2}$$

is called the *transition probability function* and is basic to the study of the structure of Markov processes. We may express the condition (3.1) as follows:

$$\Pr\{a < X_t \leq b | X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n\} = P(x_n, t_n; t, A), \tag{3.3}$$

where $A = \{\xi | a < \xi \leq b\}$. It may be proved that the probability distribution of

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$$

can be computed in terms of (3.2) and the initial distribution function of X_{t_1}. We will elaborate further on these concepts in our more detailed examination of discrete time, discrete state Markov processes (Chapter 2).

A Markov process having a finite or denumerable state space is called

a *Markov chain.* A Markov process for which all realizations or sample functions $\{X_t,\ t \in [0,\ \infty)\}$ are continuous functions is called a *diffusion process.* The Poisson process is a continuous time Markov chain and Brownian motion is a diffusion process.

(d) *Stationary Processes*

A stochastic process X_t for t in T [here T could be one of the sets $(-\infty,\ \infty)$, $[0,\ \infty)$, the set of all integers, or the set of all positive integers] is said to be *strictly stationary* if the joint distribution functions of the families of random variables

$$(X_{t_1+h},\ X_{t_2+h},\ \ldots\ldots X_{t_n+h}) \quad \text{and} \quad (X_{t_1},\ X_{t_2},\ \ldots,\ X_{t_n})$$

are the same for all $h > 0$ and arbitrary selections $t_1, t_2, \ldots, t_n$ from T. This condition asserts that in essence the process is in probabilistic equilibrium and that the particular times at which we examine the process are of no relevance. In particular, the distribution of X_t is the same for each t.

A stochastic process X_t for $t \in T$ is said to be *wide sense stationary* or *covariance stationary* if it possesses finite second moments and if $\mathrm{Cov}(X_t,\ X_{t+h}) = E(X_t X_{t+h}) - E(X_t)E(X_{t+h})$ depends only on h for all $t \in T$. A stationary process that has finite second moments is covariance stationary. There are covariance stationary processes that are not stationary.

Stationary processes are appropriate for describing many phenomena that occur in communication theory, astronomy, biology, and sometimes economics and are discussed in more detail in Chapter 9.

A Markov process is said to have stationary transition probabilities if $P(x,\ s;\ t,\ A)$ defined in (3.2) is a function only of $t - s$. Remember that $P(x,\ s;\ t,\ A)$ is a conditional probability, given the present state. Therefore, there is no reason to expect that a Markov process with stationary transition probabilities is a stationary process, and this is indeed the case.

Neither the Poisson process nor the Brownian motion process is stationary. In fact, no nonconstant process with stationary independent increments is stationary. However, if $\{X_t,\ t \in [0,\ \infty)\}$ is Brownian motion or a Poisson process, then $Z_t = X_{t+h} - X_t$ is a stationary process for any fixed $h \geq 0$.

(e) *Renewal Processes.*

A renewal process is a sequence T_k of independent and identically distributed positive random variables, representing the lifetimes of some "units." The first unit is placed in operation at time zero; it fails at time T_1 and is immediately replaced by a new unit which then fails at time $T_1 + T_2$, and so on, thus motivating the name "renewal process." The time of the nth renewal is $S_n = T_1 + \cdots + T_n$.

A renewal counting process N_t counts the number of renewals in the interval $[0, t]$. Formally,

$$N_t = n \text{ for } S_n \leq t < S_{n+1},\ n = 0, 1, 2, \ldots$$

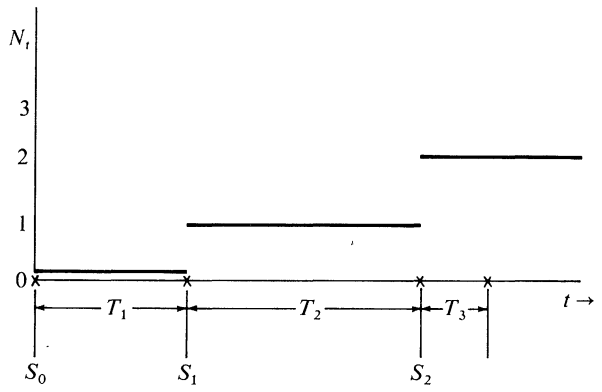


FIG. 3

Often the distinction is not made between the renewal process and the associated renewal counting process, and no real confusion results.

Renewal processes occur directly in many applied areas such as management science, economics, and biology. Of equal importance, often renewal processes may be discovered embedded in other stochastic processes that, at first glance, seem quite unrelated. Chapter 5 is devoted to renewal processes.

The Poisson process with parameter λ is a renewal counting process for which the unit lifetimes have exponential distributions with common parameter λ.

(f) *Point Processes*

Let S be a set in n-dimensional space and let $\mathscr{A}$ be a family of subsets of S. A *point process* is a stochastic process indexed by the sets $A \in \mathscr{A}$ and having the set $\{0, 1, \ldots, \infty\}$ of nonnegative integers as its state space. We think of "points" being scattered over S in some random manner, and of $N(A)$ as counting the number of points in the set A. Since $N(A)$ is a counting function there are additional requirements on each realization. For example, if A_1 and A_2 are mutually disjoint sets in $\mathscr{A}$ whose union $A_1 \cup A_2$ is also in $\mathscr{A}$, then we require

$$N(A_1 \cup A_2) = N(A_1) + N(A_2),$$

and if the empty set $\varnothing$ is in $\mathscr{A}$, then $N(\varnothing) = 0$.

Suppose S is a set in the real line (plane, 3 dimensional space) and for every subset $A \subset S$, let $V(A)$ be the length (respectively area, volume)

of A. Then $\{N(A),\ A \subset S\}$ is a homogeneous *Poisson point process* of intensity $\lambda > 0$ if:

(i) For each $A \subset S$, $N(A)$ has a Poisson distribution with parameter $\lambda V(A)$, and

(ii) For every finite collection $\{A_1, \ldots, A_n\}$ of disjoint subsets of S, the random variables $N(A_1), \ldots, N(A_n)$ are independent.

Poisson point processes arise in considering the distribution of stars or galaxies in space, the planer distribution of plants and animals, of bacteria on a slide, etc. These ideas and concepts will be further studied in Chapter 16, Volume II.

Every Poisson process $\{X_t: t \in [0, \infty)\}$ defines a Poisson point process on $S = [0, \infty)$. In fact for an interval subset $A = (s, t]$, $s < t$, we use $N(A) = X_t - X_s$.

4: Defining a Stochastic Process

The distinguishing features of a stochastic process X_t are the relationships among the random variables, X_t, $t \in T$.

These relationships are specified by giving the joint distribution function of every finite family $X_{t_1}, \ldots, X_{t_n}$ of variables of the process. For the purposes of this book, a stochastic process may be considered as well defined once its state space, index parameter, and family of joint distributions are prescribed. However, in dealing with continuous parameter processes certain difficulties arise, which we illustrate by the following example.

Let U be a r.v. uniformly distributed on [0, 1] and define X_t and Y_t as follows:

$$X_t = \begin{cases} 1 & \text{for } U = t, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$Y_t \equiv 0, \qquad (t > 0).$$

A simple computation verifies that $\{X_t\}$ and $\{Y_t\}$ have the same finite dimensional distributions. However, obviously

$$\Pr\{X_t \leq \tfrac{1}{2} \quad \text{for all } 0 \leq t \leq 1\} = 0$$

and

$$\Pr\{Y_t \leq \tfrac{1}{2} \quad \text{for all } 0 \leq t \leq 1\} = 1,$$

which is a rather disconcerting state of affairs. To pinpoint the source of the difficulty we consider the following problem.

Suppose that $\{X_t,\ 0 \leq t < \infty\}$, is a continuous parameter process, and we wish to evaluate $\Pr\{X_t \geq 0,\ 0 \leq t \leq 1\}$.

Let us consider the decreasing sequence of events

$$A_n = \{X_{t_i} \geq 0,\ t_i = i/2^n,\ i = 0, 1, \ldots, 2^n\},\ n = 1, 2, \ldots.$$

The probability of each A_n can be calculated in terms of the joint distribution function of the corresponding X_{t_i}, $i = 0, \ldots, 2^n$, and it would seem reasonable that we should take for $\Pr\{X_t \geq 0,\ 0 \leq t \leq 1\}$ the value $\lim_{n\to\infty} \Pr\{A_n\}$. However, that which seems reasonable is not necessarily free from inconsistencies. It is equally reasonable that we take $A_n' = \{X_{t_i} \geq 0,\ t_i = i/3^n,\ i = 0, 1, \ldots, 3^n\}$, and for $\Pr\{X_t \geq 0,\ 0 \leq t \leq 1\}$ the value $\lim_{n\to\infty} \Pr\{A_n'\}$, but it is by no means clear that $\lim_{n\to\infty} \Pr\{A_n\} = \lim_{n\to\infty} \Pr\{A_n'\}$, and in fact, the two limits need not be equal if no "smoothness" assumptions are made concerning the sample functions of the process. There are various sufficient conditions for the equality of the two limits: one of them is that $\lim_{\tau\to t} \Pr\{|X_t - X_\tau| \geq \varepsilon\} = 0$ for every $\varepsilon > 0$ and every t. With this condition, the problem can be formulated so that no inconsistency arises if we define $\Pr\{X_t \geq 0,\ 0 \leq t \leq 1\}$ as the common value of the two limits, and in fact, if $t_1, t_2, \ldots$ is any dense set of points in the interval $[0, 1]$, then $\Pr\{X_t \geq 0,\ 0 \leq t \leq 1\} = \lim_{n\to\infty} \Pr\{X_{t_i} \geq 0,\ i = 1, 2, \ldots, n\}$.

The nub of the matter is that while the axiom of total probability enables us to evaluate probabilities of events concerning a sequence of r.v.'s in terms of the probabilities of events involving finite and hence denumerable subsets of the sequence, the event $\{X_t \geq 0,\ 0 \leq t \leq 1\}$ involves a nondenumerable number of random variables. The details of this point are quite involved, and are well beyond the scope of the present book; the interested reader is referred to Doob,† Chapter 2. Some foundational questions which throw more light on these problems are discussed in Chapter 14.

Elementary Problems

1. Let X be a *nonnegative* discrete random variable with possible values $0, 1, 2, \ldots$. Show

$$E[X] = \sum_{n=0}^{\infty} \Pr\{X > n\} = \sum_{k=1}^{\infty} \Pr\{X \geq k\}.$$

Hint: Begin with $E[X] = \sum_{n=1}^{\infty} n \Pr\{X = n\} = \sum_{n=1}^{\infty} \sum_{k=1}^{n} \Pr\{X = n\}$.

† J. L. Doob, "Stochastic Processes," Wiley, New York, 1953.

2. Suppose a jar has n chips numbered $1, 2, \ldots, n$. A person draws a chip, returns it, draws another, returns it, and so on until he gets a chip which has been drawn before and then stops. Let X be the number of drawings required to accomplish this objective. Find the probability distribution of X.

Hint: It's easiest to first compute $\Pr\{X > k\}$.

Solution:

$$p(k) = (k-1)!\binom{n}{k-1}\frac{k-1}{n^k} \qquad \text{for} \quad k = 2, 3, \ldots, n+1.$$

3. Show that the expectation of the random variable X of Problem 2 is

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

Hint: Use Elementary Problem 1.

4. The number of accidents occurring in a factory in a week is a random variable with mean μ and variance σ^2. The numbers of individuals injured in different accidents are independently distributed each with mean ν and variance τ^2. Determine the mean and variance of the number of individuals injured in a week.

Solution: $E(\text{injuries}) = \mu\nu$; $\quad \text{Var}(\text{injuries}) = \nu^2\sigma^2 + \mu\tau^2$.

5. The following experiment is performed. An observation is made of a Poisson random variable X with parameter λ. Then a binomial event with probability p of success is repeated X times and σ successes are observed. What is the distribution of σ?

Hint: Use the generating function for the random sum of random variables,

Solution: Poisson, with parameter λp.

6. Show that the sums $S_n = X_1 + \cdots + X_n$ of independent random variables X_n with zero mean form a martingale. Assume $E[|X_k|] < \infty$ for $k = 1, 2, \ldots$.

7. Prove that every stochastic process $\{X(t);\ t = 0, 1, \ldots\}$ with independent increments is a Markov process. (Remark: This is not true of stochastic processes $\{X(t);\ -\infty < t < \infty\}$.)

8. Consider a population of n couples where a boy is born to the ith couple with probability p_i and c_i is the expected number of children born to this couple. Assume p_i is constant with time for all couples and that sexes of successive children born to a particular couple are independent r.v's. Further, assume

that no multiple births are allowed. The sex ratio is defined to be

$$S = \frac{\text{expected number of boys born in the population of } n \text{ couples}}{\text{expected number of children born in the population of } n \text{ couples}}.$$

Suppose $c_i = c$, $i = 1, 2, \ldots, n$. Find S.

Solution: $S \equiv S_0 = (\sum_{i=1}^{n} p_i)/n$.

9. If the parents of all couples decide to have children until a boy is born and then have no further children, show that

$$S = S_1 = \frac{n}{\sum_{i=1}^{n} \frac{1}{p_i}} \leq S_0 .$$

10. Suppose the parents of all couples decide that if their first child is a boy they will continue to have children until a girl is born and then have no further children. If their first child is a girl they will continue to have children until a boy is born and then have no further children. Compute S corresponding to this birth control behavior.

Solution:

$$S = S_2 = \frac{\left\{\sum_{i=1}^{n} 1/q_i - \sum_{i=1}^{n} p_i\right\}}{\sum_{i=1}^{n} [p_i q_i]^{-1} - n} \qquad \text{where} \quad q_i = 1 - p_i .$$

11. Suppose the parents of all children decide that if their first child is a boy they will continue to have children until a girl is born and then have no further children. If their first child is a girl they will have no further children. Compute S corresponding to this behavior.

Solution:

$$S = S_3 = 1 - \left[\frac{n}{\sum_{i=1}^{n} 1/q_i}\right].$$

12. Show that, depending on the value of $\{p_1, \ldots, p_n\}$, S_2 can satisfy either $S_2 < S_0$ or $S_2 > S_0$, where S_0 is the sex ratio of Elementary Problem 8 and S_2 is the sex ratio of Elementary Problem 10.

13. Suppose that a child born to the ith set of parents in a population of n sets of parents has probability p_i of a birth disorder, $i = 1, 2, \ldots, n$. Assume that the birth of one affected child deters parents from further reproduction. Let s = the number of offspring in a single family when no affected children are born. Assume that with respect to any given birth, p_i does not depend on

preceding births. Show that

$$R_1 = \frac{\text{expected number of affected children}}{\text{expected total number of children born}} = \frac{\sum_{i=1}^{n} 1 - q_i^s}{\sum_{i=1}^{n} (1 - q_i^s)/p_i}.$$

(b) Assume that the birth of two affected children (but not one) will deter parents from further reproduction. Show that under this kind of selective limitation

$$R_2 = \frac{\sum_{i=1}^{n} \{2(1 - q_i^s) - sp_i q_i^{s-1}\}}{\sum_{i-1}^{n} \{[2(1 - q_i^s)/p_i] - sq_i^{s-1}\}}.$$

Problems

The following integrals may be useful in some of the problems, and are recorded here for future reference.

The gamma function is defined by

$$\Gamma(x) = \int_0^\infty \xi^{x-1} e^{-\xi}\, d\xi, \qquad x > 0.$$

For large x, $\Gamma(x) \sim \sqrt{2\pi}\, e^{-x} x^{x+1/2}$ (Stirling's formula). When $x = n$, an integer, $\Gamma(n) = (n-1)! = (n-1)(n-2)\ldots 2 \cdot 1$.

The Beta integral is given by

$$\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 x^{p-1}(1-x)^{q-1}\, dx$$

where $p > 0$, $q > 0$.

1. Let a, b, c be independent random variables uniformly distributed on (0, 1). What is the probability that $ax^2 + bx + c$ has real roots?

Answer: $(5 + 3 \log 4)/36$.

2. For each fixed $\lambda > 0$ let X have a Poisson distribution with parameter λ. Suppose λ itself is a random variable following a gamma distribution (i.e., with density

$$f(\lambda) = \begin{cases} \dfrac{1}{\Gamma(n)} \lambda^{n-1} e^{-\lambda}, & \lambda \geq 0, \\ 0, & \lambda < 0, \end{cases}$$

where n is a fixed positive constant). Show that now

$$\Pr\{X=k\}=\frac{\Gamma(k+n)}{\Gamma(n)\Gamma(k+1)}\left(\frac{1}{2}\right)^{k+n}, \qquad k=0, 1, \ldots .$$

When n is an integer this is the negative binomial distribution with $p=\frac{1}{2}$.

3. For each given p let X have a binomial distribution with parameters p and N. Suppose N is itself binomially distributed with parameters q and M, $M \geq N$.

(a) Show analytically that X has a binomial distribution with parameters pq and M.

(b) Give a probabilistic argument for this result.

4. For each given p, let X have a binomial distribution with parameters p and N. Suppose p is distributed according to a beta distribution with parameters r and s. Find the resulting distribution of X. When is this distribution uniform on $x=0, 1, \ldots, N$?

Answer:

$$\Pr\{X=k\}=\binom{N}{k}\frac{\Gamma(r+s)\Gamma(k+r)\Gamma(N-k+s)}{\Gamma(r)\Gamma(s)\Gamma(N+r+s)};$$

$$\Pr\{X=k\}=1/(N+1) \qquad \text{when} \quad r=s=1.$$

5. (a) Suppose X is distributed according to a Poisson distribution with parameter λ. The parameter λ is itself a random variable whose distribution law is exponential with mean $=1/c$. Find the distribution of X.

(b) What if λ follows a gamma distribution of order α with scale parameter c, i.e., the density of λ is

$$c^{\alpha+1}\frac{\lambda^{\alpha}}{\Gamma(\alpha+1)}e^{-\lambda c}$$

for $\lambda>0$; 0 for $\lambda \leq 0$.

Answer:

(a) $\Pr\{X=k\}=\dfrac{c}{(c+1)^{k+1}}$;

(b) $\Pr\{X=k\}=\dfrac{\Gamma(k+\alpha+1)}{k!\Gamma(\alpha+1)}\left(\dfrac{1}{1+c}\right)^{k+\alpha+1}c^{\alpha+1}$.

6. Suppose we have N chips marked 1, 2, ..., N, respectively. We take a random sample of size $2n+1$ without replacement. Let Y be the median of the random sample. Show that the probability function of Y is

$$\Pr\{Y=k\}=\frac{\binom{k-1}{n}\binom{N-k}{n}}{\binom{N}{2n+1}} \qquad \text{for} \quad k=n+1, n+2, \ldots, N-n.$$

Verify

$$E(Y)=\frac{N+1}{2} \qquad \text{and} \qquad \operatorname{Var}(Y)=\frac{(N-2n-1)(N+1)}{8n+12}.$$

7. Suppose we have N chips, numbered 1, 2, ..., N. We take a random sample of size n without replacement. Let X be the largest number in the random sample. Show that the probability function of X is

$$\Pr\{X=k\}=\frac{\binom{k-1}{n-1}}{\binom{N}{n}} \qquad \text{for} \quad k=n, n+1, \ldots, N$$

and that

$$EX=\frac{n}{n+1}(N+1), \qquad \operatorname{Var}(X)=\frac{n(N-n)(N+1)}{(n+1)^2(n+2)}.$$

8. Let X_1 and X_2 be independent random variables with uniform distribution over the interval $[\theta-\frac{1}{2}, \theta+\frac{1}{2}]$. Show that X_1-X_2 has a distribution independent of θ and find its density function.

Answer:

$$f_{X_1-X_2}(y)=\begin{cases}1+y, & -1\leq y\leq 0,\\ 1-y, & 0\leq y<1,\\ 0, & |y|>1.\end{cases}$$

9. Let X be a *nonnegative* random variable with cumulative distribution function $F(x)=\Pr\{X\leq x\}$. Show

$$E[X]=\int_0^\infty [1-F(x)]\,dx.$$

Hint: Write $E[X]=\int_0^\infty x\,dF(x)=\int_0^\infty \left(\int_0^x dy\right) dF(x)$.

10. Let X be a *nonnegative* random variable and let

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

where c is a given constant. Express the expectation $E[X_c]$ in terms of the cumulative distribution function $F(x)=\Pr\{X\leq x\}$.

Answer: $E[X_c]=\int_0^c [1-F(x)]\,dx.$

11. Let X and Y be jointly distributed discrete random variables having possible values $0, 1, 2, \ldots$. For $|s|<1$, $|t|<1$ define the joint generating function

$$\phi_{X,Y}(s,t)=\sum_{i,j=0}^{\infty} s^i t^j \Pr\{X=i, Y=j\}$$

and the marginal generating functions

$$\phi_X(s) = \sum_{i=0}^{\infty} s^i \Pr\{X = i\}$$

$$\phi_Y(t) = \sum_{j=0}^{\infty} t^j \Pr\{Y = j\}.$$

(a) Prove that X and Y are independent if and only if

$$\phi_{X,Y}(s, t) = \phi_X(s)\phi_Y(t) \quad \text{for all} \quad s, t.$$

(b) Give an example of jointly distributed random variables X, Y which are *not* independent, but for which

$$\phi_{X,Y}(t, t) = \phi_X(t)\phi_Y(t) \quad \text{for all} \quad t.$$

(This example is pertinent because $\phi_{X,Y}(t, t)$ is the generating function of the sum $X + Y$. Thus independence is sufficient but not necessary for the generating function of a sum of random variables to be the product of the marginal generating functions.)

12. Let $A_0, A_1, \ldots, A_r$ be $r + 1$ events which can occur as outcomes of an experiment. Let p_i be the probability of the occurrence of A_i $(i = 0, 1, 2, \ldots, r)$. Suppose we perform independent trials until the event A_0 occurs k times. Let X_i be the number of occurrences of the event A_i. Show that

$$\Pr\left\{X_1 = x_1, \ldots, X_r = x_r;\ A_0 \text{ occurs for the } k\text{th time at the } \left(k + \sum_{i=1}^{r} x_i\right)\text{th trial}\right\}$$
$$= \frac{\Gamma\left(k + \sum_{i=1}^{r} x_i\right)}{\Gamma(k) \prod_{i=1}^{r} x_i!} p_0^k \prod_{i=1}^{r} p_i^{x_i}. \tag{I}$$

13. Show that the probability generating function of the *negative multinomial distribution* (I) with parameters $(k; p_0, p_1, \ldots, p_r)$ is

$$\varphi(t_1, \ldots, t_r) = p_0^k \left(1 - \sum_{i=1}^{r} t_i p_i\right)^{-k}.$$

14. Consider vector random variable $\{X_0, X_1, \ldots, X_r\}$ following a *multinomial* distribution with parameters $(n; p_0, p_1, \ldots, p_r)$, and assume that n is itself a random variable distributed as a negative binomial with parameters $(k; \rho)$. Compute the joint distribution of $X_0, \ldots, X_r$.

15. Suppose that a lot consists of $m, n_1, \ldots, n_r$ items belonging to the 0th, 1st, ..., rth classes respectively. The items are drawn one-by-one without replacement until k items of the 0th class are observed. Show that the joint distribution

of the observed frequencies $X_1, \ldots, X_r$ of the 1st, ..., rth classes is

$$\Pr\{X_1 = x_1, \ldots, X_r = x_r\} = \left\{\binom{m}{k-1}\prod_{i=1}^{r}\binom{n_i}{x_i}\Big/\binom{m+n}{k+y-1}\right\} \cdot \frac{m-(k-1)}{m+n-(k+y-1)}$$

where

$$y = \sum_{i=1}^{r} x_i \qquad \text{and} \qquad n = \sum_{i=1}^{r} n_i.$$

16. *Continuation of Problem* 15 If $m \to \infty$ and $n \to \infty$ in such a way that $m/(m+n) \to p_0$ and $n_i/(m+n) \to p_i$, $i = 1, 2, \ldots, r$, show that the distribution of Problem 15 approaches the negative multinomial.

17. The random variable X_n takes the values k/n, $k = 1, 2, \ldots, n$, each with probability $1/n$. Find its characteristic function and the limit as $n \to \infty$. Identify the random variable of the limit characteristic function.

Answer:

(a) $\varphi_n(t) = (1 - e^{it})\dfrac{1}{n} \cdot \dfrac{1}{\exp(-in^{-1}t) - 1}$,

(b) uniform (0, 1).

18. Using the central limit theorem for suitable Poisson random variables, prove that

$$\lim_{n\to\infty} e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} = \frac{1}{2}$$

***19.** The random variables X and Y have the following properties: X is positive, i.e., $P\{X > 0\} = 1$, with continuous density function $f(x)$, and $Y|X$ has a uniform distribution on $\{0, X\}$. Prove: If Y and $X - Y$ are independently distributed, then

$$f(x) = a^2 x e^{-ax}, \qquad x > 0, \quad a > 0.$$

***20.** Let U be gamma distributed with order p and let V have the beta distribution with parameters q and $p - q$ $(0 < q < p)$. Assume that U and V are independent. Show that UV is then gamma distributed with order q.

Hint:

$$\Pr\{UV \le x\} = \int_0^1 \left(\int_0^{x/\xi} e^{-\lambda}\lambda^{p-1}\,d\lambda\right)\frac{\xi^{q-1}(1-\xi)^{p-q-1}\,d\xi}{\Gamma(q)\Gamma(p-q)}.$$

Take Laplace transforms of both sides, interchange orders of integration, and then evaluate by expanding in suitable series of the form

$$(1+y)^{-\alpha-1} = \sum_{k=0}^{\infty}\binom{\alpha+k}{k}y^k.$$

*21. Let X and Y be independent, identically distributed, positive random variables with continuous density function $f(x)$. Assume, further, that $U = X - Y$ and $V = \min(X, Y)$ are independent random variables. Prove that

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{elsewhere,} \end{cases}$$

for some $\lambda > 0$. Assume $f(0) > 0$.

Hint: Show first that the joint density function of U and V is

$$f_{U,V}(u, v) = f(v)f(v + |u|).$$

Next, equate this with the product of the marginal densities for U, V.

22. Let X and Y be two independent, nonnegative integer-valued, random variables whose distribution has the property

$$\Pr\{X = x | X + Y = x + y\} = \frac{\binom{m}{x}\binom{n}{y}}{\binom{m+n}{x+y}}$$

for all nonnegative integers x and y where m and n are given positive integers. Assume that $\Pr\{X = 0\}$ and $\Pr\{Y = 0\}$ are strictly positive. Show that both X and Y have binomial distributions with the same parameter p, the other parameters being m and n, respectively.

23. (a) Let X and Y be independent random variables such that

$$\Pr\{X = i\} = f(i), \qquad \Pr\{Y = i\} = g(i),$$

where

$$f(i) > 0, \quad g(i) > 0, \quad i = 0, 1, 2, \ldots$$

and

$$\sum_{i=0}^{\infty} f(i) = \sum_{i=0}^{\infty} g(i) = 1.$$

Suppose

$$\Pr\{X = k | X + Y = l\} = \begin{cases} \binom{l}{k} p^k (1-p)^{l-k}, & 0 \leq k \leq l, \\ 0, & k > l. \end{cases}$$

Prove that

$$f(i) = e^{-\theta\alpha} \frac{(\theta\alpha)^i}{i!}, \qquad g(i) = e^{-\theta} \frac{\theta^i}{i!}, \qquad \alpha = 0, 1, 2, \ldots,$$

where $\alpha = p/(1-p)$ and $\theta > 0$ is arbitrary.

(b) Show that p is determined by the condition

$$G\left(\frac{1}{1-p}\right) = \frac{1}{f(0)}.$$

Hint: Let $F(s) = \sum f(i)s^i$, $G(s) = \sum g(i)s^i$. Establish first the relation

$$F(u)F(v) = F(vp + (1-p)u)G(vp + (1-p)u).$$

24. Let X be a nonnegative integer–valued random variable with probability generating function $f(s) = \sum_{n=0}^{\infty} a_n s^n$. After observing X, then conduct X binomial trials with probability p of success. Let Y denote the resulting number of successes.

(a) Determine the probability generating function of Y.
(b) Determine the probability generating function of X given that $Y = X$.

Solution: (a) $f(1 - p + ps)$; (b) $f(ps)/f(p)$.

25. (Continuation of Problem 24) Suppose that for every p $(0 < p < 1)$ the probability generating functions of (a) and (b) coincide. Prove that the distribution of X is Poisson, i.e., $f(s) = e^{\lambda(s-1)}$ for some $\lambda > 0$.

26. There are at least four schools of thought on the statistical distribution of stock price differences, or more generally, stochastic models for sequences of stock prices. In terms of number of followers, by far the most popular approach is that of the so-called "technical analysist", phrased in terms of short term trends, support and resistance levels, technical rebounds, and so on. Rejecting this technical viewpoint, two other schools agree that sequences of prices describe a random walk, when price changes are statistically independent of previous price history, but these schools disagree in their choice of the appropriate probability distributions. Some authors find price changes to have a normal distribution while the other group finds a distribution with "fatter tail probabilities", and perhaps even an infinite variance. Finally, a fourth group (overlapping with the preceding two) admits the random walk as a first-order approximation but notes recognizable second-order effects.

This exercise is to show a compatibility between the middle two groups. It has been noted that those that find price changes to be normal typically measure the changes over a fixed number of transactions, while those that find the larger tail probabilities typically measure price changes over a fixed time period that may contain a random number of transactions. Let Z be a price change. Use as the measure of "fatness" (and there could be dispute about this) the *coefficient of excess*

$$\gamma_2 = [m_4/(m_2)^2] - 3,$$

where m_k is the kth moment of Z about its mean.
Suppose on each transaction that the price advances by one unit, or lowers by one unit, each with equal probability. Let N be the number of transactions and write $Z = X_1 + \cdots + X_N$ where the X_n's are independent and identically distributed random variables, each equally likely to be $+1$ or -1. Compute γ_2 for Z: (a) When N is a fixed number a, and (b). When N has a Poisson distribution with mean a.

27. Consider an infinite number of urns into which we toss balls independently, in such a way that a ball falls into the kth urn with probability $1/2^k$, $k = 1, 2, 3, \ldots$. For each positive integer N, let Z_N be the number of urns which contain at

least one ball after a total of N balls have been tossed. Show that

$$E(Z_N) = \sum_{k=1}^{\infty} [1 - (1 - 1/2^k)^N],$$

and that there exist constants $C_1 > 0$ and $C_2 > 0$ such that

$$C_1 \log N \leq E(Z_N) \leq C_2 \log N \qquad \text{for all } N.$$

Hint: Verify and use the facts:

$$E(Z_N) \geq \sum_{k=1}^{\log_2 N} \left[1 - \left(1 - \frac{1}{2^k}\right)^N\right] \geq C \log_2 N$$

and

$$1 - \left(1 - \frac{1}{2^k}\right)^N \leq N \frac{1}{2^k} \quad \text{and} \quad N \sum_{\log_2 N}^{\infty} \frac{1}{2^k} \leq C_2,$$

28. Let L and R be randomly chosen interval endpoints having an arbitrary joint distribution, but, of course, $L \leq R$. Let $p(x) = \Pr\{L \leq x \leq R\}$ be the probability the interval covers the point x, and let $X = R - L$ be the length of the interval. Establish the formula $E[X] = \int_{-\infty}^{\infty} p(x)\, dx$.

29. Let N balls be thrown independently into n urns, each ball having probability $1/n$ of falling into any particular urn. Let $Z_{N,n}$ be the number of empty urns after culminating these tosses, and let $P_{N,n}(k) = \Pr(Z_{N,n} = k)$.

Define $\varphi_{N,n}(t) = \sum_{k=0}^{n} P_{N,n}(k) e^{ikt}$.

(a) Show that

$$P_{N+1,n}(k) = \left(1 - \frac{k}{n}\right) P_{N,n}(k) + \frac{k+1}{n} P_{N,n}(k+1), \quad \text{for} \quad k = 0, 1, \ldots, n.$$

(b) Show that

$$P_{N,n}(k) = \left(1 - \frac{1}{n}\right)^N P_{N,n-1}(k-1) + \sum_{i=1}^{N} \binom{N}{i} \frac{1}{n^i} \left(1 - \frac{1}{n}\right)^{N-i} P_{N-i,n-1}(k).$$

(c) Define $G_n(t, z) = \sum_{N=0}^{\infty} \varphi_{N,n}(t) \frac{n^N}{N!} z^N$. Using part (b), show that $G_n(t, z) = G_{n-1}(t, z)(e^{it} + e^z - 1)$, and conclude that

$$G_n(t, z) = (e^{it} + e^z - 1)^n, \qquad n = 0, 1, 2, \ldots$$

NOTES

A colorful and rich introduction to probability theory and its applications is found in Feller [1]. Feller's book is limited in that it deals only with discrete probabilities.

The text by Gnedenko [2] also serves as an excellent introduction.

Another useful elementary text is that by Parzen [3].

The classic treatise on the subject of stochastic processes is that by Doob [4]. Doob's book serves indispensably for all researches concerned with stochastic processes.

Another outstanding book concerned with the structure of stochastic processes is the recent translation of Dynkin [5].

REFERENCES

1. W. Feller, "An Introduction to Probability Theory and Its Applications," Vol. 1, 2nd ed. Wiley, New York, 1957.
2. B. V. Gnedenko, "Theory of Probability." Chelsea, New York, 1962.
3. E. Parzen, "Modern Probability Theory and Its Applications." Wiley, New York, 1960.
4. J. L. Doob, "Stochastic Processes." Wiley, New York, 1953.
5. E. B. Dynkin, "Theory of Markov Processes." Academic Press, New York, 1965.

Chapter 2

MARKOV CHAINS

This chapter introduces Markov chains and should be included in every first course in stochastic processes. The precise definition of a Markov process (Example c, Section 3 of Chapter 1) might be reviewed at the start.

The reader should try to construct examples illustrating the properties *accessible*, *communicate*, *aperiodic*, *recurrent*, *transient* and *irreducible* discussed in Section 4.

Section 7 is only a page but could be omitted on first reading.

1: Definitions

A discrete time Markov chain $\{X_n\}$ is a Markov stochastic process whose state space is a countable or finite set, and for which $T = (0, 1, 2, \ldots)$. We may refer to the value of X_n as the outcome of the nth trial.

It is frequently convenient to label the state space of the process by the nonnegative integers $(0, 1, 2, \ldots)$, which we will do unless the contrary is explicitly stated, and it is customary to speak of X_n being in state i if $X_n = i$.

The probability of X_{n+1} being in state j, given that X_n is in state i (called a one-step transition probability), is denoted by $P_{ij}^{n,n+1}$, i.e.,

$$P_{ij}^{n,n+1} = \Pr\{X_{n+1} = j \,|\, X_n = i\}. \tag{1.1}$$

The notation emphasizes that in general the transition probabilities are functions not only of the initial and final state, but also of the time of transition as well. When one-step transition probabilities are independent of the time variable (i.e., of the value of n), we say that the Markov process has *stationary transition probabilities* (see the close of Section 3, Chapter 1). Since the vast majority of Markov chains that we shall encounter have stationary transition probabilities, we limit our discussion primarily to such cases.

In this case, $P_{ij}^{n,n+1} = P_{ij}$ is independent of n and P_{ij} is the probability that the state value undergoes a transition from i to j in one trial. It is customary to arrange these numbers P_{ij} as a matrix, that is, an infinite square array

$$\mathbf{P} = \begin{Vmatrix} P_{00} & P_{01} & P_{02} & P_{03} & \cdots \\ P_{10} & P_{11} & P_{12} & P_{13} & \cdots \\ P_{20} & P_{21} & P_{22} & P_{23} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \\ P_{i0} & P_{i1} & P_{i2} & P_{i3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{Vmatrix}$$

and refer to $\mathbf{P} = \|P_{ij}\|$ as the Markov matrix or *transition probability matrix* of the process.

The $(i+1)$st row of $\mathbf{P}$ is the probability distribution of the values of X_{n+1} under the condition $X_n = i$. If the number of states is finite then $\mathbf{P}$ is a finite square matrix whose order (the number of rows) is equal to the number of states. Clearly, the quantities P_{ij} satisfy the conditions

$$P_{ij} \geq 0, \qquad i, j = 0, 1, 2, \ldots, \tag{1.2}$$

$$\sum_{j=0}^{\infty} P_{ij} = 1, \qquad i = 0, 1, 2, \ldots. \tag{1.3}$$

The condition (1.3) merely expresses the fact that some transition occurs at each trial. (For convenience, one says that a transition has occurred even if the state remains unchanged.)

The process is completely determined once (1.1) and the value (or more generally the probability distribution) of X_0 are specified. We shall now prove this fact.

Let $\Pr\{X_0 = i\} = p_i$. It is enough to show how to compute the quantities

$$\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\}, \tag{1.4}$$

as any probability involving $X_{j_1}, \ldots, X_{j_k}$, $j_1 < j_2 < \cdots < j_k$, may be obtained, according to the axiom of total probability, by summing terms of the form (1.4).

By the definition of conditional probabilities we obtain

$$\begin{aligned} &\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\} \\ &\quad = \Pr\{X_n = i_n \,|\, X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}\} \\ &\qquad \cdot \Pr\{X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}\}. \end{aligned} \tag{1.5}$$

Now by the definition of a Markov process,

$$\begin{aligned} &\Pr\{X_n = i_n \,|\, X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}\} \\ &\quad = \Pr\{X_n = i_n \,|\, X_{n-1} = i_{n-1}\} = P_{i_{n-1}, i_n}. \end{aligned} \tag{1.6}$$

Substituting (1.6) into (1.5) gives

$$\Pr\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\}$$
$$= P_{i_{n-1}, i_n} \Pr\{X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}\}. \tag{1.7}$$

If we proceed by induction (1.4) becomes

$$\Pr\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\} = P_{i_{n-1}, i_n} P_{i_{n-2}, i_{n-1}} \cdots P_{i_0, i_1} p_{i_0}. \tag{1.8}$$

2: *Examples of Markov Chains*

The importance of Markov chains lies in the large number of natural physical, biological, and economic phenomena that can be described by them. We now formulate several such examples.

A. SPATIALLY HOMOGENEOUS MARKOV CHAINS

Let ξ denote a discrete-valued random variable whose possible values are the nonnegative integers, $\Pr\{\xi = i\} = a_i$, $a_i \geq 0$, and $\sum_{i=1}^{\infty} a_i = 1$. Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ represent independent observations of ξ.

We shall now describe two different Markov chains connected with the sequence of ξ_i's.

In each case the state space of the process coincides with the set of nonnegative integers.

(i) Consider the process X_n, $n = 0, 1, 2, \ldots$, defined by $X_n = \xi_n$, ($X_0 = \xi_0$ prescribed). Its Markov matrix has the form

$$\mathbf{P} = \begin{Vmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ \vdots & & & & \end{Vmatrix}$$

Each row being identical plainly expresses the fact that the random variable X_{n+1} is independent of X_n.

(ii) Another important class of Markov chains arises from consideration of the successive partial sums η_n of the ξ_i, i.e.,

$$\eta_n = \xi_1 + \xi_2 + \cdots + \xi_n, \qquad n = 1, 2, \ldots$$

and, by definition, $\eta_0 = 0$. The process $X_n = \eta_n$ is readily seen to be a

Markov chain. We can easily compute its transition probability matrix as follows:

$$\begin{aligned}&\Pr\{X_{n+1}=j\,|X_n=i\}\\&\quad=\Pr\{\xi_1+\cdots+\xi_{n+1}=j\,|\xi_1+\cdots+\xi_n=i\}=\Pr\{\xi_{n+1}=j-i\}\\&\quad=\begin{cases}a_{j-i} & \text{for } j\geq i,\\ 0 & \text{for } j<i,\end{cases}\end{aligned}$$

where we have used the assumed independence of the ξ_i.

Schematically, we have

$$\mathbf{P}=\begin{Vmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & \cdots\\ 0 & a_0 & a_1 & a_2 & a_3 & \cdots\\ 0 & 0 & a_0 & a_1 & a_2 & \cdots\\ \vdots & & & & & \end{Vmatrix}. \tag{2.1}$$

If the possible values of the random variable ξ are permitted to be the positive and negative integers, then the possible values of η_n for each n will be contained among the totality of all integers. Instead of labeling the states conventionally by means of the nonnegative integers, it is more convenient to identify the state space with the totality of integers, since the probability transition matrix will then appear in a more symmetric form. The state space consists then of the values $\ldots -2, -1, 0, 1, 2, \ldots$. The transition probability matrix becomes

$$\mathbf{P}=\begin{Vmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & a_{-1} & a_0 & a_1 & a_2 & a_3 & \cdots\\ \cdots & a_{-2} & a_{-1} & a_0 & a_1 & a_2 & \cdots\\ \cdots & a_{-3} & a_{-2} & a_{-1} & a_0 & a_1 & \cdots\\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{Vmatrix},$$

where $\Pr\{\xi=k\}=a_k$, $k=0, \pm 1, \pm 2, \ldots$, and $a_k\geq 0$, $\sum_{k=-\infty}^{\infty} a_k=1$.

B. ONE-DIMENSIONAL RANDOM WALKS

In discussing random walks it is an aid to intuition to speak about the state of the system as the position of a moving "particle."

A one-dimensional random walk is a Markov chain whose state space is a finite or infinite subset $a, a+1, \ldots, b$ of the integers, in which the particle, if it is in state i, can in a single transition either stay in i or move to one of the adjacent states $i-1, i+1$. If the state space is taken

as the nonnegative integers, the transition matrix of a random walk has the form

$$\mathbf{P} = \begin{Vmatrix} r_0 & p_0 & 0 & 0 & \cdots & & \\ q_1 & r_1 & p_1 & 0 & \cdots & & \\ 0 & q_2 & r_2 & p_2 & \cdots & & \\ & \ddots & & & & & \\ & & 0 & q_i & r_i & p_i & 0 \\ & & & \ddots & & & \ddots \end{Vmatrix}, \tag{2.2}$$

where $p_i > 0$, $q_i > 0$, $r_i \geq 0$, and $q_i + r_i + p_i = 1$, $i = 1, 2, \ldots$ $(i \geq 1)$, $p_0 \geq 0$, $r_0 \geq 0$, $r_0 + p_0 = 1$. Specifically, if $X_n = i$ then, for $i \geq 1$,

$$\Pr\{X_{n+1} = i+1 \,|X_n = i\} = p_i: \qquad \Pr\{X_{n+1} = i-1 \,|X_n = i\} = q_i$$

$$\Pr\{X_{n+1} = i \,|X_n = i\} = r_i,$$

with the obvious modifications holding for $i = 0$.

The designation "random walk" seems apt since a realization of the process describes the path of a person (suitably intoxicated) moving randomly one step forward or backward.

The fortune of a player engaged in a series of contests is often depicted by a random walk process. Specifically, suppose an individual (player A) with fortune k plays a game against an infinitely rich adversary and has probability p_k of winning one unit and probability $q_k = 1 - p_k$ $(k \geq 1)$ of losing one unit in each contest (the choice of the contest at each stage may depend on his fortune), and $r_0 = 1$. The process $\{X_n\}$, where X_n represents his fortune after n contests, is clearly a random walk. Note that once the state 0 is reached (i.e., player A is wiped out), the process remains in that state. This process is also commonly known as the "gambler's ruin."

The random walk corresponding to $p_k = p$, $q_k = 1 - p = q$ for all $k \geq 1$ and $r_0 = 1$ with $p > q$ describes the situation of identical contests with a definite advantage to player A in each individual trial. We shall prove in Chapter 3 that with probability $(q/p)^{x_0}$, where x_0 represents his fortune at time 0, player A is ultimately ruined (his entire fortune is lost), while with probability $1 - (q/p)^{x_0}$, his fortune increases, in the long run, without limit. If $p < q$ then the advantage is decidedly in favor of the house, and with certainty (probability 1) player A is ultimately ruined if he persists in playing as long as he is able to. The same (i.e., certainty of ultimate ruin) is true even if the individual games are fair, that is, $p_k = q_k = \frac{1}{2}$.

If the adversary, player B, also starts with a limited fortune y and player A has an initial fortune x (let $x + y = a$), then we may again

consider the Markov chain process X_n representing player A's fortune. However, the states of the process are now restricted to the values $0, 1, 2, \ldots, a$. At any trial, $a - X_n$ is interpreted as player B's fortune. If we allow the possibility of neither player winning in a contest, the transition probability matrix takes the form

$$\mathbf{P} = \begin{Vmatrix} 1 & 0 & 0 & 0 & \cdots & \\ q_1 & r_1 & p_1 & 0 & \cdots & \\ 0 & q_2 & r_2 & p_2 & \cdots & \\ & & \cdot & \cdot & & \\ & & & q_{a-1} & r_{a-1} & p_{a-1} \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{Vmatrix} \tag{2.3}$$

Again $p_i(q_i)$, $i = 1, 2, \ldots, a-1$, denotes the probability of player A's fortune increasing (decreasing) by 1 at the subsequent trial when his present fortune is i, and r_i may be interpreted as the probability of a draw. Note that, in accordance with the Markov chain given in (2.3), when player A's fortune (the state of the process) reaches 0 or a it remains in this same state forever. We say player A is ruined when the state of the process reaches 0 and player B is ruined when the state of the process reaches a.

Random walks are not only useful in simulating situations of gambling but frequently serve as reasonable discrete approximations to physical processes describing the motion of diffusing particles. If a particle is subjected to collisions and random impulses, then its position fluctuates randomly, although the particle describes a continuous path. If the future position (i.e., its probability distribution) of the particle depends only on the present position, then the process $\{X_t\}$, where X_t is the position at time t, is Markovian. A discrete approximation to such a continuous motion corresponds to a random walk. A classical discrete version of Brownian motion (see Section 2 of Chapter 1) is provided by the symmetric random walk. By a symmetric random walk on the integers (say all the integers) we mean a Markov chain with state space the totality of all integers and whose transition probability matrix has the elements

$$P_{ij} = \begin{cases} p & \text{if } j = i+1, \\ q & \text{if } j = i-1, \\ r & \text{if } j = i, \\ 0 & \text{otherwise,} \end{cases} \qquad i, j = 0, \pm 1, \pm 2, \ldots,$$

where $p > 0$, $r \geq 0$, and $2p + r = 1$. Conventionally, "symmetric random walk" refers only to the case $r = 0$, $p = \frac{1}{2}$.

Motivated by consideration of certain physical models we are led to the study of random walks on the set of the nonnegative integers. We

classify the different processes by the nature of the zero state. Let us fix attention on the random walk described by (2.2). If $p_0 = 1$, and therefore $r_0 = 0$, we have a situation where the zero state acts like a reflecting barrier. Whenever the particle reaches the zero state, the next transition automatically returns it to state one. This corresponds to the physical process where an elastic wall exists at zero, and the particle bounces off with no after-effects.

If $p_0 = 0$ and $r_0 = 1$ then 0 acts as an absorbing barrier. Once the particle reaches zero it remains there forever. If $p_0 > 0$ and $r_0 > 0$, then 0 is a partially reflecting barrier.

When the random walk is restricted to a finite number of states S, say $0, 1, 2, \ldots, a$, then both the states 0 and a independently and in any combination may be reflecting, absorbing, or partially reflecting barriers. We have already encountered a model (gambler's ruin, involving two adversaries with finite resources) of a random walk confined to the states S where 0 and a are absorbing [see (2.3)].

A classical mathematical model of diffusion through a membrane is the famous Ehrenfest model, namely, a random walk on a finite set of states whereby the boundary states are reflecting. The random walk is restricted to the states $i = -a, -a+1, \ldots, -1, 0, 1, \ldots, a$ with transition probability matrix

$$P_{ij} = \begin{cases} \dfrac{a-i}{2a}, & \text{if } j = i+1. \\ \dfrac{a+i}{2a}, & \text{if } j = i-1. \\ 0, & \text{otherwise.} \end{cases}$$

The physical interpretation of this model is as follows. Imagine two containers containing a total of $2a$ balls. Suppose the first container, labeled A, holds k balls and the second container B holds $2a - k$ balls. A ball is selected at random (all selections are equally likely) from among the totality of the $2a$ balls and moved to the other container. Each selection generates a transition of the process. Clearly the balls fluctuate between the two containers with a drift from the one with the larger concentration of balls to the one with the smaller concentration of balls. A physical system which in the main is governed by a set of restoring forces essentially proportional to the distance from an equilibrium position may sometimes be approximated by this Ehrenfest model.

The classical symmetric random walk in n dimensions admits the following formulation. The state space is identified with the set of all integral lattice points in E^n (Euclidean n space): that is, a state is an n-tuple

$\mathbf{k} = (k_1, k_2, \ldots, k_n)$ of integers. The transition probability matrix is defined by

$$P_{\mathbf{kl}} = \begin{cases} \dfrac{1}{2n} & \text{if } \sum_{i=1}^{n} |l_i - k_i| = 1. \\ 0 & \text{otherwise.} \end{cases}$$

Analogous to the one-dimensional case, the symmetric random walk in E^n represents a discrete version of n-dimensional Brownian motion.

C. A DISCRETE QUEUEING MARKOV CHAIN

Customers arrive for service and take their place in a waiting line. During each period of time a single customer is served, provided that at least one customer is present. If no customer awaits service then during this period no service is performed. (We can imagine, for example, a taxi stand at which a cab arrives at fixed time intervals to give service. If no one is present the cab immediately departs.) During a service period new customers may arrive. We suppose the actual number of arrivals in the nth period is a random variable ξ_n whose distribution function is independent of the period and is given by

$$\Pr\{k \text{ customers arrive in a service period}\} = \Pr\{\xi_n = k\} = a_k,$$

$$k = 0, 1, \ldots, \quad a_k \geq 0 \quad \text{and} \quad \sum_{k=0}^{\infty} a_k = 1. \tag{2.4}$$

We also assume the r.v.'s ξ_n are independent. The state of the system at the start of each period is defined to be the number of customers waiting in line for service. If the present state is i then after a lapse of one period the state is

$$j = \begin{cases} i - 1 + \xi & \text{if } i \geq 1, \\ \xi & \text{if } i = 0, \end{cases} \tag{2.5}$$

where ξ is the number of new customers having arrived in this period while a single customer was served. In terms of the random variables of the process we can express (2.5) formally as

$$X_{n+1} = (X_n - 1)^+ + \xi_n, \tag{2.6}$$

where $Y^+ = \max(Y, 0)$. In view of (2.4) and (2.5) the transition probability matrix may be trivially calculated and we obtain

$$\|P_{ij}\| = \begin{Vmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & \ldots \\ a_0 & a_1 & a_2 & a_3 & a_4 & \ldots \\ 0 & a_0 & a_1 & a_2 & a_3 & \ldots \\ 0 & 0 & a_0 & a_1 & a_2 & \ldots \\ 0 & 0 & 0 & a_0 & a_1 & \ldots \end{Vmatrix} \tag{2.7}$$

It is intuitively clear that if the expected number of new customers, $\sum_{k=0}^{\infty} ka_k$, that arrive during a service period exceeds 1 then certainly with the passage of time the length of the waiting line increases without limit.

On the other hand, if $\sum_{k=0}^{\infty} ka_k < 1$ then we shall see that the length of the waiting line approaches an equilibrium (stationary state). If $\sum ka_k = 1$, a situation of gross instability develops. These statements will be formally elaborated after we have set forth the relevant theory of recurrence (see Section 5, Chapter 3).

D. INVENTORY MODEL

Consider a situation in which a commodity is stocked in order to satisfy a continuing demand. We assume that the replenishing of stock takes place at successive times $t_1, t_2, \ldots$, and we assume that the cumulative demand for the commodity over the interval (t_{n-1}, t_n) is a random variable ξ_n whose distribution function is independent of the time period,

$$\Pr\{\xi_n = k\} = a_k, \qquad k = 0, 1, 2, \ldots, \tag{2.8}$$

where $a_k \geq 0$ and $\sum_{k=0}^{\infty} a_k = 1$. The stock level is examined at the start of each period. An inventory policy is prescribed by specifying two nonnegative critical values s and $S > s$. The implementation of the inventory policy is as follows: If the available stock quantity is not greater than s then immediate procurement is done so as to bring the quantity of stock on hand to the level S. If, however, the available stock is in excess of s then no replenishment of stock is undertaken. Let X_n denote the stock on hand just prior to restocking at t_n. The states of the process $\{X_n\}$ consist of the possible values of the stock size

$$S, \quad S-1, \ldots, +1, \quad 0, \quad -1, \quad -2, \ldots,$$

where a negative value is interpreted as an unfulfilled demand for stock, which will be satisfied immediately upon restocking. According to the rules of the inventory policy, the stock levels at two consecutive periods are connected by the relation

$$X_{n+1} = \begin{cases} X_n - \xi_{n+1} & \text{if } s < X_n \leq S, \\ S - \xi_{n+1} & \text{if } X_n \leq s, \end{cases} \tag{2.9}$$

where ξ_n is the quantity of demand that arises in the nth period, based on the probability law (2.8). If we assume the ξ_n to be mutually independent, then the stock values $X_0, X_1, X_2, \ldots$ plainly constitute a Markov chain whose transition probability matrix can be calculated in accordance with the relation (2.9).

E. SUCCESS RUNS

Consider a Markov chain on the nonnegative integers with transition probability matrix of the form

$$\|P_{ij}\| = \begin{Vmatrix} p_0 & q_0 & 0 & 0 & \cdots \\ p_1 & 0 & q_1 & 0 & \cdots \\ p_2 & 0 & 0 & q_2 & \cdots \\ p_3 & 0 & 0 & 0 & \\ \vdots & & & & \end{Vmatrix}, \tag{2.10}$$

where $q_i > 0$, $p_i > 0$ and $q_i + p_i = 1$, $i = 0, 1, 2, \ldots$. The zero state plays a distinguished role here in that it can be reached in one transition from any other state, while state $i + 1$ can be reached only from state i.

This example is very easy to compute with and we will therefore frequently illustrate concepts and results in terms of it.

A special case of this transition matrix arises when one is dealing with success runs resulting from repeated trials each of which admits two possible outcomes, success (S) or failure (F). More explicitly, consider a sequence of trials with two possible outcomes (S) or (F). Moreover, suppose that in each trial, the probability of (S) is α and the probability of (F) is $\beta = 1 - \alpha$. We say a success run of length r happened at trial n if the outcomes in the preceding $r + 1$ trials, including the present trial as the last, were respectively, $F, S, S, \ldots, S$. Let us now label the present state of the process by the length of the success run currently under way. In particular, if the last trial resulted in a failure then the state is zero. Similarly, when the preceding $r + 1$ trials in order had the outcomes $F, S, S, \ldots, S$, the state variable would carry the label r. The process is clearly Markovian (since the individual trials were independent of each other) and its transition matrix has the form (2.10) where

$$p_n = \beta, \qquad n = 0, 1, 2, \ldots.$$

F. BRANCHING PROCESSES

Suppose an organism at the end of its lifetime produces a random number ξ of offspring with probability distribution

$$\Pr\{\xi = k\} = a_k, \qquad k = 0, 1, 2, \ldots, \tag{2.11}$$

where, as usual, $a_k \geq 0$ and $\sum_{k=0}^{\infty} a_k = 1$. We assume that all offspring act independently of each other and at the end of their lifetime (for

simplicity, the lifespans of all organisms are assumed to be the same) individually have progeny in accordance with the probability distribution (2.11), thus propagating their species. The process $\{X_n\}$, where X_n is the population size at the nth generation, is a Markov chain.

In fact, the only relevant knowledge regarding the distribution of $X_{n_1}, X_{n_2}, \ldots, X_{n_r}, X_n, n_1 < n_2 < \cdots < n_r < n$, is the last known population count, since the number of the offspring is a function merely of the present population size. The transition matrix is obviously given by

$$P_{ij} = \Pr\{X_{n+1} = j \,|\, X_n = i\} = \Pr\{\xi_1 + \cdots + \xi_i = j\}, \tag{2.12}$$

where the ξ's are independent observations of a random variable with probability law (2.11). The formula (2.12) may be reasoned simply as follows. In the nth generation the i individuals independently give rise to numbers of offspring $\{\xi_k\}_{k=1}^{i}$ and hence the cumulative number produced is $\xi_1 + \xi_2 + \cdots + \xi_i$.

If we use generating functions, then clearly the generating function of $\xi_1 + \xi_2 + \cdots + \xi_i$ is $[g(s)]^i$, where g is the generating function associated with ξ. (We are using the property of composition of generating functions in the case of sums of independent r.v.'s: see Section 1 of Chapter 1, page 12.) Hence, P_{ij} is simply the jth coefficient in the power series expansion of $[g(s)]^i$.

G. MARKOV CHAINS IN GENETICS

The following idealized genetics model was introduced by S. Wright to investigate the fluctuation of gene frequency under the influence of mutation and selection. We begin by describing a so-called simple haploid model of random reproduction, disregarding mutation pressures and selective forces. We assume that we are dealing with a fixed population size of $2N$ genes composed of type-a and type-A individuals. The make-up of the next generation is determined by $2N$ independent binomial trials as follows: If the parent population consists of j a-genes and $2N - j$ A-genes then each trial results in a or A with probabilities

$$p_j = \frac{j}{2N}, \qquad q_j = 1 - \frac{j}{2N},$$

respectively. Repeated selections are done with replacement. By this procedure we generate a Markov chain $\{X_n\}$ where X_n is the number of a-genes in the nth generation among a constant population size of $2N$ elements. The state space contains the $2N + 1$ values $\{0, 1, 2, \ldots, 2N\}$.

The transition probability matrix is computed according to the binomial distribution as

$$\Pr\{X_{n+1} = k \,|\, X_n = j\} = P_{jk} = \binom{2N}{k} p_j^k q_j^{2N-k} \qquad (j, k = 0, 1, \ldots, 2N). \tag{2.13}$$

For some discussion of the biological justification of these postulates we refer the reader to Fisher†.

Notice that states 0 and $2N$ are completely absorbing in the sense that once $X_n = 0$ or $(2N)$ then $X_{n+k} = 0$ or $(2N)$ respectively for all $k \geq 0$. One of the questions of interest is to determine the probability under the condition $X_0 = i$ that the population will attain fixation, i.e., a pure population composed only of a-genes or A-genes. It is also pertinent to determine the rate of approach to fixation. We will examine such questions in our general analysis of absorption probabilities.

A more realistic model takes account of mutation pressures. We assume that prior to the formation of the new generation each gene has the possibility to mutate, that is, to change into a gene of the other kind. Specifically, we assume that for each gene the mutation $\mathrm{a} \to \mathrm{A}$ occurs with probability α_1, and $\mathrm{A} \to \mathrm{a}$ occurs with probability α_2. Again we assume that the composition of the next generation is determined by $2N$ independent binomial trials. The relevant value of p_j and q_j when the parent population consists of j a-genes are now taken to be

$$\begin{aligned} p_j &= \frac{j}{2N}(1 - \alpha_1) + \left(1 - \frac{j}{2N}\right)\alpha_2, \\ q_j &= \frac{j}{2N}\alpha_1 + \left(1 - \frac{j}{2N}\right)(1 - \alpha_2). \end{aligned} \tag{2.14}$$

The rationale is as follows: We assume that the mutation pressures operate first, after which a new gene is determined by selecting at random from the population. Now the probability of selecting an a-gene after the mutation forces have acted is just $1/2N$ times the number of a-genes present: hence the average probability (averaged with respect to the possible mutations) is simply $1/2N$ times the average number of a-genes after mutation. But this average number is clearly $j(1 - \alpha_1) + (2N - j)\alpha_2$, which leads at once to (2.14).

The transition probabilities of the associated Markov chain are calculated by (2.13) using the values of p_j and q_j given in (2.14).

† R. A. Fisher, "The Genetical Theory of Natural Selection," Oxford (Clarendon) Press, London and New York, 1962.

If $\alpha_1\alpha_2 > 0$ then fixation will not occur in any state. Instead, as $n \to \infty$, the distribution function of X_n will approach a steady state distribution of a random variable ξ where $\Pr\{\xi = k\} = \pi_k$ $(k = 0, 1, 2, \ldots, 2N)$ $(\sum_{k=0}^{n} \pi_k = 1, \pi_k > 0)$. The distribution function of ξ is called the steady state gene frequency distribution.

We return to the simple random mating model and discuss the concept of a selection force operating in favor of, say, a-genes. Suppose we wish to impose a selective advantage for a-genes over A-genes so that the relative number of offspring have expectations proportional to $1 + s$ and 1, respectively, where s is small and positive. We replace $p_j = j/2N$ and $q_j = 1 - j/2N$ by

$$p_j = \frac{(1+s)j}{2N + sj}, \qquad q_j = 1 - p_j,$$

and build the next generation by binomial sampling as before. If the parent population consisted of j a-genes, then in the next generation the expected population sizes of a-genes and A-genes, respectively, are

$$2N\frac{(1+s)\,j}{2N+sj}, \qquad 2N\frac{(2N-j)}{2N+sj}.$$

The ratio of expected population size of a-genes to A-genes at the $(n+1)$th generation is

$$\frac{1+s}{1}\cdot\frac{j}{2N-j} = \left(\frac{1+s}{1}\right)\left(\frac{\text{number of a-genes in the } n\text{th generation}}{\text{number of A-genes in the } n\text{th generation}}\right)$$

which explains the meaning of selectivity.

H. GENETIC MODEL II

The gene appears to be composed of a number of subunits, say for definiteness N. When a cell containing the gene prepares to split, each subunit doubles and each of the two cells receives a gene composed of the same number of subunits as before. One or more of the subunits may be in a mutant form. When doubling occurs it is assumed that mutant units produce mutant units and nonmutant units produce nonmutant units. Moreover, the subunits are assumed to be divided between the two new genes randomly as if by drawing from an urn. We shall trace a single line of descent rather than all the population as it multiplies. To describe the history of the line we consider a Markov chain whose state space is identified with the values $0, 1, 2, \ldots, N$. Then the gene is said to be in

state i if its composition consists of i mutant subunits and $N - i$ normal subunits. The transition probabilities are computed by the formula

$$P_{ij} = \frac{\binom{2i}{j}\binom{2N-2i}{N-j}}{\binom{2N}{N}}. \tag{2.15}$$

The derivation of P_{ij} is as follows. Suppose the parent gene is in state i: then after doubling we obtain a totality of $2i$ mutant units and $2N - 2i$ normal units. The nature of the daughter gene is formed by selecting an arbitrary N units from this collection. In accordance with the hypergeometric probability law, the probability that the daughter gene is in state j is given by (2.15).

The states $j = 1, 2, \ldots, N - 1$ are called mixed, and states 0 and N will be called pure. State N is of interest in that a gene all of whose subunits are mutant may cause the death of its possessor, while state 0 implies that a gene of this type will produce no more of the mutant form. We will later determine the explicit probabilities that, starting from state i, the gene ultimately fixes in state 0 or state N.

3: Transition Probability Matrices of a Markov Chain

A Markov chain is completely defined by its one-step transition probability matrix and the specification of a probability distribution on the state of the process at time 0. The analysis of a Markov chain concerns mainly the calculation of the probabilities of the possible realizations of the process. Central in these calculations are the n-step transition probability matrices, $\mathbf{P}^{(n)} = \|P_{ij}^n\|$. Here P_{ij}^n denotes the probability that the process goes from state i to state j in n transitions. Formally,

$$P_{ij}^n = \Pr\{X_{n+m} = j \,|\, X_m = i\}. \tag{3.1}$$

Observe that we are dealing only with temporally homogeneous processes having stationary transition probabilities, since otherwise the left-hand side of (3.1) would also depend on m.

The Markovian assumption allows us to express (3.1) immediately in terms of $\|P_{ij}\|$ as stated in the following theorem.

Theorem 3.1. *If the one-step transition probability matrix of a Markov chain is* $\mathbf{P} = \|P_{ij}\|$, *then*

$$P_{ij}^n = \sum_{k=0}^{\infty} P_{ik}^r P_{kj}^s \tag{3.2}$$

for any fixed pair of nonnegative integers r and s satisfying $r + s = n$, *where we define*

$$P_{ij}^0 = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

From the theory of matrices (see the appendix), we recognize relation (3.2) as just the formula for matrix multiplication, so that $\mathbf{P}^{(n)} = \mathbf{P}^n$; in other words, the numbers P_{ij}^n may be regarded as the entries in the matrix $\mathbf{P}^n$, the nth power of $\mathbf{P}$.

Proof. We carry out the argument in the case $n = 2$. The event of going from state i to state j in two transitions can be realized in the mutually exclusive ways of going to some intermediate state k ($k = 0, 1, 2, \ldots$) in the first transition and then going from state k to state j in the second transition.

Because of the Markovian assumption the probability of the second transition is P_{kj}, and that of the first transition is clearly P_{ik}. If we use the law of total probabilities Eq. (3.2) follows. The argument in the general case is identical.

If the probability of the process initially being in state j is p_j, i.e., the distribution law of X_0 is $\Pr\{X_0 = j\} = p_j$, then the probability of the process being in state k at time n is

$$p_k^{(n)} = \sum_{j=0}^{\infty} p_j P_{jk}^n = \Pr\{X_n = k\}. \tag{3.3}$$

Besides determining the joint probability distributions of the process for all time, usually a formidable task, it is frequently of interest to find the asymptotic behavior of P_{ij}^n as $n \to \infty$. One might expect that the influence of the initial state recedes in time and that consequently, as $n \to \infty$, P_{ij}^n approaches a limit which is independent of i. In order to analyze precisely the asymptotic behavior of the process we need to introduce some principles of classifying states of a Markov chain.

4: Classification of States of a Markov Chain

State j is said to be *accessible* from state i if for some integer $n \geq 0$, $P_{ij}^n > 0$: i.e., state j is accessible from state i if there is positive probability that in a finite number of transitions state j can be reached starting from state i. Two states i and j, each accessible to the other, are said to

communicate and we write $i \leftrightarrow j$. If two states i and j do not communicate, then either

$$P_{ij}^n = 0 \qquad \text{for all} \quad n \geq 0$$

or

$$P_{ji}^n = 0 \qquad \text{for all} \quad n \geq 0$$

or both relations are true. The concept of communication is an equivalence relation.

(i) $i \leftrightarrow i$ (reflexivity), a consequence of the definition of $P_{ij}^0 = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$.

(ii) If $i \leftrightarrow j$, then $j \leftrightarrow i$ (symmetry), from the definition of communication.

(iii) If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$ (transitivity).

The proof of transitivity proceeds as follows: $i \leftrightarrow j$ and $j \leftrightarrow k$ imply that there exist integers n and m such that $P_{ij}^n > 0$ and $P_{jk}^m > 0$. Consequently by (3.2) and the nonnegativity of each P_{rs}^t, we conclude that

$$P_{ik}^{n+m} = \sum_{r=0}^{\infty} P_{ir}^n P_{rk}^m \geq P_{ij}^n P_{jk}^m > 0.$$

A similar argument shows the existence of an integer v such that $P_{ki}^v > 0$, as desired.

We can now partition the totality of states into equivalence classes. The states in an equivalence class are those which communicate with each other. It may be possible, starting in one class, to enter some other class with positive probability; if so, however, it is clearly not possible to return to the initial class, or else the two classes would together form a single class. We say that the Markov chain is *irreducible* if the equivalence relation induces only one class. In other words, a process is irreducible if all states communicate with each other.

To illustrate this concept, consider the transition probability matrix

$$\mathbf{P} = \left\| \begin{array}{cc:ccc} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \end{array} \right\| = \left\| \begin{array}{cc} \mathbf{P}_1 & 0 \\ 0 & \mathbf{P}_2 \end{array} \right\|,$$

where $\mathbf{P}_1$ is an abbreviation for the matrix formed from the initial two rows and columns of $\mathbf{P}$, and similarly for $\mathbf{P}_2$. This Markov chain clearly divides into the two classes composed of states $\{1, 2\}$ and states $\{3, 4, 5\}$.

If the state of X_0 lies in the first class, then the state of the system thereafter remains in this class and for all purposes the relevant transition matrix is $\mathbf{P}_1$. Similarly, if the initial state belongs to the second class, then the relevant transition matrix is $\mathbf{P}_2$. This is a situation where we have two completely unrelated processes labeled together.

In the random walk model with transition matrix

$$\mathbf{P} = \begin{array}{c} \\ \left\| \begin{matrix} 1 & 0 & 0 & 0\cdots\cdots & 0 & 0 & 0 \\ q & 0 & p & 0\cdots\cdots & 0 & 0 & 0 \\ 0 & q & 0 & p\cdots\cdots & 0 & 0 & 0 \\ \vdots & & & & \vdots & \vdots & \vdots \\ 0 & \cdots\cdots\cdots & & & q & 0 & p \\ 0 & \cdots\cdots\cdots\cdots & & & 0 & 0 & 1 \end{matrix} \right\| \end{array} \begin{array}{c} \textit{states} \\ 0 \\ 1 \\ 2 \\ \vdots \\ a-1 \\ a \end{array}$$

we have the three classes $\{0\}$, $\{1, 2, \ldots, a-1\}$, and $\{a\}$. This is an example where it is possible to reach the first class or third class from the second class, but it is not possible to return to the second class from either the first or the third class.

Direct inspection shows that the queueing Markov chain example C of Section 2 is irreducible when $a_k > 0$ for all k. Under the same condition we may easily verify that the inventory model (Example D) is irreducible: the success run Markov chain model (Example E) under the condition $q_i > 0$, $p_i > 0$, $(i = 0, 1, \ldots)$ is also irreducible.

PERIODICITY OF A MARKOV CHAIN

We define the period of state i, written $d(i)$, to be the greatest common divisor (g.c.d.) of all integers $n \geq 1$ for which $P_{ii}^n > 0$. [If $P_{ii}^n = 0$ for all $n \geq 1$ define $d(i) = 0$.] If in the random walk [see (2.2)] all $r_i = 0$, then every state has period two. If for a single state, i_0, $r_{i_0} > 0$, then every state now has period one, since regardless of the initial state j the system can reach state i_0 and remain in this state any length of time before returning to state j.

In a finite Markov chain of n states with transition matrix

$$\mathbf{P} = \begin{array}{c} \overbrace{\qquad\qquad\qquad\qquad}^{n} \\ \left\| \begin{matrix} 0 & 1 & 0 & 0 & \cdots\cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots\cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & \cdots\cdots\cdots\cdots & & & 1 \\ 1 & 0 & 0 & \cdots\cdots\cdots & & 0 \end{matrix} \right\| \end{array}$$

each state has period n.

We state, without proof, three basic properties of the period of a state. (In this connection, consult Problems 2—4 of this chapter.)

Theorem 4.1. *If* $i \leftrightarrow j$ *then* $d(i) = d(j)$.

This assertion shows that the period is a constant in each class of communicating states.

Theorem 4.2. *If state* i *has period* $d(i)$ *then there exists an integer* N *depending on* i *such that for all integers* $n \geq N$

$$P_{ii}^{nd(i)} > 0.$$

This asserts that a return to state i can occur at all sufficiently large multiples of the period $d(i)$.

Corollary 4.1. *If* $P_{ji}^{m} > 0$, *then* $P_{ji}^{m+nd(i)} > 0$ *for all* n (*a positive integer*) *sufficiently large.*

A Markov chain in which each state has period one is called *aperiodic*. The vast majority of Markov chain processes we deal with are aperiodic. Random walks usually typify the periodic cases arising in practice. Results will be developed for the aperiodic case and the modified conclusions for the general case will be stated usually without proof. The industrious reader can easily supply the required formal proof.

5: *Recurrence*

Consider an arbitrary, but fixed, state i. We define for each integer $n \geq 1$,

$$f_{ii}^{n} = \Pr\{X_n = i, X_v \neq i, v = 1, 2, \ldots, n-1 \,|\, X_0 = i\}.$$

In other words, f_{ii}^{n} is the probability that, starting from state i, the first return to state i occurs at the nth transition. Clearly $f_{ii}^{1} = P_{ii}$ and f_{ii}^{n} may be calculated recursively according to

$$P_{ii}^{n} = \sum_{k=0}^{n} f_{ii}^{k} P_{ii}^{n-k}, \qquad n \geq 1, \tag{5.1}$$

where we define $f_{ii}^{0} = 0$ for all i. Equation (5.1) is derived by decomposing the event from which P_{ii}^{n} is computed according to the time of the first return to state i. Indeed, consider all the possible realizations of the process for which $X_0 = i$, $X_n = i$ and the first return to state i occurs at the kth transition. Call this event E_k. The events E_k ($k = 1, 2, \ldots, n$) are clearly mutually exclusive. The probability of the event that the first return is at the kth transition is by definition f_{ii}^{k}. In the remaining $n - k$

transitions, we are dealing only with those realizations for which $X_n \neq i$. Using the Markov property, we have

$$\begin{aligned}\Pr\{E_k\} &= \Pr\{\text{first return is at } k\text{th transition} \mid X_0 = i\} \Pr\{X_n = i \mid X_k = i\} \\ &= f_{ii}^k P_{ii}^{n-k}, \qquad 1 \leq k \leq n,\end{aligned}$$

(recall that $P_{ii}^0 = 1$). Hence

$$\Pr\{X_n = i \mid X_0 = i\} = \sum_{k=1}^{n} \Pr\{E_k\} = \sum_{k=1}^{n} f_{ii}^k P_{ii}^{n-k} = \sum_{k=0}^{n} f_{ii}^k P_{ii}^{n-k},$$

since by definition $f_{ii}^0 = 0$.

We next introduce the related generating functions.

Definition. The generating function $P_{ij}(s)$ of the sequence $\{P_{ij}^n\}$ is

$$P_{ij}(s) = \sum_{n=0}^{\infty} P_{ij}^n s^n \qquad \text{for} \quad |s| < 1. \tag{5.2}$$

In a similar manner we define the generating function of the sequence $\{f_{ij}^n\}$ (for the definition of $\{f_{ij}^n\}$ when $i \neq j$, see immediately below Eq. (5.9))

$$F_{ij}(s) = \sum_{n=0}^{\infty} f_{ij}^n s^n \qquad \text{for} \quad |s| < 1. \tag{5.3}$$

Recall the property (see page 12 of Chapter 1)† that, if

$$A(s) = \sum_{k=0}^{\infty} a_k s^k \qquad \text{and} \qquad B(s) = \sum_{l=0}^{\infty} b_l s^l, \tag{5.4}$$

then

$$A(s)B(s) = C(s) = \sum_{r=0}^{\infty} c_r s^r, \qquad \text{for} \quad |s| < 1, \tag{5.5}$$

where

$$c_r = a_0 b_r + a_1 b_{r-1} + \ldots + a_r b_0 . \tag{5.6}$$

If we identify the a_k's with the f_{ii}^k's and the b_l's with the P_{ii}^l's, then comparing (5.1) with (5.6) we obtain

$$F_{ii}(s) P_{ii}(s) = P_{ii}(s) - 1 \qquad \text{for} \quad |s| < 1 \tag{5.7}$$

or

$$P_{ii}(s) = \frac{1}{1 - F_{ii}(s)} \qquad \text{for} \quad |s| < 1. \tag{5.8}$$

† $$\begin{aligned}A(s)B(s) &= \left(\sum_{k=0}^{\infty} a_k s^k\right)\left(\sum_{l=0}^{\infty} b_l s^l\right) \\ &= a_0 b_0 + (a_1 b_0 + b_1 a_0)s + (a_2 b_0 + a_1 b_1 + a_0 b_2)s^2 + \cdots \\ &= \sum_{k=0}^{\infty} s^k \left(\sum_{j=0}^{k} a_j b_{k-j}\right) = \sum_{k=0}^{\infty} c_k s^k.\end{aligned}$$

Subtracting the constant 1 in (5.7) is necessary since (5.1) is not valid for $n = 0$.

By an argument analogous to that which led to (5.1), we obtain

$$P_{ij}^n = \sum_{k=0}^{n} f_{ij}^k P_{jj}^{n-k}, \qquad i \neq j,\ n \geq 0, \tag{5.9}$$

where f_{ij}^k is the probability that first passage from state i to state j occurs at the kth transition. Again we define $f_{ij}^0 = 0$ for all i and j. It follows from (5.9), if we refer to (5.5), that

$$P_{ij}(s) = F_{ij}(s) P_{jj}(s) \qquad \text{for} \quad |s| < 1. \tag{5.10}$$

We say a state i is *recurrent* if and only if $\sum_{n=1}^{\infty} f_{ii}^n = 1$. This says that a state i is recurrent if and only if, starting from state i, the probability of returning to state i after some finite length of time is one. A nonrecurrent state is said to be *transient*. We will prove a theorem that relates the recurrence or nonrecurrence of a state to the behavior of the n-step transition probabilities P_{ii}^n. Before proving the theorem, we need the following:

Lemma 5.1. (Abel).

(a) *If* $\sum_{k=0}^{\infty} a_k$ *converges, then*

$$\lim_{s \to 1-} \sum_{k=0}^{\infty} a_k s^k = \sum_{k=0}^{\infty} a_k = a \tag{5.11}$$

($\lim_{s \to 1-}$ means that s approaches 1 from values less than 1).

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

$$\sum_{k=0}^{\infty} a_k = \lim_{N \to \infty} \sum_{k=0}^{N} a_k = a.$$

Proof. (a) We will show that

$$\lim_{s \to 1-} \left| \sum_{k=0}^{\infty} a_k (s^k - 1) \right| = 0. \tag{5.12}$$

Since $\sum_{k=0}^{\infty} a_k$ converges, for any $\varepsilon > 0$ we can find an $N(\varepsilon)$ such that $|\sum_{k=N}^{N'} a_k| < \varepsilon/4$ for all $N' \geq N$. Choose such an N. Then write

$$\left| \sum_{k=0}^{\infty} a_k (s^k - 1) \right| = \left| \sum_{k=0}^{N} a_k (s^k - 1) + \sum_{k=N+1}^{\infty} a_k (s^k - 1) \right| \tag{5.13}$$

$$\leq \left| \sum_{k=0}^{N} a_k (s^k - 1) \right| + \left| \sum_{k=N+1}^{\infty} a_k (s^k - 1) \right|.$$

Now, for $0 \leq s < 1$

$$\left| \sum_{k=0}^{N} a_k(s^k - 1) \right| \leq MN|s^N - 1|, \tag{5.14}$$

where $M = \max_{0 \leq k \leq N} |a_k| < \infty$, so that for s sufficiently close to 1 we have

$$\left| \sum_{k=0}^{N} a_k(s^k - 1) \right| < \varepsilon/2.$$

To estimate $\sum_{k=N+1}^{\infty} a_k(s^k - 1)$ we sum by parts. This gives

$$\begin{aligned} \left| \sum_{k=N+1}^{\infty} a_k(s^k - 1) \right| &= \left| \sum_{k=N+1}^{\infty} (A_k - A_{k+1})(s^k - 1) \right| \\ &= \left| A_{N+1}(s^{N+1} - 1) + \sum_{k=N+2}^{\infty} A_k(s^k - s^{k-1}) \right|, \end{aligned} \tag{5.15}$$

where

$$A_k = \sum_{r=k}^{\infty} a_r.$$

Obviously, (5.15) is bounded by

$$\frac{\varepsilon}{4} |(s^{N+1} - 1)| + \frac{\varepsilon}{4} s^{N+1} \leq \frac{\varepsilon}{2}.$$

Putting these estimates together, we have

$$\left| \sum_{k=0}^{\infty} a_k(s^k - 1) \right| < \varepsilon,$$

provided s is sufficiently close to 1. ■

(b) Since $\sum_{k=0}^{\infty} a_k s^k \leq \sum_{k=0}^{\infty} a_k$ for $0 < s < 1$, the case $a = \infty$ is obvious. If $a < \infty$, then by our hypothesis

$$\sum_{k=0}^{\infty} a_k s^k < a < \infty \qquad \text{for} \quad 0 < s < 1.$$

Hence,

$$\sum_{k=0}^{n} a_k \leq a \qquad \text{for all} \quad n.$$

Since $\sum_{k=0}^{n} a_k$ is a bounded monotone increasing function of n it has a finite limit, call it a'. But by part (a) of this lemma we may conclude that, $a = a'$. ■

With this lemma we easily prove

Theorem 5.1. *A state i is recurrent if and only if*

$$\sum_{n=1}^{\infty} P_{ii}^{n} = \infty.$$

Proof. Assume i is recurrent, that is, $\sum_{n=1}^{\infty} f_{ii}^{n} = 1$. Then by Lemma 5.1(a)

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

Thus from (5.8)

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

Appealing to Lemma 5.1(b), we have

$$\sum_{n=0}^{\infty} P_{ii}^{n} = \infty,$$

the desired result. To prove sufficiency, assume that state i is transient, that is, $\sum_{n=1}^{\infty} f_{ii}^{n} < 1$. Using Lemma 5.1(a) and consulting (5.8), we infer that

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

Now appealing to Lemma 5.1(b), we have the result that $\sum_{n=1}^{\infty} P_{ii}^{n} < \infty$, contradicting our hypothesis and proving sufficiency. ■

As an immediate consequence of Theorem 5.1 we obtain

Corollary 5.1. *If $i \leftrightarrow j$ and if i is recurrent then j is recurrent.*

Proof. Since $i \leftrightarrow j$ there exists m, $n \geq 1$ such that

$$P_{ij}^{n} > 0 \qquad \text{and} \qquad P_{ji}^{m} > 0.$$

Let $v > 0$. We obtain, by the usual argument (see page 60), $P_{jj}^{m+n+v} \geq P_{ji}^{m} P_{ii}^{v} P_{ij}^{n}$ and, on summing,

$$\sum_{v=0}^{\infty} P_{jj}^{m+n+v} \geq \sum_{v=0}^{\infty} P_{ji}^{m} P_{ii}^{v} P_{ij}^{n} = P_{ji}^{m} P_{ij}^{n} \sum_{v=0}^{\infty} P_{ii}^{v}.$$

Hence if $\sum_{v=0}^{\infty} P_{ii}^{v}$ diverges, then $\sum_{v=0}^{\infty} P_{jj}^{v}$ also diverges. ■

This corollary proves that recurrence, like periodicity, is a class property: that is, all states in an equivalence class are either recurrent or nonrecurrent.

Remark. The expected number of returns to state i, given $X_0 = i$, is

$\sum_{n=1}^{\infty} P_{ii}^n$. Thus, Theorem 5.1 states that a state i is recurrent if and only if the expected number of returns is infinite.

6: *Examples of Recurrent Markov Chains*

Example 1. Consider first the one-dimensional random walk on the positive and negative integers, where at each transition the particle moves with probability p one unit to the right and with probability q one unit to the left ($p + q = 1$). Hence

$$P_{00}^{2n+1} = 0, \qquad n = 0, 1, 2, \ldots, \qquad \text{and} \qquad P_{00}^{2n} = \binom{2n}{n} p^n q^n = \frac{(2n)!}{n!n!} p^n q^n. \tag{6.1}$$

We appeal now to Stirling's formula,

$$n! \sim n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi}. \tag{6.2}$$

Applying (6.2) to (6.1) we obtain

$$P_{00}^{2n} \sim \frac{(pq)^n 2^{2n}}{\sqrt{\pi n}} = \frac{(4pq)^n}{\sqrt{\pi n}}.$$

It is readily verified that $p(1-p) = pq \leq \frac{1}{4}$ with equality holding if and only if $p = q = \frac{1}{2}$. Hence $\sum_{n=0}^{\infty} P_{00}^n = \infty$ if and only if $p = \frac{1}{2}$. Therefore, from Theorem 5.1, the one-dimensional random walk is recurrent if and only if $p = q = \frac{1}{2}$. Remember that recurrence is a class property. Intuitively, if $p \neq q$ there is positive probability that a particle initially at the origin will drift to $+\infty$ if $p > q$ (to $-\infty$ if $p < q$) without ever returning to the origin.

Example 2. We look now at the two-dimensional random walk in the full infinite plane. Let the probabilities of a transition one unit to the right, left, up, or down all be equal to $\frac{1}{4}$. We proceed to investigate recurrence of the state represented by the origin. Consider all paths whereby we move i units to the right, i units to the left, j units down, and j units up, where $2i + 2j = 2n$. We compute that

$$\begin{aligned} P_{00}^{2n+1} &= 0, && n = 0, 1, 2, \ldots, \\ P_{00}^{2n} &= \sum_{i,j,i+j=n} \frac{(2n)!}{i!i!j!j!} \left(\frac{1}{4}\right)^{2n}, && n = 1, 2, 3, \ldots. \end{aligned} \tag{6.3}$$

[The terms of (6.3) result from applying the multinomial distribution.] Multiplying numerator and denominator of (6.3) by $(n!)^2$ we obtain

$$P_{00}^{2n} = \left(\frac{1}{4}\right)^{2n} \binom{2n}{n} \sum_{i=0}^{n} \binom{n}{i}\binom{n}{n-i}.$$

But

$$\sum_{i=0}^{n} \binom{n}{i}\binom{n}{n-i} = \binom{2n}{n}$$

Hence,

$$P_{00}^{2n} = \left(\frac{1}{4}\right)^{2n}\binom{2n}{n}^2.$$

Using Stirling's formula (6.2) again we obtain

$$P_{00}^{2n} \sim \frac{1}{\pi n}.$$

Hence, $\sum_{n=0}^{\infty} P_{00}^{n} = \infty$ and the state represented by the origin is again a recurrent state.

Example 3. We consider the symmetric random walk in three dimensions. By reasoning similar to the above we obtain

$$P_{00}^{2n+1} = 0, \qquad n = 0, 1, 2, \ldots,$$

$$P_{00}^{2n} = \sum_{i,j,0\le i+j\le n} \frac{(2n)!}{i!i!j!j!(n-i-j)!\,(n-i-j)!}\left(\frac{1}{6}\right)^{2n}. \tag{6.4}$$

Multiplying numerator and denominator by $(n!)^2$ and factoring out a term $(\frac{1}{2})^{2n}$ gives

$$P_{00}^{2n} = \frac{1}{2^{2n}}\binom{2n}{n}\sum_{i,j,0\le i+j\le n}\left[\frac{n!}{i!j!(n-i-j)!}\right]^2\left(\frac{1}{3}\right)^{2n}, \tag{6.5}$$

$$P_{00}^{2n} \le c_n \frac{1}{2^{2n}}\binom{2n}{n}\frac{1}{3^n}, \tag{6.6}$$

where

$$c_n = \max_{i,j,0\le i+j\le n}\left[\frac{n!}{i!j!(n-i-j)!}\right]. \tag{6.7}$$

Observe that we have used the fact that

$$\sum_{i,j,0\le i+j\le n} \frac{n!}{i!j!(n-i-j)!}\left(\frac{1}{3}\right)^n = 1. \tag{6.8}$$

For large n the value of c_n is attained for $i = j \sim n/3$. We show this result as follows. Let i_0 and j_0 be the values of i and j which maximize the terms

$$\frac{n!}{i!j!(n-i-j)!} \qquad \text{where } 0 \le i+j \le n.$$

We may immediately write the four inequalities

$$\frac{n!}{j_0!(i_0-1)!(n-j_0-i_0+1)!} \leq \frac{n!}{j_0!i_0!(n-j_0-i_0)!},$$

$$\frac{n!}{j_0!(i_0+1)!(n-j_0-i_0-1)!} \leq \frac{n!}{j_0!i_0!(n-j_0-i_0)!},$$

$$\frac{n!}{(j_0-1)!i_0!(n-j_0-i_0+1)!} \leq \frac{n!}{j_0!i_0!(n-j_0-i_0)!},$$

$$\frac{n!}{(j_0+1)!i_0!(n-j_0-i_0-1)!} \leq \frac{n!}{j_0!i_0!(n-j_0-i_0)!}.$$

These inequalities reduce to

$$n-i_0-1 \leq 2j_0 \leq n-i_0+1,$$

$$n-j_0-1 \leq 2i_0 \leq n-j_0+1.$$

Hence for large n, $i_0 \sim n/3$ and $j_0 \sim n/3$. Inserting $i=j=n/3$ in (6.7), we obtain for (6.6)

$$P_{00}^{2n} \leq \frac{n!}{(n/3)!(n/3)!(n/3)!(2)^{2n}3^n}\binom{2n}{n}. \tag{6.9}$$

If we use Stirling's formula the right-hand side of the inequality (6.9) is asymptotic to

$$\frac{3\sqrt{3}}{2\pi^{3/2}\,n^{3/2}}.$$

But if we sum these terms, we obtain

$$\sum_{n=1}^{\infty} \frac{3\sqrt{3}}{2\pi^{3/2}\,n^{3/2}} < \infty.$$

Hence $\sum_{n=1}^{\infty} P_{00}^n < \infty$ and by Theorem 5.1 the state represented by 0 is a transient state. Now, since recurrence is a class property and all states communicate we see that for the one- and two-dimensional symmetric random walks the particle will return with certainty to any state that it once occupied. However, in the three-dimensional symmetric random walk there is positive probability that once the particle leaves a state it never returns.

Example 4. Consider now the Markov chain which represents the success runs of binomial trials. The transition probability matrix is

$$\begin{Vmatrix} p_0 & 1-p_0 & 0 & 0 & \cdots & \\ p_1 & 0 & 1-p_1 & 0 & \cdots & \\ p_2 & 0 & 0 & 1-p_2 & \cdots & \\ \vdots & \vdots & & & & \\ p_r & 0 & \cdots & & 1-p_r & 0\cdots \\ \vdots & & & & \vdots & \vdots \end{Vmatrix} \qquad (0<p_i<1).$$

The states of this Markov chains all belong to the same equivalence class (any state can be reached from any other state). Since recurrence is a class property (see Corollary 5.1), we will investigate recurrence for the zeroth state. We compute that

$$\begin{aligned} f_{00}^1 &= p_0 = 1-(1-p_0), \\ f_{00}^n &= \left(\prod_{i=0}^{n-2}(1-p_i)\right)p_{n-1} \qquad \text{for} \quad n>1. \end{aligned} \tag{6.10}$$

Rewriting (6.10) we obtain

$$f_{00}^n = \prod_{i=0}^{n-2}(1-p_i)[1-(1-p_{n-1})], \qquad n>1,$$

$$f_{00}^n = (1-p_0)(1-p_1)\cdots(1-p_{n-2}) - (1-p_0)(1-p_1)\cdots(1-p_{n-1}), \qquad n>1.$$

Let

$$u_n = \begin{cases} (1-p_0)(1-p_1)\cdots(1-p_n), & n\geq 0, \\ 1, & n=-1. \end{cases}$$

Then if we sum the f_{00}^n's we have

$$\sum_{n=1}^{m+1} f_{00}^n = \sum_{n=1}^{m+1}(u_{n-2}-u_{n-1}) = (1-u_0)+(u_0-u_1)+\cdots+(u_{m-1}-u_m)$$

or

$$\sum_{n=1}^{m+1} f_{00}^n = 1-u_m. \tag{6.11}$$

To complete our argument we need the following:

Lemma 6.1. *If* $0<p_i<1$, $i=0,1,2,\ldots$, *then* $u_m=\prod_{i=0}^{m}(1-p_i)\to 0$ *as* $m\to\infty$ *if and only if* $\sum_{i=0}^{\infty}p_i=\infty$.

Proof. Assume $\sum_{i=0}^{\infty}p_i=\infty$. Since the series expansion for $\exp(-p_i)$ is an alternating series with terms decreasing in absolute value, we can

write

$$1 - p_i < 1 - p_i + \frac{p_i^2}{2!} - \frac{p_i^3}{3!} + \cdots = \exp(-p_i), \qquad i = 0, 1, 2, \ldots . \tag{6.12}$$

Since (6.12) holds for all i, we obtain $\prod_{i=0}^{m}(1-p_i) < \exp(-\sum_{i=0}^{m} p_i)$. But, by assumption,

$$\lim_{m\to\infty} \sum_{i=0}^{m} p_i = \infty; \qquad \text{hence} \quad \lim_{m\to\infty} \prod_{i=0}^{m} (1-p_i) = 0.$$

To prove necessity observe that from a straightforward induction

$$\prod_{i=j}^{m} (1-p_i) > (1 - p_j - p_{j+1} - \cdots - p_m)$$

for any j and all $m = j+1, j+2, \ldots$. Assume now that $\sum_{i=1}^{\infty} p_i < \infty$; then $0 < \sum_{i=j}^{\infty} p_i < 1$ for some $j > 1$. Thus

$$\lim_{m\to\infty} \prod_{i=j}^{m} (1-p_i) > \lim_{m\to\infty} (1 - \sum_{i=j}^{m} p_i) > 0,$$

which contradicts $u_m \to 0$. ■

Returning to (6.11) and applying Lemma 6.1, we deduce that $\sum_{n=1}^{\infty} f_{00}^n = 1$ if and only if $\sum_{i=0}^{\infty} p_i = \infty$, or state 0 is recurrent if and only if the sum of the p_i's diverges.

We insert parenthetically the remark that, given any set $\{a_1, a_2, \ldots\}$, such that $a_i > 0$ and $\sum_{i=1}^{\infty} a_i \leq 1$, we can exhibit a set of p_i's for which $f_{00}^n = a_n$ in the Markov chain discussed above. We let

$$f_{00}^1 = p_0 = a_1,$$

$$f_{00}^2 = (1 - p_0)p_1 = a_2,$$

and then determine that

$$p_1 = \frac{a_2}{1 - a_1}.$$

Set

$$f_{00}^3 = (1 - p_0)(1 - p_1)p_2 = a_3,$$

and this implies that

$$p_2 = \frac{a_3}{1 - a_1 - a_2}.$$

Proceeding in this manner we can derive an explicit set of p_i's satisfying $0 \leq p_i < 1$.

7: *More on Recurrence*

The next theorem shows that if recurrence is certain for a specified state, then the state will be occupied infinitely often with probability 1. We define

$$Q_{ii} = \Pr\left\{\begin{matrix}\text{a particle starting in state } i \text{ re-}\\ \text{turns infinitely often to state } i\end{matrix}\right\}$$

Theorem 7.1. *State i is recurrent or transient according to whether $Q_{ii} = 1$ or 0, respectively.*

Proof. Let Q_{ii}^N be defined as

$$Q_{ii}^N = \Pr\left\{\begin{matrix}\text{a particle starting in state } i \text{ re-}\\ \text{turns to state } i \text{ at least } N \text{ times}\end{matrix}\right\}$$

We can write

$$Q_{ii}^N = \sum_{k=1}^{\infty} f_{ii}^k Q_{ii}^{N-1} = Q_{ii}^{N-1} f_{ii}^*, \qquad \text{where} \quad f_{ii}^* = \sum_{k=1}^{\infty} f_{ii}^k.$$

The validity of this formula rests on decomposing the event of Q_{ii}^N according to the first return time. Proceeding recursively, we obtain

$$Q_{ii}^N = f_{ii}^* Q_{ii}^{N-1} = (f_{ii}^*)^2 Q_{ii}^{N-2} = \ldots = [f_{ii}^*]^{N-1} Q_{ii}^1.$$

But $Q_{ii}^1 = f_{ii}^*$ by definition. Hence

$$Q_{ii}^N = [f_{ii}^*]^N.$$

Since $\lim_{N\to\infty} Q_{ii}^N = Q_{ii}$, we have $Q_{ii} = 1$ or 0 according to whether $f_{ii}^* = 1$ or < 1, respectively, or equivalently, according to whether state i is recurrent or transient.

Theorem 7.2. *If $i \leftrightarrow j$ and the class is recurrent, then*

$$f_{ij}^* = \sum_{n=1}^{\infty} f_{ij}^n = 1.$$

We omit the simple proof.

We define the symbol Q_{ij} to be

$$\Pr\{\text{particle starting in state } i \text{ visits state } j \text{ infinitely often}\}.$$

An immediate consequence of Theorem 7.2 is

Corollary 7.1. *If $i \leftrightarrow j$ and the class is recurrent, then $Q_{ij} = 1$.*

Proof. It is easy to see that

$$Q_{ij} = f_{ij}^* Q_{jj}.$$

Since j is a recurrent state, by Theorem 7.1, $Q_{jj} = 1$. By Theorem 7.2, $f_{ij}^* = 1$, hence $Q_{ij} = 1$. ■

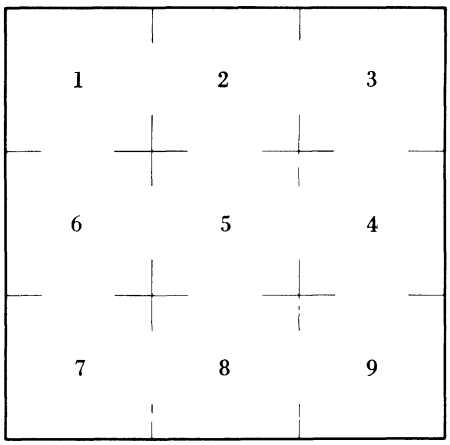
Elementary Problems

1. Determine the transition matrix $\|P_{jk}\|$ for the following Markov chains:

(a) Consider a sequence of tosses of a coin with the probability of "heads" p. At time n (after n tosses of the coin) the state of the process is the number of heads in the n tosses minus the number of tails.

(b) N black balls and N white balls are placed in two urns so that each urn contains N balls. At each step one ball is selected at random from each urn and the two balls interchange. The state of the system is the number of white balls in the first urn.

(c) A white rat is put into the maze shown. The rat moves through the compartments at random, i.e., if there are k ways to leave a compartment he



chooses each of these with probability $1/k$. He makes one change of compartment at each instant of time. The state of the system is the number of the compartment the rat is in.

Solutions:

(a) $$P_{jk} = \Pr\left\{\begin{matrix}\text{number of heads minus number of tails} \\ = k \text{ after } n+1 \text{ tosses} \end{matrix} \middle| \begin{matrix}\text{number of heads} \\ \text{minus number of tails} = j \text{ after } n \text{ tosses}\end{matrix}\right\}$$

$$= \begin{cases} p & \text{if } k = j+1, \\ 1-p & \text{if } k = j-1, \\ 0 & \text{otherwise,} \end{cases}$$

independent of n.

(b) $P_{jk} = \Pr\left\{\begin{matrix} k \text{ white balls in first urn after } n+1 \text{ interchanges} \\ \text{white balls in first urn after } n \text{ interchanges} \end{matrix} \middle| j \right\}$

$$= \begin{cases} \left(\frac{j}{N}\right)^2 & \text{if } k = j-1, \quad j = 1, 2, \ldots, N, \\ 2\left(\frac{j}{N}\right)\left(\frac{N-j}{N}\right) & \text{if } k = j, \quad j = 0, 1, \ldots, N, \\ \left(1 - \frac{j}{N}\right)^2 & \text{if } k = j+1, \quad j = 0, 1, \ldots, N-1, \\ 0 & \text{otherwise,} \end{cases}$$

independent of n.

2. (a) Consider two urns A and B containing a total of N balls. An experiment is performed in which a ball is selected at random (all selections equally likely) at time t ($t = 1, 2, \ldots$) from among the totality of N balls. Then an urn is selected at random (A is chosen with probability p and B is chosen with probability q) and the ball previously drawn is placed in this urn. The state of the system at each trial is represented by the number of balls in A. Determine the transition matrix for this Markov chain.

(b) Assume that at time t there are exactly k balls in A. At time $t+1$ an urn is selected at random in proportion to its contents (i.e., A is chosen with probability k/N and B is chosen with probability $(N-k)/N$). Then a ball is selected from A with probability p or from B with probability q and placed in the previously chosen urn. Determine the transition matrix for this Markov chain.

(c) Now assume that at time $t+1$ a ball and an urn are chosen with probability depending on the contents of the urn (i.e., a ball is chosen from A with probability k/N or from B with probability $(N-k)/N$. Urn A is chosen with probability k/N or urn B is chosen with probability $(N-k)/N$). Determine the transition matrix of the Markov chain with states represented by the contents of A.

(d) Determine the equivalence classes in (a), (b), and (c).

Solution:

(a) $$P_{ik} = \begin{cases} \frac{N-i}{N}p & \text{if } k = i+1, \\ \frac{i}{N}p + \frac{N-i}{N}q & \text{if } k = i, \quad i = 0, 1, 2, \ldots, N. \\ \frac{i}{N}q & \text{if } k = i-1, \\ 0 & \text{otherwise,} \end{cases}$$

One equivalence class: $\{0, 1, 2, \ldots, N\}$.

$$\text{(b) } P_{ik} = \begin{cases} \dfrac{i}{N}q & \text{if } k = i+1, \\ \dfrac{i}{N}p + \dfrac{N-i}{N}q & \text{if } k = i, \qquad i = 1, 2, \ldots, N-1. \\ \dfrac{N-i}{N}p & \text{if } k = i-1, \\ 0 & \text{otherwise,} \end{cases}$$

$P_{ii} = 1$ if $i = 0$ and $i = N$. Equivalence classes are $\{0\}, \{N\}, \{1, 2, \ldots, N-1\}$.

$$\text{(c) } P_{ik} = \begin{cases} \dfrac{i^2}{N^2} + \dfrac{(N-i)^2}{N^2} & \text{if } k = i, \\ \dfrac{i(N-i)}{N^2} & \text{if } k = i+1 \text{ or } k = i-1, \\ 0 & \text{otherwise.} \end{cases}$$

Equivalence classes are $\{0\}, \{1, 2, \ldots, N-1\}, \{N\}$.

3. (a) A psychological subject can make one of two responses A_1 and A_2. Associated with these responses are a set of N stimuli $\{S_1, S_2, \ldots, S_N\}$. Each stimulus is conditioned to one of the responses. A single stimulus is sampled at random (all possibilities equally likely) and the subject responds according to the stimulus sampled. Reinforcement occurs at each trial with probability π $(0 < \pi < 1)$ independent of the previous history of the process. When reinforcement occurs, the stimulus sampled does not alter its conditioning state. In the contrary event the stimulus becomes conditioned to the other response. Consider the Markov chain whose state variable is the number of stimuli conditioned to response A_1. Determine the transition probability matrix of this M.C.

(b) A subject S can make one of three responses A_0, A_1, and A_2. The A_0 response corresponds to a guessing state. If S makes response A_1, the experiment reinforces the subject with probability π_1 and at the next trial S will make the same response. If no reinforcement occurs (probability $1 - \pi_1$), then at the next trial S passes to the guessing state. Similarly π_2 is the probability of reinforcement for response A_2. Again the subject remains in this state if reinforced and otherwise passes to the guessing state. When S is in the guessing state, he stays there for the next trial with probability $1 - c$ and with probabilities $c/2$ and $c/2$ makes responses A_1 and A_2 respectively. Consider the Markov chain of the state of the subject and determine its transition probability matrix.

Solutions: (a) $P_{ii} = \pi$; $P_{i,i+1} = ((N-i)/N)(1-\pi)$, $P_{i,i-1} = (i/N)(1-\pi)$, $P_{ij} = 0$ otherwise $(i, j = 1, 2, \ldots, N)$.

(b) $P_{00} = 1 - c$, $P_{0,1} = P_{0,2} = c/2$; $P_{10} = 1 - \pi_1$, $P_{11} = \pi_1$; $P_{20} = 1 - \pi_2$, $P_{22} = \pi_2$.

4. Determine the classes and the periodicity of the various states for a Markov chain with transition probability matrix

$$\text{(a)} \begin{Vmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{Vmatrix}, \qquad \text{(b)} \begin{Vmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \end{Vmatrix}.$$

5. Consider repeated independent trials of two outcomes S (success) or F (failure) with probabilities p and q, respectively. Determine the distribution of the number of trials required for the first occurrence of the event SF (i.e., success followed by failure). Do the same for the event SSF and SFS.

6. The following sequential approach to estimating the size of a finite population, perhaps a wildlife population such as fish, is proposed. A member of the population is sampled at random, tagged and returned. Another is sampled, tagged and returned and so on, until a member is drawn that has been drawn before. When this occurs, say at trial T, we stop, (and possibly begin anew with a new kind of tag.) Based on the observed value of T we want to estimate the population size N.

Let X_n be the number of most recent successive untagged members observed without seeing a tagged one. Then $X_n = n$ for $n = 0, \ldots, T-1$ but $X_T = 0$, so that T is the first passage time $T = \min\{n \geq 1 \colon X_n = 0\}$.

(a) For any fixed N, we claim that (X_n) is a success runs Markov chain. Specify the probabilities p_n and q_n.

(b) Compute $\Pr[T = t | X_0 = 0]$ for $t = 2, \ldots, N$. (See Elementary Problem 2 of Chapter 1.)

7. A component of a computer has an active life, measured in discrete units, that is a random variable T where $\Pr[T = k] = a_k$ for $k = 1, 2, \ldots$. Suppose one starts with a fresh component and each component is replaced by a new component upon failure. Let X_n be the age of the component in service at time n. Then (X_n) is a success runs Markov chain.

(a) Compute the probabilities p_i and q_i.

A "planned replacement" policy specifies replacing the component upon failure, or at time N, whichever occurs first. Then the time to replacement is $T^* = \min\{T, N\}$ where $T = \min\{n \geq 1 \colon X_n = 0\}$.

(b) Compute $E[T^*]$ (See Elementary Problems 1, 2, and 3 of Chapter 1.)

8. Unknown to public health officials, a person with a highly contagious disease enters the population. During each period he either infects a new person which occurs with probability p, or his symptoms appear and he is discovered by public health officials, which occurs with probability $1 - p$. Compute the probability distribution of the number of infected but undiscovered people in the population at the time of first discovery of a carrier. Assume each infective behaves like the first.

Problems

1. Every stochastic $n \times n$ matrix corresponds to a Markov chain for which it is the one-step transition matrix. (By "Stochastic matrix" we mean $\mathbf{P} = \|P_{ij}\|$ with $0 \leq P_{ij} \leq 1$ and $\sum_j P_{ij} = 1$.) However, not every stochastic $n \times n$ matrix is the two-step transition matrix of a Markov chain. In particular, show that a 2×2 stochastic matrix is the two-step transition matrix of a Markov chain if and only if the sum of its principal diagonal terms is greater than or equal to 1.

2. Let $n_1, n_2, \ldots, n_k$ be positive integers with greatest common divisor d. Show that there exists a positive integer M such that $m \geq M$ implies there exist nonnegative integers $\{c_j\}_{j=1}^k$ such that

$$md = \sum_{j=1}^{k} c_j n_j.$$

(This result is needed for Problem 4 below.)

Hint: Let $A = \{n | n = c_1 n_1 + \cdots + c_k n_k, \{c_i\}$ nonnegative integers$\}$

Let $B = \left\{ b_1 n_1 + \cdots + b_j n_j | n_1, n_2, \ldots, n_j \in A, \text{ and } b_1, \ldots, b_j \text{ are positive or negative integers} \right\}$.

Let d' be the smallest positive integer in B and prove that d' is a common divisor of all integers in A. Then show that d' is the greatest common divisor of all integers in A. Hence $d' = d$. Rearrange the terms in the representation $d = a_1 n_1 + \cdots + a_l n_l$ so that the terms with positive coefficients are written first. Thus $d = N_1 - N_2$ with $N_1 \in A$ and $N_2 \in A$. Let M be the positive integer, $M = N_2^2/d$. Every integer $m \geq M$ can be written as $m = M + k = N_2^2/d + k$, $(k = 0, 1, 2, \ldots)$, and $k = \delta N_2/d + b$ where $0 \leq b < N_2/d$ and $\delta = j$ when $j(N_2/d) \leq k < (j+1)N_2/d, j = 0, 1, 2, \ldots,$ so $md = N_2^2 + (\delta N_2/d + b)d = N_2(N_2 + \delta - b) + bN_1$.

3. Prove Theorem 4.1.

Hint: Let $P_{ii}^s > 0$, $P_{jj}^{n+s+m} \geq P_{ji}^n P_{ii}^s P_{ij}^m > 0$ for some $m > 0$ and $n > 0$. Since also $P_{ii}^{2s} > 0$, we have $P_{jj}^{n+2s+m} > 0$. Thus $d(j)$ divides $(n + 2s + m) - (n + s + m) = s$.

4. Prove Theorem 4.2 and Corollary 4.1.

Hint: By Problem 2 there exists N such that if $n \geq N$

$$P_{ii}^{nd(i)} = P_{ii}^{(c_1 n_1 + \cdots c_k n_k)}.$$

5. Given a finite aperiodic irreducible Markov chain, prove that for some n all terms of P^n are positive.

6. If j is a transient state prove that for all i

$$\sum_{n=1}^{\infty} P_{ij}^n < \infty.$$

Hint: Use relation (5.10).

7. Let a Markov chain contain r states. Prove the following:

(a) If a state k can be reached from j, then it can be reached in $r-1$ steps or less.

(b) If j is a recurrent state, there exists $\alpha\ (0<\alpha<1)$ such that for $n>r$ the probability that first return to state j occurs after n transitions is $\leq \alpha^n$.

8. Consider a sequence of Bernoulli trials $X_1, X_2, X_3, \ldots$, where $X_n = 1$ or 0.

Assume

$$\Pr\{X_n = 1 | X_1, X_2, \ldots, X_{n-1}\} \geq \alpha > 0, \qquad n = 1, 2, \ldots.$$

Prove that

(a) $\Pr\{X_n = 1 \text{ for some } n\} = 1$,

(b) $\Pr\{X_n = 1 \text{ infinitely often}\} = 1$.

9. Let

$$\mathbf{P} = \begin{Vmatrix} 1-a & a \\ b & 1-b \end{Vmatrix}, \qquad 0<a,\ b<1.$$

Prove

$$\mathbf{P}^n = \frac{1}{a+b}\begin{Vmatrix} b & a \\ b & a \end{Vmatrix} + \frac{(1-a-b)^n}{a+b}\begin{Vmatrix} a & -a \\ -b & b \end{Vmatrix}.$$

10. Consider a random walk on the integers such that $P_{i,i+1} = p$, $P_{i,i-1} = q$ for all integer $i\ (0<p<1, p+q=1)$. Determine P_{00}^n.

Answer: $$P_{00}^{2m} = \binom{2m}{m} p^m q^m, \qquad P_{00}^{2m+1} = 0.$$

11. (Continuation) Find the generating function of $u_n = P_{00}^n$, i.e., determine

$$P(x) = \sum_{n=0}^{\infty} u_n x^n.$$

Hint: Use the identity $\binom{2n}{n} = (-1)^n \binom{-\frac{1}{2}}{n} 2^{2n}$ where we define

$$\binom{a}{n} = \frac{a(a-1)\cdots(a-n+1)}{n!} \qquad \text{for any real } a.$$

Answer: $$P(x) = (1-4pqx^2)^{-1/2}.$$

12. (Continuation) Determine the generating function of the recurrence time from state 0 to state 0.

Answer: $$F(x) = 1 - \sqrt{(1-4pqx^2)}.$$

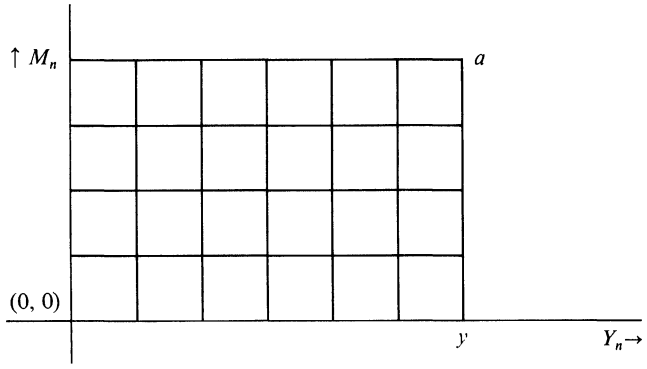
13. (Continuation) What is the probability of eventual return to the origin?

14. Suppose 2 distinguishable fair coins are tossed simultaneously and repeatedly. An account of the tallies of heads and tails are recorded. Consider the event E_n that at the nth toss the cumulative number of heads on both tallies are equal. Relate the event E_n to the recurrence time of a given state for a symmetric random walk on the integers.

15. Suppose $X_1, X_2, \ldots$ are independent with $\Pr\{X_k = +1\} = p$, $\Pr\{X_k = -1\} = q = 1 - p$ where $p \geq q$. With $S_0 = 0$, set $S_n = X_1 + \cdots + X_n$, $M_n = \max\{S_k\colon 0 \leq k \leq n\}$ and $Y_n = M_n - S_n$. If $T(a) = \min\{n\colon S_n = a\}$, show

$$\Pr\{\max_{0 \leq k \leq T(a)} Y_k < y\} = \begin{cases} \left(\dfrac{y}{1+y}\right)^a & \text{if } p = q = \dfrac{1}{2} \\ \left[\dfrac{\dfrac{p}{q} - \left(\dfrac{p}{q}\right)^{y+1}}{1 - \left(\dfrac{p}{q}\right)^{y+1}}\right]^a & \text{if } p \neq q. \end{cases}$$

Hint: The bivariate process (M_n, Y_n) is a random walk on the positive lattice. What is the probability that this random walk leaves the rectangle



at the top?

16. (Continuation). Adding to the notation of Problem 15, let τ be the first time the partial sums S_n deviate y units from their maximum to date. That is, let $\tau = \min\{n\colon Y_n = y\}$. Show that $M_\tau = S_\tau + y$ has a geometric distribution $\Pr\{M_\tau \geq a\} = \theta^a$ for $a = 0, 1, \ldots$ and determine θ.

Hint: $M_\tau \geq a$ if and only if $\max_{0 \leq k \leq T(a)} Y_k < y$.

NOTES

Some aspects of the theory of Markov chains can be found in the last half of Feller [1].

The book by Kemeny and Snell [2] contains several enticing examples of Markov chains drawn from psychology, sociology, economics, biology, and elsewhere.

The most advanced treatment devoted to the analysis of the structure of Markov chains is that of Chung [3].

REFERENCES

1. W. Feller, "An Introduction to Probability Theory and Its Applications," Vol. 1, 2nd ed. Wiley, New York, 1957.
2. J. G. Kemeny and J. L. Snell, "Finite Markov Chains." Van Nostrand, Princeton, New Jersey, 1960.
3. K. L. Chung, "Markov Chains with Stationary Transition Probabilities." Springer-Verlag, Berlin, 1960.

Chapter 3

THE BASIC LIMIT THEOREM OF MARKOV CHAINS AND APPLICATIONS

The content of Sections 1–4 of this chapter is part of the standard apparatus of Markov chains that should be covered in every introductory course. However, the reader may wish to defer Section 2, a proof of the discrete renewal theorem, until Chapter 5 where renewal theory is covered in full generality.

The examples of Sections 5 and 6 are classical in the area of stochastic queueing models. Perhaps one of these sections might be skipped on first reading.

1: Discrete Renewal Equation

A key tool in the analysis of Markov chains is furnished by the following theorem.

Theorem 1.1. *Let* $\{a_k\}, \{u_k\}, \{b_k\}$ *be sequences indexed by* $k = 0, \pm 1, \pm 2, \ldots$. *Suppose that* $a_k \geq 0$, $\sum a_k = 1$, $\sum |k|\, a_k < \infty$, $\sum k\, a_k > 0$, $\sum |b_k| < \infty$, *and that the greatest common divisor of the integers* k *for which* $a_k > 0$ *is* 1. *If the renewal equation*

$$u_n - \sum_{k=-\infty}^{\infty} a_{n-k} u_k = b_n \qquad \text{for} \quad n = 0, \pm 1, \pm 2, \ldots$$

is satisfied by a bounded sequence $\{u_n\}$ *of real numbers, then* $\lim_{n\to\infty} u_n$ *and* $\lim_{n\to-\infty} u_n$ *exist. Furthermore, if*

$$\lim_{n\to-\infty} u_n = 0, \qquad \text{then} \qquad \lim_{n\to\infty} u_n = \frac{\sum_{k=-\infty}^{\infty} b_k}{\sum_{k=-\infty}^{\infty} k a_k} \tag{1.1}$$

In case $\sum_{k=-\infty}^{\infty} k a_k = \infty$, *the limit relations are still valid provided we interpret*

$$\frac{\sum_{k=-\infty}^{\infty} b_k}{\sum_{k=-\infty}^{\infty} k a_k} = 0.$$

The proof of this theorem in its general form as stated is beyond the scope of this book. Actually we will make use of this theorem only for the case where $\{a_k\}$, $\{u_k\}$, $\{b_k\}$ vanish for negative values of k, and $b_k \geq 0$. A proof of the theorem for this case is given in Section 2 below.

Remark 1.1. In the case where $a_{-k} = 0$, $b_{-k} = 0$, and $u_{-k} = 0$, for $k > 0$ the renewal equation becomes

$$u_n - \sum_{k=0}^{n} a_{n-k} u_k = b_n \qquad \text{for} \quad n = 0, 1, 2, \ldots.$$

Remark 1.2. (Reason for the term "renewal equation.") Consider a light bulb whose lifetime, measured in discrete units, is a random variable ξ, where

$$\Pr\{\xi = k\} = a_k \qquad \text{for} \quad k = 0, 1, 2, \ldots, \quad a_k > 0, \quad \sum_{k=0}^{\infty} a_k = 1.$$

Let each bulb be replaced by a new one when it burns out. Suppose the first bulb lasts until time ξ_1, the second bulb until time $\xi_1 + \xi_2$, and the nth bulb until time $\sum_{i=1}^{n} \xi_i$, where the ξ_i are independent identically distributed random variables each distributed as ξ. Let u_n denote the expected number of renewals (replacements) up to time n. If the first replacement occurs at time k then the expected number of replacements in the remaining time up to n is u_{n-k}, and summing over all possible values for k, we obtain

$$\begin{aligned} u_n &= \sum_{k=0}^{n} (1 + u_{n-k})\, a_k + 0 \sum_{k=n+1}^{\infty} a_k \\ &= \sum_{k=0}^{n} u_{n-k}\, a_k + \sum_{k=0}^{n} a_k \\ &= \sum_{k=0}^{n} a_{n-k}\, u_k + b_n, \end{aligned} \tag{1.2}$$

where

$$\sum_{k=0}^{n} a_k = b_n.$$

The reasoning behind (1.2) goes as follows. The term $1 + u_{n-k}$ is the expected number of replacements in time n if the first bulb fails at time k $(0 \leq k \leq n)$, the probability of this event being a_k. The second sum is the probability that the first bulb lasts a duration exceeding n time units. Taking account of the regenerative nature of the process, we may clearly evaluate u_n by decomposing the possible realizations by the event of the time of the first replacement.

The following theorem, the ergodic theorem for this particular case, describes the limiting behavior of P_{ij}^n as $n \to \infty$ for all i and j in the case of an aperiodic recurrent Markov chain. The proof is a simple application of the basic renewal limit theorem.

Theorem 1.2. (The basic limit theorem of Markov chains.)

(a) *Consider a recurrent irreducible aperiodic Markov chain. Let* $P_{ii}^n =$ *the probability of entering state* i *at the* nth *transition,* $n = 0, 1, 2, \ldots$, *given that* $X(0) = i$ (*the initial state is* i). *By our earlier convention* $P_{ii}^0 = 1$. *Let* $f_{ii}^n =$ *probability of first returning to state* i *at the* nth *transition,* $n = 0, 1, 2, \ldots$, where $f_{ii}^0 = 0$. *Thus*

$$P_{ii}^n - \sum_{k=0}^{n} f_{ii}^{n-k} P_{ii}^k = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{if } n > 0. \end{cases}$$

[This is formula (5.1) of Chapter 2 derived earlier.] *Then*

$$\lim_{n \to \infty} P_{ii}^n = \frac{1}{\sum_{n=0}^{\infty} n f_{ii}^n}.$$

(b) *Under the same conditions as in* (a), $\lim_{n \to \infty} P_{ji}^n = \lim_{n \to \infty} P_{ii}^n$.

Proof. (a) Identify

$$u_n = P_{ii}^n, \quad n \geq 0; \qquad u_n = 0, \quad n < 0;$$
$$a_n = f_{ii}^n, \quad n \geq 0; \qquad a_n = 0, \quad n < 0;$$
$$b_n = \begin{cases} 1 & n = 0, \\ 0 & n \neq 0; \end{cases}$$

and then apply Theorem 1.1.

(b) We use the recursion relation

$$P_{ji}^n = \sum_{v=0}^{n} f_{ji}^v P_{ii}^{n-v} \qquad i \neq j, \quad n \geq 0$$

[cf. formula (5.9) of Chapter 2]. More generally, let

$$y_n = \sum_{k=0}^{n} a_{n-k} x_k,$$

where

$$a_m \geq 0, \qquad \sum_{m=0}^{\infty} a_m = 1, \qquad \lim_{k \to \infty} x_k = c.$$

Under these circumstances we prove that $\lim_{n \to \infty} y_n = c$. In fact,

$$y_n - c = \sum_{k=0}^{n} a_{n-k} x_k - c \sum_{m=0}^{\infty} a_m = \sum_{k=0}^{n} a_{n-k}(x_k - c) - c \sum_{m=n+1}^{\infty} a_m.$$

For $\varepsilon > 0$ prescribed we determine $K(\varepsilon)$ so that $|x_k - c| < \varepsilon/3$ for all $k \geq K(\varepsilon)$.

$$y_n - c = \sum_{k=0}^{K(\varepsilon)} a_{n-k}(x_k - c) + \sum_{k=K(\varepsilon)+1}^{n} a_{n-k}(x_k - c) - c \sum_{m=n+1}^{\infty} a_m$$

and so

$$|y_n - c| \leq M \sum_{k=0}^{K(\varepsilon)} a_{n-k} + \frac{\varepsilon}{3} \sum_{h=K(\varepsilon)+1}^{n} a_{n-h} + |c| \sum_{m=n+1}^{\infty} a_m,$$

where

$$M = \max_{k \geq 0} |x_k - c|.$$

We choose $N(\varepsilon)$ so that $|c| \sum_{m=n+1}^{\infty} a_m < \varepsilon/3$ and

$$\sum_{k=0}^{K(\varepsilon)} a_{n-k} \equiv \sum_{m=n-K(\varepsilon)}^{n} a_m < \frac{\varepsilon}{3M} \qquad \text{for} \quad n \geq N(\varepsilon).$$

Then

$$|y_n - c| \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \qquad \text{for} \quad n \geq N(\varepsilon).$$

Now, setting

$$y_n = P_{ji}^n, \qquad a_n = f_{ji}^n, \qquad x_n = P_{ii}^n,$$

we have the desired result.

Remark 1.3. Let C be a recurrent class. Then $P_{ij}^n = 0$ for $i \in C$, $j \notin C$, and every n. Hence, once in C it is not possible to leave C. It follows that the submatrix $\|P_{ij}\|$, $i, j \in C$, is a transition probability matrix and the associated Markov chain is irreducible and recurrent. The limit theorem, therefore, applies verbatim to any aperiodic recurrent class.

Remark 1.4. If $a_n \to a$ as $n \to \infty$, it can be proved by elementary methods that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} a_k = a. \tag{1.3}$$

Thus, if i is a member of a recurrent aperiodic class,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} P_{ii}^m = \frac{1}{\sum_{n=0}^{\infty} n f_{ii}^n} = \frac{1}{m_i}, \tag{1.4}$$

where m_i is the mean recurrence time.

If i is a member of a recurrent periodic class with period d, one can show (see Problem 4 of Chapter 2) that $P_{ii}^m = 0$ if m is not a multiple of d (i.e., if $m \neq nd$ for any n), and that

$$\lim_{n\to\infty} P_{ii}^{nd} = \frac{d}{m_i}.$$

These last two results are easily combined with (1.3) to show that (1.4) also holds in the periodic case.

If $\lim_{n\to\infty} P_{ii}^n = \pi_i > 0$ for one i in an aperiodic recurrent class, then $\pi_j > 0$ for all j in the class of i. (The proof of this fact follows the method of Corollary 5.1 of Chapter 2 and will be omitted.) In this case, we call the class *positive recurrent* or *strongly ergodic*. If each $\pi_i = 0$ and the class is recurrent we speak of the class as *null recurrent* or *weakly ergodic*.

Theorem 1.3. *In a positive recurrent aperiodic class with states* $j = 0, 1, 2, \ldots,$

$$\lim_{n\to\infty} P_{jj}^n = \pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, \qquad \sum_{i=0}^{\infty} \pi_i = 1$$

and the π*'s are uniquely determined by the set of equations*

$$\pi_i \geq 0, \qquad \sum_{i=0}^{\infty} \pi_i = 1, \qquad \text{and} \qquad \pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}. \tag{1.5}$$

Any set $(\pi_i)_{i=0}^{\infty}$ satisfying (1.5) is called a *stationary probability distribution* of the Markov chain. We will expand on this concept in Chapter 11.

Proof. For every n and M, $1 = \sum_{j=0}^{\infty} P_{ij}^n \geq \sum_{j=0}^{M} P_{ij}^n$. Letting $n \to \infty$, and using Theorem 1.2, we obtain $1 \geq \sum_{j=0}^{M} \pi_j$ for every M. Thus, $\sum_{j=0}^{\infty} \pi_j \leq 1$. Now $P_{ij}^{n+1} \geq \sum_{k=0}^{M} P_{ik}^n P_{kj}$; if we let $n \to \infty$, we obtain $\pi_j \geq \sum_{k=0}^{M} \pi_k P_{kj}$. Next, since the left-hand side is independent of M, $M \to \infty$ gives

$$\pi_j \geq \sum_{k=0}^{\infty} \pi_k P_{kj}. \tag{1.6}$$

Multiplying by P_{ji}, then summing on j and using (1.6), yields $\pi_j \geq \sum_{k=0}^{\infty} \pi_k P_{kj}^2$ and then generally $\pi_j \geq \sum_{k=0}^{\infty} \pi_k P_{kj}^n$ for any n. Suppose strict inequality holds for some j. Adding these inequalities with respect to j, we have

$$\sum_{j=0}^{\infty} \pi_j > \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \pi_k P_{kj}^n = \sum_{k=0}^{\infty} \pi_k \sum_{j=0}^{\infty} P_{kj}^n = \sum_{k=0}^{\infty} \pi_k,$$

a contradiction. Thus, $\pi_j = \sum_{k=0}^{\infty} \pi_k P_{kj}^n$ for all n. Letting $n \to \infty$, since $\sum \pi_k$ converges and P_{kj}^n is uniformly bounded, we conclude that

$$\pi_j = \sum_{k=0}^{\infty} \pi_k \lim_{n\to\infty} P_{kj}^n = \pi_j \sum_{k=0}^{\infty} \pi_k \qquad \text{for every } j.$$

Thus, $\sum_{k=0}^{\infty} \pi_k = 1$ since $\pi_j > 0$ by positive recurrence.

Suppose $x = \{x_n\}$ satisfies the relations (1.5). Then

$$x_k = \sum_{j=0}^{\infty} x_j P_{jk} = \sum_{j=0}^{\infty} x_j P_{jk}^n,$$

and if we let $n \to \infty$ as before,

$$x_k = \sum_{j=0}^{\infty} x_j \lim_{n\to\infty} P_{jk}^n = \pi_k \sum_{j=0}^{\infty} x_j = \pi_k. \quad \blacksquare$$

Example. Consider the class of random walks whose transition matrices are given by

$$\mathbf{P} = \|P_{ij}\| = \begin{Vmatrix} 0 & 1 & 0 & \cdots\cdots \\ q_1 & 0 & p_1 & \cdots\cdots \\ 0 & q_2 & 0 & p_2 \cdots \\ \vdots & & & \end{Vmatrix}$$

(c.f. Example B, Chapter 2). This Markov chain has period 2. Nevertheless we investigate the existence of a stationary probability distribution, i.e., we wish to determine the positive solutions of

$$x_i = \sum_{j=0}^{\infty} x_j P_{ji} = p_{i-1}\, x_{i-1} + q_{i+1} x_{i+1}, \qquad i = 0, 1, \ldots, \tag{1.7}$$

under the normalization

$$\sum_{i=0}^{\infty} x_i = 1,$$

where $p_{-1} = 0$ and $p_0 = 1$, and thus $x_0 = q_1 x_1$. Using Equation (1.7) for $i = 1$, we could determine x_2 in terms of x_0. Equation (1.7) for $i = 2$ determines x_3 in terms of x_0, etc. It is immediately verified that

$$x_i = \frac{p_{i-1}\, p_{i-2} \cdots p_1}{q_i q_{i-1} \cdots q_1} x_0 = x_0 \prod_{k=0}^{i-1} \frac{p_k}{q_{k+1}}, \qquad i \geq 1,$$

is a solution of (1.7), with x_0 still to be determined. Now since

$$1 = x_0 + \sum_{i=1}^{\infty} x_0 \prod_{k=0}^{i-1} \frac{p_k}{q_{k+1}},$$

we have

$$x_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \prod_{k=0}^{i-1} \frac{p_k}{q_{k+1}}}$$

and so

$$x_0 > 0 \quad \text{if and only if} \quad \sum_{i=1}^{\infty} \prod_{k=0}^{i-1} \frac{p_k}{q_{k+1}} < \infty .$$

In particular, if $p_k = p$ and $q_k = q = 1 - p$ for $k \geq 1$, the series

$$\sum_{i=1}^{\infty} \prod_{k=0}^{i-1} \frac{p_k}{q_{k+1}} = \frac{1}{p} \sum_{i=1}^{\infty} \left(\frac{p}{q}\right)^i$$

converges only when $p < q$.

2: *Proof of Theorem 1.1*

We prove Theorem 1.1 for the case where a_k, b_k, u_k all vanish when $k < 0$; b_k, $a_k \geq 0$ and $a_1 > 0$, $\sum_{k=0}^{\infty} a_k = 1$. The renewal equation then becomes

$$u_n - \sum_{k=0}^{n} a_{n-k} u_k = b_n \qquad \text{for} \quad n = 0, 1, 2, \ldots$$

or equivalently

$$u_n - \sum_{k=0}^{n} a_k u_{n-k} = b_n , \qquad n = 0, 1, 2, \ldots . \tag{2.1}$$

It is easily established inductively (by considering successive equations) that $u_k \geq 0$ for all k.

Since $\{u_n\}$ is by hypothesis a bounded sequence, $\lambda = \limsup_{n\to\infty} u_n$ is finite. Let $n_1 < n_2 < \cdots$ denote a subsequence for which $\lim_{j\to\infty} u_{n_j} = \lambda$. We prove that $\lim_{j\to\infty} u_{n_j - 1} = \lambda$ using the condition $a_1 > 0$. Suppose, to the contrary, that the last relation is not valid. It then follows from the definition of λ that there exists $\lambda' < \lambda$ such that $u_{n_j - 1} < \lambda'$ for an infinite number of j. We put $\varepsilon = [a_1(\lambda - \lambda')]/4$, $M = \sup_{n\geq 0} u_n$, and determine N such that

$$\sum_{k=0}^{n} a_k > 1 - \frac{\varepsilon}{M} \qquad \text{if} \quad n \geq N. \tag{2.2}$$

Let j be chosen so large that $n_j \geq N$ and

$$u_{n_j} > \lambda - \varepsilon, \qquad u_{n_j - 1} < \lambda' < \lambda, \qquad 0 \leq b_{n_j} < \varepsilon,$$
$$\text{and} \qquad u_n < \lambda + \varepsilon \qquad \text{for all} \quad n \geq n_j - N. \tag{2.3}$$

This is all possible by the very definition of λ and the determination of λ'.

From (2.1), (2.2), and (2.3), we have

$$u_{n_j} \leq \sum_{k=0}^{n_j} a_k u_{n_j-k} + \varepsilon < \sum_{k=0}^{N} a_k u_{n_j-k} + M \sum_{k=N+1}^{n_j} a_k + \varepsilon$$

$$< \sum_{k=0}^{N} a_k u_{n_j-k} + 2\varepsilon \qquad \text{[use (2.2) and (2.3)]}$$

$$< (a_0 + a_2 + a_3 + \cdots + a_{N-1} + a_N)(\lambda + \varepsilon) + a_1\lambda' + 2\varepsilon \qquad \text{[use (2.3)]}$$

$$\leq (1 - a_1)(\lambda + \varepsilon) + a_1\lambda' + 2\varepsilon < \lambda + 3\varepsilon - a_1(\lambda - \lambda') = \lambda - \varepsilon,$$

the last line resulting by the choice of ε. But this contradicts the first inequality in (2.3), and so $\lim_{j\to\infty} u_{n_j-1} = \lambda$.

Repeating the argument, we find that, for any integer $d \geq 0$,

$$\lim_{j\to\infty} u_{n_j-d} = \lambda. \tag{2.4}$$

Next, let $r_n = a_{n+1} + a_{n+2} + \cdots$; evidently $\sum_{k=0}^{\infty} k a_k = \sum_{n=0}^{\infty} r_n$, which is verified by summation by parts. (We do not postulate the convergence of the series $\sum r_n$.) Further, $a_1 = r_0 - r_1$, $a_2 = r_1 - r_2$, etc. Substituting into (2.1), we find

$$r_0 u_n + r_1 u_{n-1} + \cdots + r_n u_0 = r_0 u_{n-1} + r_1 u_{n-2} + \cdots + r_{n-1} u_0 + b_n, \qquad n = 1, 2, \ldots.$$

Setting $A_n = r_0 u_n + \cdots + r_n u_0$, we may write this as

$$A_n = A_{n-1} + b_n, \qquad n = 1, 2, \ldots,$$

where $A_0 = r_0 u_0 = (1 - a_0)u_0 = b_0$. It follows that $A_n = \sum_{i=0}^{n} b_i$. Now, since $r_n \geq 0$ and $u_n \geq 0$ for all n, we obtain for any fixed $N > 0$ and $j > 0$

$$r_0 u_{n_j} + r_1 u_{n_j-1} + \cdots + r_N u_{n_j-N} \leq A_{n_j} = \sum_{n=0}^{n_j} b_n.$$

Letting $j \to \infty$ leads to the relation $(r_0 + \cdots + r_N)\lambda \leq \sum_{n=0}^{\infty} b_n$ or equivalently $\lambda \leq \sum_0^{\infty} b_n \left(\sum_0^N r_n\right)^{-1}$.

Since $N > 0$ is arbitrary, it follows that

$$\lambda \leq \frac{\sum_{n=0}^{\infty} b_n}{\sum_{n=0}^{\infty} r_n}. \tag{2.5}$$

Since $u_k \geq 0$ for all k, this proves the theorem in the case $\sum_{n=0}^{\infty} r_n = \infty$, for then $\lambda = \lim_{n\to\infty} u_n = 0$ as is clear from (2.5).

If $\sum_{n=0}^{\infty} r_n < \infty$, let $\mu = \liminf_{n\to\infty} u_n$. Reasoning as in the case of lim sup, we deduce that if $\lim_{j\to\infty} u_{n_j} = \mu$, then $\lim_{j\to\infty} u_{n_j-d} = \mu$ for each integer $d \geq 0$. We set $\sum_{n=N+1}^{\infty} r_n = g(N)$; then plainly $\lim_{N\to\infty} g(N) = 0$, and

$$\sum_{n=0}^{n_j} b_n \leq r_0 u_{n_j} + r_1 u_{n_j-1} + \cdots + r_N u_{n_j-N} + g(N)\cdot M.$$

Letting $j \to \infty$, we conclude that $\sum_{n=0}^{\infty} b_n \leq (r_0 + \cdots + r_N)\mu + g(N)M$.

Now, taking the limit as $N \to \infty$, we find

$$\sum_{n=0}^{\infty} b_n \leq \mu \sum_{n=0}^{\infty} r_n \qquad \text{or} \qquad \mu \geq \frac{\sum_{n=0}^{\infty} b_n}{\sum_{n=0}^{\infty} r_n}. \tag{2.6}$$

But (2.5) and (2.6), in conjunction, yield $\mu \geq \lambda$. On the other hand, by their definition $\mu \leq \lambda$. Thus $\mu = \lambda$, which means that $\lim_{n\to\infty} u_n$ exists and its value is

$$\lim_{n\to\infty} u_n = \frac{\sum_{n=0}^{\infty} b_n}{\sum_{n=0}^{\infty} r_n}.$$

For the case where $a_1 = 0$ but the greatest common divisor of those m for which $a_m > 0$ is 1, the proof can be carried through with the aid of Corollary 4.1 of Chapter 2 combined with the method above.

3: *Absorption Probabilities*

We have previously established (see Problem 6, Chapter 2) that if j is a transient state, then $P_{ij}^n \to 0$, and that if i, j are in the same aperiodic recurrent class, then $P_{ij}^n \to \pi_j \geq 0$. If i, j are in the same periodic recurrent class, the same conclusion holds if we replace P_{ij}^n by $n^{-1}\sum_{m=1}^{n} P_{ij}^m$. In order to complete the discussion of the limiting behavior of P_{ij}^n, it remains to consider the case where i is transient and j is recurrent.

If T is the set of all transient states, then consider

$$x_i^1 = \sum_{j\in T} P_{ij} \leq 1, \qquad i \in T,$$

and define recursively

$$x_i^n = \sum_{j\in T} P_{ij} x_j^{n-1}, \qquad n \geq 2, \quad i \in T.$$

Observe that x_i^n is just the probability that, starting from i, the state of the process stays in T for the next n transitions. Since $x_i^n \leq 1$ for all

$n \geq 1$ (they are probabilities), we may prove by induction that x_i^n is nonincreasing as a function of n. In fact

$$x_i^2 = \sum_{j \in T} P_{ij} x_j^1 \leq \sum_{j \in T} P_{ij} = x_i^1.$$

Now assuming that $x_j^n \leq x_j^{n-1}$ for all $j \in T$ we have

$$0 \leq x_i^{n+1} = \sum_{j \in T} P_{ij} x_j^n \leq \sum_{j \in T} P_{ij} x_j^{n-1} = x_i^n.$$

Therefore, $x_i^n \downarrow x_i$ i.e., x_i^n decreases to some limit x_i, and

$$x_i = \sum_{j \in T} P_{ij} x_j, \qquad i \in T. \tag{3.1}$$

It follows that if the only bounded solution of this set of equations is the zero vector $(0, 0, \ldots)$, then starting from any transient state absorption into a recurrent class occurs with probability one. In fact, it is clear that x_i $(i \in T)$ is the probability of never being absorbed into a recurrent class, starting from state i. Since this sequence is a bounded solution of (3.1) it follows that x_i is zero for all i.

Remark 3.1. If there are only a finite number of states, M, then there are no null states and not all states can be transient. In fact, since $\sum_{j=0}^{M-1} P_{ij}^n = 1$ for all n, it cannot happen that $\lim_{n \to \infty} P_{ij}^n = 0$ for all j.

The same argument restricted to recurrent classes shows that there are no null states. Let $C, C_1, C_2, \ldots$ denote recurrent classes. We define $\pi_i(C)$ as the probability that the process will be ultimately absorbed into the recurrent class C if the initial state is the transient state i. (Recall that once the process enters a recurrent class, it never leaves it.)

Let $\pi_i^n(C) =$ probability that the process will enter and thus be absorbed in C for the first time at the nth transition, given that the initial state is $i \in T$. Then

$$\begin{aligned} \pi_i(C) &= \sum_{n=1}^{\infty} \pi_i^n(C) \leq 1, \\ \pi_i^1(C) &= \sum_{j \in C} P_{ij}, \end{aligned} \tag{3.2}$$

$$\pi_i^n(C) = \sum_{j \in T} P_{ij} \pi_j^{n-1}(C), \qquad n \geq 2. \tag{3.3}$$

Rewriting (3.2) using (3.3) gives

$$\begin{aligned} \pi_i(C) = \pi_i^1(C) + \sum_{n=2}^{\infty} \pi_i^n(C) &= \pi_i^1(C) + \sum_{n=2}^{\infty} \sum_{j \in T} P_{ij} \pi_j^{n-1}(C) \\ &= \pi_i^1(C) + \sum_{j \in T} P_{ij} \sum_{n=2}^{\infty} \pi_j^{n-1}(C), \\ \pi_i(C) &= \pi_i^1(C) + \sum_{j \in T} P_{ij} \pi_j(C), \qquad i \in T. \end{aligned} \tag{3.4}$$

Assuming the only *bounded* solution of the homogeneous set of equations

$$w_i = \sum_{j \in T} P_{ij} w_j, \qquad i \in T,$$

is the zero vector, then $\{\pi_i(C)\}$ is determined as the unique bounded solution of the system of equations (3.4). Moreover, either $\pi_i^1(C) > 0$ for some $i \in T$ or $\pi_i(C) = 0$ for every $i \in T$ and hence $\pi_i^n(C) = 0$ for all n.

Theorem 3.1. *Let $j \in C$ (C an aperiodic recurrent class). Then for $i \in T$, we have*

$$\lim_{n\to\infty} P_{ij}^n = \pi_i(C) \lim_{n\to\infty} P_{jj}^n = \pi_i(C)\pi_j .$$

Proof. Clearly $\pi_i^n(C) = \sum_{k \in C} \pi_{ik}^n(C)$ where $\pi_{ik}^n(C)$ represents the probability starting from state i of being absorbed at the nth transition into class C at state k. We have

$$\pi_i(C) = \sum_{v=1}^{\infty} \sum_{k \in C} \pi_{ik}^v(C) \leq 1.$$

Therefore for any $\varepsilon > 0$ there exists a finite number of states $C' \subset C$ and an integer $N(\varepsilon) = N$ such that

$$\left|\pi_i(C) - \sum_{v=1}^{n} \sum_{k \in C'} \pi_{ik}^v(C)\right| < \varepsilon, \qquad \text{i.e.,} \quad \left|\sum_{v=1}^{\infty} \sum_{k \in C} \pi_{ik}^v - \sum_{v=1}^{n} \sum_{k \in C'} \pi_{ik}^v\right| < \varepsilon \tag{3.5}$$

for $n > N(\varepsilon)$. [Here we have abbreviated π_{ik}^v for $\pi_{ik}^v(C)$.]

For $j \in C$ consider

$$P_{ij}^n - \sum_{v=1}^{n} \sum_{k \in C'} \pi_{ik}^v \pi_j .$$

Now by the usual recursion argument, which involves decomposing the events by the time of first entering some state in C, we obtain

$$P_{ij}^n = \sum_{v=1}^{n} \sum_{k \in C} \pi_{ik}^v P_{kj}^{n-v}, \qquad i \in T, \quad j \in C.$$

Combining these relations, we have

$$\begin{aligned}
\left|P_{ij}^n - \left(\sum_{v=1}^{n} \sum_{k \in C'} \pi_{ik}^v\right)\pi_j\right| &= \left|\sum_{v=1}^{n} \sum_{k \in C'} \pi_{ik}^v (P_{kj}^{n-v} - \pi_j) + \sum_{v=1}^{n} \sum_{k \in C, k \notin C'} \pi_{ik}^v P_{kj}^{n-v}\right| \\
&\leq \left|\sum_{v=1}^{N} \sum_{k \in C'} \pi_{ik}^v (P_{kj}^{n-v} - \pi_j)\right| \\
&\quad + \left|\sum_{v=N+1}^{n} \sum_{k \in C'} \pi_{ik}^v (P_{kj}^{n-v} - \pi_j)\right| + \sum_{v=1}^{n} \sum_{k \in C, k \notin C'} \pi_{ik}^v P_{kj}^{n-v}.
\end{aligned}$$

But $P_{kj}^{n-\nu} \leq 1$, $|P_{kj}^{n-\nu} - \pi_j| \leq 2$, and $\lim_{n\to\infty} P_{kj}^{n-\nu} = \pi_j$ if C is aperiodic and $k \in C'$. Therefore, there exists $N' > N$ such that for $n > N'$, $|P_{kj}^{n-N} - \pi_j| < \varepsilon$ $(k \in C')$, so that for $n > N'$

$$\left| P_{ij}^n - \left(\sum_{\nu=1}^{n} \sum_{k\in C'} \pi_{ik}^{\nu} \right) \pi_j \right| \leq \varepsilon + 2 \sum_{\nu=N+1}^{n} \sum_{k\in C'} \pi_{ik}^{\nu} + \sum_{\nu=1}^{n} \sum_{k\in C, k\notin C'} \pi_{ik}^{\nu}.$$

However, the choice of N and C' assures us that the right-hand side is $\leq 4\varepsilon$. Then appealing to (3.5) and the above result, we obtain

$$|P_{ij}^n - \pi_i(C)\pi_j| \leq 4\varepsilon + \varepsilon\pi_j \qquad \text{for} \quad n > N',$$

and therefore

$$\lim_{n\to\infty} P_{ij}^n = \pi_i(C)\pi_j. \quad \blacksquare$$

If C is periodic and $j \in C$, a similar proof may be used to show

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P_{ij}^m = \pi_i(C)\pi_j.$$

We emphasize the fact that if i is a transient state and j is a recurrent state, then the limit of P_{ij}^n depends on both i and j. This is in sharp contrast with the case where i and j belong to the same recurrent class.

Example. (The gambler's ruin on $n+1$ states).

$$\begin{Vmatrix} 1 & 0 & 0 & 0 & \dots & & & \\ q & 0 & p & 0 & \dots & & & \\ 0 & q & 0 & p & \dots & & & \\ \vdots & & & & & & & \\ & & & & \dots & q & 0 & p \\ & & & & & 0 & 0 & 1 \end{Vmatrix}$$

We shall calculate $u_i = \pi_i(C_0)$ and $v_i = \pi_i(C_n)$, the probabilities that starting from i the process ultimately enters the absorbing (and therefore recurrent) states 0 and n, respectively. The system of equations (3.4) becomes

$$\begin{aligned} u_1 &= q \qquad + pu_2, \\ u_i &= qu_{i-1} + pu_{i+1}, \qquad 2 \leq i \leq n-2. \\ u_{n-1} &= qu_{n-2}, \end{aligned} \tag{3.6}$$

These are $n-1$ nonhomogeneous equations in $n-1$ unknowns. We try a solution of the form $u_r = x^r$. Substituting in the middle equations and removing common factors leads to

$$px^2 + q = x.$$

There are two solutions, $x = 1$ and $x = q/p$. Thus the quantities $u_r = A + B(q/p)^r$, $r = 1, 2, \ldots, n-1$, satisfy the middle equations of (3.6) for any values of A and B. We now determine A and B so that the first and last equations are fulfilled. (If $q = p$, the solution $x = 1$ is a double root of $px^2 + q = x$, and one then has to replace $(q/p)^r$ by r.) In the case $q \neq p$ this leads to the conditions

$$A + B\frac{q}{p} = q + p\left(A + B\frac{q^2}{p^2}\right)$$

or, simplifying,

$$A = 1 - B$$

and

$$A + B\left(\frac{q}{p}\right)^{n-1} = q\left(A + B\left(\frac{q}{p}\right)^{n-2}\right) \qquad \text{or} \qquad p^n A + q^n B = 0.$$

Solving, we get

$$A = \frac{q^n}{q^n - p^n}, \qquad B = \frac{-p^n}{q^n - p^n}.$$

Combining, we have

$$u_r = \frac{(q/p)^n - (q/p)^r}{(q/p)^n - 1} \qquad \text{if} \quad \frac{q}{p} \neq 1.$$

If $q = p$, we find similarly that $A = 1$, $B = -1/n$ so that

$$u_r = \frac{n - r}{n} \qquad \text{when} \quad p = q.$$

A similar calculation shows that

$$v_i = 1 - u_i,$$

which is to be expected, since it is evident that absorption into one of the classes C_0, C_n is certain.

Consider the gambler's ruin with an infinitely rich adversary. The equations for the probability of the gambler's ruin (absorption into 0) become

$$\begin{aligned} u_1 &= q \quad + pu_2, \\ u_i &= qu_{i-1} + pu_{i+1}, \qquad i \geq 2. \end{aligned} \tag{3.7}$$

Again we find

$$u_i = A + B\left(\frac{q}{p}\right)^i \qquad (q \neq p) \qquad \text{and} \qquad u_i = A + Bi \qquad (q = p = \tfrac{1}{2})$$

If $q \geq p$ then the condition that u_i is bounded requires that $B=0$ and the first equation of (3.7) shows that $u_i \equiv 1$. If $q<p$ we find that $u_i=(q/p)^i$. In fact, a simple passage to the limit from the finite state gambler's ruin yields $u_1 = q/p$ and then it readily follows that $u_i=(q/p)^i$.

4: *Criteria for Recurrence*

We prove two theorems which will be useful in determining whether a given Markov chain is recurrent or transient and then we apply them to several examples.

Theorem 4.1. *Let $\mathfrak{P}$ be an irreducible Markov chain whose state space is labeled by the nonnegative integers. Then a necessary and sufficient condition that $\mathfrak{P}$ be transient (i.e., each state is a transient state) is that the system of equations*

$$\sum_{j=0}^{\infty} P_{ij}y_j = y_i, \qquad i \neq 0, \tag{4.1}$$

have a bounded nonconstant solution.

Proof. Let the transition matrix for $\mathfrak{P}$ be

$$\mathbf{P} = \|P_{ij}\| = \begin{Vmatrix} P_{00} & P_{01} & \cdots \\ P_{10} & P_{11} & \cdots \\ \cdots\cdots & \cdots\cdots & \cdots \end{Vmatrix}$$

and associate with it the new transition matrix

$$\tilde{\mathbf{P}} = \|\tilde{P}_{ij}\| = \begin{Vmatrix} 1 & 0 & 0 & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ P_{20} & P_{21} & P_{22} & \cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots \end{Vmatrix} \tag{4.2}$$

in which the zero state has been converted into an absorbing barrier while the transition probabilities governing the motion among the other states are unchanged. We denote the Markov chain with transition probability matrix (4.2) by $\tilde{\mathfrak{P}}$.

For the necessity, we shall assume that the process is transient and then exhibit a nonconstant bounded solution of (4.1).

Let $f_{i0}^* =$ probability of entering state 0 in some finite time, given that i is the initial state. Since the process $\mathfrak{P}$ is transient $f_{j0}^* < 1$ for some $j \neq 0$ or otherwise state 0 would be recurrent. (Prove this. Remember that all

states in an irreducible Markov chain are simultaneously recurrent or nonrecurrent.) For the process $\tilde{\mathfrak{P}}$ clearly $\tilde{\pi}_0(C_0)=1$, $\tilde{\pi}_j(C_0)=f_{j0}^*<1$ for some $j\neq 0$, and $\tilde{\pi}_i(C_0)=\sum_{j=0}^{\infty}\tilde{P}_{ij}\tilde{\pi}_j(C_0)$ for all i. Hence $\tilde{\pi}_i(C_0)=\sum_{j=0}^{\infty}P_{ij}\tilde{\pi}_j(C_0)$ for $i\neq 0$ and thus $y_j=\tilde{\pi}_j(C_0)$ $(j=0,1,2,\ldots)$ is the desired bounded nonconstant solution.

Now assume that we have a bounded solution $\{y_i\}$ of (4.1). Then

$$\sum_{j=0}^{\infty}\tilde{P}_{ij}y_j=y_i \qquad \text{for all } i\geq 0,$$

and iterating we have for all $i\geq 0$ and all $n\geq 1$

$$\sum_{j=0}^{\infty}\tilde{P}_{ij}^n y_j=y_i.$$

If the chain is recurrent, then

$$\lim_{n\to\infty}\tilde{P}_{i0}^n=1,$$

and

$$\sum_{j\neq 0}\tilde{P}_{ij}^n y_j\leq M(1-\tilde{P}_{i0}^n)\to 0 \quad \text{as} \quad n\to\infty$$

where M is a bound for $\{y_j\}$. Hence

$$y_i=\sum_{j\neq 0}\tilde{P}_{ij}^n y_j+\tilde{P}_{i0}^n y_0\to y_0.$$

Thus $y_i=y_0$ for all i and $\{y_j\}$ is constant.

Theorem 4.2. *In an irreducible Markov chain a sufficient condition for recurrence is that there exists a sequence $\{y_i\}$ such that*

$$\sum_{j=0}^{\infty}P_{ij}y_j\leq y_i \qquad \text{for} \quad i\neq 0 \quad \text{with} \quad y_i\to\infty. \tag{4.3}$$

Proof. Using the same notation as in the previous theorem, we have

$$\sum_{j=0}^{\infty}\tilde{P}_{ij}y_j\leq y_i \qquad \text{for all } i.$$

Since $z_i=y_i+b$ satisfies (4.3), we may assume $y_i>0$ for all $i\geq 0$. Iterating the preceding inequality, we have

$$\sum_{j=0}^{\infty}\tilde{P}_{ij}^m y_j\leq y_i.$$

Given $\varepsilon > 0$ we choose $M(\varepsilon)$ such that $1/y_i \leq \varepsilon$ for $i \geq M(\varepsilon)$. Now

$$\sum_{j=0}^{M-1} \tilde{P}_{ij}^m y_j + \sum_{j=M}^{\infty} \tilde{P}_{ij}^m y_j \leq y_i$$

and so

$$\sum_{j=0}^{M-1} \tilde{P}_{ij}^m y_j + \min_{r \geq M} \{y_r\} \sum_{j=M}^{\infty} \tilde{P}_{ij}^m \leq y_i .$$

Since

$$\sum_{j=0}^{\infty} \tilde{P}_{ij}^m = 1$$

we have

$$\sum_{j=0}^{M-1} \tilde{P}_{ij}^m y_j + \min_{r \geq M} \{y_r\} \left(1 - \sum_{j=0}^{M-1} \tilde{P}_{ij}^m\right) \leq y_i .$$

As observed in the proof of the preceding theorem,

$$\lim_{n \to \infty} \tilde{P}_{ij}^n = 0 \qquad \text{for} \quad j \neq 0.$$

Thus, passing to the limit as $m \to \infty$, we obtain for each fixed i

$$\tilde{\pi}_i(C_0) y_0 + \min_{r \geq M} \{y_r\}(1 - \tilde{\pi}_i(C_0)) \leq y_i$$

or

$$1 - \tilde{\pi}_i(C_0) \leq \frac{1}{\min\limits_{r \geq M} \{y_r\}} (y_i - \tilde{\pi}_i(C_0) y_0) \leq \varepsilon K,$$

where

$$K = y_i - \tilde{\pi}_i(C_0) y_0 .$$

Since ε was arbitrary and $\tilde{\pi}_i(C_0) \leq 1$ we have $\tilde{\pi}_i(C_0) = 1$ for each i, proving the original process recurrent. ■

5: A Queueing Example

Let us consider the queueing model discussed in Chapter 2 (Example C). The transition matrix is

$$\|P_{ij}\| = \begin{Vmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ 0 & a_0 & a_1 & a_2 & \dots \\ 0 & 0 & a_0 & a_1 & \dots \end{Vmatrix}, \qquad \text{where} \quad a_k > 0 \quad \text{and} \quad \sum_{k=0}^{\infty} a_k = 1.$$

(Actually, in the subsequent analysis, we only use the properties $0 < a_0 < 1$ and $a_0 + a_1 < 1$ which guarantee that this Markov chain is irreducible.) If $\sum_{k=0}^{\infty} k a_k > 1$ we show that the system of equations $\sum_{j=0}^{\infty} P_{ij} y_j = y_i$, $i \neq 0$, admits a nonconstant bounded solution, and so by Theorem 4.1 the process will be transient. Letting $y_j = \xi^j$, the above system of equations takes the form

$$\sum_{j=0}^{\infty} P_{ij}\xi^j = \sum_{j=i-1}^{\infty} a_{j-i+1}\xi^j = \xi^i$$

or

$$\sum_{j=i-1}^{\infty} a_{j-i+1}\xi^{j-i+1} = \xi = \sum_{k=0}^{\infty} a_k \xi^k = f(\xi), \qquad i \neq 0.$$

Now $f(0) = a_0 > 0$ and $f(1) = \sum_{k=0}^{\infty} a_k = 1$, so that if $f'(1) = \sum_{k=0}^{\infty} k a_k > 1$ then there exists a ξ_0, $0 < \xi_0 < 1$, such that $f(\xi_0) = \xi_0$. This is easily seen from Fig. 1. The vector $y_j = \xi_0^j$, $i = 0, 1, \ldots$, is the desired bounded solution and is clearly nonconstant. If $\sum k a_k \leq 1$ then, applying Theorem 4.2 above with $y_j = j$, we have, provided $i \neq 0$,

$$\begin{aligned}\sum_{j=0}^{\infty} P_{ij} j &= \sum_{j=i-1}^{\infty} a_{j-i+1} j \\ &= \sum_{j=i-1}^{\infty} a_{j-i+1}(j-i+1) + i - 1 \\ &= \sum_{k=0}^{\infty} k a_k - 1 + i \\ &\leq i.\end{aligned}$$

Therefore if $\sum k a_k \leq 1$, the process is recurrent.

In order to ascertain whether the process $\mathfrak{P}$ is null recurrent or positive recurrent we first deal with the following auxiliary problem of some independent interest.

Let $X_1, X_2, X_3, \ldots$ denote a sequence of independent identically distributed random variables taking on the values $-1, 0, 1, 2, \ldots$ with

$$\Pr\{X_i = k\} = b_k, \qquad k = -1, 0, 1, 2, \ldots, \quad b_{-1} > 0,$$

and let $S_n = X_1 + X_2 + \cdots + X_n$. Define Z as the value of n for which S_n first becomes negative and suppose that

$$\Pr\{Z = k\} = \gamma_k, \qquad k = 1, 2, 3, \ldots. \tag{5.1}$$

Let

$$U(s) = \sum_{k=0}^{\infty} \gamma_k s^k \qquad (\gamma_0 = 0) \tag{5.2}$$

denote the generating function of (5.1). If $T_n^{(r)} = r + S_n$ (r is a nonnegative integer), let $Z^{(r)}$ be the random variable equal to the first value of n for which $T_n^{(r)} < 0$. Since each $X_i \geq -1$ we infer easily that $Z^{(r)} = Z_1 + Z_2$

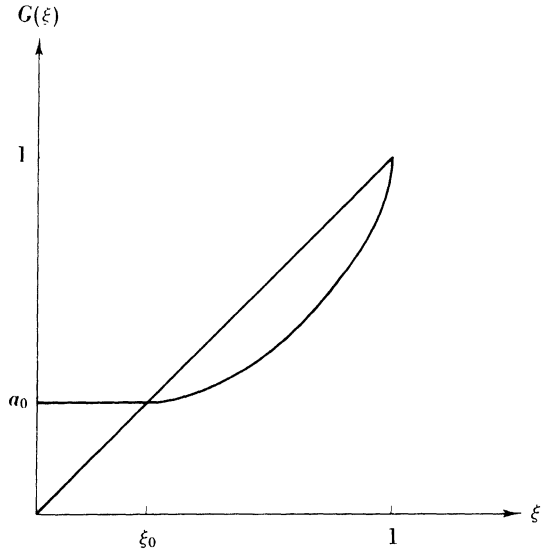


FIG. 1

$+ \cdots + Z_{r+1}$, where the Z_i are independent and identically distributed according to (5.1). The generating function of $Z^{(r)}$ is clearly $[U(s)]^{r+1}$. Let $\gamma_m^{(r+1)}$ denote the coefficient of s^m in $U(s)^{r+1}$.

Finally, we set

$$G(s) = \frac{b_{-1}}{s} + b_0 + b_1 s + b_2 s^2 + \cdots.$$

Our aim is to determine $U(s)$ in terms of $G(s)$. To this end we write the usual renewal relations

$$\gamma_1 = b_{-1}, \qquad \gamma_k = \sum_{j=0}^{\infty} b_j \gamma_{k-1}^{(j+1)}, \qquad k \geq 2. \tag{5.3}$$

The first of these relations is obvious. As for the second, the event $\{S_n \geq 0, n = 1, \ldots, k-1; S_k = -1\}$ is the union of the disjoint events $\{X_1 = j; X_2 + \cdots + X_n + j \geq 0, n = 2, \ldots, k-1; X_2 + \cdots + X_k + j = -1\}$, $j = 0, 1, \ldots$, whose probabilities are clearly equal to $b_j \gamma_{k-1}^{(j+1)}$, since the X_i are independent and identically distributed. By the law of total

probabilities (5.3) results. Passing to generating functions on the basis of (5.3), we have

$$\begin{aligned} U(s) &= b_{-1}s + \sum_{n=2}^{\infty}\Big(\sum_{j=0}^{\infty} b_j \gamma_{n-1}^{(j+1)}\Big)s^n \\ &= b_{-1}s + s\sum_{j=0}^{\infty} b_j\Big(\sum_{n=2}^{\infty}\gamma_{n-1}^{(j+1)}s^{n-1}\Big) \\ &= b_{-1}s + s\sum_{j=0}^{\infty} b_j[U(s)]^{j+1} \\ &= b_{-1}s + sU(s)\Big[G(U(s)) - \frac{b_{-1}}{U(s)}\Big] \qquad \text{for} \quad 0 < s \leq 1 \\ &= sU(s)G(U(s)). \end{aligned}$$

Now $U(s)$ is continuous and strictly increasing for $s \in [0, 1]$, while $U(0) = 0$. Hence, for $0 < s \leq 1$, $U(s)$ satisfies $G(U(s)) = 1/s$. But

$$G''(s) = \frac{2b_{-1}}{s^3} + 2b_2 + 6b_3 s + 12b_4 s^2 + \cdots > 0 \qquad \text{for} \quad s > 0,$$

so that $G(s)$ is a convex function, while from the definition of $G(s)$ it follows that $\lim_{s\downarrow 0} G(s) = +\infty$ and $G(1) = 1$. A glance at the figure below shows that the equation $G(x) = 1/s$ can have at most two positive solutions for each $s \in [0, 1]$. Since $\lim_{s\downarrow 0} U(s) = 0$ and $U(s)$ is strictly increasing in $[0, 1]$, we see that $U(s)$ must be the smaller of the two solutions of $G(x) = 1/s$, if there are two.

We now investigate the conditions under which $\sum_{k=0}^{\infty} \gamma_k = 1$ or <1; the following two cases arise.

Case 1. $G'(1) > 0$. $G'(1) > 0$ is equivalent to $b_{-1} < \sum_{n=0}^{\infty} nb_n$ and Fig.2 clearly shows that $U(1) = \sum_{k=0}^{\infty} \gamma_k = \xi_0 < 1$. Therefore the probability of the event $\{S_n \geq 0 \text{ for all } n\}$ is strictly positive.

Case 2. $G'(1) \leq 0$. $G'(1) \leq 0$ is equivalent to $b_{-1} \geq \sum_{n=0}^{\infty} nb_n$ and here we have $\sum_{k=0}^{\infty} \gamma_k = U(1) = 1$. Now, for $0 < s \leq 1$, $G'(U(s))U'(s) = -1/s^2$ so that in Case 2, $U(s) \to 1$ when $s \to 1$ (see Figure 3). This implies that if $G'(1) < 0$, i.e., if

$$b_{-1} > \sum_{n=0}^{\infty} nb_n,$$

then

$$E(Z) = \sum_{n=0}^{\infty} n\gamma_n = U'(1) = \frac{-1}{G'(1)} < \infty,$$

and if $G'(1) = 0$, i.e., if

$$b_{-1} = \sum_{n=0}^{\infty} nb_n,$$

then

$$E(Z) = \sum_{n=0}^{\infty} n\gamma_n = U'(1) = \infty.$$

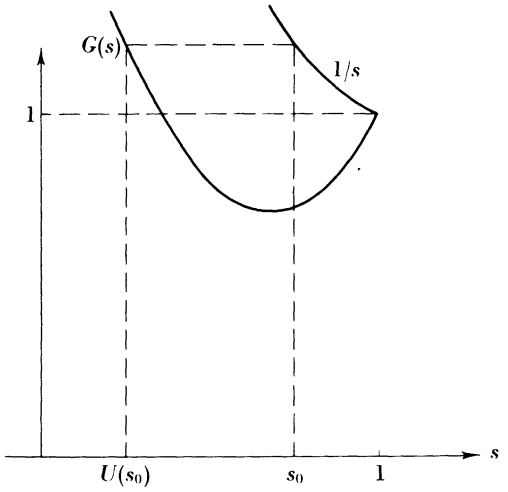


FIG. 2

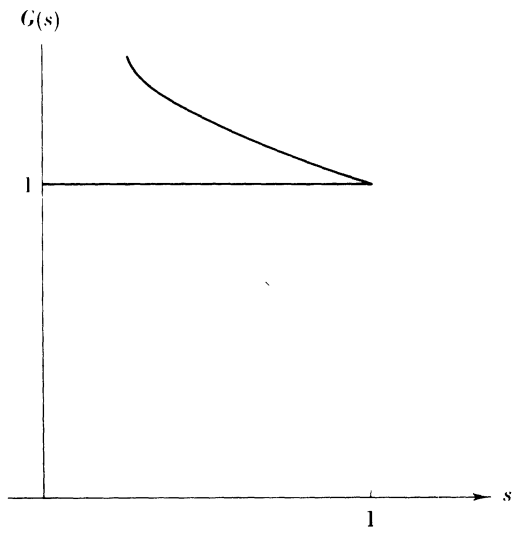


FIG. 3

Referring again to the queueing process, we identify the b's and the a's by $a_k = b_{k-1}$ and let Z_{ij} = length of time (number of transitions) required, starting in state i, to reach state $j < i$ for the first time. A little reflection reveals that $Z_{i,i-1}$ is precisely the random variable Z examined above whose generating function $U(s)$ was determined. Since

$$\sum_{i=0}^{\infty} a_i = 1,$$

we have

$$\left(b_{-1} > \sum_{n=0}^{\infty} nb_n\right) \leftrightarrow \left(a_0 > \sum_{n=0}^{\infty} na_{n+1}\right) \leftrightarrow \left(1 > \sum_{n=0}^{\infty} na_n\right)$$

and similarly

$$\left(b_{-1} = \sum_{n=0}^{\infty} nb_n\right) \leftrightarrow \left(a_0 = \sum_{n=0}^{\infty} na_{n+1}\right) \leftrightarrow \left(1 = \sum_{n=0}^{\infty} na_n\right).$$

Hence $E(Z_{i,i-1}) = \mu < \infty$ if $\sum_{n=0}^{\infty} na_n < 1$ and $E(Z_{i,i-1}) = \mu = \infty$ if $\sum na_n = 1$. Since we are permitted to move back only one step at a time (the process is "continuous" in this respect), we have

$$Z_{i,j} = Z_{i,i-1} + Z_{i-1,i-2} + \cdots + Z_{j+1,j}, \qquad j < i,$$

and therefore $E(Z_{i,j}) = (i-j)\mu$ and in particular $E(Z_{i,0}) = i\mu$.

Let us now consider the mean recurrence time of state zero. We notice first that the probability of the recurrence time equalling 1 is just a_0, the transition probability P_{00}. Now, the sample functions which start at 0 and first return to 0 in two or more transitions can be grouped according to the state i occupied at the first transition. The average recurrence time for such a group is precisely 1 plus the average time required to reach state 0 from state i. This decomposition, in conjunction with the Markov property, yields the following expression for the average recurrence time:

$$\begin{aligned}\sum_{n=0}^{\infty} nf_{00}^{n} &= E(\text{recurrence time}) \\ &= a_0 + \sum_{i=1}^{\infty} a_i[E(Z_{i,0}) + 1] = 1 + \sum_{i=1}^{\infty} a_i E(Z_{i,0}) \\ &= 1 + \sum_{i=1}^{\infty} i\mu a_i = 1 + \mu \sum_{i=0}^{\infty} ia_i.\end{aligned}$$

Thus

$$\sum_{n=0}^{\infty} nf_{00}^{n} < \infty \qquad \text{if} \quad \mu < \infty$$

i.e., provided

$$\sum_{i=0}^{\infty} ia_i < 1,$$

and

$$\sum_{n=0}^{\infty} nf_{00}^{n} = \infty \qquad \text{if } \mu = \infty,$$

or what is the same if

$$\sum_{i=0}^{\infty} ia_i = 1.$$

Summing up, we have

$$\begin{aligned} &\sum_{n=0}^{\infty} na_n < 1 \Rightarrow \text{positive recurrent,} \\ &\sum_{n=0}^{\infty} na_n = 1 \Rightarrow \text{null recurrent,} \end{aligned} \tag{5.4}$$

and

$$\sum_{n=0}^{\infty} na_n > 1 \Rightarrow \text{transient.}$$

These conclusions are rather intuitive. The expression $\sum_{n=0}^{\infty} na_n$ is the mean number of customers arriving during a service period. Thus, if $\sum_{n=0}^{\infty} na_n > 1$, then on an average more people arrive than are served in each period. Therefore, we could expect the waiting line to grow beyond all bounds. On the other hand, if $\sum_{n=0}^{\infty} na_n < 1$ then the state of the process approaches a stationary state. The evaluation of the stationary distribution is rather complicated (see Chapter 18).

6: *Another Queueing Model*

The state of the process is the length of the waiting line where, in each unit of time, one person arrives and k persons are served with probability $a_k > 0$, $k = 0, 1, 2, \ldots$ if there are at least k in the waiting line. The transition probability matrix may be easily evaluated as

$$\|P_{ij}\| = \begin{Vmatrix} \sum_{i=1}^{\infty} a_i & a_0 & 0 & 0 & \ldots \\ \sum_{i=2}^{\infty} a_i & a_1 & a_0 & 0 & \\ \sum_{i=3}^{\infty} a_i & a_2 & a_1 & a_0 & \end{Vmatrix}$$

We show that there exists a stationary distribution if $\sum_{k=0}^{\infty} ka_k > 1$ so that in this case the process is positive recurrent. A stationary distribution is expected to exist since the average number of customers served is $\sum ka_k$ while a single new customer arrives.

Consider the equations $\sum_{i=0}^{\infty} \xi_i P_{ij} = \xi_j$ and let $\xi_i = \xi^i$. Then

$$\sum_{i=j-1}^{\infty} \xi^i a_{i-j+1} = \xi^j \quad \text{for } j \geq 1; \quad \text{i.e., } \sum_{i=j-1}^{\infty} \xi^{i-j+1} a_{i-j+1} = \xi,$$

which by a change of variable reduces to

$$\sum_{k=0}^{\infty} a_k \xi^k = \xi.$$

If ξ $(0 < \xi < 1)$ satisfies these equations, then for $j = 0$ we have

$$\begin{aligned}
\sum_{i=0}^{\infty} P_{i0} \xi^i &= \sum_{i=0}^{\infty} \left(\sum_{k=i+1}^{\infty} a_k \right) \xi^i \\
&= \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} a_k \xi^i \qquad \text{(rearranging the order of summation)} \\
&= \sum_{k=1}^{\infty} a_k \left(\frac{1-\xi^k}{1-\xi} \right) = \frac{1}{1-\xi} \left(1 - a_0 - \sum_{k=1}^{\infty} a_k \xi^k \right) \\
&= \frac{1}{1-\xi} (1 - a_0 - (\xi - a_0)) = 1,
\end{aligned}$$

so that the equations are satisfied for $j = 0$ as well.

We consider $f(\xi) = \sum_{k=0}^{\infty} a_k \xi^k$. Since $f(0) = a_0 > 0$ and $f(1) = 1$, if $f'(1) = \sum_{k=0}^{\infty} ka_k > 1$ then there exists ξ_0 satisfying $0 < \xi_0 < 1$ and $f(\xi_0) = \xi_0$ (see Fig. 4). The values $\pi_i = (1 - \xi_0)\xi_0^i$, which sum to 1, are

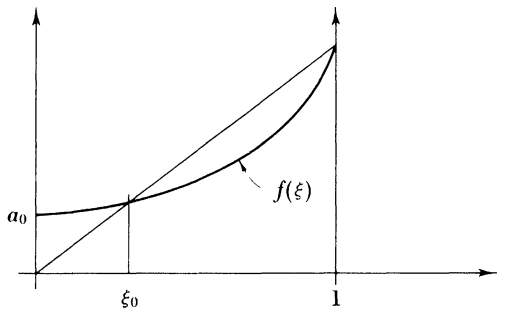


FIG. 4

therefore the stationary probabilities of the state of the process. In particular, the long-run probability that the line is empty is $1 - \xi_0$.

Now the system of equations $\sum_{i=0}^{\infty} \xi_i P_{ij} = \xi_j$, $j \neq 0$, is identical with the system $\sum_{j=0}^{\infty} \tilde{P}_{ij}\xi_j = \xi_i$, $i \neq 0$, where the $\tilde{P}_{ij}$ are now the elements of the matrix from Section 5. In the case $\sum_{k=0}^{\infty} ka_k \leq 1$ the $\tilde{P}$ process (as we have seen) is recurrent and hence the latter system has no non-constant bounded solution. Therefore, if $\sum ka_k \leq 1$ the system $\sum_{j=0}^{\infty} \eta_i P_{ij} = \eta_j$ admits no bounded solution and therefore, in particular, no stationary distribution exists so that the process is either null recurrent or transient. We now prove that the system of equations

$$\sum_{j=0}^{\infty} P_{ij}y_j = y_i, \qquad i \neq 0, \tag{6.1}$$

has a nonconstant bounded solution if and only if $\sum ka_k < 1$, so that the process will be transient if and only if $\sum ka_k < 1$, and therefore must be null recurrent in the case $\sum_{k=0}^{\infty} ka_k = 1$. Since (6.1) admits a constant solution we may let $y_0 = 0$. Then (6.1) reduces to

$$\begin{aligned} a_2y_0 + a_1y_1 + a_0y_2 &= y_1 \\ a_3y_0 + a_2y_1 + a_1y_2 + a_0y_3 &= y_2 \\ &\cdots\cdots\cdots\cdots \\ a_{n+1}y_0 + a_ny_1 + \quad \cdots \quad + a_1y_n + a_0y_{n+1} &= y_n \\ &\cdots\cdots\cdots\cdots \end{aligned}$$

Multiplying the ith equation by s^{i+1} and summing, we obtain, after letting

$$Y(s) = \sum_{k=0}^{\infty} y_k s^k, \qquad A(s) = \sum_{k=0}^{\infty} a_k s^k,$$

and recognizing the convolution product, that

$$Y(s)A(s) - sa_0y_1 = sY(s) \qquad \text{or} \qquad Y(s) = \frac{sa_0y_1}{A(s) - s}, \tag{6.2}$$

provided $A(s) \neq s$. Since $A(0) = a_0$ and $A(1) = 1$, $A(s) = s$ for some s such that $0 < s < 1$ if $A'(1) = \sum_{k=0}^{\infty} ka_k > 1$. Therefore, $Y(s)$ cannot have bounded coefficients in this case since $Y(s)$ would then converge for every $s \in [0, 1)$. This implies that for $\sum_{k=0}^{\infty} ka_k > 1$ the process is recurrent.

From the strict convexity of $A(s)$, i.e., $A''(s) > 0$, it follows that $A(s) \neq s$ for $0 \leq s < 1$ if $A'(1) = \sum ka_k \leq 1$ (see Fig. 5). Consider the case $\sum ka_k \leq 1$:

$$A(s)-s=(1-s)\left[1-\frac{1-A(s)}{1-s}\right]$$

$$=(1-s)\left[1-(1-A(s))\sum_{k=0}^{\infty}s^k\right]$$

$$=(1-s)\left[1-\sum_{n=0}^{\infty}\left(1-\sum_{i=0}^{n}a_i\right)s^n\right]$$

(rearranging orders of summations)

$$=(1-s)\left[1-\sum_{n=0}^{\infty}\left(\sum_{i=n+1}^{\infty}a_i\right)s^n\right]$$

$$=(1-s)[1-W(s)],\qquad W(s)=\sum_{n=0}^{\infty}w_n s^n,$$

where

$$w_n=\sum_{i=n+1}^{\infty}a_i>0$$

and

$$\sum_{n=0}^{\infty}w_n=\sum_{n=0}^{\infty}\left(\sum_{i=n+1}^{\infty}a_i\right)=\sum_{k=0}^{\infty}ka_k\leq 1.$$

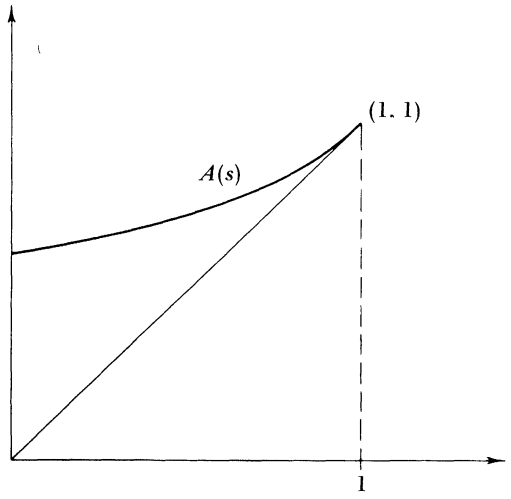


FIG. 5

Then

$$Y(s) = \frac{sa_0y_1}{(1-s)[1-W(s)]}$$

$$= \frac{sa_0y_1}{1-s}(1 + W(s) + (W(s))^2 + \cdots)$$

$$= sa_0y_1 \frac{U(s)}{1-s} \qquad \text{where} \quad u_n \geq 0, \quad U(s) = \sum_{n=0}^{\infty} u_n s^n = \sum_{k=0}^{\infty} [W(s)]^k$$

$$= sa_0y_1V(s) \qquad \text{where} \quad v_n = \sum_{k=0}^{n} u_k, \quad V(s) = \sum_{n=0}^{\infty} v_n s^n,$$

i.e.,

$$V(s) = \frac{U(s)}{1-s}$$

Now

$$\left(W(1) = \sum ka_k < 1\right) \leftrightarrow \left(U(1) = \sum_{k=0}^{\infty} u_k < \infty\right),$$

since $U(1) = 1 + W(1) + (W(1))^2 + \cdots$, which is a convergent geometric series. Clearly $v_1 < v_2 < \cdots \to U(1)$, so that $Y(s) = sa_0y_1V(s)$ has bounded coefficients in its power series expansion if and only if $\sum ka_k < 1$. Therefore if $\sum ka_k < 1$ we may take $y_1 \neq 0, y_k = a_0y_1v_{k-1}$, and, retracing steps to Eq. (6.2) and equating coefficients, obtain a bounded nonconstant solution of (6.1). This implies that the process is transient. If $\sum ka_k = 1$ any solution of (6.1) is necessarily unbounded, implying the process is recurrent. To sum up, if

$\sum ka_k < 1$, the process is transient;

$\sum ka_k = 1$, the process is null recurrent;

$\sum ka_k > 1$, the process is positive recurrent.

7: *Random Walk*

We apply the recurrence criteria of Section 4 to the random walk induced by the Markov matrix

$$\|P_{ij}\| = \begin{Vmatrix} r_0 & p_0 & 0 & 0 & \cdots \\ q_1 & r_1 & p_1 & 0 & \cdots \\ 0 & q_2 & r_2 & p_2 & \cdots \end{Vmatrix}$$

Let

$$\pi_0 = 1, \qquad \pi_n = \frac{p_0 p_1 \cdots p_{n-1}}{q_1 q_2 \cdots q_n}.$$

For the case $r_i \equiv 0$ it was shown (see the example of Section 1) that the random walk process at hand has a stationary distribution if and only if $\sum_{n=0}^{\infty} \pi_n < \infty$. Now consider the system of equations

$$\sum_{j=0}^{\infty} P_{ij} y_j = y_i, \qquad i \neq 0,$$

or

$$\begin{aligned} q_1 y_0 + r_1 y_1 + p_1 y_2 &= y_1, \\ &\cdots\cdots\cdots\cdots \\ q_n y_{n-1} + r_n y_n + p_n y_{n+1} &= y_n. \\ &\cdots\cdots\cdots\cdots \end{aligned}$$

Inspection shows that the solutions span a two dimensional linear space. We can prescribe y_0 and y_1 arbitrarily and then all the other y_i are determined by these equations. Trivially $y_i \equiv 1$ is a solution. We show that $y_0 = 0, y_n = \sum_{i=0}^{n-1} 1/p_i \pi_i$, $n \geq 1$, is also a solution. For the first equation

$$q_1 y_0 + r_1 y_1 + p_1 y_2 = r_1 \left(\frac{1}{p_0}\right) + p_1 \left(\frac{1}{p_0} + \frac{q_1}{p_1 p_0}\right) = \frac{1}{p_0} = y_1.$$

For the nth equation we must show

$$q_n \left(\sum_{i=0}^{n-2} \frac{1}{p_i \pi_i}\right) + r_n \sum_{i=0}^{n-1} \frac{1}{p_i \pi_i} + p_n \sum_{i=0}^{n} \frac{1}{p_i \pi_i} = \sum_{i=0}^{n-1} \frac{1}{p_i \pi_i}.$$

Since $p_n + r_n + q_n = 1$ it suffices to verify that

$$q_n \sum_{i=0}^{n-2} \frac{1}{p_i \pi_i} + p_n \sum_{i=0}^{n} \frac{1}{p_i \pi_i} = (p_n + q_n) \sum_{i=0}^{n-1} \frac{1}{p_i \pi_i}.$$

But the left-hand side is just

$$(q_n + p_n) \sum_{i=0}^{n-1} \frac{1}{p_i \pi_i} - q_n \frac{1}{p_{n-1} \pi_{n-1}} + p_n \frac{1}{p_n \pi_n},$$

while

$$-q_n \frac{1}{p_{n-1} \pi_{n-1}} = \frac{-1}{(p_{n-1}/q_n) \pi_{n-1}} = \frac{-1}{\pi_n}$$

by the definition of π_n, which proves the assertion. Since the two solutions $(y_i \equiv 1)$ and $(y_n = \sum_{i=1}^{n-1} 1/p_i \pi_i)$ are clearly independent, the general

solution is $z_n = \alpha + \beta y_n$, and a nonconstant bounded solution of $\sum_{j=0}^{\infty} P_{ij} z_j = z_i$, $i \neq 0$, exists if and only if the y_n are bounded, i.e., $\sum_{i=0}^{\infty} 1/p_i\pi_i < \infty$.
Therefore, we have

$$\sum_{i=0}^{\infty} \frac{1}{p_i\pi_i} = \infty \Rightarrow \text{recurrent},$$

$$\sum_{i=0}^{\infty} \frac{1}{p_i\pi_i} = \infty \quad \text{and} \quad \sum_{i=0}^{\infty} \pi_i = \infty \Rightarrow \text{null recurrent},$$

$$\sum_{i=0}^{\infty} \frac{1}{p_i\pi_i} = \infty \quad \text{and} \quad \sum_{i=0}^{\infty} \pi_i < \infty \Rightarrow \text{positive recurrent},$$

$$\sum_{i=0}^{\infty} \frac{1}{p_i\pi_i} < \infty \Rightarrow \text{transient}.$$

Elementary Problems

1. A matrix $P = \|P_{ij}\|_{i,j=1}^{\infty}$ is called stochastic if

(i) $P_{ij} \geq 0$ for all i and $j = 1, 2, \ldots$

and

(ii) $\sum_{j=1}^{\infty} P_{ij} = 1$ for all $i = 1, 2, \ldots$.

A matrix P is called doubly stochastic if in addition to (i) and (ii) also

$$\sum_{i=1}^{\infty} P_{ij} = 1 \quad \text{for all } j = 1, 2, \ldots .$$

Prove that if a finite irreducible Markov chain has a doubly stochastic transition probability matrix, then all the stationary probabilities are equal.

2. A Markov chain on states $\{0, 1, 2, 3, 4, 5\}$ has transition probability matrix

$$\text{(a)} \begin{Vmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{5} & \frac{4}{5} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{Vmatrix}, \qquad \text{(b)} \begin{Vmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{8} & \frac{7}{8} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{8} & \frac{3}{8} & 0 \\ \frac{1}{3} & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{Vmatrix}$$

Find all classes. Compute the limiting probabilities $\lim_{n\to\infty} P_{5i}^n$ for $i = 0, 1, 2, 3, 4, 5$.

3. Consider a Gambler's ruin with initial fortune a and b ($a > 10, b > 10$) for players I and II, respectively. Let $p\,(1-p)$ be the probability that I wins

(loses) from II one unit each game. What is the probability that player I will achieve a fortune $a+b-3$ before his fortune dwindles to 5?

4. Let Y_n be the sum of n independent rolls of a fair die. Find

$$\lim_{n\to\infty} \Pr\{Y_n \text{ is a multiple of } 13\}.$$

5. Consider a Markov chain with transition probability matrix

$$P=\begin{bmatrix} p_0 & p_1 & p_2 & \cdots & p_m \\ p_m & p_0 & p_1 & \cdots & p_{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ p_1 & p_2 & p_3 & \cdots & p_0 \end{bmatrix}$$

where $0<p_0<1$ and $p_0+p_1+\cdots+p_m=1$. Determine $\lim_{n\to\infty} P_{ij}^n$, the stationary distribution.

6. Members of an indefinitely large population are either immune to a given disease or are susceptible to it. Let X_n be the number of susceptible members in the population at time period n and suppose $X_0=0$ and that in the absence of an epidemic $X_{n+1}=X_n+1$. Thus, in the absence of the disease, the number of susceptibles in the population increases in time, possibly owing to individuals losing their immunity, or to the introduction of new susceptible members to the population.

But in each period there is a constant but unknown probability p of a pandemic disease. When the disease occurs all susceptibles are stricken. The disease is non-lethal and confers immunity, so that if T is the first time of disease occurrence, then $X_T=0$.

Compute the stationary distribution for (X_n).

7. An airline reservation system has two computers only one of which is in operation at any given time. A computer may break down on any given day with probability p. There is a single repair facility which takes 2 days to restore a computer to normal. The facilities are such that only one computer at a time can be dealt with. Form a Markov chain by taking as states the pairs (x, y) where x is the number of machines in operating condition at the end of a day and y is 1 if a day's labor has been expended on a machine not yet repaired and 0 otherwise. The transition matrix is

$$P=\begin{array}{c} \text{To} \\ \text{From} \\ \text{State} \\ \downarrow \\ (2,0) \\ (1,0) \\ (1,1) \\ (0,1) \end{array}\begin{array}{c} \text{State}\to \\ (2,0) \\ (1,0) \\ (1,1) \\ (0,1) \end{array}\begin{pmatrix} q & p & 0 & 0 \\ 0 & 0 & q & p \\ q & p & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

where $p+q=1$. Find the stationary distribution in terms of p and q.

8. Consider a production line where each item has probability p of being defective. Assume that the condition of a particular item (defective or nondefective) does not depend on the condition of other items. The following sampling plan is used:

Initially every item is sampled as it is produced; this continues until i consecutive nondefective items are found. Then the sampling plan calls for sampling only one out of every r items at random until a defective one is found. When this happens the plan calls for reverting to 100% sampling until i consecutive nondefective items are found, etc.

State E_k $(k = 0, 1, \ldots, i)$ denotes that k consecutive nondefective items have been found in the 100% sampling portion of the plan, while state E_{i+1} denotes that the plan is in the second stage (sampling one out of r) and one or more nondefective items have been sampled in this stage. (Time m is considered to follow the mth observation for any m.) Then the sequence of states is a Markov chain with

$$P_{jk} = \Pr\left\{\begin{matrix}\text{in state } E_k \text{ after } m+1 \text{ observations} \\ \text{after } m \text{ observations}\end{matrix} \middle| \text{in state } E_j\right\}$$

$$= \begin{cases} p & \text{if } k = 0, \quad j = 0, 1, \ldots, i, i+1, \\ 1-p & \text{if } k = j+1, \quad j = 0, 1, \ldots, i \text{ or } k = j = i+1, \\ 0 & \text{otherwise,} \end{cases}$$

for all m.

(a) Determine the stationary distribution.

(b) Determine the long run fraction of items that are inspected.

(c) Determine the average outgoing quality (AOQ), the long run fraction of defective items in the output of the sampling plan.

9. Sociologists often assume that the social classes of successive generations in a family can be regarded as a Markov chain. Thus, the occupation of a son is assumed to depend only on his father's occupation and not on his grandfather's. Suppose that such a model is appropriate and that the transition probability matrix is given by

		Son's Class		
		Lower	Middle	Upper
Father's Class	Lower	.40	.50	.10
	Middle	.05	.70	.25
	Upper	.05	.50	.45

For such a population, what fraction of people are middle class in the long run?

10. Suppose that the weather on any day depends on the weather conditions for the previous two days. To be exact, suppose that if it was sunny today and yesterday, then it will be sunny tomorrow with probability .8; if it was sunny today but cloudy yesterday, then it will be sunny tomorrow with probability .6; if it was cloudy today but sunny yesterday, then it will be sunny tomorrow

with probability .4; if it was cloudy for the last two days, then it will be sunny tomorrow with probability .1.

Such a model can be transformed into a Markov chain provided we say that the state at any time is determined by the weather conditions during both that day and the previous day. We say the process is in

State (S, S) if it was sunny both today and yesterday,
State (S, C) if it was sunny yesterday but cloudy today,
State (C, S) if it was cloudy yesterday but sunny today,
State (C, C) if it was cloudy both today and yesterday.

Then the transition probability matrix is

		Today's State			
		(S, S)	(S, C)	(C, S)	(C, C)
Yesterday's State	(S, S)	.8	.2		
	(S, C)			.4	.6
	(C, S)	.6	.4		
	(C, C)			.1	.9

(a) Find the stationary distribution of the Markov chain.
(b) On what fraction of days in the long run is it sunny?

11. Consider a regular $2r+1$ polygon consisting of vertices $V_1, V_2, \ldots, V_{2r+1}$. Suppose that at each point V_k there is a nonnegative mass w_k^1 where $w_1^1 + \cdots + w_{2r+1}^1 = 1$. Obtain new masses $w_1^2, \ldots, w_{2r+1}^2$ by replacing the old mass at k by the arithmetic mean of neighboring masses, i.e.

$$w_k^2 = \tfrac{1}{2}(w_{k-1}^1 + w_{k+1}^1).$$

Do this transformation n times. Determine $\lim_{n\to\infty} w_k^n$.

Solution:

$1/(2r+1)$, independent of k.

12. Consider a light bulb whose life, measured in discrete units, is a random variable X where $\Pr[X = k] = p_k$ for $k = 1, 2, \ldots$. If one starts with a fresh bulb and if each bulb is replaced by a new one when it burns out then u_n, the expected number of replacements up to time n, solves the equation

$$u_n = F_X(n) + \sum_{k=1}^{n} p_k u_{n-k}, \qquad n = 1, 2, \ldots$$

where $F_X(n) = \sum_{k \le n} p_k$.

In a large building it is often cheaper, on a per bulb basis, to replace all the bulbs, failed or not, than it is to replace a single bulb, due to economies of scale. A "block replacement policy" is a function of the block period N and calls for replacing bulbs as they fail during periods 0, 1, ..., $N-1$, and then replacing all bulbs, failed or not, in period N. If C_1 is the per bulb block replacement cost

and C_2 is the per bulb failure replacement cost it can be shown that the long run per bulb time average cost of such a policy is $[C_1 + C_2 u_{N-1}]/N$, which is the expected cost over a replacement cycle divided by the length of the cycle. (This result is formally proved in Chapter 5.)

(a) Based on intuition, note that u_n, the expected renewals up to time n, cannot converge but should grow unboundedly. What condition in Theorem 1.1 is violated in the renewal equation for u_n?

(b) Derive a renewal equation for $v_n = \Pr\{\text{a replacement is needed at time } n\}$. Note $v_n = u_n - u_{n-1}$ for $n = 1, 2, \ldots (u_0 = 0)$.

(c) If $p_1 = .4$, $p_2 = .3$, $p_3 = .2$ and $p_4 = .1$ compute and plot v_n for $n = 1, 2, \ldots, 10$. Compute $u_n = v_1 + \cdots + v_n$.

(d) If $C_1 = \$1$ and $C_2 = \$2$, determine the value for N that yields minimum cost.

Solution:

(b) $v_n = p_n + \sum_{k=1}^{n} p_{n-k} v_k, \qquad v_0 = p_0 = 0.$

(c)
$v_1 = .4000$ $\quad$ $v_6 = .4991$
$v_2 = .4600$ $\quad$ $v_7 = .5013$
$v_3 = .5040$ $\quad$ $v_8 = .5005$
$v_4 = .5196$ $\quad$ $v_9 = .4994$
$v_5 = .4910$ $\quad$ $v_{10} = .5002$

(d) $N^* = 2$.

Problems

1. Consider the following random walk:

$$\begin{aligned} P_{i,i+1} &= p && \text{with} && 0 < p < 1, \\ P_{i,i-1} &= q = 1 - p && \text{for} && i = 1, 2, \ldots, r-1, \\ P_{0,0} &= P_{r,r} = 1. \end{aligned}$$

Find $d(k) = E[\text{time to absorption into states } 0 \text{ or } r \mid \text{initial state is } k]$.

Answer:

$$\begin{aligned} d(k) &= \frac{k}{q-p} - \frac{r}{q-p}\frac{(1-(q/p)^k)}{1-(q/p)^r} && \text{if} \quad p \neq \tfrac{1}{2}, \\ &= k(r-k) && \text{if} \quad p = \tfrac{1}{2}. \end{aligned}$$

2. Let $\mathbf{P} = \|P_{ij}\|$ be the transition probability matrix of an irreducible Markov chain and suppose $\mathbf{P}$ is idempotent (i.e., $\mathbf{P}^2 = \mathbf{P}$). Prove that $P_{ij} = P_{jj}$ for all i and j and that the Markov chain is aperiodic.

Hint: Use Theorem 1.2 for the averages $(1/m)\sum_{m=1}^{n} P_{ij}^m$.

3. Consider a finite Markov chain $\mathfrak{M}$ on the state space $\{0, 1, 2, \ldots, N\}$ with transition probability matrix $\mathbf{P} = \|P_{ij}\|_{i,j=0}^{N}$ consisting of three classes $\{0\}$,

$\{1, 2, \ldots, N-1\}$ and $\{N\}$ where 0 and N are absorbing states, both accessible from $k = 1, \ldots, N-1$, and $\{1, 2, \ldots, N-1\}$ is a transient class. Let k be a state satisfying $0 < k < N$. We define an auxiliary process $\tilde{\mathfrak{M}}$ called "the return process" by altering the first and last row of $\mathbf{P}$ so that $\tilde{P}_{0k} = \tilde{P}_{Nk} = 1$ and leave the other rows unchanged. The return process $\tilde{\mathfrak{M}}$ is clearly irreducible. Prove that the expected time until absorption u_k with initial state k in the $\mathfrak{M}$ process equals $1/(\pi_0 + \pi_N) - 1$ where $\pi_0 + \pi_N$ is the stationary probability of being in state 0 or N for the $\tilde{\mathfrak{M}}$ process.

Hint: Use the relation between stationary probabilities and expected recurrence times to states.

4. Consider a discrete time Markov chain with states $0, 1, \ldots, N$ whose matrix has elements

$$P_{ij} = \begin{cases} \mu_i, & j = i-1, \\ \lambda_i, & j = i+1, \\ 1-\lambda_i-\mu_i, & j = i, \\ 0, & |j-i| > 1, \end{cases} \qquad i, j = 0, 1, \ldots, N.$$

Suppose that $\mu_0 = \lambda_0 = \mu_N = \lambda_N = 0$, and all other μ_i's and λ_i's are positive, and that the initial state of the process is k. Determine the absorption probabilities at 0 and N.

Answer: Define

$$\rho_0 = 1, \qquad \rho_i = \frac{\mu_1\mu_2 \cdot \cdots \cdot \mu_i}{\lambda_1\lambda_2 \cdot \cdots \cdot \lambda_i};$$

$$\Pr\{\text{absorption at } 0\} = 1 - \Pr\{\text{absorption at } N\} = \frac{\sum_{i=k}^{N-1} \rho_i}{\sum_{i=0}^{N-1} \rho_i}.$$

5. Under the conditions of Problem 4, determine the expected time until absorption.

6. Consider a Markov chain with the $N+1$ states $0, 1, \ldots, N$ and transition probabilities

$$P_{ij} = \binom{N}{j} \pi_i^j (1-\pi_i)^{N-j}, \qquad 0 \leq i, \quad j \leq N,$$

$$\pi_i = \frac{1 - e^{-2ai/N}}{1 - e^{-2a}}, \qquad a > 0.$$

Note that 0 and N are absorbing states. Verify that $\exp(-2aX_t)$ is a martingale [or, what is equivalent, prove the identity $E(\exp(-2aX_{t+1})|X_t) = \exp(-2aX_t)$], where X_t is the state at time t ($t = 0, 1, 2, \ldots$). Using this property show that the probability $P_N(k)$ of absorption into state N starting at state k is given by

$$P_N(k) = \frac{1 - e^{-2ak}}{1 - e^{-2aN}}.$$

Hint: Use the fact that absorption into one of the states 0 or N in finite time occurs with certainty and that the relations

$$E(\exp(-2aX_0)) = E(\exp(-2aX_n)) = P_N(k)\exp(-2aN) + (1 - P_N(k))$$

hold (justify this).

7. Consider a finite population (of fixed size N) of individuals of possible types A and a undergoing the following growth process. At instants of time $t_1 < t_2 < t_3 < \cdots$, one individual dies and is replaced by another of type A or a. If just before a replacement time t_n there are j A's and $N-j$ a's present, we postulate that the probability that an A individual dies is $j\mu_1/B_j$ and that an a individual dies is $(N-j)\mu_2/B_j$ where $B_j = \mu_1 j + \mu_2(N-j)$. The rationale of this model is predicated on the following structure: Generally a type A individual has chance $\mu_1/(\mu_1+\mu_2)$ of dying at each epoch t_n and an a individual has chance $\mu_2/(\mu_1+\mu_2)$ of dying at time t_n. (μ_1/μ_2 can be interpreted as the selective advantage of A types over a types.) Taking account of the sizes of the population it is plausible to assign the probabilities $\mu_1 j/B_j$ and $(\mu_2(N-j)/B_j)$ to the events that the replaced individual is of type A and type a, respectively. We assume no difference in the birth pattern of the two types and so the new individual is taken to be A with probability j/N and a with probability $(N-j)/N$. Consider the Markov chain $\{X_n\}$, where X_n is the number of A types at time t_n ($n = 1, 2, \ldots$) with transition probabilities

$$P_{j,j-1} = \frac{\mu_1 j(N-j)}{B_j N}, \qquad P_{j,j+1} = \frac{\mu_2(N-j)j}{B_j N},$$

$$P_{jj} = 1 - P_{j,j-1} - P_{j,j+1}, \qquad P_{ij} = 0, \qquad \text{for } |i-j| > 1.$$

Find the probability that the population is eventually all of type a, given k A's and $(N-k)$ a's initially.

Hint: Show that the equations that determine the absorption probabilities can be reduced to a corresponding system of equations for absorption probabilities of a gambler's ruin random walk.

Answer:

$$\Pr\{\text{all } a\text{'s eventually left}\} = \frac{(\mu_1/\mu_2)^N - (\mu_1/\mu_2)^k}{(\mu_1/\mu_2)^N - 1}, \qquad \mu_1 \neq \mu_2,$$

$$= 1 - \frac{k}{N}, \qquad \mu_1 = \mu_2.$$

8. Let $\mathbf{P}$ be a 3×3 Markov matrix and define $\mu(\mathbf{P}) = \max_{i_1,i_2,j}[P_{i_1,j} - P_{i_2,j}]$ Show that $\mu(\mathbf{P}) = 1$ if and only if $\mathbf{P}$ has the form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & p & q \\ r & s & t \end{pmatrix} \qquad (p, q \geq 0, p+q=1; \quad r, s, t \geq 0, r+s+t=1)$$

or any matrix obtained from this one by interchanging rows and/or columns.

*9. If $\mathbf{P}$ is a finite Markov matrix, we define $\mu(\mathbf{P}) = \max_{i_1,i_2,j}(P_{i_1,j} - P_{i_2,j})$. Suppose $P_1, P_2, \ldots, P_k$ are 3×3 transition matrices of irreducible aperiodic Markov chains. Assume furthermore that for any set of integers α_i $(1 \leq \alpha_i \leq k)$, $i = 1, 2, \ldots, m$, $\prod_{i=1}^{m} \mathbf{P}_{\alpha_i}$ is also the matrix of an aperiodic irreducible Markov chain. Prove that, for every $\varepsilon > 0$, there exists an $M(\varepsilon)$ such that $m > M$ implies

$$\mu\left(\prod_{i=1}^{m} \mathbf{P}_{\alpha_i}\right) < \varepsilon \qquad \text{for any set } \alpha_i \ (1 \leq \alpha_i \leq k) \qquad i = 1, 2, \ldots, m.$$

*10. If i is a recurrent state and X_k represents the state of the Markov chain at time k, then show that

$$\lim_{N \to \infty} \Pr\{X_k \neq i \quad \text{for} \quad n+1 \leq k \leq n+N | X_0 = i\} = 0.$$

If i is a positive recurrent state prove that the convergence in the above equation is uniform with respect to n.

*11. Generalized Pólya Urn Scheme. In an urn containing a white and b black balls we select a ball at random. If a white ball is selected we return it and add α white and β black to the urn and if a black ball is selected we return it and add γ white and δ black, where $\alpha + \beta = \gamma + \delta$. The process is repeated. Let X_n be the number of selections that are white among the first n repetitions.

(i) If $P_{n,k} = \Pr\{X_n = k\}$ and $\varphi_n(x) = \sum_{k=0}^{n} P_{n,k} x^k$ establish the identity

$$\varphi_n(x) = \frac{(\alpha - \gamma)(x^2 - x)}{(n-1)(\alpha+\beta) + a + b} \varphi'_{n-1}(x) + \frac{\{x[(n-1)\gamma + a] + b + (n-1)\delta\}}{(n-1)(\alpha+\beta) + a + b} \varphi_{n-1}(x).$$

(ii) Prove the limit relation $E(X_n/n) \to \gamma/(\beta+\gamma)$ as $n \to \infty$.

Hint: Show that

$$\varphi'_n(1) = (\alpha - \gamma) \sum_{k=1}^{n} \frac{\varphi'_{k-1}(1)}{(k-1)(\alpha+\beta) + a + b} + \sum_{k=1}^{n} \frac{a + (k-1)\gamma}{(k-1)(\alpha+\beta) + a + b}$$

and deduce from this that $\varphi'_n(1)/n \to \gamma/(\beta+\gamma)$.

*12. Under the conditions of Problem 11 prove

$$\lim_{n \to \infty} E\left[\left(\frac{X_n}{n}\right)^2\right] = \left(\frac{\gamma}{\beta+\gamma}\right)^2 \qquad \text{as} \quad n \to \infty$$

Hint: Determine a recursion relation for $\varphi''_n(1)$ as in (i) above.

*13. Under the conditions of Problem 11 show that $X_n/n \to \gamma/(\beta+\gamma)$ in probability as $n \to \infty$.

*14. Consider an irreducible Markov chain with a finite set of states $\{1, 2, \ldots, N\}$. Let $\|P_{ij}\|$ be the transition probability matrix of the Markov chain and denote by $\{\pi_j\}$ the stationary distribution of the process. Let $\|P_{ij}^{(m)}\|$ denote the m-step transition probability matrix. Let $\varphi(x)$ be a concave function on $x \geq 0$ and define

$$E_m = \sum_{j=1}^{N} \pi_j \varphi(P_{jl}^{(m)}) \qquad \text{with } l \text{ fixed.}$$

Prove that E_m is a nondecreasing function of m, i.e., $E_{m+1} \geq E_m$ for all $m \geq 1$.

Hint: Use Jensen's inequality.

*15. Assume state 0 is positive recurrent. We take the initial state to be 0. Let $\{W_n\}$ $(n = 1, 2, \ldots)$ denote successive recurrence times which are of course independent and identically distributed random variables with finite mean and with a generating function $F(t) = \sum_{k=1}^{\infty} t^k \Pr\{W_1 = k\}$ $(|t| < 1)$. Define Y_n as the time of the last visit to state 0 before the time n. Show that

$$\sum_{n=0}^{\infty} t^n \sum_{j=0}^{n} x^j \Pr\{Y_n = j\} = \frac{(1 - F(t))}{(1-t)(1 - F(xt))}.$$

Hint: Prove and use the relation $\Pr\{Y_n = j\} = \Pr\{W_1 + \ldots + W_{N_n} = j\} \cdot q_{n-j}$ where $q_i = \Pr\{W_1 > i\}$ and N_n is the number of visits to state 0 in the first n trials.

16. Fix the decreasing sequence of nonnegative numbers $1 = b_0 \geq b_1 \geq \cdots$ and consider the Markov chain having transition probabilities

$$P_{ij} = \begin{cases} \dfrac{b_j}{b_i}(\beta_i - \beta_{i+1}) & j \leq i \\ \dfrac{\beta_{i+1}}{\beta_i} & j = i+1 \\ 0 & \text{elsewhere,} \end{cases}$$

where $\beta_n = b_n/(b_1 + \cdots + b_n)$. Show that $P_{00}^n = 1/\sigma_n$ where $\sigma_n = b_1 + \cdots + b_n$. Thus the chain is transient if and only if $\sum \dfrac{1}{\sigma_n} < \infty$.

NOTES

The content of Sections 1–4 is part of the standard apparatus of Markov chains that is included in most books on the subject.

The examples of Section 5 are classical in the area of stochastic queueing models. For further refinements see, e.g., Takacs [1].

REFERENCE

1. L. Takacs, "Introduction to the Theory of Queues." Oxford Univ. Press, London and New York, 1962.

Chapter 4

CLASSICAL EXAMPLES OF CONTINUOUS TIME MARKOV CHAINS

Poisson and birth and death processes play a fundamental role in the theory and applications that embrace queueing and inventory models, population growth, engineering systems, etc. This chapter should be studied in every introductory course.

1: General Pure Birth Processes and Poisson Processes

The previous chapters were devoted to an elaboration of the basic concepts and methods of discrete time Markov chains. In this chapter we present a brief discussion of several important examples of continuous time, discrete state, Markov processes.

Specifically, we deal here with a family of random variables $\{X(t); 0 \leq t < \infty\}$ where the possible values of $X(t)$ are the nonnegative integers. We shall restrict attention to the case where $\{X(t)\}$ is a Markov process with stationary transition probabilities. Thus, the transition probability function for $t > 0$,

$$P_{ij}(t) = \Pr\{X(t+u) = j | X(u) = i\}, \qquad i, j = 0, 1, 2, \ldots, \tag{1.1}$$

is independent of $u \geq 0$.

It is usually more natural in investigating particular stochastic models based on physical phenomena to prescribe the so-called infinitesimal probabilities relating to the process and then derive from them an explicit expression for the transition probability function.

For the case at hand, we will postulate the form of $P_{ij}(h)$ for h small and, using the Markov property, we will derive a system of differential equations satisfied by $P_{ij}(t)$ for all $t > 0$. The solution of these equations

under suitable boundary conditions gives $P_{ij}(t)$. We recall that the Poisson process introduced in Section 2, Chapter 1 was in fact treated from just this point of view.

By way of introduction to the general pure birth process we review briefly the axioms characterizing the Poisson process.

A. POSTULATES FOR THE POISSON PROCESS

The Poisson process has been considered in Section 2, Chapter 1, where it was shown that it could be defined by a few simple postulates. In order to define more general processes of a similar kind, let us point out various further properties that the Poisson process possesses. In particular, it is a Markov process on the nonnegative integers which has the following properties:

(i) $\Pr\{X(t+h) - X(t) = 1 | X(t) = x\} = \lambda h + o(h)$
as $h \downarrow 0$ $(x = 0, 1, 2, \ldots)$.

The precise interpretation of (i) is the relationship

$$\lim_{h \to 0+} \frac{\Pr\{X(t+h) - X(t) = 1 | X(t) = x\}}{h} = \lambda.$$

The $o(h)$ symbol means that if we divide this term by h then its value tends to zero as h tends to zero. Notice that the right-hand side is independent of x.

(ii) $\Pr\{X(t+h) - X(t) = 0 | X(t) = x\} = 1 - \lambda h + o(h)$ as $h \downarrow 0$.

(iii) $X(0) = 0$.

These properties are easily verified by direct computation, since the explicit formulas for all the relevant probabilities are available.

B. EXAMPLES OF POISSON PROCESSES

(a) An illustrative example of the Poisson process is that of fishing. Let the random variable $X(t)$ denote the number of fish caught in the time interval $[0, t]$. Suppose that the number of fish available is very large, that the enthusiast stands no better chance of catching fish than the rest of us, and that as many fish are likely to nibble at one instant of time as at another. Under these "ideal" conditions, the process $\{X(t); t \geq 0\}$ may be considered to be a Poisson process. This example serves to point up the Markov property (the chance of catching a fish does not depend upon the number

caught) and the "no premium for waiting" property, which is the most distinctive property possessed by the Poisson process. It means that the fisherman who has just arrived at the pier has as good a chance of catching a fish in the next instant of time as he who has been waiting for a bite for four hours without success.

(b) A less imaginative example is afforded by problems arising in the theory of counters. If $X(t)$ is the number of radioactive disintegrations detected by a Geiger counter in the time interval $[0, t]$, the process is Poisson as long as the half-life of the substance is large relative to t. This provision ensures that the chance for a disintegration per unit of time may be considered as constant over time.

(c) Poisson processes arise naturally in many models of queueing phenomena. In these examples most attention is placed upon the times at which $X(t)$ (= length of queue at time t) jumps rather than upon the values of $X(t)$ themselves. The fishing example (a) is of course a special waiting time example.

C. PURE BIRTH PROCESS

A natural generalization of the Poisson process is to permit the chance of an event occurring at a given instant of time to depend upon the number of events which have already occurred. An example of this phenomenon is the reproduction of living organisms (and hence the name of the process), in which under certain conditions—sufficient food, no mortality, no migration, etc.—the probability of a birth at a given instant is proportional (directly) to the population size at that time. This example is known as the Yule process.

Consider a sequence of positive numbers, $\{\lambda_k\}$. We define a pure birth process as a Markov process satisfying the postulates:

(i) $\Pr\{X(t+h) - X(t) = 1 | X(t) = k\} = \lambda_k h + o_{1,k}(h)$, $(h \to 0+)$,

(ii) $\Pr\{X(t+h) - X(t) = 0 | X(t) = k\} = 1 - \lambda_k h + o_{2,k}(h)$,

(iii) $\Pr\{X(t+h) - X(t) < 0 | X(t) = k\} = 0$, $(k \geq 0)$.

As a matter of convenience we often add the postulate

(iv) $X(0) = 0$.

With this postulate $X(t)$ does not denote the population size but, rather, the number of births in the time interval $[0, t]$.

Note that the left sides of (i) and (ii) are just $P_{k,k+1}(h)$ and $P_{k,k}(h)$, respectively (owing to stationarity), so that $o_{1,k}(h)$ and $o_{2,k}(h)$ do not depend upon t.

We define $P_n(t) = \Pr\{X(t) = n\}$, assuming $X(0) = 0$.

In exactly the same way as for the Poisson process, we may derive a system of differential equations satisfied by $P_n(t)$ for $t \geq 0$, namely

$$\begin{aligned} P_0'(t) &= -\lambda_0 P_0(t), \\ P_n'(t) &= -\lambda_n P_n(t) + \lambda_{n-1} P_{n-1}(t), \qquad n \geq 1, \end{aligned} \tag{1.2}$$

with boundary conditions

$$P_0(0) = 1, \qquad P_n(0) = 0, \qquad n > 0.$$

Indeed, if $h > 0$, $n \geq 1$, then by invoking the law of total probabilities, the Markov property, and postulate (iii) we obtain

$$\begin{aligned} P_n(t+h) &= \sum_{k=0}^{\infty} P_k(t) \Pr\{X(t+h) = n | X(t) = k\} \\ &= \sum_{k=0}^{\infty} P_k(t) \Pr\{X(t+h) - X(t) = n-k | X(t) = k\} \\ &= \sum_{k=0}^{n} P_k(t) \Pr\{X(t+h) - X(t) = n-k | X(t) = k\}. \end{aligned}$$

Now for $k = 0, 1, \ldots, n-2$ we have

$$\begin{aligned} &\Pr\{X(t+h) - X(t) = n-k | X(t) = k\} \\ &\quad \leq \Pr\{X(t+h) - X(t) \geq 2 | X(t) = k\} \\ &\quad = o_{1,k}(h) + o_{2,k}(h) \end{aligned}$$

or

$$\Pr\{X(t+h) - X(t) = n-k | X(t) = k\} = o_{3,n,k}(h) \qquad k = 0, \ldots n-2.$$

Thus

$$\begin{aligned} P_n(t+h) &= P_n(t)\,[1 - \lambda_n h + o_{2,n}(h)] \\ &\quad + P_{n-1}(t)\,[\lambda_{n-1} h + o_{1,n-1}(h)] \\ &\quad + \sum_{k=0}^{n-2} P_k(t) o_{3,n,k}(h) \end{aligned}$$

or

$$\begin{aligned} P_n(t+h) - P_n(t) &= P_n(t)[-\lambda_n h + o_{2,n}(h)] \\ &\quad + P_{n-1}(t)[\lambda_{n-1} h + o_{1,n-1}(h)] + o_n(h), \end{aligned} \tag{1.3}$$

where, clearly, $\lim_{h \downarrow 0} o_n(h)/h = 0$ uniformly in $t \geq 0$ since $o_n(h)$ is bounded by the finite sum $\sum_{k=0}^{n-2} o_{3,n,k}(h)$ which does not depend on t.

Dividing by h and passing to the limit $h \downarrow 0$, we obtain the validity of the relations (1.2) where on the left-hand side we should, to be precise, write the right-hand derivative. However, with a little more care we can derive the same relation involving the left-hand derivative. In fact, from (1.3) we see at once that the $P_n(t)$ are continuous functions of t. Replacing

t by $t - h$ in (1.3), dividing by h, and passing to the limit $h \downarrow 0$, we find that each $P_n(t)$ has a left derivative which also satisfies Eq. (1.2).

The first equation of (1.2) can be solved immediately and yields

$$P_0(t) = \exp(-\lambda_0 t) > 0.$$

Define T_k as the time between the kth and the $(k+1)$st birth, so that

$$P_n(t) = \Pr\{\sum_{i=0}^{n-1} T_i \leq t < \sum_{i=0}^{n} T_i\}.$$

The random variables T_k are called the "waiting times" between births, and

$$S_k = \sum_{i=0}^{k-1} T_i = \text{the time at which the } k\text{th birth occurs.}$$

We have already seen that $P_0(t) = \exp(-\lambda_0 t)$. Therefore,

$$\Pr\{T_0 \leq z\} = 1 - \Pr\{X(z) = 0\} = 1 - \exp(-\lambda_0 z),$$

i.e., T_0 has an exponential distribution with parameter λ_0. It may be deduced from postulates (i)–(iv) that T_k, $k > 0$, also has an exponential distribution with parameter λ_k and that the T_i's are mutually independent (see Chapter 14 of Volume II, where a formal proof of this fact is given). Therefore, the characteristic function of S_n is given by

$$\varphi_n(w) = E\{\exp(iwS_n)\} = \prod_{k=0}^{n-1} E(\exp(iwT_k)) = \prod_{k=0}^{n-1} \frac{\lambda_k}{\lambda_k - iw}. \tag{1.4}$$

In the case of the Poisson process where $\lambda_k = \lambda$ for all k, we recognize from (1.4) that S_n is distributed according to a gamma distribution of order n with mean n/λ.

For a specific set of $\lambda_k \geq 0$ we may solve each equation of (1.2) by means of the integrating factor $\exp(\lambda_k t)$, obtaining

$$P_k(t) = \lambda_{k-1} \exp(-\lambda_k t) \int_0^t \exp(\lambda_k x)\, P_{k-1}(x)\, dx, \qquad k = 1, 2, \ldots,$$

which makes it clear that all $P_k(t) \geq 0$.

But there is still a possibility that

$$\sum_{n=0}^{\infty} P_n(t) < 1.$$

To assure the validity of the process, i.e., to determine criteria in order that $\sum_{n=0}^{\infty} P_n(t) = 1$ for all t, we must restrict the λ_k according to the following

$$\sum_{n=0}^{\infty} P_n(t) = 1 \leftrightarrow \sum_{n=0}^{\infty} \frac{1}{\lambda_n} = \infty. \tag{1.5}$$

The proof of this is given in Feller's book† and so is omitted here. The intuitive argument for this result is as follows: The time T_k between consecutive births is shown below to be exponentially distributed with a corresponding parameter λ_k. Therefore, the quantity $\sum_n 1/\lambda_n$ equals the expected time before the population becomes infinite. By comparison $1 - \sum_{n=0}^{\infty} P_n(t)$ is the probability that $X(t) = \infty$.

If $\sum \lambda_n^{-1} < \infty$ then the expected time for the population to become infinite is finite. It is then plausible that for all $t > 0$ the probability that $X(t) = \infty$ is positive.

D. THE YULE PROCESS

The Yule process is an example of a pure birth process that arises in physics and biology. Assume that each member in a population has a probability $\beta h + o(h)$ of giving birth to a new member in an interval of time length h ($\beta > 0$). Furthermore assume that there are $X(0) = N$ members present at time 0. Assuming independence and no interaction among members of the population, the binomial theorem gives

$$\begin{aligned}&\Pr\{X(t+h) - X(t) = 1 | X(t) = n\}\\ &= \binom{n}{1}[\beta h + o(h)][1 - \beta h + o(h)]^{n-1} = n\beta h + o_n(h),\end{aligned}$$

i.e., in this example $\lambda_n = n\beta$. The system of equations (1.2) in the case that $N = 1$ becomes

$$P_n'(t) = -\beta[nP_n(t) - (n-1)P_{n-1}(t)], \qquad n = 1, 2, \ldots,$$

under the initial conditions

$$P_1(0) = 1, \qquad P_n(0) = 0, \qquad n = 2, 3, \ldots$$

Its solution is

$$P_n(t) = e^{-\beta t}(1 - e^{-\beta t})^{n-1} \qquad n \geq 1,$$

as may be verified directly.

The generating function may be determined easily by summing a geometric series. We have

$$\begin{aligned}f(s) &= \sum_{n=1}^{\infty} P_n(t)s^n\\ &= se^{-\beta t}\sum_{n=1}^{\infty}[(1 - e^{-\beta t})s]^{n-1} = \frac{se^{-\beta t}}{1 - (1 - e^{-\beta t})s}.\end{aligned}$$

† W. Feller, "An Introduction to Probability Theory and Its Applications," Vol. 1, 2nd ed. p. 406. Wiley, New York, 1957.

Let us return to the general case in which there are $X(0) = N$ members present at time 0. Since we have assumed independence and no interaction among the members, we may view this population as the sum of N independent Yule processes, each beginning with a single member. Thus, if we let

$$P_{Nn}(t) = \Pr\{X(t) = n | X(0) = N\}$$

and

$$f_N(s) = \sum_{n=N}^{\infty} P_{Nn}(t)s^n \tag{1.6}$$

we have

$$\begin{aligned} f_N(s) &= [f(s)]^N \\ &= \left[\frac{se^{-\beta t}}{1-(1-e^{-\beta t})s}\right]^N \\ &= (se^{-\beta t})^N \sum_{m=0}^{\infty} \binom{m+N-1}{m}(1-e^{-\beta t})^m s^m \\ &= \sum_{n=N}^{\infty} \binom{n-1}{n-N}(e^{-\beta t})^N(1-e^{-\beta t})^{n-N}s^n, \end{aligned}$$

where we have used the binomial series $(1-x)^{-N} = \sum_{m=0}^{\infty} \binom{m+N-1}{m}x^m$. According to (1.6), the coefficient of s^n in this expression must be $P_{Nn}(t)$. That is

$$P_{Nn}(t) = \binom{n-1}{n-N}e^{-N\beta t}(1-e^{-\beta t})^{n-N} \qquad \text{for} \quad n = N, N+1, \ldots \tag{1.7}$$

2: More about Poisson Processes

In the previous section we derived the Poisson process from a set of assumptions that are approximated well in many practical situations. This process is often referred to as the completely random process, as it distributes points "at random" over the infinite interval $[0, \infty)$ in much the same way that the uniform distribution distributes points over a finite interval. In particular, the probability of an observation falling in a subinterval is a function of its length only and the number of events occurring in two disjoint time intervals are independent random variables.

Let us now examine the Poisson process a little more closely.

A. CHARACTERISTIC FUNCTION AND WAITING TIMES

We may write the characteristic function of $X(t)$ in a Poisson process as

$$\varphi_t(w) = E\{e^{iwX(t)}\} = \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n e^{iwn}}{n!} = \exp[\lambda t(e^{iw} - 1)]$$

Thus

$$E(X(t)) = \lambda t, \qquad \mathrm{Var}(X(t)) = \lambda t.$$

In our discussion of the pure birth process we showed that

$$\Pr\{T_0 \leq z\} = 1 - \exp(-\lambda_0 z)$$

and mentioned that T_k follows an exponential distribution with parameter λ_k and that the T_k's are independent. For the Poisson process, however, $\lambda_k = \lambda$ for all k, so that the result becomes

Theorem 2.1. *The waiting times T_k are independent and identically distributed following an exponential distribution with parameter λ.*

The rigorous proof of this theorem will follow from the more general considerations of Chapter 14.

The definition of the process requires more than is present for the validity of this theorem. We need to assume that the time until the next change of $X(t)$ follows the same distribution laws from any start of measured time, not just if we measure from a previous change. This is simply the statement that

$$\Pr\{X(t_0 + \tau) - X(t_0) > 0\} = 1 - e^{-\lambda\tau},$$

which was derived in Section 1. This property can also be obtained in a more direct manner. Let $F(x) = \Pr\{X(t_0 + x) - X(t_0) > 0\}$ where t_0 is some time, depending perhaps on the history of the process up to that time, whose specification does not affect this probability.† Then

$$\begin{aligned} F(x+y) &= \Pr\{X(t_0 + x + y) - X(t_0) > 0\} \\ &= \Pr\{X(t_0 + y) - X(t_0) > 0\} + \Pr\{X(t_0 + y) - X(t_0) = 0\} \\ &\quad \times \Pr\{X(t_0 + x + y) - X(t_0 + y) > 0 | X(t_0 + y) - X(t_0) = 0\}. \end{aligned}$$

From the definition of $F(x)$, the independence of the increments of the

† The interpretation of this seemingly vague phrase will be given precision in our discussion of the concept of " Markov time "; see Chapter 14.

Poisson process, and the fact (which enters as an initial assumption in defining the Poisson process) that

$$\Pr\{X(t_0 + x) - X(t_0) > 0\}$$

is independent of t_0, we obtain the functional equation

$$F(x+y) = F(y) + [1 - F(y)]F(x).$$

The fact that this property characterizes the exponential distribution is the content of the following theorem.

Theorem 2.2. *If $F(x)$ is a distribution such that $F(0) = 0$ and $F(x) < 1$ for some $x > 0$, then $F(x)$ is an exponential distribution if and only if*

$$F(x+y) - F(y) = F(x)[1 - F(y)] \qquad \text{for all} \quad x, y \geq 0. \tag{*}$$

Proof. That the exponential distribution satisfies the condition follows directly by substitution. To show the converse, set $G(x) = 1 - F(x)$; then the condition (*) becomes

$$G(x+y) = G(x)G(y). \tag{2.1}$$

Obviously, $G(0) = 1$, $G(x)$ is nonincreasing, and for some $x > 0$, $G(x) > 0$. Suppose that $G(x_0) = 0$ for some $x_0 > 0$. From Eq. (2.1) it immediately follows that $G(x_0) = [G(x_0/n)]^n$ for every integer $n > 0$; hence $G(x_0/n) = 0$. But then (2.1) shows that $G(x) = 0$ for $x > x_0/n$. Since n is arbitrary, $G(x) = 0$ for all $x > 0$, contrary to hypothesis. Thus $G(x) > 0$ for every $x > 0$. Now for any integers m, $n > 0$ we deduce easily from (2.1) that $G(m/n) = [G(1)]^{m/n}$. Since $G(x)$ and $[G(1)]^x$ are both nonincreasing functions which coincide whenever x is rational, and $[G(1)]^x$ is continuous, it follows that $G(x) = [G(1)]^x = \exp(x \log G(1))$ for all $x > 0$. But $F(x)$ is a distribution, and so

$$\lim_{x \to \infty} G(x) = 1 - \lim_{x \to \infty} F(x) = 0,$$

which implies that $G(1) < 1$. Hence $G(x) = e^{-\lambda x}$, where

$$\lambda = -\log G(1) > 0. \quad \blacksquare$$

Another proof assuming that G is differentiable goes as follows: Observe that (2.1) implies

$$G'(x+y) = \frac{\partial}{\partial x} G(x+y) = G'(x)G(y),$$

$$G(x)G'(y) = \frac{\partial}{\partial y} G(x+y) = G'(x+y),$$

and therefore

$$G'(x) = aG(x), \tag{2.2}$$

where $a = G'(y_0)/G(y_0)$ for some y_0 where $G(y_0) \neq 0$. The solution of Eq. (2.2) is $G(x) = Ae^{ax}$ and $A = 1$ since $G(0) = 1 - F(0) = 1$. The parameter a is negative since $G(x) < 1$ for some $x > 0$.

B. UNIFORM DISTRIBUTION

The class of distributions that are connected with the Poisson process does not stop with the Poisson and exponential distributions. We shall show how the uniform and binomial distributions also arise.

Consider the times $\{S_i\}$ at which changes of $X(t)$ occur, i.e.,

$$S_i = \sum_{k=0}^{i-1} T_k.$$

We have the following result.

Theorem 2.3. *For any numbers s_i satisfying $0 \leq s_1 \leq s_2 \leq \cdots \leq s_n \leq t$,*

$$\Pr\{S_i \leq s_i,\ i = 1, \ldots, n \mid X(t) = n\} = \frac{n!}{t^n} \int_0^{s_1} \cdots \int_{x_{n-2}}^{s_{n-1}} \int_{x_{n-1}}^{s_n} dx_n \cdots dx_1,$$

which is the distribution of the order statistics from a sample of n observations taken from the uniform distribution on $[0, t]$.†

Proof. The proof is an easy consequence of Theorem 2.1. In fact

$$\begin{aligned}
&\Pr\{S_1 \leq s_1, S_2 \leq s_2, \cdots, S_n \leq s_n,\ X(t) = n\} \\
&\quad = \Pr\{T_0 \leq s_1,\ T_0 + T_1 \leq s_2, \ldots,\ T_0 + \cdots + T_{n-1} \leq s_n,\ T_0 + \cdots + T_n > t\} \\
&\quad = \int_0^{s_1} \int_0^{s_2 - t_1} \int_0^{s_3 - (t_1 + t_2)} \cdots \int_0^{s_n - (t_1 + \cdots + t_{n-1})} \int_{t - (t_1 + \cdots + t_n)}^{\infty} \lambda^{n+1} e^{-\lambda(t_1 + \cdots + t_{n+1})} \\
&\quad \times dt_{n+1} \cdot \cdots \cdot dt_1
\end{aligned}$$

† This means the following. Take n independent observations of a random variable which is uniformly distributed over the interval $[0, t]$. Let $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ denote these observations arranged in increasing order. Then the joint distribution of $Y_1, \ldots, Y_n$ is precisely the expression in the assertion of the theorem. The proof of this fact is quite simple, but a more complete discussion will be presented in Chapter 13 of Volume II.

$$= \lambda^{n+1} \int_0^{s_1} \int_0^{s_2-t_1} \int_0^{s_3-(t_1+t_2)} \cdots \int_0^{s_n-(t_1+\cdots+t_{n-1})} e^{-\lambda(t_1+\cdots+t_n)}$$

$$\times \left[-\frac{1}{\lambda} \exp(-\lambda t_{n+1}) \right]_{t-(t_1+\cdots+t_n)}^{\infty} dt_n \cdot \cdots \cdot dt_1$$

$$= \lambda^n e^{-\lambda t} \int_0^{s_1} \int_0^{s_2-t_1} \int_0^{s_3-(t_1+t_2)} \cdots \int_0^{s_n-(t_1+\cdots+t_{n-1})} dt_n \cdot \cdots \cdot dt_1 .$$

If we introduce the new variables

$$\begin{aligned} u_n &= t_1 + \cdots + t_n \\ u_{n-1} &= t_1 + \cdots + t_{n-1} \\ &\vdots \\ u_1 &= t_1 , \end{aligned}$$

the last expression becomes

$$\lambda^n e^{-\lambda t} \int_0^{s_1} \int_{u_1}^{s_2} \int_{u_2}^{s_3} \cdots \int_{u_{n-1}}^{s_n} du_n \cdot \cdots \cdot du_1 .$$

But

$$\Pr\{X(t) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!};$$

hence

$$\begin{aligned} &\Pr\{S_1 \le s_1, S_2 \le s_2, \ldots, S_n \le s_n | X(t) = n\} \\ &\quad = \frac{\Pr\{S_1 \le s_1, \ldots, S_n \le s_n, X(t) = n\}}{\Pr\{X(t) = n\}} \\ &\quad = \frac{n!}{t^n} \int_0^{s_1} \int_{u_1}^{s_2} \cdots \int_{u_{n-1}}^{s_n} du_n \cdot \cdots \cdot du_1 . \end{aligned}$$

C. BINOMIAL DISTRIBUTION

It follows from the properties of a Poisson process that for $u < t$ and $k < n$,

$$\begin{aligned} &\Pr\{X(u) = k | X(t) = n\} = \\ &\qquad \Pr\{X(u) = k,\ X(t) - X(u) = n - k\} / \Pr\{X(t) = n\} \qquad (2.3) \\ &\quad = \frac{(e^{-\lambda u} u^k / k!)[e^{-\lambda(t-u)}(t-u)^{n-k}/(n-k)!]}{e^{-\lambda t}(t^n/n!)} = \binom{n}{k} \frac{u^k (t-u)^{n-k}}{t^n} \end{aligned}$$

A second example in which the binomial distribution plays a part may be given by considering two independent Poisson processes $X_1(t)$ and $X_2(t)$ with parameters λ_1 and λ_2.

$$\Pr\{X_1(t) = k | X_1(t) + X_2(t) = n\} = \frac{\Pr\{X_1(t) = k,\ X_2(t) = n - k\}}{\Pr\{X_1(t) + X_2(t) = n\}}$$

$$= \frac{[\exp(-\lambda_1 t)(\lambda_1 t)^k/k!][\exp(-\lambda_2 t)(\lambda_2 t)^{n-k}/(n-k)!]}{\exp[-(\lambda_1 + \lambda_2)t](\lambda_1 + \lambda_2)^n t^n/n!}$$

$$= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k}$$

3: A Counter Model

An interesting application of the Poisson process is the following problem. Electrical pulses with random amplitudes X_i arrive at random times t_i (i.e., according to a Poisson process) at a detector whose output for each pulse at time t is

$$X_i \exp[-\alpha(t - t_i)]_+ = \begin{cases} 0 & \text{for } t < t_i, \\ X_i \exp[-\alpha(t - t_i)] & \text{for } t > t_i; \end{cases}$$

that is, the amplitude impressed on the detector when the pulse arrives is X_i and its effect thereafter decays at an exponential rate. The detector is linear (i.e., additive) so if N_t pulses occur during the time epoch $[0, t]$ the output at time t is

$$\eta(t) = \sum_{i=1}^{N_t} X_i \exp[-\alpha(t - t_i)]_+.$$

A typical realization of this process has the shape shown in Fig. 1. We would like to know the distribution function of $\eta(t)$ for each t, or, equivalently, its characteristic function $\varphi_t(w)$.

We assume that the X_i are identically and independently distributed positive random variables with density function $h(x)$ and characteristic function

$$\psi(s) = \int_0^\infty e^{isx} h(x)\, dx.$$

Set

$$R(v; t) = \Pr\{\eta(t) \leq v\} = \sum_{n=0}^{\infty} \Pr\{\eta(t) \leq v | N_t = n\} \Pr\{N_t = n\}. \quad (3.1)$$

Of course $\Pr\{N_t = n\} = [(\lambda t)^n e^{-\lambda t}]/n!$, where λ is the intensity parameter of the Poisson process describing the arrival times of the pulses. From the result of Theorem 2.3 we know that, conditioned by the event $N_t = n$, i.e.,

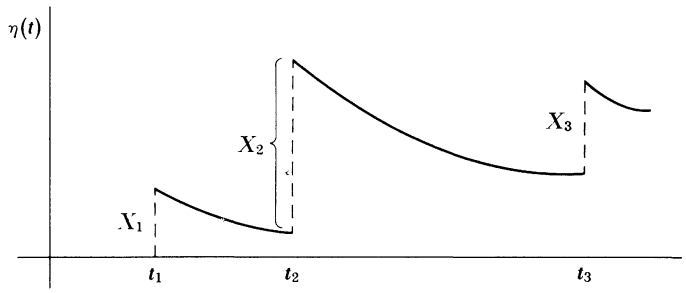


FIG. 1

that n pulses have arrived during the time interval $(0, t)$, the t_i are distributed like ordered observations from a uniform distribution on $(0, t)$. Let τ_i $(i = 1, 2, \ldots, N_t)$ denote independent uniformly distributed [on $(0, t)$] random variables whose values arranged in increasing order of magnitude are the t_j's.

Now let $Z_1, \ldots, Z_n$ be independent random variables, whose distributions are identical with that of the X_i, and which are also independent of the $\{\tau_i\}$. Consider the sum

$$\sum_{i=1}^{n} Z_i \exp[-\alpha(t - \tau_i)]_+ .$$

We define new random variables $Z_1', \ldots, Z_n'$ in accordance with

$$\begin{aligned}
Z_1' &= Z_j \quad \text{when} \quad \tau_j = \min(\tau_1, \ldots, \tau_n) = t_1 , \\
Z_2' &= Z_j \quad \text{when} \quad \tau_j \text{ is the second smallest among the } \{\tau_i\} = t_2 , \\
&\vdots \qquad\quad \vdots \\
Z_n' &= Z_j \quad \text{when} \quad \tau_j = \max(\tau_1, \ldots, \tau_n) = t_n .
\end{aligned}$$

The ambiguity that occurs when two or more of the τ_i's are equal causes no trouble, as the probability of this event is zero. Then

$$\sum_{i=1}^{n} Z_i \exp[-\alpha(t - \tau_i)]_+ = \sum_{i=1}^{n} Z_i' \exp[-\alpha(t - t_i)]_+ ,$$

since the two sums differ only by a random rearrangement. Now since the Z_i are independent, identically distributed, and also independent of the τ_i, it is easily verified that the Z_i' are independent, their distributions coincide with the common distribution of the Z_i, and they are also independent of

the τ_i. Being independent of the τ_i, the families $\{Z_i\}$ and $\{Z_i'\}$ are clearly independent of the t_i.

Since the Z_i' have all the properties required of the X_i, we can take

$$\eta(t) = \sum_{i=1}^{n} Z_i' \exp[-\alpha(t - t_i)]_+ = \sum_{i=1}^{n} Z_i \exp[-\alpha(t - \tau_i)]_+.$$

Let

$$Y_t(i) = Z_i \exp[-\alpha(t - \tau_i)]_+;$$

clearly for fixed t the $Y_t(i)$, $i = 1, \ldots, n$ are independent and identically distributed random variables. Now define

$$\theta_t(s) = \int_0^\infty e^{isy} g_t(y, k)\, dy,$$

the characteristic function of $Y_t(k)$ where $g_t(y; k)$ is the density function of $Y_t(k)$. Since τ_k is uniformly distributed on $(0, t)$ and τ_k and Z_k are independent, we have

$$\begin{aligned}
\int_0^y g_t(u, k)\, du &= \Pr\{Y_t(k) \leq y\} \\
&= \Pr\{Z_k \exp[-\alpha(t - \tau_k)]_+ \leq y\} \\
&= \int_0^t \Pr\{Z_k \exp[-\alpha(t - \tau_k)]_+ \leq y | \tau_k = u\} \frac{du}{t} \\
&= \int_0^t \Pr\{Z_k \leq y e^{\alpha(t-u)}\} \frac{du}{t} \\
&= \frac{1}{t} \int_0^t H(y e^{\alpha(t-u)})\, du,
\end{aligned} \tag{3.2}$$

where H is the cumulative distribution function corresponding to the density h. Differentiating (3.2) gives

$$g_t(y; k) = \frac{1}{t} \int_0^t h(y e^{\alpha(t-u)}) e^{\alpha(t-u)}\, du.$$

Therefore

$$\theta_t(s) = \int_0^\infty e^{isy} g_t(y; k)\, dy = \frac{1}{t} \int_0^t e^{\alpha(t-u)} \left(\int_0^\infty e^{isy} h(y e^{\alpha(t-u)})\, dy \right) du$$

$$= \frac{1}{t} \int_0^t du \int_0^\infty \exp[is(e^{-\alpha(t-u)} z)] h(z)\, dz \qquad \text{(if we make the change of variables } y e^{\alpha(t-u)} = z\text{)}$$

$$= \frac{1}{t}\int_0^t \psi(se^{-\alpha t}e^{\alpha u})\,du \qquad \text{(by the definition of } \psi\text{)}$$

$$= \frac{1}{t}\int_0^t \psi(se^{-\alpha v})\,dv \qquad \text{(if we put } v = t - u\text{)}.$$

It follows that if $r(x; t)$ is the density function of $R(x, t)$ then

$$\varphi_t(w) = \int_0^\infty e^{iwx} r(x; t)\,dx$$

$$= \sum_{n=0}^\infty \left(\int_0^\infty e^{iwx}\frac{d}{dx}\Pr\{\eta(t) \le x | N_t = n\}\,dx\right) e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

[using (3.1)]

$$= \sum_{n=0}^\infty e^{-\lambda t}\frac{(\lambda t)^n}{n!}[\theta_t(w)]^n \qquad \text{(where we use the independence of the } Y_t(k)\text{, given the value of } N_t\text{)}$$

$$= \sum_{n=0}^\infty \frac{e^{-\lambda t}}{n!}\left(\lambda\int_0^t \psi(we^{-\alpha v})\,dv\right)^n$$

$$= \exp -\left\{\lambda\int_0^t [1 - \psi(we^{-\alpha v})]\,dv\right\}.$$

By differentiating with respect to w we may compute moments of $\eta(t)$. For example,

$$E(\eta(t)) = (-i)\frac{d}{dw}\varphi_t(w)\Big|_{w=0} = \lambda(-i)\psi'(0)\cdot\int_0^t e^{-\alpha v}\,dv$$

$$= \lambda\cdot E(X_k)\frac{1 - e^{-\alpha t}}{\alpha}.$$

4: Birth and Death Processes

One of the obvious generalizations of the pure birth processes discussed in Section 1 is to permit $X(t)$ to decrease as well as increase, for example, by the death of members. Thus if at time t the process is in state n it may, after a random waiting time, move to either of the neighboring states $n + 1$ or $n - 1$. The resulting "birth and death processes" can then be regarded

as the continuous time analogs of random walks (Example B, Section 2, Chapter 2).

A. POSTULATES

As in the case of the pure birth processes we assume that $X(t)$ is a Markov process on the states 0, 1, 2, ... and that its transition probabilities $P_{ij}(t)$ are stationary, i.e.,

$$P_{ij}(t) = \Pr\{X(t+s) = j | X(s) = i\}.$$

In addition we assume that the $P_{ij}(t)$ satisfy

1. $P_{i,i+1}(h) = \lambda_i h + o(h)$ as $h \downarrow 0$, $i \geq 0$
2. $P_{i,i-1}(h) = \mu_i h + o(h)$ as $h \downarrow 0$, $i \geq 1$
3. $P_{i,i}(h) = 1 - (\lambda_i + \mu_i)h + o(h)$ as $h \downarrow 0$, $i \geq 0$
4. $P_{ij}(0) = \delta_{ij}$.
5. $\mu_0 = 0$, $\lambda_0 > 0$, $\mu_i, \lambda_i > 0$, $i = 1, 2, \ldots$.

The $o(h)$ in each case may depend on i. The matrix

$$\mathbf{A} = \begin{Vmatrix} -\lambda_0 & \lambda_0 & 0 & 0 \ldots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 \ldots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 \ldots \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) \ldots \\ \vdots & \vdots & \vdots & \vdots \end{Vmatrix} \tag{4.1}$$

is called the *infinitesimal generator* of the process. The parameters λ_i and μ_i are called, respectively, the infinitesimal birth and death rates. In Postulates 1 and 2 we are assuming that if the process starts in state i, then in a small interval of time the probabilities of the population increasing or decreasing by 1 are essentially proportional to the length of the interval. Sometimes a transition from zero to some ignored state is allowed (see Section 7).

Since the $P_{ij}(t)$ are probabilities we have $P_{ij}(t) \geq 0$ and

$$\sum_{j=0}^{\infty} P_{ij}(t) = 1. \tag{4.2}$$

Using the Markovian property of the process we may also derive the Chapman–Kolmogorov equation

$$P_{ij}(t+s) = \sum_{k=0}^{\infty} P_{ik}(t) P_{kj}(s). \tag{4.3}$$

This equation states that in order to move from state i to state j in time $t+s$, $X(t)$ moves to some state k in time t and then from k to j in the remaining time s. This is the continuous time analog of formula (3.2) of Chapter 2.

So far we have mentioned only the transition probabilities $P_{ij}(t)$. In order to obtain the probability that $X(t) = n$ we must specify where the process starts or more generally the probability distribution for the initial state. We then have

$$\Pr(X(t) = n) = \sum_{i=0}^{\infty} q_i P_{in}(t),$$

where

$$q_i = \Pr\{X(0) = i\}.$$

B. WAITING TIMES

With the aid of the above assumptions we may calculate the distribution of the random variable T_i which is the waiting time of $X(t)$ in state i; that is, given the process in state i, what is the distribution of the time T_i until it first leaves state i? Letting

$$\Pr(T_i \geq t) = G_i(t)$$

it follows easily by the Markov property that as $h \downarrow 0$

$$G_i(t+h) = G_i(t)G_i(h) = G_i(t)(P_{ii}(h) + o(h)) = G_i(t)[1 - (\lambda_i + \mu_i)h] + o(h)$$

or

$$\frac{G_i(t+h) - G_i(t)}{h} = -(\lambda_i + \mu_i)G_i(t) + o(1),$$

so that

$$G_i'(t) = -(\lambda_i + \mu_i)G_i(t). \tag{4.4}$$

If we use the condition $G_i(0) = 1$ the solution of this equation is

$$G_i(t) = \exp[-(\lambda_i + \mu_i)t],$$

i.e., T_i follows an exponential distribution with mean $(\lambda_i + \mu_i)^{-1}$. The proof presented above is not quite complete, since we have used the intuitive relationship

$$G_i(h) = P_{ii}(h) + o(h)$$

without a formal proof. A rigorous proof of (4.4) will be given in Chapter 14 of Volume II.

According to Postulates 1 and 2, during a time duration of length h a transition occurs from state i to $i+1$ with probability $\lambda_i h + o(h)$ and from state i to $i-1$ with probability $\mu_i h + o(h)$. It follows intuitively that, given that a transition occurs at time t, the probability this transition is to state $i+1$ is $\lambda_i(\mu_i + \lambda_i)^{-1}$ and to state $i-1$ is $\mu_i(\mu_i + \lambda_i)^{-1}$. The rigorous demonstration of this result is beyond the scope of this book; however, comments on this problem and its intrinsic subtleties will be given later (see Chapter 14 of Volume II).

The description of the motion of $X(t)$ is as follows: The process sojourns in a given state i for a random length of time whose distribution function is an exponential distribution with parameter $(\lambda_i + \mu_i)$. When leaving state i the process enters either state $i+1$ or state $i-1$ with probabilities $\lambda_i(\mu_i+\lambda_i)^{-1}$ and $\mu_i(\mu_i+\lambda_i)^{-1}$, respectively. The motion is analogous to that of a random walk except that transitions occur at random times rather than at fixed time periods.

The traditional procedure for constructing birth and death processes is to prescribe the birth and death parameters $\{\lambda_i, \mu_i\}_{i=0}^{\infty}$ and to build the path structure by utilizing the above description concerning the waiting times and the conditional transition probabilities of the various states. We determine realizations of the process as follows. Suppose $X(0) = i$; the particle spends a random length of time (exponentially distributed with parameter $\lambda_i + \mu_i$) in state i and subsequently moves with probability $\lambda_i/(\mu_i+\lambda_i)$ to state $i+1$ and with probability $\mu_i/(\lambda_i+\mu_i)$ to state $i-1$. Next the particle sojourns a random length of time in the new state and then moves to one of its neighboring states, and so on. More specifically, we observe a value t_1 from the exponential distribution with parameter $(\mu_i+\lambda_i)$ which fixes the initial sojourn time in state i. Then we toss a coin with probability of heads $p_i = \lambda_i/(\lambda_i+\mu_i)$. If heads (tails) appears we move the particle to state $i+1$ $(i-1)$. In state $i+1$ we observe a value t_2 from the exponential distribution with parameter $(\lambda_{i+1}+\mu_{i+1})$ which fixes the sojourn time in the second state visited. If the particle at the first transition enters state $i-1$, the subsequent sojourn time t_2' is an observation from the exponential distribution with parameter $(\lambda_{i-1}+\mu_{i-1})$. After completing the second wait, a binomial trial is performed which chooses the next state to be visited, etc.

A typical outcome of these sampling procedures determines a realization. Its form could be

$$X(t) = \begin{cases} i, & 0 < t < t_1, \\ i+1, & t_1 < t < t_1 + t_2, \\ i, & t_1 + t_2 < t < t_1 + t_2 + t_3, \\ \vdots & \\ \cdot & \end{cases}$$

Thus by sampling from exponential and binomial distributions appropriately, we construct typical sample paths of the process. Now it is possible to assign to this set of paths (realizations of the process) a probability measure in a consistent way so that $P_{ij}(t)$ is determined satisfying (4.2), (4.3), and the infinitesimal relations (p. 132). This result is rather deep and its rigorous discussion is beyond the level of this book. The process obtained in this manner is called the minimal process associated with the matrix **A**.

The above construction of the minimal process is fundamental since the infinitesimal parameters need not determine a unique stochastic process obeying (4.2), (4.3), and the postulates of page 132. In fact there could be several Markov processes which possess the same infinitesimal generator. This whole subject is rather complicated and we refer the reader to Chung.† In the special case of birth and death processes, a sufficient condition that there exists a unique Markov process with transition probability function $P_{ij}(t)$ for which the infinitesimal relations, (4.2) and (4.3) hold is that

$$\sum_{n=0}^{\infty} \pi_n \sum_{k=0}^{n} \frac{1}{\lambda_k \pi_k} = \infty, \tag{4.5}$$

where

$$\pi_0 = 1, \qquad \pi_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}, \qquad n = 1, 2, \ldots .$$

In most practical examples of birth and death processes the condition (4.5) is met and the birth and death process associated with the prescribed parameters is uniquely determined.

5: *Differential Equations of Birth and Death Processes*

As in the case of the pure birth and Poisson processes the transition probabilities $P_{ij}(t)$ satisfy a system of differential equations known as the *backward Kolmogorov differential equations.* These are given by

$$\begin{aligned} P'_{0j}(t) &= -\lambda_0 P_{0j}(t) + \lambda_0 P_{1j}(t), \\ P'_{ij}(t) &= \mu_i P_{i-1,j}(t) - (\lambda_i + \mu_i) P_{ij}(t) + \lambda_i P_{i+1,j}(t), \qquad i \geq 1, \end{aligned} \tag{5.1}$$

and the boundary condition $P_{ij}(0) = \delta_{ij}$.

To derive these we have from Eq. (4.3)

$$\begin{aligned} P_{ij}(t+h) &= \sum_{k=0}^{\infty} P_{ik}(h) P_{kj}(t) \qquad (5.2) \\ &= P_{i,i-1}(h) P_{i-1,j}(t) + P_{i,i}(h) P_{ij}(t) + P_{i,i+1}(h) P_{i+1,j}(t) \\ &\qquad + \textstyle\sum'_k P_{ik}(h) P_{kj}(t), \end{aligned}$$

†K. L. Chung; "Markov chains with stationary transition probabilities." Springer-Verlag, Berlin, 1960.

where the last summation is over all $k \neq i-1, i, i+1$. Using postulates 1, 2, and 3 of Section 4 we obtain

$$\begin{aligned}\textstyle\sum_k' P_{ik}(h)P_{kj}(t) &\leq \textstyle\sum_k' P_{ik}(h) \\ &= 1-[P_{i,i}(h)+P_{i,i-1}(h)+P_{i,i+1}(h)] \\ &= 1-[1-(\lambda_i+\mu_i)h+o(h)+\mu_i h+o(h)+\lambda_i h+o(h)] \\ &= o(h),\end{aligned}$$

so that

$$P_{ij}(t+h) = \mu_i h P_{i-1,j}(t) + (1-(\lambda_i+\mu_i)h)P_{ij}(t) + \lambda_i h P_{i+1,j}(t) + o(h).$$

Transposing the term $P_{ij}(t)$ to the left-hand side and dividing the equation by h, we obtain, after letting $h \downarrow 0$,

$$P'_{ij}(t) = \mu_i P_{i-1,j}(t) - (\lambda_i+\mu_i)P_{ij}(t) + \lambda_i P_{i+1,j}(t).$$

The above analysis is a special case of the derivation of the backward differential equations given in Chapter 14.

The backward equations are deduced by decomposing the time interval $(0, t+h)$, where h is positive and small, into the two periods

$$(0, h), \quad (h, t+h),$$

and examining the transitions in each period separately.

The equations (5.1) feature the initial state as the variable.

A different result arises from splitting the time interval $(0, t+h)$ into the two periods.

$$(0, t), \quad (t, t+h)$$

and adapting the preceding analysis. In this viewpoint, under more stringent conditions, we can derive a further system of differential equations

$$\begin{aligned}P'_{i0}(t) &= -\lambda_0 P_{i,0}(t) + \mu_1 P_{i,1}(t), \\ P'_{ij}(t) &= \lambda_{j-1}P_{i,j-1}(t) - (\lambda_j+\mu_j)P_{ij}(t) + \mu_{j+1}P_{i,j+1}(t), \qquad j \geq 1,\end{aligned} \tag{5.3}$$

with the same initial condition $P_{ij}(0) = \delta_{ij}$. These are known as the *forward Kolmogorov differential equations*. To do this we interchange t and h in Eq. (5.2) and under *stronger assumptions* in addition to Postulates 1, 2, and 3 it can be shown that the last term is again $o(h)$. The remainder of the argument is the same as before. The usefulness of the differential equations will become apparent in the examples which we study below.

A sufficient condition that (5.3) hold is that $(P_{kj}(h))/h = o(1)$ for $k \neq j$, $j-1, j+1$ where the $o(1)$ term apart from tending to zero is uniformly bounded with respect to k for fixed j as $h \to 0$. In this case it can easily be proved that $\sum_k' P_{ik}(t)P_{kj}(h) = o(h)$.

Before proceeding with some examples we discuss briefly the behavior of $P_{ij}(t)$ as t becomes large. It can be proved that the limits

$$\lim_{t\to\infty} P_{ij}(t) = p_j \tag{5.4}$$

exist and are independent of the initial state i and also that they satisfy the equations

$$\begin{aligned} -\lambda_0 p_0 + \mu_1 p_1 &= 0, \\ \lambda_{j-1}p_{j-1} - (\lambda_j + \mu_j)p_j + \mu_{j+1}p_{j+1} &= 0, \qquad j \geq 1. \end{aligned} \tag{5.5}$$

These equations are simply (5.3) where the left-hand side is set equal to zero. The convergence of $\sum_j p_j$ follows since $\sum_j P_{ij}(t) = 1$. If $\sum_j p_j = 1$ then the sequence $\{p_j\}$ is called a "stationary distribution." The reason for this is that p_j also satisfy

$$p_j = \sum_{i=0}^{\infty} p_i P_{ij}(t), \tag{5.6}$$

which tells us that if the process starts in state i with probability p_i then at any given time t it will be in state i with the same probability p_i. The proof of (5.6) follows from (4.3) and (5.4) if we let $t\uparrow\infty$ and use the fact that $\sum_{i=0}^{\infty} p_i < \infty$. The solution to (5.5) is obtained by induction. Letting

$$\pi_0 = 1, \qquad \pi_j = \frac{\lambda_0\lambda_1\cdot\cdots\cdot\lambda_{j-1}}{\mu_1\mu_2\cdot\cdots\cdot\mu_j}, \quad j \geq 1,$$

we have $p_1 = \lambda_0\,\mu_1^{-1}p_0 = \pi_1 p_0$. Assuming that $p_k = \pi_k p_0$ for $k = 1, \ldots, j$ we obtain

$$\begin{aligned} \mu_{j+1}p_{j+1} &= (\lambda_j + \mu_j)\pi_j p_0 - \lambda_{j-1}\pi_{j-1}p_0 \\ &= \lambda_j\pi_j p_0 + (\mu_j\pi_j - \lambda_{j-1}\pi_{j-1})p_0 \\ &= \lambda_j\pi_j p_0, \end{aligned}$$

and finally

$$p_{j+1} = \pi_{j+1}p_0.$$

In order that the sequence $\{p_j\}$ define a distribution we must have $\sum p_j = 1$. If $\sum\pi_k < \infty$ we see in this case that

$$p_j = \frac{\pi_j}{\sum\pi_k}, \qquad j = 0, 1, 2, \ldots.$$

If $\sum\pi_k = \infty$ then necessarily $p_0 = 0$ and the p_j are all zero. Hence, we do not have a limiting stationary distribution.

6: *Examples of Birth and Death Processes*

Example 1. *Linear Growth with Immigration.* A birth and death process is called a linear growth process if $\lambda_n = \lambda n + a$ and $\mu_n = \mu n$ with $\lambda > 0$, $\mu > 0$, and $a > 0$. Such processes occur naturally in the study of biological

reproduction and population growth. If the state n describes the current population size, then the average instantaneous rate of growth is $\lambda n + a$. Similarly, the probability of the state of the process decreasing by one after the elapse of a small duration of time is $\mu nt + o(t)$. The factor λn represents the natural growth of the population owing to its current size while the second factor a may be interpreted as the infinitesimal rate of increase of the population due to an external source such as immigration. The component μn which gives the mean infinitesimal death rate of the present population possesses the obvious interpretation.

If we substitute the above values of λ_n and μ_n in (5.3) we obtain

$$\begin{aligned} P'_{i0}(t) &= -aP_{i0}(t) + \mu P_{i1}(t), \\ P'_{ij}(t) &= (\lambda(j-1) + a)P_{i,j-1}(t) - ((\lambda+\mu)j + a)P_{ij}(t) \\ &\qquad + \mu(j+1)P_{i,j+1}(t), \qquad j \geq 1. \end{aligned}$$

Now if we multiply the jth equation by j and sum, it follows that the expected value

$$EX(t) = M(t) = \sum_{j=1}^{\infty} jP_{ij}(t)$$

satisfies the differential equation

$$M'(t) = a + (\lambda - \mu)M(t),$$

with initial condition $M(0) = i$, if $X(0) = i$. The solution of this equation is

$$M(t) = at + i \qquad \text{if} \quad \lambda = \mu,$$

and

$$M(t) = \frac{a}{\lambda - \mu}\{e^{(\lambda-\mu)t} - 1\} + ie^{(\lambda-\mu)t} \qquad \text{if} \quad \lambda \neq \mu.$$

The second moment or variance may be calculated in a similar way. It is interesting to note that $M(t) \to \infty$ as $t \to \infty$ if $\lambda \geq \mu$, while if $\lambda < \mu$ the mean population size for large t is approximately

$$\frac{a}{\mu - \lambda}.$$

Example 2. *Queueing.* A queueing process is a process in which customers arrive at some designated place where a service of some kind is being rendered, for example, at the teller's window in a bank or beside the cashier at a supermarket. It is assumed that the time between arrivals, or inter-arrival time, and the time that is spent in providing service for a given customer are governed by probabilistic laws. The length of the queue at a given time t is represented by $X(t)$.

If we let $\lambda_i = \lambda$ for all i in the general birth and death process, the resulting process is a special simple case of a continuous time queueing process. The state of the system is then interpreted as the length of a queue for which the times between arrivals of the customers are independent random variables with an exponential distribution of parameter λ and for which the duration of the service time of the current customer is a random variable with an exponential distribution whose parameter, μ_n, may depend on the length of the line. At the completion of each service the line decreases by 1 and with each new arrival the line increases by 1. The classical case of a single-server queue corresponds to $\mu_i = \mu$, $i \geq 1$, i.e., each service follows the same exponential distribution with parameter μ independent of the length of the waiting line.

The classical telephone trunking model can be formulated as a queueing birth and death process with infinitely many servers, each of whose service time distribution has the same parameter μ, so that $\mu_i = i\mu$, $i \geq 1$. The rationale underlying this specification goes as follows: Suppose the queue consists of i individual customers; then since the number of servers is unlimited each customer is simultaneously receiving service. Now the length of service of each is independent of the others and distributed exponentially with parameter μ. It follows that the probability distribution of the time until at least one of the customers completes service (i.e., the length of time until the waiting line decreases by 1) is also exponentially distributed, but is now of parameter $i\mu$ (the student should prove this).

Besides the two special cases mentioned above it is possible to consider numerous other queueing models by appropriate specifications of the parameters μ_k. For example, a queue with n servers, each of whose service time has an exponential distribution with the same parameter μ, would correspond to $\mu_k = k\mu$ for $1 \leq k \leq n$, $\mu_i = n\mu$ for $i \geq n$.

For the single-server process with $\lambda < \mu$ the stationary distribution is easily calculated. In fact, in this case

$$\pi_n = \frac{\lambda_0 \lambda_1 \cdot \cdots \cdot \lambda_{n-1}}{\mu_1 \mu_2 \cdot \cdots \cdot \mu_n} = \left(\frac{\lambda}{\mu}\right)^n,$$

which, when normalized, results in

$$p_n = \frac{\mu - \lambda}{\mu} \left(\frac{\lambda}{\mu}\right)^n, \qquad n \geq 0,$$

i.e., a geometric distribution with mean $\lambda(\mu - \lambda)^{-1}$.

This gives us the answer to many problems involving stationarity. If the process has been going on a long time and $\lambda < \mu$, the probability of being served immediately upon arrival is

$$p_0 = \left(1 - \frac{\lambda}{\mu}\right).$$

We can also calculate the distribution of waiting time in the stationary case when $\lambda < \mu$. If an arriving customer finds n people in front of him, his total waiting time T, including his own service time, is the sum of the service times of himself and those ahead, all distributed exponentially with parameter μ, and since the service times are independent of the queue size, T has a gamma distribution of order $n+1$ with scale parameter μ

$$\Pr\{T \leq t | n \text{ ahead}\} = \int_0^t \frac{\mu^{n+1} \tau^n e^{-\mu\tau}}{\Gamma(n+1)} d\tau. \tag{6.1}$$

By the law of total probabilities, we have

$$\Pr\{T \leq t\} = \sum_{n=0}^{\infty} \Pr\{T \leq t | n \text{ ahead}\} \cdot \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right),$$

since $(\lambda/\mu)^n(1 - \lambda/\mu)$ is the probability that in the stationary case a customer on arrival will find n ahead in line. Now, substituting from (6.1), we obtain

$$\begin{aligned}
\Pr\{T \leq t\} &= \sum_{n=0}^{\infty} \int_0^t \frac{\mu^{n+1} \tau^n e^{-\mu\tau}}{\Gamma(n+1)} \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) d\tau \\
&= \int_0^t \mu e^{-\mu\tau} \left(1 - \frac{\lambda}{\mu}\right) \sum_{n=0}^{\infty} \frac{\tau^n \lambda^n}{\Gamma(n+1)} d\tau \\
&= \int_0^t \left(1 - \frac{\lambda}{\mu}\right) \mu \exp\left\{-\tau\mu \left(1 - \frac{\lambda}{\mu}\right)\right\} d\tau \\
&= 1 - \exp\left[-t\mu\left(1 - \frac{\lambda}{\mu}\right)\right],
\end{aligned}$$

which is also an exponential distribution.

If we wish to answer nonstationary questions, it is essential to determine $P_{ij}(t)$ for all t. This is a much harder problem but it has been solved. The details of this solution are beyond the scope of this book and we refer the interested student to any of the advanced books on queuing theory listed in the references.

For the telephone trunking problem with $\lambda_n = \lambda$ and $\mu_n = n\mu$ it is easily seen that

$$p_n = \frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!},$$

which is the familiar Poisson distribution with mean λ/μ. As in Example 1, it is easy to show that

$$M(t) = \sum_{j=0}^{\infty} jP_{ij}(t)$$

satisfies the equation

$$M'(t) = \lambda - \mu M(t),$$

whose solution is

$$M(t) = \frac{\lambda}{\mu}(1 - e^{-\mu t}) + ie^{-\mu t}.$$

If we let $t \to \infty$, then $M(t) \to \lambda/\mu$, which is the mean value of the stationary distribution given above.

Example 3. *Some Genetic Models.* Consider a population consisting of N individuals which are either of gene type a or gene type A. The state of the process $X(t)$ represents the number of a-individuals at time t. We assume that the probability that the state changes during the time interval $(t, t+h)$ is $\lambda h + o(h)$ independent of the values of $X(t)$ and that the probability of two or more changes occurring in a time interval h is $o(h)$.

The changes in the population structure are effected as follows. An individual is to be replaced by another chosen randomly from the population; i.e., if $X(t) = j$ then an a-type is selected to be replaced with probability j/N and an A-type with probability $1 - j/N$. We refer to this stage as death. Next, birth takes place by the following rule. Another selection is made randomly from the population to determine the type of the new individual replacing the one that died. The model introduces mutation pressures which admit the possibility that the type of the new individual may be altered upon birth. Specifically, let γ_1 denote the probability that an a-type mutates to an A-type and let γ_2 denote the probability of an A-type mutating to an a-type.

The probability that the new individual added to the population is of type a is

$$\frac{j}{N}(1 - \gamma_1) + \left(1 - \frac{j}{N}\right)\gamma_2. \tag{6.2}$$

We deduce this formula as follows: The probability that we select an a-type

and no mutation occurs is $(j/N)(1-\gamma_1)$. Moreover, the final type may be an a-type if we select an A-type which subsequently mutates into an a-type. The probability of this contingency is $(1-j/N)\gamma_2$. The combination of these two possibilities gives (6.2).

We assert that the conditional probability that $X(t+)-X(t)=1$, when a change of state occurs, is

$$\left(1-\frac{j}{N}\right)\left[\frac{j}{N}(1-\gamma_1)+\left(1-\frac{j}{N}\right)\gamma_2\right], \qquad \text{where} \quad X(t)=j. \tag{6.3}$$

In fact, the a-type population size can increase only if an A-type dies (is replaced). This probability is $1-(j/N)$. The second factor is the probability that the new individual is of type a as in (6.2).

In a similar way we find that the conditional probability that $X(t+)-X(t)=-1$ when a change of state occurs is

$$\frac{j}{N}\left[\left(1-\frac{j}{N}\right)(1-\gamma_2)+\frac{j}{N}\gamma_1\right], \qquad \text{where} \quad X(t)=j.$$

The stochastic process described is thus a birth and death process with a finite number of states† whose infinitesimal birth and death rates are

$$\lambda_j=\lambda\left(1-\frac{j}{N}\right)\left[\frac{j}{N}(1-\gamma_1)+\left(1-\frac{j}{N}\right)\gamma_2\right]$$

and

$$\mu_j=\lambda\frac{j}{N}\left[\frac{j}{N}\gamma_1+\left(1-\frac{j}{N}\right)(1-\gamma_2)\right],$$

respectively corresponding to an a-type population size j, $0\leq j\leq N$.

Although these parameters seem rather complicated, it is interesting to see what happens to the stationary measure $\{\pi_k\}_{k=0}^N$ if we let the population size $N\to\infty$ and the probabilities of mutation per individual γ_1 and γ_2 tend to zero in such a way that $\gamma_1 N\to\kappa_1$ and $\gamma_2 N\to\kappa_2$, where $0<\kappa_1$, $\kappa_2<\infty$. At the same time we shall transform the state of the process to the interval [0, 1] by defining new states j/N, i.e., the fraction of

† The definition of birth and death processes was given for an infinite number of states. The adjustments in the definitions and analyses for the case of a finite number of states is straightforward and even simpler and left to the reader.

a-types in the population. To examine the stationary density at a fixed fraction x, where $0 < x < 1$, we shall evaluate π_k as $k \to \infty$ in such a way that $k = [xN]$, where $[xN]$ is the greatest integer less than or equal to xN.

Keeping these relations in mind we write

$$\lambda_j = \frac{\lambda(N-j)}{N^2}(1-\gamma_1-\gamma_2)j\left(1+\frac{a}{j}\right), \qquad \text{where} \quad a = \frac{N\gamma_2}{1-\gamma_1-\gamma_2},$$

and

$$\mu_j = \frac{\lambda(N-j)}{N^2}(1-\gamma_1-\gamma_2)j\left(1+\frac{b}{N-j}\right), \qquad \text{where} \quad b = \frac{N\gamma_1}{1-\gamma_1-\gamma_2}.$$

Then

$$\begin{aligned}\log \pi_k &= \sum_{j=0}^{k-1} \log \lambda_j - \sum_{j=1}^{k} \log \mu_j \\ &= \sum_{j=1}^{k-1} \log\left(1+\frac{a}{j}\right) - \sum_{j=1}^{k-1} \log\left(1+\frac{b}{N-j}\right) + \log Na \\ &\quad - \log(N-k)k\left(1+\frac{b}{N-k}\right).\end{aligned}$$

Now using the expression

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots, \qquad |x| < 1,$$

it is possible to write

$$\sum_{j=1}^{k-1} \log\left(1+\frac{a}{j}\right) = a\sum_{j=1}^{k-1}\frac{1}{j} + c_k,$$

where c_k approaches a finite limit as $k \to \infty$. Therefore, using the relation

$$\sum_{j=1}^{k-1}\frac{1}{j} \sim \log k \qquad \text{as} \quad k \to \infty,$$

we have

$$\sum_{j=1}^{k-1} \log\left(1+\frac{a}{j}\right) \sim \log k^a + c_k \qquad \text{as} \quad k \to \infty.$$

In a similar way we obtain

$$\sum_{j=1}^{k-1} \log\left(1+\frac{b}{N-j}\right) \sim \log \frac{N^b}{(N-k)^b} + d_k \qquad \text{as} \quad k \to \infty,$$

where d_k approaches a finite limit as $k \to \infty$. Using the above relations we have

$$\log \pi_k \sim \log\left(C_k \frac{k^a(N-k)^b Na}{N^b(N-k)k}\right) \qquad \text{as} \quad k \to \infty, \tag{6.4}$$

where $\log C_k = c_k + d_k$, which approaches a limit, say C, as $k \to \infty$. Notice that $a \to \kappa_2$ and $b \to \kappa_1$ as $N \to \infty$. Since $k = [Nx]$ we have, for $N \to \infty$,

$$\pi_k \sim C\kappa_2 N^{\kappa_2 - 1} x^{\kappa_2 - 1}(1-x)^{\kappa_1 - 1}.$$

Now from (6.4) we have

$$\pi_k \sim aC_k k^{a-1}\left(1 - \frac{k}{N}\right)^{b-1}.$$

Therefore

$$\frac{1}{N^a}\sum_{k=0}^{N-1} \pi_k \sim \frac{a}{N}\sum_{k=0}^{N-1} C_k \left(\frac{k}{N}\right)^{a-1}\left(1-\frac{k}{N}\right)^{b-1}.$$

Since $C_k \to C$ as k tends to ∞ we recognize the right-hand side as the Riemann sum approximation of

$$\kappa_2 C \int_0^1 x^{\kappa_2 - 1}(1-x)^{\kappa_1 - 1}\, dx.$$

Thus

$$\sum_{i=0}^{N} \pi_i \sim N^{\kappa_2}\kappa_2 C \int_0^1 x^{\kappa_2 - 1}(1-x)^{\kappa_1 - 1}\, dx,$$

so that the resulting density on $[0, 1]$ is

$$\frac{\pi_k}{\sum \pi_i} \sim \frac{1}{N}\frac{x^{\kappa_2 - 1}(1-x)^{\kappa_1 - 1}}{\int_0^1 x^{\kappa_2 - 1}(1-x)^{\kappa_1 - 1}\, dx} = \frac{x^{\kappa_2 - 1}(1-x)^{\kappa_1 - 1}\, dx}{\int_0^1 x^{\kappa_2 - 1}(1-x)^{\kappa_1 - 1}\, dx},$$

since $dx \sim 1/N$. This is a beta distribution with parameters κ_1 and κ_2.

Example 4. *Logistic Process.* Suppose we consider a population whose size $X(t)$ ranges between two fixed integers N_1 and N_2 ($N_1 < N_2$) for all $t \geq 0$. We assume that the birth and death rates per individual at time t are given by

$$\lambda = \alpha(N_2 - X(t)) \qquad \text{and} \qquad \mu = \beta(X(t) - N_1),$$

and that the individual members of the population act independently of each other. The resulting birth and death rates for the population then become

$$\lambda_n = \alpha n(N_2 - n) \qquad \text{and} \qquad \mu_n = \beta n(n - N_1).$$

To see this we observe that if the population size $X(t)$ is n, then each of the n individuals has an infinitesimal birth rate λ so that $\lambda_n = \alpha n(N_2 - n)$. The same rationale applies in the interpretation of the μ_n.

Under such conditions one would expect the process to fluctuate between the two constants N_1 and N_2, since, for example, if $X(t)$ is near N_2 the death rate is high and the birth rate low and then $X(t)$ will tend toward N_1. Ultimately the process should display stationary fluctuations between the two limits N_1 and N_2.

The stationary distribution in this case is

$$p_{N_1+m} = \frac{c}{N_1+m}\binom{N_2-N_1}{m}\left(\frac{\alpha}{\beta}\right)^m, \qquad m = 0, 1, 2, \ldots, N_2 - N_1,$$

where c is an appropriate constant determined so that $\sum_m p_{N_1+m} = 1$. To see this we observe that

$$\begin{aligned}\pi_{N_1+m} &= \frac{\lambda_{N_1}\lambda_{N_1+1}\cdots\lambda_{N_1+m}}{\mu_{N_1+1}\mu_{N_1+2}\cdots\mu_{N_1+m}} \\ &= \frac{\alpha^m N_1(N_1+1)\cdot\cdots\cdot(N_1+m-1)(N_2-N_1)\cdot\cdots\cdot(N_2-N_1-m+1)}{\beta^m(N_1+1)\cdot\cdots\cdot(N_1+m)\,m!} \\ &= \frac{N_1}{N_1+m}\binom{N_2-N_1}{m}\left(\frac{\alpha}{\beta}\right)^m.\end{aligned}$$

7: *Birth and Death Processes with Absorbing States*

It is of importance to treat the case of birth and death processes where $\lambda_0 = 0$. This stipulation converts the zero state into an absorbing state. When a transition occurs from state 1, the particle moves to state 2 with probability $\lambda_1/(\lambda_1 + \mu_1)$ or it is trapped in state 0 with probability $\mu_1/(\lambda_1 + \mu_1)$. An important example of a birth and death process where 0 acts as an absorbing state is the linear growth process without immigration (cf. Example 1 of Section 6). In this case $\lambda_n = n\lambda$ and $\mu_n = n\mu$. Since growth of the population results exclusively from the existing population it is clear that when the population size becomes 0 it remains zero thereafter, i.e., 0 is an absorbing state.

A. PROBABILITY OF ABSORPTION INTO STATE 0

It is of interest to compute the probability of absorption into state 0 starting from state i ($i \geq 1$). This is not, *a priori*, a certain event since conceivably the particle (i.e., state variable) may wander forever among the states $(1, 2, \ldots)$ or possibly drift to infinity.

Let u_i $(i = 1, 2, \ldots)$ denote the probability of absorption into state 0 from the initial state i. We can write a recursion formula for u_i by considering the possible states after the first transition. We know that the first transition entails the movements

$$i \to i+1 \qquad \text{with probability} \quad \frac{\lambda_i}{\mu_i + \lambda_i},$$

$$i \to i-1 \qquad \text{with probability} \quad \frac{\mu_i}{\mu_i + \lambda_i}.$$

We directly obtain

$$u_i = \frac{\lambda_i}{\mu_i + \lambda_i} u_{i+1} + \frac{\mu_i}{\mu_i + \lambda_i} u_{i-1}, \qquad i \geq 1, \tag{7.1}$$

where $u_0 = 1$. Another method for deriving (7.1) is to consider the "embedded random walk" associated with a given birth and death process. Specifically we examine the birth and death process only at the transition times. The discrete time Markov chain generated in this manner is denoted by $\{Y_n\}_{n=0}^{\infty}$, where $Y_0 = X_0$ is the initial state and Y_n $(n \geq 1)$ is the state at the nth transition. Obviously, the transition probability matrix has the form

$$\mathbf{P} = \begin{Vmatrix} 1 & 0 & 0 & 0 & \ldots \\ q_1 & 0 & p_1 & 0 & \ldots \\ 0 & q_2 & 0 & p_2 & \ldots \\ \vdots & \vdots & & & \end{Vmatrix},$$

where

$$p_i = \frac{\lambda_i}{\lambda_i + \mu_i} = 1 - q_i \qquad (i \geq 1).$$

The probability of absorption into state 0 for the embedded random walk is the same as for the birth and death process since both processes execute the same transitions.

We turn to the task of solving (7.1) subject to the conditions $u_0 = 1$ and $0 \leq u_i \leq 1$ $(i \geq 1)$. Rewriting (7.1) we have

$$(u_{i+1} - u_i) = \frac{\mu_i}{\lambda_i}(u_i - u_{i-1}), \qquad i \geq 1.$$

Defining $v_i = u_{i+1} - u_i$, we obtain

$$v_i = \frac{\mu_i}{\lambda_i} v_{i-1}, \qquad i \geq 1.$$

Iteration of the last relation yields the formula

$$u_{i+1} - u_i = v_i = \left(\prod_{j=1}^{i} \frac{\mu_j}{\lambda_j}\right) v_0, \qquad i \geq 1.$$

Summing these equations from $i = 1$ to $i = m$ we have

$$u_{m+1} - u_1 = (u_1 - 1) \sum_{i=1}^{m} \left(\prod_{j=1}^{i} \frac{\mu_j}{\lambda_j}\right), \qquad m \geq 1. \tag{7.2}$$

Since u_m, by its very meaning, is bounded by 1 we see that if

$$\sum_{i=1}^{\infty} \left(\prod_{j=1}^{i} \frac{\mu_j}{\lambda_j}\right) = \infty \tag{7.3}$$

then necessarily $u_1 = 1$ and $u_m = 1$ for all $m \geq 2$. In other words, if (7.3) holds then ultimate absorption into state 0 is certain from any initial state. Suppose $0 < u_1 < 1$; then, of course,

$$\sum_{i=1}^{\infty} \left(\prod_{j=1}^{i} \frac{\mu_j}{\lambda_j}\right) < \infty.$$

Obviously, u_m is decreasing in m since passing from state m to state 0 requires entering the intermediate states in the intervening time. Furthermore, we claim that $u_m \to 0$ as $m \to \infty$. If we assume the contrary i.e., $u_m \geq \alpha > 0$ $(m \geq 1)$, a simple probabilistic argument implies that $u_m \equiv 1$ $(m \geq 1)$. (The student should supply a formal proof.) Now letting $m \to \infty$ in (7.2) permits us to solve for u_1; thus

$$u_1 = \frac{\sum_{i=1}^{\infty} \left(\prod_{j=1}^{i} \mu_j/\lambda_j\right)}{1 + \sum_{i=1}^{\infty} \left(\prod_{j=1}^{i} \mu_j/\lambda_j\right)}.$$

and in addition we have

$$u_{m+1} = \frac{\sum_{i=m+1}^{\infty} \left(\prod_{j=1}^{i} \mu_j/\lambda_j\right)}{1 + \sum_{i=1}^{\infty} \left(\prod_{j=1}^{i} \mu_j/\lambda_j\right)}, \qquad m \geq 1.$$

In the special example of a linear growth birth and death process where $\mu_n = n\mu$ and $\lambda_n = n\lambda$, a direct calculation yields

$$\begin{aligned} u_m &= \left(\frac{\mu}{\lambda}\right)^m \quad \text{when} \quad \mu < \lambda \\ u_m &= 1 \quad \text{when} \quad \mu \geq \lambda \end{aligned} \qquad (m \geq 1).$$

B. MEAN TIME UNTIL ABSORPTION

Consider the problem of determining the mean time until absorption, starting from state m.

We assume that condition (7.3) holds so that absorption is certain. Notice that we cannot reduce our problem to a consideration of the embedded random walk since the actual time spent in each state is relevant for the calculation of the mean absorption time.

Let ω_i be the mean absorption time starting from state i (this could be infinite). Considering the possible states following the first transition and recalling the fact that the mean waiting time in state i is $(\lambda_i + \mu_i)^{-1}$ (it is actually exponentially distributed with parameter $\lambda_i + \mu_i$), we deduce the recursion relation

$$\omega_i = \frac{1}{\lambda_i + \mu_i} + \frac{\lambda_i}{\lambda_i + \mu_i}\omega_{i+1} + \frac{\mu_i}{\lambda_i + \mu_i}\omega_{i-1}, \qquad i \geq 1, \tag{7.4}$$

where by convention $\omega_0 = 0$. Letting $z_i = \omega_i - \omega_{i+1}$ and rearranging (7.4) leads to

$$z_i = \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i} z_{i-1}, \qquad i \geq 1. \tag{7.5}$$

Iterating this relation gives

$$z_m = \frac{1}{\lambda_m} + \frac{\mu_m}{\lambda_m}\frac{1}{\lambda_{m-1}} + \frac{\mu_m \mu_{m-1}}{\lambda_m \lambda_{m-1}} z_{m-2}$$

and finally

$$z_m = \sum_{i=1}^{m} \frac{1}{\lambda_i} \prod_{j=i+1}^{m} \frac{\mu_j}{\lambda_j} + \left(\prod_{j=1}^{m} \frac{\mu_j}{\lambda_j}\right) z_0 .$$

(The product $\prod_{m+1}^{m} \mu_j/\lambda_j$ is interpreted as 1.)

In terms of ω_m we have

$$\omega_m - \omega_{m+1} = \sum_{i=1}^{m} \frac{1}{\lambda_i} \prod_{j=i+1}^{m} \frac{\mu_j}{\lambda_j} - \omega_1 \prod_{j=1}^{m} \frac{\mu_j}{\lambda_j}, \qquad m \geq 1. \tag{7.6}$$

It is more convenient to write

$$\sum_{i=1}^{m} \frac{1}{\lambda_i} \prod_{j=i+1}^{m} \frac{\mu_j}{\lambda_j} = \prod_{j=1}^{m} \frac{\mu_j}{\lambda_j} \sum_{i=1}^{m} \rho_i, \tag{7.7}$$

where

$$\rho_i = \frac{\lambda_1 \lambda_2 \cdot \cdots \cdot \lambda_{i-1}}{\mu_1 \mu_2 \cdot \cdots \cdot \mu_i} .$$

Then in terms of (7.7), the relation (7.6) becomes

$$\left(\prod_{j=1}^{m}\frac{\lambda_j}{\mu_j}\right)(\omega_m-\omega_{m+1})=\sum_{i=1}^{m}\rho_i-\omega_1\,. \tag{7.8}$$

Note that if $\sum_{i=1}^{\infty}\rho_i=\infty$, inspection of (7.8) reveals that necessarily $\omega_1=\infty$. Indeed, it is probabilistically evident that $\omega_m<\omega_{m+1}$ for all m and this property would be violated for m large if we assume to the contrary that ω_1 is finite.

Now suppose $\sum_{i=1}^{\infty}\rho_i<\infty$; then letting $m\to\infty$ in (7.8) gives

$$\omega_1=\sum_{i=1}^{\infty}\rho_i-\lim_{m\to\infty}\left(\prod_{j=1}^{m}\frac{\lambda_j}{\mu_j}\right)(\omega_m-\omega_{m+1}).$$

It is more involved but still possible to prove that

$$\lim_{m\to\infty}\left(\prod_{j=1}^{m}\frac{\lambda_j}{\mu_j}\right)(\omega_m-\omega_{m+1})=0$$

and then indeed

$$\omega_1=\sum_{i=1}^{\infty}\rho_i\,.$$

We summarize the discussion of this section in the following theorem:

Theorem 7.1. *Consider a birth and death process with birth and death parameters λ_n and μ_n, $n\geq 1$, where $\lambda_0=0$ so that 0 is an absorbing state.*

The probability of absorption into state 0 from the initial state m is

$$\begin{cases}\dfrac{\sum_{i=m}^{\infty}\left(\prod_{j=1}^{i}\mu_j/\lambda_j\right)}{1+\sum_{i=1}^{\infty}\left(\prod_{j=1}^{i}\mu_j/\lambda_j\right)} & \text{if } \sum_{i=1}^{\infty}\left(\prod_{j=1}^{i}\frac{\mu_j}{\lambda_j}\right)<\infty,\\[2ex] 1 & \text{if } \sum_{i=1}^{\infty}\left(\prod_{j=1}^{i}\frac{\mu_j}{\lambda_j}\right)=\infty.\end{cases} \tag{7.9}$$

The mean time to absorption is

$$\begin{cases}\infty & \text{if } \sum_{i=1}^{\infty}\rho_i=\infty\,,\\[2ex] \sum_{i=1}^{\infty}\rho_i+\sum_{r=1}^{m-1}\left(\prod_{k=1}^{r}\frac{\mu_k}{\lambda_k}\right)\sum_{j=r+1}^{\infty}\rho_j & \text{if } \sum_{i=1}^{\infty}\rho_i<\infty\,,\end{cases} \tag{7.10}$$

where $\rho_i=(\lambda_1\lambda_2\cdot\cdots\cdot\lambda_{i-1})/(\mu_1\mu_2\cdot\cdots\cdot\mu_i)$.

For the example of the linear growth birth and death process ($\lambda_n = n\lambda$, $\mu_n = n\mu$, and $\mu > \lambda$) the mean time ω_1 to absorption from state 1 is

$$\sum_{i=1}^{\infty} \rho_i = \frac{1}{\mu}\sum_{i=1}^{\infty}\frac{1}{i}\left(\frac{\lambda}{\mu}\right)^{i-1} = \frac{1}{\lambda}\sum_{i=0}^{\infty}\int_0^{\lambda/u} \xi^i\, d\xi = \frac{1}{\lambda}\int_0^{\lambda/u}\frac{1}{1-\xi}\, d\xi = -\frac{1}{\lambda}\log\left(1-\frac{\lambda}{\mu}\right). \tag{7.11}$$

8: *Finite State Continuous Time Markov Chains*

A continuous time Markov chain X_t ($t > 0$) is a Markov process on the states 0, 1, 2, ... We assume as usual that the transition probabilities are stationary, i.e.,

$$P_{ij}(t) = \Pr\{X_{t+s} = j | X_s = i\}. \tag{8.1}$$

In this section we consider only the case where the state space S is finite, labeled as $\{0, 1, 2, \ldots, N\}$. Some aspects of the general, infinite state, continuous time, Markov chain are discussed in the following chapter.

The Markovian property asserts that $P_{ij}(t)$ satisfies

(a) $P_{ij}(t) \geq 0$,

(b) $\sum_{j=0}^{N} P_{ij}(t) = 1, \qquad i, j \in S$

(c) $P_{ik}(s+t) = \sum_{j=0}^{N} P_{ij}(s)P_{jk}(t) \qquad t, s \geq 0$ (Chapman–Kolmogorov relation),

and we postulate in addition that

(d) $\lim_{t\to 0+} P_{ij}(t) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$

holds.

If $\mathbf{P}(t)$ denotes the matrix $\|P_{ij}(t)\|_{i,j=0}^{N}$ then property (c) can be written compactly in matrix notation as

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s), \qquad t, s \geq 0. \tag{8.2}$$

Property (d) asserts that $\mathbf{P}(t)$ is continuous at $t = 0$ since the fact $\mathbf{P}(0) = \mathbf{I}$ (= identity matrix) is implied by (8.2). It follows simply from (8.2) that $\mathbf{P}(t)$ is continuous for all $t > 0$. In fact if $s = h > 0$ in (8.2) then because of (d) we have

$$\lim_{h\to 0+} \mathbf{P}(t+h) = \mathbf{P}(t) \lim_{h\to 0+} \mathbf{P}(h) = \mathbf{P}(t)\mathbf{I} = \mathbf{P}(t). \tag{8.3}$$

On the other hand, for $t > 0$ and $0 < h < t$ we write (8.2) in the form

$$\mathbf{P}(t) = \mathbf{P}(t-h)\mathbf{P}(h). \tag{8.4}$$

But $\mathbf{P}(h)$ is near the identity when h is sufficiently small and so $\mathbf{P}(h)^{-1}$ [the inverse of $\mathbf{P}(h)$] exists and also approaches the identity $\mathbf{I}$. Therefore

$$\mathbf{P}(t) = \mathbf{P}(t) \lim_{h\to 0+} (\mathbf{P}(h))^{-1} = \lim_{h\to 0+} \mathbf{P}(t-h). \tag{8.5}$$

The limit relations (8.3) and (8.5) together show that $\mathbf{P}(t)$ is continuous. It is proved in Theorems 1.1 and 1.2 of Chapter 14 for the general, infinite state, continuous time, Markov chain that

$$\begin{aligned} \lim_{h\to 0+} \frac{1 - P_{ii}(h)}{h} &= q_i, \\ \lim_{h\to 0+} \frac{P_{ij}(h)}{h} &= q_{ij}, \qquad i \neq j. \end{aligned} \tag{8.6}$$

exist, where $0 \leq q_{ij} < \infty$ $(i \neq j)$ and $0 \leq q_i \leq \infty$, i.e., q_{ij} $(i \neq j)$ is always finite and q_i is defined but could be infinite. The possibility $q_i = \infty$ cannot occur in the case of a finite state, continuous time, Markov chain. In fact, starting with the relation

$$1 = P_{ii}(h) + \sum_{j=0, j\neq i}^{N} P_{ij}(h),$$

dividing by h, and letting h decrease to zero yields directly the relation

$$q_i = \sum_{j=0, j\neq i}^{N} q_{ij},$$

which shows that q_i is indeed finite.

Assuming that (8.6) has been verified we now derive an explicit expression for $P_{ij}(t)$ in terms of the infinitesimal matrix

$$\mathbf{A} = \begin{Vmatrix} -q_0 & q_{01} \cdots q_{0N} \\ q_{10} & -q_1 \cdots q_{1N} \\ \vdots & \\ q_{N0} & q_{N1} \cdots -q_N \end{Vmatrix}.$$

The limit relations (8.6) can be expressed concisely in matrix form:

$$\lim_{h\to 0+} \frac{\mathbf{P}(h) - \mathbf{I}}{h} = \mathbf{A}. \tag{8.7}$$

With the aid of this formula and referring to (8.2) we have

$$\frac{\mathbf{P}(t+h) - \mathbf{P}(t)}{h} = \frac{\mathbf{P}(t)[\mathbf{P}(h) - \mathbf{I}]}{h} = \frac{\mathbf{P}(h) - \mathbf{I}}{h}\mathbf{P}(t). \tag{8.8}$$

The limit on the right exists and this leads to the matrix differential equation

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{A} = \mathbf{A}\mathbf{P}(t), \tag{8.9}$$

where $\mathbf{P}'(t)$ denotes the matrix whose elements are $P'_{ij}(t)$.

The existence of $P'_{ij}(t)$ is obviously an immediate consequence of (8.7) and (8.8).

Equations (8.9) can be solved under the initial condition $\mathbf{P}(0) = \mathbf{I}$ by the standard methods of systems of ordinary differential equations† to yield the formula

$$\mathbf{P}(t) = e^{\mathbf{A}t} = \mathbf{I} + \sum_{n=1}^{\infty} \frac{\mathbf{A}^n t^n}{n!}. \tag{8.10}$$

In practical terms we determine the eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_N$ of $\mathbf{A}$ and a complete system of associated right eigenvectors $\mathbf{u}^{(0)}, \ldots, \mathbf{u}^{(N)}$ when possible (see the Appendix at the close of the book). Then we have the representation

$$\mathbf{P}(t) = \mathbf{U}\mathbf{\Lambda}(t)\mathbf{U}^{-1}, \tag{8.11}$$

where $\mathbf{U}$ is the matrix whose column vectors are, respectively, $\mathbf{u}^{(0)}, \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(N)}$ and $\mathbf{\Lambda}(t)$ is the diagonal matrix

$$\mathbf{\Lambda}(t) = \begin{Vmatrix} \exp(\lambda_0 t) & 0 & \cdots & 0 \\ 0 & \exp(\lambda_1 t) & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp(\lambda_N t) \end{Vmatrix}.$$

The rows of the matrix $\mathbf{U}^{-1}$ can also be identified as a complete system of left eigenvectors normalized to be biorthogonal to the $\{\mathbf{u}^{(l)}\}_{l=0}^{N}$.

Applications of (8.10) or (8.11) are implicit in Elementary Problems 7 and 13 of this chapter.

Elementary Problems

1. Let X_1 and X_2 be independent exponentially distributed random variables with parameters λ_1 and λ_2 so that

$$\Pr\{X_i > t\} = \exp\{-\lambda_i t\} \quad \text{for} \quad t \geq 0.$$

† E. A. Coddington, and N. Levinson, "Theory of Ordinary Differential Equations," Chapter 3. McGraw-Hill, New York, 1955.

Let

$$N = \begin{cases} 1 & \text{if } X_1 < X_2, \\ 2 & \text{if } X_2 \leq X_1, \end{cases}$$
$$U = \min\{X_1, X_2\} = X_N,$$
$$V = \max\{X_1, X_2\},$$

and

$$W = V - U = |X_1 - X_2|.$$

Show

(a) $\Pr\{N = 1\} = \lambda_1/(\lambda_1 + \lambda_2)$ and $\Pr\{N = 2\} = \lambda_2/(\lambda_1 + \lambda_2)$.
(b) $\Pr\{U > t\} = \exp\{-(\lambda_1 + \lambda_2)t\}$ for $t \geq 0$.
(c) N and U are independent random variables.
(d) $\Pr\{W > t | N = 1\} = \exp\{-\lambda_2 t\}$ and
$\Pr\{W > t | N = 2\} = \exp\{-\lambda_1 t\}$ for $t \geq 0$.
(e) U and $W = V - U$ are independent random variables.

2. Assume a device fails when a cumulative effect of k shocks occur. If the shocks happen according to a Poisson process with parameter λ find the density function for the life T of the device.

Solution:

$$f(t) = \begin{cases} \dfrac{\lambda^k t^{k-1} e^{-\lambda t}}{\Gamma(k)}, & t > 0, \\ 0, & t \leq 0. \end{cases}$$

3. Let $\{X(t), t \geq 0\}$ be a Poisson process with intensity parameter λ. Suppose each arrival is "registered" with probability p, independent of other arrivals. Let $\{Y(t), t \geq 0\}$ be the process of "registered" arrivals. Prove that $Y(t)$ is a Poisson process with parameter λp.

4. Let $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$ be independent Poisson processes with parameters λ_1 and λ_2, respectively. Define $Z_1(t) = X(t) + Y(t)$, $Z_2(t) = X(t) - Y(t)$, $Z_3(t) = X(t) + k$, k a positive integer. Determine which of the above processes are Poisson and find λ.

5. Messages arrive at a telegraph office in accordance with the laws of a Poisson process with mean rate of 3 messages per hour.
(a) What is the probability that no message will have arrived during the morning hours (8 to 12)?
(b) What is the distribution of the time at which the first afternoon message arrives?

6. Let $X(t)$ be a homogeneous Poisson process with parameter λ. Determine the covariance between $X(t)$ and $X(t+\tau)$, $t > 0$ and $\tau > 0$, i.e., compute $E[(X(t) - E(X(t))(X(t+\tau) - EX(t+\tau))]$.

7. A continuous time Markov chain has two states labeled 0 and 1. The waiting time in state 0 is exponentially distributed with parameter $\lambda > 0$. The waiting time in state 1 follows an exponential distribution with parameter $\mu > 0$. Compute the probability $P_{00}(t)$ of being in state 0 at time t starting at time 0 in state 0.

Solution:

$$P_{00}(t) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)t}.$$

8. In Elementary Problem 7 let $\lambda = \mu$ and define $N(t)$ to be the number of times the system has changed states in time $t \geq 0$. Find the probability distribution of $N(t)$.

Solution:

$$\Pr\{N(t) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

9. Let $X(t)$ be a pure birth continuous time Markov chain. Assume that

$$\Pr\{\text{an event happens in } (t, t+h) | X(t) = \text{odd}\} = \lambda_1 h + o(h),$$
$$\Pr\{\text{an event happens in } (t, t+h) | X(t) = \text{even}\} = \lambda_2 h + o(h),$$

where $o(h)/h \to 0$ as $h \downarrow 0$. Take $X(0) = 0$. Find the following probabilities:

$$P_1(t) = \Pr\{X(t) = \text{odd}\}, \qquad P_2(t) = \Pr\{X(t) = \text{even}\}.$$

Hint: Derive the differential equations

$$P_1'(t) = -\lambda_1 P_1(t) + \lambda_2 P_2(t), \qquad P_2'(t) = \lambda_1 P_1(t) - \lambda_2 P_2(t)$$

and solve them.

Solution:

$$P_1(t) = \frac{\lambda_2}{\lambda_1+\lambda_2}(1 - \exp\{-(\lambda_1+\lambda_2)t\});$$

$$P_2(t) = \frac{\lambda_1}{\lambda_1+\lambda_2} + \frac{\lambda_2}{\lambda_1+\lambda_2} \exp\{-(\lambda_1+\lambda_2)t\}.$$

10. Under the conditions of Elementary Problem 9 determine $E[X(t)]$.

Solution:

$$EX(t) = \frac{2\lambda_1\lambda_2}{\lambda_1+\lambda_2} t + \frac{(\lambda_1-\lambda_2)\lambda_2}{(\lambda_1+\lambda_2)^2} [\exp\{-(\lambda_1+\lambda_2)t\} - 1].$$

11. Suppose $g(t)$ is the conditional rate of failure of an article at time t, given that it has not failed up to time t, i.e., $\Pr\{\text{failure in time } (t, t+h) | \text{no failure up to time } t\} = g(t)h + o(h)$ as $h \downarrow 0$. Assume that $g(t)$ is positive and continuous on $(0, \infty)$. Find an expression for $F(t) = \Pr\{\text{failure at some time } \tau, \tau < t\}$ in terms of $g(\cdot)$.

Hint: Derive a differential equation for $F(t)$.

Solution: $F(t) = 1 - \exp[-\int_0^t g(\tau)\, d\tau]$.

12. Consider a variable time Poisson process, i.e., the occurrence of an event E during the time duration $(t, t+h)$ is independent of the number of previous occurrences of E and its probability is $\lambda(t)h + o(h)$ $(h \to 0)$. (Note that λ may now depend on t.)

(a) Prove that the probability of no occurrence of E during the time duration $[0, s]$ is

$$\exp\left(-\int_0^s \lambda(\xi)\, d\xi\right).$$

(b) Prove that the probability of k occurrences of E during the time duration $[0, s]$ is

$$\frac{1}{k!}\left(\int_0^s \lambda(\xi)\, d\xi\right)^k \exp\left(-\int_0^s \lambda(\xi)\, d\xi\right).$$

13. There are two transatlantic cables each of which can handle one telegraph message at a time. The time-to-breakdown for each has the same exponential distribution with parameter λ. The time to repair for each cable has the same exponential distribution with parameter μ. Given that at time 0 both cables are in working condition, find the probability that, if at time t two messages arrive simultaneously, they will find both cables operative.

Hint: This is a three-state continuous time, Markov chain.

Solution:

$$\frac{\mu^2}{(\lambda+\mu)^2} + \frac{\lambda^2 e^{-2(\lambda+\mu)t}}{(\lambda+\mu)^2} + \frac{2\lambda\mu}{(\lambda+\mu)^2} e^{-(\lambda+\mu)t}.$$

14. Consider the linear growth birth and death process $X(t)$ with parameters λ, μ and $a = 0$. Assume $X(0) = 1$. Find the distribution of the number of living individuals at the time of the first death.

Solution:

$$\Pr\{\text{number } k \text{ of births before the first death}\} = (\mu/(\mu+\lambda))(\lambda/(\lambda+\mu))^k.$$

15. Find the stationary distribution for the linear growth birth and death process when $\lambda < \mu$ (Example 1 of Section 6).

Solution:

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \frac{(a/\lambda)((a/\lambda)+1)\cdot\cdots\cdot((a/\lambda)+n-1)}{n!}\left(1-\frac{\lambda}{\mu}\right)^{a/\lambda}.$$

16. A telephone exchange has m channels. Calls arrive in the pattern of a Poisson process with parameter λ; they are accepted if there is an empty channel, otherwise they are lost. The duration of each call is a r.v. whose distri-

bution function is exponential with parameter μ. The lifetimes of separate calls are independent random variables. Find the stationary probabilities of the number of busy channels.

Solution:

$$p_n = \frac{(\lambda/\mu)^n(1/n!)}{\sum_{k=0}^{m}(\lambda/\mu)^k(1/k!)}, \qquad n = 0, 1, 2, \ldots, m.$$

17. We start observing a radioactive atom at time 0. It will decay and cease to be radioactive at a time t, $t > 0$, determined by the distribution

$$F(\tau) = \begin{cases} 0, & \tau < 0 \\ 1 - e^{-\lambda\tau}, & \tau \geq 0 \end{cases} = \Pr\{t \leq \tau\}.$$

Consider the state of the atom at time t as a random variable

$$x_t = \begin{cases} 0 & \text{if the atom is radioactive at time } t, \\ 1 & \text{if the atom is not radioactive at time } t. \end{cases}$$

The variables $\{x_t\}$ define a stochastic process.

Suppose that at time 0 we begin observing N independent radioactive atoms, represented in the above sense by x_t^i, $i = 1, 2, \ldots, N$. Let $X_t = \sum_{i=1}^{N} x_t^i$. Then $\{X_t\}$ is also a stochastic process. Show that for $t \ll 1/\lambda$ (t negligibly small compared with $1/\lambda$) and sufficiently large N, $\{X_t\}$ is very closely approximated by a Poisson process $Y(t)$ with parameter $\lambda N t$.

18. Suppose that in Problem 17 the approximation $t \ll \lambda^{-1}$ cannot be made. (i) Is the process a process with independent increments? (ii) Is it stationary? (iii) Does it have stationary transition probabilities? (iv) Is it a Markov process?

Solution: (i) Yes, (ii) no, (iii) yes, (iv) yes.

19. This problem attempts to relate the properties of the life history of a colonizing species to its chances for success, or more precisely, to the length of time it persists before going extinct.

Let $Z(t)$ be the population size at time t. We suppose $(Z(t);\ t \geq 0)$ evolves as a birth and death process with *individual* birth rate λ and *individual* death rate μ. By this we mean that each individual alive at time t gives birth to a new individual during the interval $(t, t + \Delta t)$ with probability (approximately) $\lambda(\Delta t)$, and dies during that interval with probability $\mu(\Delta t)$.

We want to construct a model that can be used to estimate the mean survival time of a population of such individuals, and we want the model to reflect the fact that all populations are limited in their maximum size by the carrying capacity of the environment, which we assume to be K individuals. Since all individuals have a chance of dying, all populations will surely go extinct if given enough time. We want to build into the model the properties of exponential growth (on the average) for small populations, as well as the ceiling K, beyond which the population cannot normally grow. There are any number of ways of

approaching population size K and staying there at equilibrium. We will take the simple case where the birth parameters are

$$\lambda_i = \begin{cases} \lambda i & \text{for } i = 0,\ldots, K-1 \\ 0 & \text{for } i \geq K, \end{cases}$$

and the death parameters are $\mu_i = \mu i$ for $i = 0, 1,\ldots$.

In this model, compute the expected time to extinction, given the population begins with a single individual.

20. Suppose that we have a mechanism which can fail in two ways. Let the probability of the first type failure in the interval $(t, t+h)$ be $\lambda_1 h + o(h)$ and the probability of the second type failure in the interval $(t, t+h)$ be $\lambda_2 h + o(\mathrm{h})$. Upon failure, repair is performed whose duration is distributed as an exponetial random variable with parameter depending upon the type failure. Let μ_1 and μ_2 denote the respective parameters. Compute the probability that the mechanism is working at time t.

Solution:

$$\mathbf{P}(t) = e^{\mathbf{Q}t} \qquad \text{where} \quad \mathbf{Q} = \begin{Vmatrix} -(\lambda+\lambda_2) & \lambda_1 & \lambda_2 \\ \mu_1 & -\mu_1 & 0 \\ \mu_2 & 0 & -\mu_2 \end{Vmatrix}.$$

21. Compare the $M/M/1$ system for a first-come first-served queue discipline with one of last-come first-served type (for example, articles for service are taken from the top of a stack). How do the queue size, waiting time, and busy period distribution differ, if at all?

Solution: Queue size and busy period do not differ but the waiting time distributions differ. Why?

22. (Queuing with Balking) Customers, with independent and identically distributed service times, arrive at a counter in the manner of a Poisson process with parameter λ. A customer who finds the server busy joins the queue with probability $p(0 < p < 1)$. The service time distribution is exponential with parameter μ.

Formulate this model as a birth and death process.

Solution: $\lambda_n = \lambda p$; $\mu_n = \mu$.

23. Consider the $M/M/1$ system with queue discipline of last-come first-served type. Let $X(t)$ be the queue size at time t. Show that the process $\{X(t);\ t \geq 0\}$ is a birth and death process and determine its parameters.

Solution: $\lambda_n = \lambda$, $\mu_n = \mu$.

24. Under the condition that $X(0) = N = 1$, determine the mean and variance of the Yule process.

Solution:

$$E[X(t)] = e^{\lambda t}, \qquad \mathrm{Var}\,[X(t)] = e^{2\lambda t}(1 - e^{-\lambda t}).$$

Problems

1. Let $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$ be two independent Poisson processes with parameters λ_1 and λ_2, respectively. Define

$$Z(t) = X(t) - Y(t), \qquad t \geq 0.$$

This is a stochastic process whose state space consists of all the integers (positive, negative, and zero). Let

$$P_n(t) = \Pr\{Z(t) = n\}, \qquad n = 0, \pm 1, \pm 2, \ldots.$$

Establish the formula

$$\sum_{n=-\infty}^{\infty} P_n(t) z^n = \exp(-(\lambda_1 + \lambda_2)t) \exp(\lambda_1 zt + (\lambda_2/z)t), \qquad |z| \neq 0.$$

Compute $E(Z(t))$ and $E(Z(t)^2)$.

Answer: $E(Z(t))^2 = (\lambda_1 + \lambda_2)t + (\lambda_1 - \lambda_2)^2 t^2$.

2. Consider two independent Poisson processes $X(t)$ and $Y(t)$ where $E(X(t)) = \lambda t$ and $E(Y(t)) = \mu t$. Let two successive events of the $X(t)$ process occur at T and $T' > T$ so that $X(t) = X(T)$ for $T \leq t < T'$ and $X(T') = X(T) + 1$. Define $N = Y(T') - Y(T) =$ the number of events of the $Y(t)$ process in the time interval (T, T'). Find the distribution of N.

Answer:

$$\Pr\{N = m\} = \frac{\lambda}{\lambda + \mu}\left(\frac{\mu}{\lambda + \mu}\right)^m, \qquad m = 0, 1, 2, \ldots.$$

3. Consider a pure death process where $\mu_n = n\mu$ for $n = 1, 2, \ldots$, i.e., $P\{X(t + h) = j | X(t) = k\} = 0$ for $j > k$ and t and k positive. Assume an initial population of size i. Find $P_n(t) = P\{X(t) = n\}$, $EX(t)$, and Var $X(t)$.

Answer:

$$P_n(t) = \binom{i}{n} e^{-n\mu t} (1 - e^{-\mu t})^{(i-n)t},$$

$$EX(t) = ie^{-\mu t},$$

$$\text{Var } X(t) = ie^{-\mu t}(1 - e^{-\mu t}).$$

4. Consider a Yule process with parameter β and initial state $N = 1$. Suppose the first individual is also subject to death, with the probability of death in the interval t to $t + h$, given that the individual is living at time t, being $\mu h + o(h)$. Compute the distribution of the number of offspring due to a single individual and his descendants at the time of death of the original parent.

Answer: Probability of a total of n offspring originating from a specified individual and of his line of descendants at the time of his death is

$$\int_0^\infty e^{-\beta t}(1-e^{-\beta t})^n \mu e^{-\mu t}\,dt = \frac{\mu}{\beta}\frac{\Gamma((\mu/\beta)+1)\Gamma(n+1)}{\Gamma(n+(\mu/\beta)+2)}.$$

5. Let $(X(t), Y(t))$ describe a stochastic process in two-dimensional space where $X(t)$ is a Poisson process with parameter λ_1 and $Y(t)$ is a Poisson process independent of $X(t)$ with parameter λ_2. Given that the process is in the state (x_0, y_0) at time $t=0$, $x_0+y_0<z$, what is the probability that it will intersect the line $x+y=z$ at the point (x, y)?

Answer:

$$\begin{cases} \binom{z-x_0-y_0}{x-x_0}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{x-x_0}\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{y-y_0} & \text{for } x\geq x_0, y\geq y_0, \\ 0 & \text{otherwise.} \end{cases}$$

6. Consider a Poisson process with parameter λ. Let T be the time required to observe the first event, and let $N(T/\kappa)$ be the number of events in the next T/κ units of time. Find the first two moments of $N(T/\kappa)T$.

Answer:

$$E\left\{N\left(\frac{T}{\kappa}\right)T\right\}=\frac{2}{\lambda\kappa}; \qquad E\left\{\left(N\left(\frac{T}{\kappa}\right)T\right)^2\right\}=\frac{6}{\lambda^2\kappa}+\frac{24}{\lambda^2\kappa^2}.$$

7. Consider n independent objects (such as light bulbs) whose failure time (i.e., lifetime) is a random variable exponentially distributed with density function $f(x, \theta)=\theta^{-1}\exp(-x/\theta)$, $x>0$; 0 for $x\leq 0$ (θ is a positive parameter). The observations of lifetime become available in order of failure. Let

$$X_{1,n}\leq X_{2,n}\leq\cdots\leq X_{r,n}$$

denote the lifetimes of the first r objects that fail. Determine the joint density function of $X_{i,n}$, $i=1, 2, \ldots, r$.

Answer:

$$f(x_1, x_2, \ldots, x_r)=r!\binom{n}{r}\frac{1}{\theta^r}\exp\left(-\frac{x_1+x_2+\cdots+x_{r-1}+(n-r+1)x_r}{\theta}\right).$$

8. In the preceding problem define $Y_{1,n}=X_{1,n}$ and

$$Y_{i,n}=X_{i,n}-X_{i-1,n} \qquad \text{for } 2\leq i\leq r.$$

Prove that $Y_{i,n}$ are mutually independent and find the distribution function of each.

Answer:

$$\Pr\{Y_{i,n}\leq y\}=1-\exp\left(-\frac{n-i+1}{\theta}y\right).$$

9. Consider a Poisson process of parameter λ. Given that n events happen in time t, find the density function of the time of the occurrence of the rth event $(r < n)$.

Answer:

$$p(x) = \begin{cases} \dfrac{n!}{(r-1)!(n-r)!} \dfrac{x^{r-1}}{t^r} \left(1 - \dfrac{x}{t}\right)^{n-r} & 0 < x < t, \\ 0, & \text{otherwise.} \end{cases}$$

10. Let $\mathfrak{M}$ be a continuous time birth and death process where $\lambda_n = \lambda > 0$, $n \geq 0$, $\mu_0 = 0$, $\mu_n > 0$, $n \geq 1$. Let $\pi = \sum_n \pi_n < \infty$, where $\pi_n = \lambda^n/(\mu_1\mu_2 \cdot \cdots \cdot \mu_n)$ so that π_i/π is the stationary distribution of the process. Suppose the initial state is a r.v. whose distribution is the stationary distribution of the process. Prove that the number of deaths in $[0, t]$ has a Poisson distribution with parameter λt.

Hint: Let $a_k(t)$ be the probability that the number of deaths by time t is k. Derive the differential equation

$$a_k'(t) = -\lambda a_k(t) + \lambda a_{k-1}(t), \qquad k = 1, 2, \ldots.$$

11. The following defines one concept of a multivariate Poisson process in two dimensions. Let $(X(t), Y(t))$ be defined by $X(t) = \alpha(t) + \gamma(t)$, $Y(t) = \beta(t) + \gamma(t)$, where $\alpha(t)$, $\beta(t)$, and $\gamma(t)$ are three independent Poisson processes with parameters λ_1, λ_2, and λ_3, respectively. Find the generating function of the distribution of $(X(t), Y(t))$.

Answer:

$$\sum \Pr\{X(t) = i,\ Y(t) = j\} x^i y^j = \exp\{t(\lambda_1 x + \lambda_2 y + \lambda_3 xy - \lambda_1 - \lambda_2 - \lambda_3\}.$$

12. Consider a Yule process $\{N_t, t \geq 0\}$ with birthrate λ and initial population of size 1. Find the distribution function of $N_t(x) =$ number of members of the population at time t of age less than or equal to x.

Hint: Condition on the value of N_{t-x}.

Solution:

$$\Pr\{N_t(x) = k\} = \frac{e^{-\lambda t}(1 - e^{-\lambda x})^k}{[1 - e^{-\lambda x} + e^{-\lambda t}]^{k+1}}.$$

13. Let $\{X_i(t); t \geq 0\}$, $i = 1, 2$, be two independent Yule processes with the same parameter λ. Let $X_i(0) = n_i$, $i = 1, 2$. Determine the conditional distribution of $X_1(t)$ given $X_1(t) + X_2(t) = N$ $(N \geq n_1 + n_2)$.

Answer:

$$\Pr\{X_1(t) = k | X_1(t) + X_2(t) = N\} = \frac{\binom{k-1}{n_1-1}\binom{N-k-1}{n_2-1}}{\binom{N-1}{n_1+n_2-1}} \quad \text{for} \quad k \geq n_1.$$

14. Continuation of Problem 13.
Prove the limit distribution relation

$$\lim_{t\to\infty} \Pr\left\{\frac{X_1(t)}{X_1(t)+X_2(t)} \le x\right\} = \frac{(n_1+n_2-1)!}{(n_1-1)!\,(n_2-1)!}\int_0^x y^{n_1-1}(1-y)^{n_2-1}\,dy.$$

Hint: Let $N\to\infty$ and $k\to\infty$ in such a way that $k/N\to y$ $(0<y<1)$. Then with the aid of Sterling's approximation establish the asymptotic relation

$$\lim_{\substack{\frac{k}{N}\to y,\ N\to\infty \\ k\to\infty}} \frac{\binom{k-1}{n_1-1}\binom{N-k-1}{n_2-1}}{\binom{N-1}{n_1+n_2-1}} = \frac{(n_1+n_2-1)!}{(n_1-1)!\,(n_2-1)!}y^{n_1-1}(1-y)^{n_2-1}$$

Use this to show that

$$\lim_{t\to\infty} \Pr\left\{y \le \frac{X_1(t)}{X_1(t)+X_2(t)} \le y+h\right\}$$
$$= \frac{(n_2+n_1-1)!}{(n_1-1)!\,(n_2-1)!}hy^{n_1-1}(1-y)^{n_2-1} + o(h).$$

15. A system is composed of N identical components; each independently operates a random length of time until failure. Suppose the failure time distribution is exponential with parameter λ. When a component fails it undergoes repair. The repair time is random, with distribution function exponential with parameter μ. The system is said to be in state n at time t if there are exactly n components under repair at time t. This process is a birth and death process. Determine its infinitesimal parameters.

16. In Problem 15, suppose that initially all N components are operative. Find the distribution $F(t)$ of the first time that there are two inoperative components.

Answer: The Laplace transform $\varphi(s)$ of $F(t)$ is

$$\varphi(s) = \frac{N(N-1)\lambda^2}{s^2+s[(2N-1)\lambda+\mu]+N(N-1)\lambda^2}.$$

In the case $\lambda=\mu$,

$$1-F(t) = \frac{\sqrt{N}(N-1)}{2}\{\exp[(-N+\sqrt{N})\lambda t] - \exp[(-N-\sqrt{N})\lambda t]\}.$$

17. Consider the following continuous version of the Ehrenfest model (see page 51, Chapter 2). We have $2N$ balls labeled 1, 2, 3, ..., $2N$. At time 0 each ball is equally likely to be placed in one of two urns. Subsequently, the balls independently undergo displacement randomly in time from one urn to the other by the following rules. A ball has a probability $\frac{1}{2}h+o(h)$ of changing urns during the time interval $(t, t+h)$ and probability $1-(h/2)+o(h)$ of remaining in the same urn during that interval. The movements over disjoint intervals of time are independent. Let $X(t)$ denote the number of balls in urn I

at time t. Set

$$P_{jk}(t) = \Pr\{X(t) = k | X(0) = j\}, \qquad j, k = 0, 1, \ldots, 2N.$$

Establish the formula

$$g(t, s) = \sum_{k=0}^{2N} P_{jk}(t)s^k = 2^{-2N}[1 - e^{-t} + (1 + e^{-t})s]^j[1 + e^{-t} + (1 - e^{-t})s]^{2N-j}.$$

Hint: Define the random variables

$$X_i(t) = \begin{cases} 1 & \text{if the } i\text{th ball is in urn I at time } t, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for} \quad i = 1, 2, \ldots, 2N.$$

Then

$$X(t) = \sum_{i=1}^{2N} X_i(t).$$

Show that

$$\Pr\{X_i(t) = X_i(0)\} = \frac{1 + e^{-t}}{2}, \qquad i = 1, 2, \ldots, 2N.$$

18. For a linear growth birth and death process $X(t)$ with $\lambda = \mu$ (Example 1, Section 6), prove that

$$u(t) = \Pr\{X(t) = 0 | X(0) = 1\}$$

satisfies the integral equation

$$u(t) = \frac{1}{2}\int_0^t 2\lambda e^{-2\lambda\tau}\, d\tau + \frac{1}{2}\int_0^t 2\lambda e^{-2\lambda\tau}[u(t-\tau)]^2 d\tau.$$

Hint: Note that the waiting time to the first event (birth or death) is exponentially distributed with parameter 2λ.

19. (Continuation of Problem 18) Show that $u(t)$ satisfies the Ricatti differential equation

$$u'(t) + 2\lambda u(t) = \lambda + \lambda u^2(t), \qquad u(0) = 0.$$

20. (Continuation of Problem 19) Find $u(t)$.

Answer:

$$u(t) = \frac{\lambda t}{1 + \lambda t}.$$

21. (Continuation of Problem 20). Determine $\Pr\{X(t) = 0 | X(0) = 1, X(T) = 0\}$ for $0 < t < T$.

22. Consider a birth and death process with infinitesimal parameter λ_n, μ_n. Show that the expected length of time for reaching state $r + 1$ starting from state 0 is

$$\sum_{n=0}^{r} \frac{1}{\lambda_n \pi_n} \sum_{k=0}^{n} \pi_k$$

For the definition of π_n see Eq. (4.5).

Hint: Let T_n^* denote the elapsed time of first entering state $n+1$ starting from state n. Derive a recursion relation for $E(T_n^*)$.

23. The following problem arises in molecular biology. The surface of a bacterium is supposed to consist of several sites at which a foreign molecule may become attached if it is of the right composition. A molecule of this composition will be called acceptable. We consider a particular site and postulate that molecules arrive at the site according to a Poisson process with parameter μ. Among these molecules a proportion β is acceptable. Unacceptable molecules stay at the site for a length of time which is exponentially distributed with parameter λ. While at the site they prevent further attachments there. An acceptable molecule "fixes" the site preventing any further attachments. What is the probability that the site in question has not been fixed by time t?

Hint: Set the problem up as a three-state continuous time Markov chain for the site in question.

Answer: $$\frac{\beta\mu}{s_2 - s_1}\left[\left(1+\frac{\lambda}{s_1}\right)e^{s_1 t} - \left(1+\frac{\lambda}{s_2}\right)e^{s_2 t}\right]$$

where s_1, s_2 are the roots of $s^2 + s(\lambda+\mu) + \mu\beta\lambda = 0$.

24. Consider an infinitely many-server queue with an exponential service time distribution with parameter μ. Suppose customers arrive in batches with the interarrival time following an exponential distribution with parameter λ. The number of arrivals in each batch is assumed to follow the geometric distribution with parameter ρ $(0 < \rho < 1)$, i.e., Pr{number of arrivals in a batch has size k} $= \rho^{k-1}(1-\rho)$ $(k = 1, 2, \ldots)$.
Formulate this process as a continuous time Markov chain and determine explicitly the infinitesimal matrix of the process.

25. Show for the $M/M/1$ queueing process in a stationary state that the distribution of time between successive departures has the same (exponential) distribution as the interarrival time distribution (see also Problem 10).

26. Let $\{X_i(t); t \geq 0\}$ $i = 1, 2$ be two independent Poisson processes with parameters λ_1 and λ_2 respectively. Let $X_1(0) = m$, $X_2(0) = N - 1$, and $m < N$.
(a) Determine the probability that the X_2 process reaches N before the X_1 process does.
(b) Solve the same problem for $X_2(0) = n$ where $n < N$.

Answer: (*b*):

$$\sum_{r=0}^{N-m-1}\binom{N-n+r-1}{r}p^r q^{N-n}, \qquad p = \frac{\lambda_1}{\lambda_1+\lambda_2}$$

27. The following two birth and death processes (cf. Section 4, Chapter 4) can be viewed as models for queueing with balking.

(a) First consider a birth and death process with parameters

$$\lambda_n = \lambda q^n, \quad 0 < q < 1, \quad \lambda > 0 \quad (n = 0, 1, 2, \ldots),$$
$$\mu_n = \mu, \quad \mu > 0,$$
$$\mu_0 = 0.$$

(b) Let the parameters be

$$\lambda_n = \frac{\lambda}{n+1}, \quad \mu_n = \mu \quad (n = 1, 2, \ldots),$$
$$\mu_0 = 0.$$

Determine the stationary distribution in each case.

Answer: (a) $p_m = p_0(\lambda/\mu)^m q^{m(m-1)/2}$ for $m \geq 1$. (b) $p_m = p_0(\lambda/\mu)^m(1/m!)$ for $m \geq 0$, whence $p_0 = e^{-\lambda/\mu}$.

28. Show for the $M/M/s$ system that the stationary queue size distribution $\{p_n, n = 0, 1, 2, \ldots\}$ is given by

$$p_0 = \left\{ \frac{(s\rho)^s}{s!(1-\rho)} + \sum_{i=0}^{s-1} \frac{(s\rho)^i}{i!} \right\}^{-1}$$

$$p_n = \begin{cases} p_0 \dfrac{(s\rho)^n}{n!}, & 1 \leq n \leq s, \\ p_0 \rho^n \dfrac{s^s}{s!}, & s < n < \infty, \end{cases}$$

where $\rho = \lambda/s\mu < 1$. Let $Q = \max(n - s, 0)$ $(n = 0, 1, 2, \ldots)$ be the size of the queue not including those being served. Show that

(i) $$\gamma = \Pr\{Q = 0\} = \frac{\sum_{i=0}^{s} (s\rho)^i/i!}{\sum_{i=0}^{s} [(s\rho)^i/i!] + [(s\rho)^s \rho/s!(1-\rho)]};$$

(ii) $E(Q) = (1 - \gamma)/(1 - \rho)$.

29. A system is composed of N machines. At most $M \leq N$ can be operating at any one time; the rest are "spares". When a machine is operating, it operates a random length of time until failure. Suppose this failure time is exponentially distributed with parameter μ.

When a machine fails it undergoes repair. At most R machines can be "in repair" at any one time. The repair time is exponentially distributed with parameter λ. Thus a machine can be in any of four states: (i) Operating, (ii) "Up", but not operating, i.e., a spare, (iii) In repair, (iv) Waiting for repair. There are a total of N machines in the system. At most M can be operating. At most R can be in repair.

Let $X(t)$ be the number of machines "up" at time t, either operating or spare. Then, (we assume) the number operating is min $\{X(t), M\}$ and the number of spares is max $\{0, X(t) - M\}$. Let $Y(t) = N - X(t)$ be the number of machines

" down ". Then the number in repair is $\min\{Y(t), R\}$ and the number waiting for repair is $\max\{0, Y(t) - R\}$. The above formulas permit to determine the number of machines in any category, once $X(t)$ is known.

$X(t)$ is a birth and death process.

(a) Determine the birth and death parameters, λ_i and μ_i, $i = 0, \ldots, N$.
(b) In the following special cases, determine π_j, the stationary probability that $X(t) = j$.
(a) $R = M = N$. (b) $R = 1$, $M = N$.

30. (Continuation of Problem 29) (a) Determine the stationary distribution in the case $R < M = N$. (b) Let $\mathbf{X}_{N,R}(\infty)$ denote a random variable having the stationary distribution in (a). Let $N \to \infty$, $R \to \infty$ but such that $N/R \to 1 + \alpha$ where $\alpha > 0$ is fixed. Determine normalizing constants a_N and b_N and a limiting distribution Φ for which

$$\lim_{N \to \infty} \Pr\left\{\frac{X_{N,R}(\infty) - a_N}{b_N} \leq x\right\} = \Phi(x).$$

(There are two cases (i) $\alpha > \lambda/\mu$ and (ii) $\alpha < \lambda/\mu$).

31. Consider a pure birth process having infinitesimal parameters $\lambda_n = \lambda n^2$, where $\lambda > 0$ is fixed. Given that at time 0 there is a single particle, determine

$$P_\infty(t) = 1 - \sum_{k=1}^{\infty} P_k(t).$$

32. Let $X(t)$ be a Yule process starting at $X(0) = N$ and having birth rate β. Show

$$\Pr\{X(t) \geq n | X(0) = N\} = \sum_{k=n-N}^{n-1} \binom{n-1}{k} p^k q^{n-1-k}$$

where $q = 1 - p = e^{-\beta t}$.

NOTES

Poisson and birth and death processes play a fundamental role in the theory and applications that embrace queueing and inventory models, population growth, engineering systems, etc. Elementary discussions on Poisson and related processes can be found in all textbooks on stochastic processes.

The literature of queueing theory is voluminous. An elegant monograph reviewing this theory and its applications is that of Cox and Smith [1].

We also direct the student to the advanced books by Takács [2] and Riordan [3].

A compendium of results on queueing theory is contained in Saaty [4]. This reference also includes an extensive bibliography.

Applications to congestion theory and telephone trunking problems can be found in Siski [5].

Some special mathematical aspects of queueing theory are developed in the monograph by Beneš [6].

REFERENCES

1. D. R. Cox and W. L. Smith, "Queues." Methuen, London, 1961.
2. L. Takács, "Introduction to the Theory of Queues." Oxford Univ. Press, London and New York, 1962.
3. J. Riordan, "Stochastic Service Systems." Wiley, New York, 1962.
4. T. L. Saaty, "Elements of Queueing Theory with Applications." McGraw-Hill, New York, 1961.
5. E. Syski, "Congestion Theory." Wiley, New York, 1960.
6. V. E. Beneš, "General Stochastic Processes in the Theory of Queues." Addison-Wesley, Reading, Massachusetts, 1963.

Chapter 5

RENEWAL PROCESSES

Renewal theory began with the study of stochastic systems whose evolution through time was interspersed with renewals, times when, in a statistical sense, the process began anew. Today, the subject is viewed as the general study of functions of independent, identically distributed, nonnegative random variables representing the successive intervals between renewals. The results are applicable in a wide variety of both theoretical and practical probability models.

The first six sections of this chapter are vital and should be included in every introductory course. Sections 7 and 8 are not difficult and will round out a basic knowledge of renewal theory if time permits their study. Section 9 offers a glimpse of a topic of some recent interest and may be omitted at first reading.

1: Definition of a Renewal Process and Related Concepts

A *renewal (counting) process* $\{N(t),\ t \geq 0\}$ is a nonnegative integer-valued stochastic process that registers the successive occurrences of an event during the time interval $(0, t]$, where the time durations between consecutive "events" are *positive*, *independent*, *identically distributed*, random variables (i.i.d.r.v.). Let the successive occurence times between events be $\{X_k\}_{k=1}^{\infty}$ (often representing the lifetimes of some units successively placed into service) such that X_i is the elapsed time from the $(i-1)$st event until the occurrence of the ith event. We write

$$F(x) = \Pr\{X_k \leq x\}, \qquad k = 1, 2, 3, \ldots,$$

for the common probability distribution of $\{X_k\}$. A basic stipulation for renewal processes is $F(0) = 0$, signifying that X_k are *positive* random variables. We refer to

$$S_n = X_1 + X_2 + \cdots + X_n, \qquad n \geq 1 \qquad (S_0 = 0, \text{ by convention}) \quad (1.1)$$

as the *waiting time* until the occurrence of the nth event. Note formally that

$$N(t) = \text{number of indices } n \text{ for which } 0 < S_n \leq t. \quad (1.2)$$

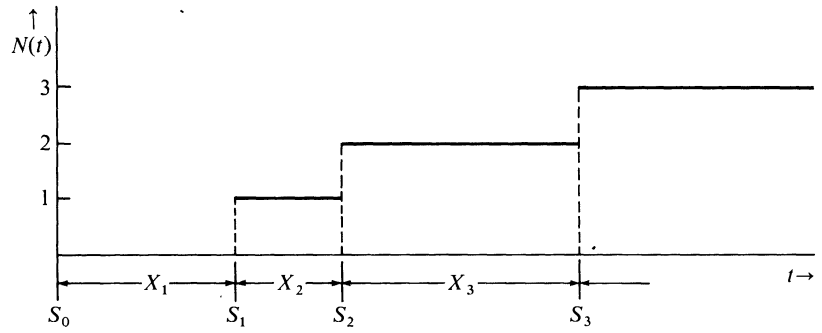


FIG. 1. The relation between the process of i.i.d.r.v's $\{X_n\}$ and the renewal counting process $N(t)$.

In common practice the counting process $\{N(t),\ t \geq 0\}$ and the partial sum process $\{S_n,\ n \geq 0\}$ are interchangeably called the "renewal process". The prototype physical renewal model involves successive replacements of light bulbs. A bulb is installed for service at time 0, fails at time X_1, and is then exchanged for a fresh bulb. The second bulb fails at time $X_1 + X_2$ and is replaced by a third bulb. In general, the nth bulb burns out at time $\sum_{i=1}^{n} X_i$ and is immediately replaced, etc. It is natural to assume that the successive lifetimes are statistically independent, with probabilistically identical characteristics in that

$$\Pr\{X_k \leq x\} = F(x).$$

Manifestly, $N(t)$ in this process records the number of renewals (light-bulb replacements) up to time t.

The principal objective of renewal theory is to derive properties of certain random variables associated with $\{N(t)\}$ and $\{S_n\}$ from knowledge of the interoccurrence distribution F. For example, it is of significance and relevance to compute the expected number of renewals for the time duration $(0, t]$:

$$E[N(t)] = M(t) \qquad \text{called the } \textit{renewal function}.$$

For this end, several pertinent relationships and formulas are worth recording. In principle, the probability law of $S_n = X_1 + \cdots + X_n$ can be calculated in accordance with the convolution formula

$$\Pr\{S_n \leq x\} = F_n(x),$$

where $F_1(x) = F(x)$ is assumed known or prescribed, and then

$$F_n(x) = \int_0^\infty F_{n-1}(x-y)\,dF(y) = \int_0^x F_{n-1}(x-y)\,dF(y).\dagger$$

We highlighted (1.2) earlier as the connecting link between the process $\{S_n\}$ and $\{N(t)\}$. The relation (1.2) can be expressed in the form

$$N(t) \geq k \qquad \text{if and only if} \quad S_k \leq t. \tag{1.3}$$

It follows instantly that

$$\begin{aligned} \Pr\{N(t) \geq k\} &= \Pr\{S_k \leq t\} \\ &= F_k(t), \qquad t \geq 0, \quad k = 1, 2, \ldots, \end{aligned} \tag{1.4}$$

and consequently

$$\begin{aligned} \Pr\{N(t) = k\} &= \Pr\{N(t) \geq k\} - \Pr\{N(t) \geq k+1\} \\ &= F_k(t) - F_{k+1}(t), \qquad t \geq 0, \quad k = 1, 2, \ldots. \end{aligned} \tag{1.5}$$

From the definition and taking cognizance of (1.4) and (1.5) we obtain

$$\begin{aligned} M(t) = E[N(t)] &= \sum_{k=1}^\infty k \Pr\{N(t) = k\} \\ &= \sum_{k=1}^\infty \Pr\{N(t) \geq k\} = \sum_{k=1}^\infty \Pr\{S_k \leq t\} = \sum_{k=1}^\infty F_k(t) \end{aligned}$$

(the third equality results after summation by parts. Consult also Elementary Problem 1, Chapter 1.) The convergence of the series

$$M(t) = \sum_{k=1}^\infty F_k(t) \tag{1.6}$$

will be verified formally in Section 4.

Problems 3, 16, and 17 concern the variance and higher moments of $N(t)$.

There are a number of other random variables of interest. Three of these are: the *excess life* (also called the excess random variable), the

† Here, and in what follows, our convention is to include the right endpoint in an integral over an interval, and omit the left. That is, $\int_a^b h(x)\,dG(x) = \int_{a+}^{b+} h(x)\,dG(x)$.

current life (also called the age random variable) and the *total life*, defined, respectively, by

$$\gamma_t = S_{N(t)+1} - t \qquad (\textit{excess or residual lifetime})$$
$$\delta_t = t - S_{N(t)} \qquad (\textit{current life or age random variable})$$
$$\beta_t = \gamma_t + \delta_t \qquad (\textit{total life}).$$

A pictorial description of these random variables is given in Figure 2.

The fundamental significance of these random functions in the theory of renewal processes and their relevance for applications will be amply developed throughout this chapter.

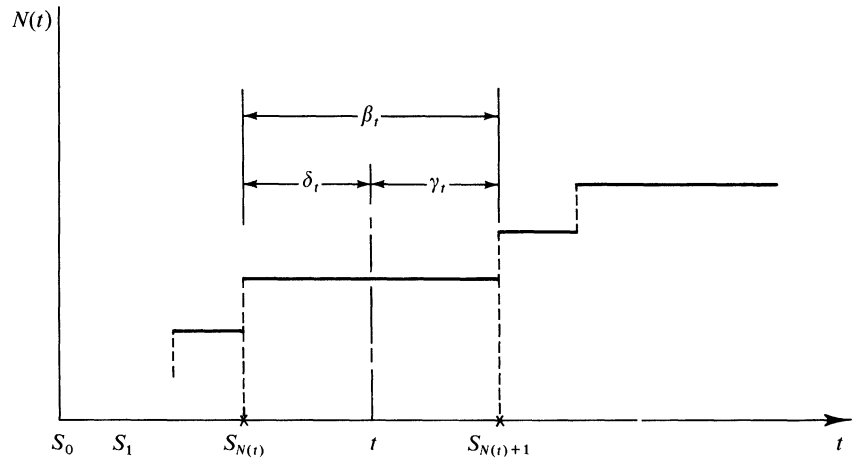


FIG. 2. The exess life γ_t, the current life δ_t, and the total life β_t.

2: *Some Examples of Renewal Processes*

The listing below points out the wide scope and diverse contexts in which renewal processes arise. Several of the examples will be studied in depth in later sections.

(a) *Poisson Processes*

A Poisson process $\{N(t), t \geq 0\}$ with parameter λ is a renewal counting process having the exponential interoccurrence distribution

$$F(x) = 1 - e^{-\lambda x}, \qquad x \geq 0,$$

as established in Theorem 2.1 of Chapter 4. This particular renewal process possesses a host of special features highlighted later in Section 3.

(b) *Counter Processes*

The times between successive electrical impulses or signals impinging on a recording device (counter) are often assumed to form a renewal process. Most physically realizable counters lock for some duration immediately upon registering an impulse and will not record impulses arriving during this dead period. Impulses are recorded only when the counter is freed (i.e., unlocked). Under quite reasonable assumptions, the sequence of events of the times of recorded impulses forms a renewal process, but it should be emphasized that the renewal process of recorded impulses is a *secondary* renewal process derived from the original renewal process comprised of the totality of all arriving impulses. Sections 3 and 7 elaborate the theory of counters concentrating on two kinds of locking mechanisms, the so-called counters of Types I and II.

(c) *Traffic Flow*

The distances between successive cars on an indefinitely long single-lane highway are often assumed to form a renewal process. So also are the time durations between consecutive cars passing a fixed location.

(d) *Renewal Processes Associated with Queues*

In a single-server queueing process there are imbedded many natural renewal processes. We cite two examples:

(i) If customer arrival times form a renewal process, then the times of the start of successive busy periods generate a second renewal process.

(ii) For the situation where the input process (the arrival pattern of customers) is Poisson, then the successive moments when the server passes from a busy to a free state determine a renewal process.

(e) *Inventory Systems*

In the analysis of most inventory processes it is customary to assume that the pattern of demands forms a renewal process. Most of the standard inventory policies induce renewal sequences, e.g., the times of replenishment of stock (see Section 8).

(f) *Renewal Processes Connected to Sums of Independent Random Variables*

(i) Let $\xi_1, \xi_2, \xi_3, \ldots$ be a sequence of real-valued (not necessarily positive) i.i.d. random variables. Assume $E(\xi_i) \geq 0$. Consider the

process of partial sums $U_0 = 0$, $U_n = \xi_1 + \xi_2 + \cdots + \xi_n$, $n \geq 1$, whose values range over the real line. Define

$$S_1 = \inf\{n : U_n > 0\} \qquad \text{and} \qquad S_k = \inf\{n : U_n > U_{S_{k-1}}\}, \quad k = 2, 3, \ldots,$$

$$X_1 = S_1 \qquad \text{and} \qquad X_k = S_k - S_{k-1}, \qquad k = 2, 3, \ldots .$$

It is clear that the sequence $X_1, X_2, X_3, \ldots$ constitutes a sequence of positive independent and identically distributed integer-valued random variables. The renewal process $S_m = X_1 + \cdots + X_m$, $m \geq 1$, can be interpreted as the successive times (indices) where U_n exceeds its previous maximum. In words, these form the sequence of successive new maxima.

(ii) The sequence $\{U_{S_m} - U_{S_{m-1}}\}_{m=1}^{\infty}$ also provides a sequence of i.i.d. positive random variables which therefore generate a renewal process. This example is of importance in the theory of fluctuations of sums of independent random variables (see Chapter 17 of Volume II.)

(g) *Renewal Processes in Markov Chains*

Let $Z_0, Z_1, \ldots$ be a recurrent Markov chain. Suppose $Z_0 = i$ and consider the times (elapsed number of generations) between successive visits to state i. Specifically,

$$X_1 = \min\{n > 0 : Z_n = i\},$$

and

$$X_{k+1} = \min\{n > X_k : Z_n = i\} - X_k, \qquad k = 1, 2, \ldots.$$

Since each of these times is computed from the same starting state i, the Markov property guarantees that $X_1, X_2, \ldots$ are i.i.d. and thus $\{X_k\}$ generates a renewal process. This fact permitted the application of renewal theory in Chapter 3 to prove the basic limit theorem for Markov chains.

(h) *Natural Embedded Renewal Processes*

Natural embedded renewal processes can be found in many diverse fields of applied probability including branching processes, insurance risk models, phenomena of population growth, evolutionary genetic mechanisms, engineering systems, econometric structures, and elsewhere.

3: *More on Some Special Renewal Processes*

A. THE POISSON PROCESS VIEWED AS A RENEWAL PROCESS

As mentioned earlier, the Poisson process with parameter λ is a renewal process whose interoccurrence times have the exponential distribution $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$. The memoryless property of the exponential distribution (see p.125) serves decisively in yielding the explicit computation of a number of functionals of the Poisson renewal process.

(i) *The Renewal Function*

Since $N(t)$ has a Poisson distribution,

$$\Pr\{N(t) = k\} = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \qquad k = 0, 1, \ldots,$$

and

$$M(t) = E[N(t)] = \lambda t.$$

(ii) *Excess Life*

Observe that the excess life at time t exceeds x if and only if there are no renewals in the interval $(t, t + x]$ (Figure 3). This event has the same probability as that of no renewals in the interval $(0, x]$, since a Poisson process has stationary independent increments. In formal terms, we have

$$\begin{aligned} \Pr\{\gamma_t > x\} &= \Pr\{N(t + x) - N(t) = 0\} \\ &= \Pr\{N(x) = 0\} = e^{-\lambda x}. \end{aligned} \tag{3.1}$$

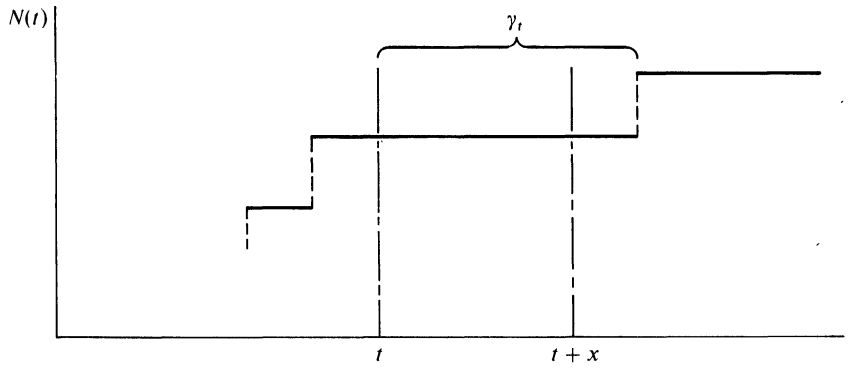


FIG. 3.

Thus, in a Poisson process, the excess life possesses the same exponential distribution

$$\Pr\{\gamma_t \leq x\} = 1 - e^{-\lambda x}, \qquad x \geq 0 \tag{3.2}$$

as every life, another manifestation of the memoryless property of the exponential distribution.

Equation (3.2) can be written in the less explicit form

$$\Pr\{\gamma_t > x\} = 1 - F(x), \qquad x \geq 0, \quad t > 0. \tag{3.3}$$

We will see in Section 8 that this identity characterizes the Poisson process among all renewal processes.

(iii) *Current Life*

The current life δ_t, of course, cannot exceed t, while for $x < t$ the current life exceeds x if and only if there are no renewals in $(t - x, t]$, which again has probability $e^{-\lambda x}$. Thus the current life follows the truncated exponential distribution

$$\Pr\{\delta_t \leq x\} = \begin{cases} 1 - e^{-\lambda x}, & \text{for } 0 \leq x < t, \\ 1, & \text{for } t \leq x. \end{cases} \tag{3.4}$$

It is convenient for later purposes to represent this formula in the guise

$$\Pr\{\delta_t \leq x\} = \begin{cases} F(x), & x < t, \\ 1, & x \geq t. \end{cases} \tag{3.5}$$

It will be shown in Section 8 that this property of δ_t also characterizes the Poisson process.

(iv) *Mean Total Life*

Using the evaluation of Problem 9 of Chapter 1 for the mean of a nonnegative random variable, we have

$$\begin{aligned} E[\beta_t] &= E[\gamma_t] + E[\delta_t] \\ &= \frac{1}{\lambda} + \int_0^t \Pr\{\delta_t > x\}\, dx \\ &= \frac{1}{\lambda} + \int_0^t e^{-\lambda x}\, dx \\ &= \frac{1}{\lambda} + \frac{1}{\lambda}(1 - e^{-\lambda t}). \end{aligned}$$

Observe that the mean total life is significantly larger than the mean life $1/\lambda = E[X_k]$ of any particular renewal interval. A more striking expression of this phenomenon is especially revealed when t is large, where the process has been in operation for a long duration. Then the mean total life $E(\beta_t)$ is approximately twice the mean life. These facts appear at first paradoxical.

Let us reexamine the manner of the definition of the total life β_t with a view to explaining on an intuitive basis the above seeming discrepancy. First, an arbitary time point t is fixed. Then β_t measures the length of the renewal interval containing the point t. Such a procedure will tend with higher likelihood to favor a lengthy renewal interval rather than one of short duration. The phenomenon is known as *length-biased sampling* and occurs, well disguised, in a number of sampling situations.

(v) *Joint Distribution of γ_t and δ_t*

The joint distribution of γ_t and δ_t is determined in the same manner as the marginals. In fact, for any $x > 0$ and $0 < y < t$, the event $\{\gamma_t > x, \delta_t > y\}$ occurs if and only if there are no renewals in the interval $(t-y, t+x]$, which has probability $e^{-\lambda(x+y)}$. Thus

$$\Pr\{\gamma_t > x, \delta_t > y\} = \begin{cases} e^{-\lambda(x+y)}, & \text{if } x > 0, \quad 0 < y < t, \\ 0, & \text{if } y \geq t. \end{cases} \tag{3.6}$$

Observe that for the Poisson process γ_t and δ_t are independent, since their joint distribution factors as the product of their marginal distributions. That this property also characterizes the Poisson process among renewal processes is incorporated as Problem 25 at the close of the chapter.

B. REPLACEMENT MODELS

Let $X_1, X_2, \ldots$ represent the lifetimes of items (light bulbs, transistor cards, machines, etc.) that are successively placed in service, the next item commencing service immediately following the failure of the previous one. We stipulate that $\{X_i\}$ are independent and identically distributed positive random variables with finite mean $\mu = E[X_k]$. Since each item lasts, on the average, μ time units, we would expect to replace items over the long run at a mean rate of $1/\mu$ per unit time. That is, we would expect

$$\frac{1}{t} M(t) \to \frac{1}{\mu}, \qquad \text{as} \quad t \to \infty. \tag{3.7}$$

This is indeed the case, as will be demonstrated in Section 4.

In the long run, any replacement strategy that substitutes items prior to their failure will use more than $1/\mu$ items per unit time. Nonetheless, where there is some benefit in avoiding failure in service, and where units deteriorate, in some sense, with age, there may be an advantage in considering alternative replacement strategies.

An *age-replacement policy* calls for replacing an item upon failure or upon reaching age T, whichever occurs first. Arguing intuitively, we would expect the long-run fraction of failure replacements, items that fail before age T, will be $F(T)$, and the corresponding fraction of (conceivably less expensive) planned replacements will be $1 - F(T)$. A renewal interval for this modified age-replacement policy obviously follows a distribution law

$$F_T(x) = \begin{cases} F(x), & \text{for } x < T, \\ 1, & \text{for } x \geq T, \end{cases}$$

and the mean renewal duration is

$$\mu_T = \int_0^\infty \{1 - F_T(x)\}\, dx = \int_0^T \{1 - F(x)\}\, dx < \mu.$$

The same reasoning that led to (3.7) indicates that the long-run mean replacement rate is increased to $1/\mu_T$.

Now, let $Y_1, Y_2, \ldots$ denote the times between *actual successive failures*. The random variable Y_1 is composed of a random number of time periods of length T (corresponding to replacements not associated with failures), plus a last time period in which the distribution is that of a failure conditioned on failure before age T; that is, Y_1 has the distribution of $NT + Z$, where

$$\Pr\{N \geq k\} = \{1 - F(T)\}^k, \qquad k = 0, 1, \ldots,$$

and

$$\Pr\{Z \leq z\} = F(z)/F(T), \qquad 0 \leq z \leq T.$$

Hence,

$$E[Y_1] = \frac{1}{F(T)} \{T[1 - F(T)] + \int_0^T (F(T) - F(x))\, dx\}$$

$$= \frac{1}{F(T)} \int_0^T \{1 - F(x)\}\, dx.$$

The sequence of random variables for interoccurrence times of the bona fide failures $\{Y_i\}$ generates a renewal process whose mean rate of failures

per unit time in the long run is $1/E[Y_1]$. This inference again relies on a parallel reasoning to that leading to (3.7). Depending on F, the modified failure rate $1/E(Y_1)$ may possibly yield a lower failure rate than $1/\mu$, the rate when replacements are made only upon failure.

Block replacement policies are suggested when there are a number of units functioning in parallel at any one time, and where, because of economies of scale, it costs less per unit to refurbish all units simultaneously than to substitute for items individually. A *block replacement policy* calls for replacing items individually upon failure and all items at the block times $T, 2T, 3T, \ldots$. Accordingly, there is exactly one planned or block replacement every T units of time and, on the average, $M(T)$ failure replacements. Thus the total long-run replacements per unit time is $\{1 + M(T)\}/T$, which may be compared to the corresponding figure for the other replacement strategies.

C. COUNTER MODELS

A counter is a device for detecting and registering instantaneous pulse-type signals. Familiar examples are the Geiger–Muller counter, which measures atmospheric cosmic radiation or source radiation, such as that emitted by a mass of radium. Another example is the electron multiplier.

All physically realizable counters are imperfect, incapable of detecting all signals that enter their detection chambers. After a particle or signal is registered, a counter must recuperate or renew itself in preparation for the next arrival. Signals arriving during the readjustment period, called *dead time* or *locked time*, are lost. We must distinguish between the *arriving* particles and the *recorded* particles. The experimenter observes only the particles recorded; from this he desires to infer the properties of the arrival process.

The particles or signals are assumed to arrive according to a renewal process with interarrival times $X_1, X_2, \ldots$, where $\Pr\{X_k \leq x\} = F(x)$. Counter models are distinguished in the nature of the locking time mechanisms. We describe the two most common types.

Type I Counters

A particle arrives at time 0 and locks the counter for a dead time duration Y_1. The first particle to be registered is the first particle to arrive after time Y_1. With the registration of the particle, the counter is blocked for a time length, say Y_2. The next particle to be registered is that of the first arrival once the counter is freed. This process is repeated, where the successive locking times, denoted by $Y_1, Y_2, Y_3, \ldots$, are assumed independent with common distribution $\Pr\{Y_k \leq y\} = G(y)$, and independent of the arrival process $\{X_k\}$.

Let Z_1 denote the elapsed time until the first signal is registered (not counting the one at the origin) and let Z_n, $n = 2, 3, \ldots$, be the elapsed time between the $(n-1)$st and nth registrations. Since the process starts afresh following each registration, $\{Z_k\}$ constitutes a renewal process. The action is diagrammed in Figure 4.

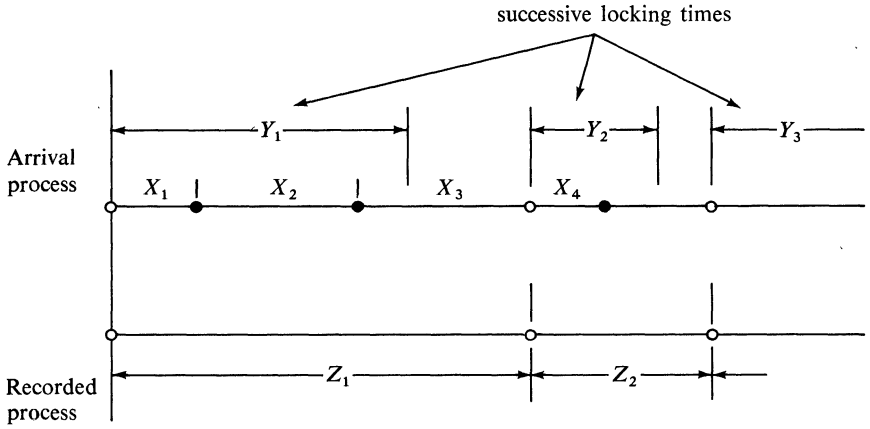


FIG. 4. A Type I Counter. ● denotes a lost signal and ○, a recorded signal.

Inspection of Figure 4 reveals that Z_1 is Y_1 plus the excess life at Y_1, or

$$Z_1 = Y_1 + \gamma_{Y_1} = S_{N(Y_1)+1},$$

where $S_k = X_1 + X_2 + \cdots + X_k$.

Since the $\{X_n\}$ and $\{Y_n\}$ processes are independent, invoking the law of total probability we obtain

$$\Pr\{Z_1 \leq z\} = \int_0^z \Pr\{y + \gamma_y \leq z \mid Y_1 = y\}\, dG(y)$$

$$= \int_0^z \{1 - A_{z-y}(y)\}\, dG(y),$$

where $A_x(t) = \Pr\{\gamma_t > x\}$.

An explicit formula for $A_x(t)$ is available (later) in Equation (6.1), so that the distribution of the time duration between counts in a Type I counter is completely specified.

In the long run, the mean rate per unit of time of recorded particles will be $1/E[Z_1]$, while the similar rate for arriving particles is $1/E[X_1]$. We will show later that the long-run fraction of recorded particles among the totality of all arriving particles is $E[X_1]/E[Z_1]$.

When the arrival process is Poisson, with mean λ, the memoryless property of the interoccurrence exponential distribution tells us that γ_t follows an exponential distribution, independent of t. Thus in this Poisson case

$$\Pr\{Z_1 \leq z\} = \int_0^z G(z-y)\lambda e^{-\lambda y}\, dy. \tag{3.8}$$

Type II Counters

Here the locking mechanism is more complicated. As before, an incoming signal is registered if and only if it arrives when the counter is free. Previously, however, only recorded particles induced the counter to lock. For Type II counters, every arriving signal can prolong the dead period of the counter, the associated locking times being added concurrently. For example, suppose the first particle locks the counter for a time duration σ_1 and a second pulse arrives at time $\tau < \sigma_1$ and independently engenders a locking time of extent σ_2; then the counter is next free at time σ_1 or $\tau + \sigma_2$, whichever occurs last assuming no additional particle arrivals prior to then. A typical realization of the process is diagrammed in Figure 5.

As with the Type I counter, let Z_n be the time between the $(n-1)$st and nth *recorded* pulses. Again $\{Z_k\}$ is a renewal process.

This counter process is quite difficult to analyze in general. We present a few results under the assumption that the arrival pattern is Poisson with rate λ.

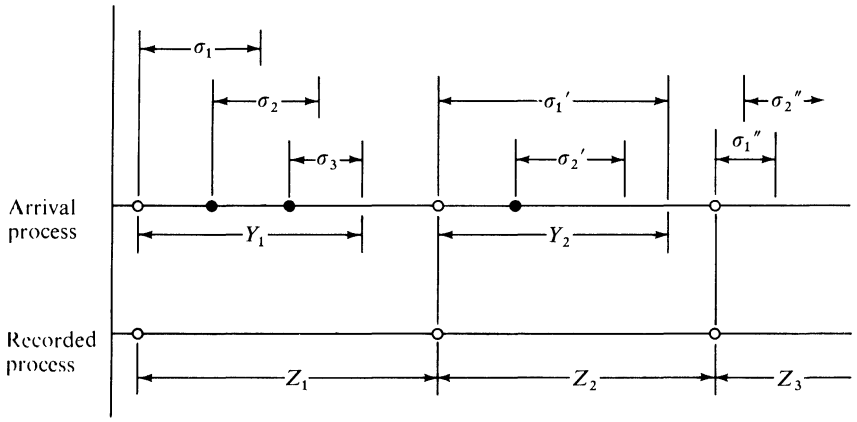


FIG. 5. A Type II counter. ● denotes a lost signal and ○, a recorded signal.

Let $p(t)$ be the probability that the counter is free at time t. We claim

$$p(t) = \exp\left\{-\lambda \int_0^t [1 - G(y)]\, dy\right\}, \tag{3.9}$$

where $G(y) = \Pr\{\sigma_k \leq y\}$. To derive this formula, recall from Theorem 2.3 of Chapter 4 that, given n occurrences of a Poisson process in the interval $(0, t]$, the distribution of the occurrence times is the same as that of n *independent* random variables taken from a uniform distribution on $(0, t]$. The counter is free at time t if and only if all dead periods (i.e., locking times) engendered by these n signals have terminated before time t. The probability is $G(t - y)$ that a dead period commencing at time y will end before time t. Conditional that a signal impinges during the time interval $(0, t]$, its actual arrival time has a uniform distribution. The requirement that its induced dead period is culminated prior to time t therefore has probability $\int_0^t G(t - y)\, dy/t$. Since the locking times are assumed independent and independent of the arrival process, we have

$$\Pr\{\text{counter free at time } t | n \text{ signals in } (0, t]\} = \left\{\int_0^t G(t - y)\frac{1}{t}\, dy\right\}^n.$$

But the number of signals arriving during the time interval $(0, t]$ has a Poisson distribution with mean λt. Invoking the law of total probabilities, we obtain

$$\begin{aligned} p(t) &= \sum_{j=0}^{\infty} \left\{\int_0^t G(t - y)\frac{1}{t}\, dy\right\}^j \frac{(\lambda t)^j e^{-\lambda t}}{j!} \\ &= \exp\left\{-\lambda t\left[1 - \int_0^t G(t - y)\frac{1}{t}\, dy\right]\right\} \\ &= \exp\left\{-\lambda \int_0^t [1 - G(y)]\, dy\right\}, \end{aligned}$$

and formula (3.9) is confirmed.

Continuing with our assumption of a Poisson stream of signals, $p(t)$ can be related to $M_R(t)$, the mean number of recorded signals in $(0, t]$. We claim

$$\frac{dM_R(t)}{dt} = \lambda p(t). \tag{3.10}$$

To prove this contention the following facts are relevant:

The probability of a signal appearing in the interval $(t, t+h]$ is $\lambda h + o(h)$. By definition, the probability of finding the counter free during this same short time period is $p(t) + o(h)$. Therefore, up to terms of order $o(h)$, $\lambda h p(t)$ is the probability of a recorded arrival in $(t, t+h]$. Since the probability of more than one signal in an interval $(t, t+h]$ is of order less than h as $h \downarrow 0$, we have

$$M_R(t+h) = M_R(t) + \lambda h p(t) + o(h),$$

where $o(h)$ incorporates all the negligible terms. Then

$$\frac{dM_R(t)}{dt} = \lim_{h \downarrow 0} \frac{M_R(t+h) - M_R(t)}{h} = \lambda p(t),$$

which gives (3.10).

Note that $M_R(0) = 0$, by its meaning. Now combining (3.9) and (3.10) and integrating yields

$$M_R(t) = \int_0^t \lambda \exp\left\{-\lambda \int_0^s [1 - G(y)]\, dy\right\} ds. \tag{3.11}$$

4: *Renewal Equations and the Elementary Renewal Theorem*

A. THE RENEWAL FUNCTION

In Section 1 we derived a formula for the mean number of counts in $(0, t]$. Explicitly,

$$M(t) = E[N(t)] = \sum_{j=1}^{\infty} F_j(t), \tag{4.1}$$

where

$$F_j(t) = \Pr\{S_j \leq t\}, \qquad t \geq 0.$$

Because of its far-reaching significance and relevance beyond its interpretation as a mean number of counts, $M(t)$ has been ascribed the special name, *the renewal function.*†

Our initial task will be to show that $M(t)$ for each $t > 0$ is finite. To this end, first we infer, on the basis of the definition and monotonic nature of $F_j(x)$, the inequality

$$F_n(t) = \int_0^t F_{n-m}(t-\xi)\, dF_m(\xi) \leq F_{n-m}(t) F_m(t), \qquad 1 \leq m \leq n-1.$$

† Recent authors tend to include the renewal that takes place at the origin and define the renewal function to be $1 + M(t) = E[1 + N(t)] = \sum_{n=0}^{\infty} F_n(t)$ where $F_0(t) = 1$ for $t \geq 0$ and 0 elsewhere. While this definition slightly simplifies some formulas, it calls more heavily on the Lebesgue–Stieltjes theory of integration than does the more traditional definition that we use.

In particular, for any integers k, r, and n, we have

$$F_{nr+k}(t) \leq F_{(n-1)r+k}(t)F_r(t).$$

Direct iteration leads to the relations

$$F_{nr+k}(t) \leq [F_r(t)]^n F_k(t), \qquad 0 \leq k \leq r-1. \tag{4.2}$$

In light of (4.2), we see that the series $M(t) = \sum_{k=1}^{\infty} F_k(t)$ for any t, where $F_r(t) < 1$, converges, indeed at least geometrically fast.

Since X_i are positive random variables, so that $F(0+) = 0$, and accordingly $F(t_0) < 1$ for some positive t_0, we infer inductively that for each $t > 0$ there must exist r fulfilling $F_r(t) < 1$. This fact is intuitive from probability considerations and equivalent to the statement that the partial sums S_n, sums of positive i.i.d.r.v.'s, increase to infinity with probability one.

Two important consequences emerge from the preceding discussion, which are now displayed for easy reference:

For any fixed t,

$$F_n(t) \to 0, \qquad \text{as} \quad n \to \infty \text{ (at least geometrically fast)}, \tag{4.3}$$

and

$$M(t) < \infty, \qquad \text{for all} \quad t.$$

Manifestly, $M(t)$ by its meaning (or from inspection of the formula (4.1)) is a nondecreasing function of t. Moreover, it readily can be checked that $M(t)$ is continuous from the right (since each $F_k(t)$ has this property and the series converges uniformly on finite intervals). Thus $M(t)$ is endowed with the characteristics of a distribution function apart from the exception that $M(\infty) = \lim_{t\to\infty} M(t) \neq 1$ (actually, $M(\infty) = \infty$). Nonetheless, it will be meaningful to write expressions of the type $\int a(t-y)\,dM(y)$, to be interpreted in a manner parallel to $\int a(t-y)\,dF(y)$, where F is a distribution function. In particular, in the common case in which M is differentiable with $m(t) = dM(t)/dt$, the integral $\int a(t-y)\,dM(y)$ reduces to $\int a(t-y)m(y)\,dy$.

At this point it is useful to generalize the notion of convolution to apply to any two increasing functions. Let A and B be nondecreasing functions, continuous from the right, with $A(0) = B(0) = 0$. *Define the convolution, denoted* $A * B$, by

$$A * B(t) = \int_0^t B(t-y)\,dA(y), \qquad t \geq 0. \tag{4.4}$$

Since $B(0) = 0$, we have $B(t-y) = \int_0^{t-y} dB(z)$. Inserting this in (4.4) and changing the order of integration produces

$$\begin{aligned} A * B(t) &= \int_0^t \left\{ \int_0^{t-y} dB(z) \right\} dA(y) \\ &= \int_0^t \left\{ \int_0^{t-z} dA(y) \right\} dB(z) \\ &= B * A(t), \end{aligned}$$

so that $*$ is a commutative operation.

We next show that the renewal function $M(t)$ satisfies the equation

$$M(t) = F(t) + \int_0^t M(t-y)\, dF(y), \qquad t \geq 0,$$

or, in convolution notation,

$$M(t) = F(t) + F * M(t), \qquad t \geq 0. \tag{4.5}$$

This identity will be validated invoking the *renewal argument*, which proceeds by conditioning on the time X_1 of the first renewal and counting the expected number of renewals thereafter. Manifestly, the probabilistic structure of events begins anew after the moment X_1, and consequently

$$E[N(t)|X_1 = x] = \begin{cases} 0, & \text{if } x > t, \\ 1 + M(t-x), & \text{if } x \leq t. \end{cases}$$

In words, there are no renewals in $(0, t]$ if the first lifetime X_1 exceeds t. On the other hand, where $X_1 = x < t$ there is the renewal engendered at time x plus, on the average, $M(t-x)$ further renewals occurring during the time interval extending from x to t. Applying the law of total probability yields

$$\begin{aligned} M(t) &= E[N(t)] \\ &= \int_0^t E[N(t)|X_1 = x]\, dF(x) \\ &= \int_0^t \{1 + M(t-x)\}\, dF(x) \\ &= F(t) + \int_0^t M(t-x)\, dF(x), \end{aligned}$$

and relation (4.5) is established.

Much of the power of renewal theory derives from the preceding method of reasoning that views the dynamic process starting anew at the occurrence of the first "event."

Renewal Equations

An integral equation of the form

$$A(t) = a(t) + \int_0^t A(t-x)\, dF(x), \qquad t \geq 0, \tag{4.6}$$

is called a renewal equation. The prescribed (or known) functions are $a(t)$ and the distribution function $F(t)$, while the undetermined (or unknown) quantity is $A(t)$.

Without ambiguity, we will employ the notation $B * c(t)$ for convolution of a function $c(t)$ (assumed reasonably smooth and bounded on finite intervals) with an increasing right-continuous function $B(t)$, $B(0) = 0$, to stand for

$$B * c(t) = \int_0^t c(t-\tau)\, dB(\tau). \tag{4.7}$$

Where $B'(t) = b(t)$ exists, then (4.7) reduces to

$$B * c(t) = \int_0^t c(t-\tau) b(\tau)\, d\tau,$$

provided this integral is well defined in the ordinary sense. Some elementary properties of $B * c$ are listed for ready reference leaving their straightforward validations to the student:

$$\begin{aligned} &\text{(i)} \quad \max_{0\leq t\leq T} |(B * c)(t)| \leq \max_{0\leq t\leq T} |c(t)| \cdot B(T) \text{ (consult also (4.11) below);} \\ &\text{(ii)} \quad B * c_1 + B * c_2 = B * (c_1 + c_2); \\ &\text{(iii)} \quad \text{If } B_1 \text{ and } B_2 \text{ are increasing, then} \\ &\qquad\qquad B_1 * (B_2 * c) = (B_1 * B_2) * c. \end{aligned} \tag{4.8}$$

Comparing (4.6) with (4.5) reveals that the renewal function $M(t)$ satisfies a renewal equation in which $a(t) = F(t)$. The following theorem affirms that the solution of an arbitrary renewal equation can be represented in terms of the renewal function.

Theorem 4.1. *Suppose a is a bounded function. There exists one and only one function A bounded on finite intervals that satisfies*

$$A(t) = a(t) + \int_0^t A(t-y)\, dF(y). \tag{4.9}$$

This function is

$$A(t) = a(t) + \int_0^t a(t-x)\, dM(x), \tag{4.10}$$

where $M(t) = \sum_{k=1}^{\infty} F_k(t)$ *is the renewal function.*

Proof. We verify first that A specified by (4.10) fulfills the requisite boundedness properties and indeed solves (4.9). Because a is a bounded function and M is nondecreasing and finite, for every T, it follows that

$$\begin{aligned}\sup_{0\le t\le T} |A(t)| &\le \sup_{0\le t\le T} |a(t)| + \int_0^T \left\{ \sup_{0\le y\le T} |a(y)| \right\} dM(x) \\ &= \sup_{0\le t\le T} |a(t)|\{1 + M(T)\} < \infty, \end{aligned} \tag{4.11}$$

establishing that the expression (4.10) is bounded on finite intervals. To check that $A(t)$ of (4.10) satisfies (4.9), we have

$$\begin{aligned} A(t) &= a(t) + M * a(t) \\ &= a(t) + \left(\sum_{k=1}^{\infty} F_k \right) * a(t) \\ &= a(t) + F * a(t) + \sum_{k=2}^{\infty} F_k * a(t) \\ &= a(t) + F * \left\{ a(t) + \left(\sum_{k=1}^{\infty} F_k \right) * a(t) \right\} \qquad \text{(since } F_k = F * F_{k-1} \text{ and using (4.8))} \\ &= a(t) + F * A(t). \end{aligned}$$

To complete the proof of Theorem 4.1 it remains to certify the uniqueness of A. This is done by showing that any solution of the renewal equation (4.9), bounded in finite intervals, is represented by (4.10).

Note for this end that the renewal equation (4.9) is suited to successive approximations by repeatedly substituting the expression for $A(t)$ into the right-hand side of (4.9) and expanding appropriately. We carry out this program using the convolution notation. Accordingly, we write (4.9) in the abbreviated form

$$A = a + F * A$$

and substitute for A on the right to get

$$\begin{aligned} A &= a + F * (a + F * A) \\ &= a + F * a + F * (F * A) \\ &= a + F * a + F_2 * A \qquad \text{(recall the identification } F_2 = F * F). \end{aligned}$$

Reliance on the properties of (4.8) has been tacitly implemented. We iterate this procedure, securing the equation

$$\begin{aligned} A &= a + F * a + F_2 * (a + F * A) \\ &= a + F * a + F_2 * a + F_3 * A = \cdots \\ &= a + \left(\sum_{k=1}^{n-1} F_k\right) * a + F_n * A. \end{aligned}$$

Next, observe that

$$\begin{aligned} |F_n * A(t)| &= \left| \int_0^t A(t-y)\, dF_n(y) \right| \\ &\leq \left\{ \sup_{0 \leq y \leq t} |A(t-y)| \right\} \times F_n(t). \end{aligned}$$

Since A is assumed bounded in finite intervals, and $\lim_{n\to\infty} F_n(t) = 0$ (consult (4.3)), it follows that $\lim_{n\to\infty} |F_n * A(t)| = 0$ for every fixed t. Similarly, since a is bounded, we obtain

$$\lim_{n\to\infty} \left(\sum_{k=1}^{n-1} F_k\right) * a(t) = \left(\sum_{k=1}^{\infty} F_k\right) * a(t) = M * a(t).$$

Thus

$$\begin{aligned} A(t) &= a(t) + \lim_{n\to\infty} \left\{ \sum_{k=1}^{n-1} F_k * a(t) + F_n * A(t) \right\} \\ &= a(t) + M * a(t), \end{aligned}$$

and the general solution A of (4.9) acquires the representation (4.10). The uniqueness proof of Theorem 4.1 is complete. ■

With the help of Theorem 4.1 we will prove the important relation

$$\begin{aligned} E[S_{N(t)+1}] &= E[X_1 + X_2 + \cdots + X_{N(t)+1}] \\ &= E[X_1] \cdot E[N(t) + 1] \\ &= E[X_1] \cdot [M(t) + 1] \end{aligned} \tag{4.12}$$

At first glance, this identity perhaps resembles an identity derived in Chapter 1 affirming the property that if $X_1, X_2, X_3, \ldots$ are independent identically distributed random variables and N is an integer-valued random variable independent of the X_i's, the equation $E[X_1 + \cdots + X_N] = E[X_1] \cdot E[N]$ prevails provided all mean values exist. The crucial difference in the present context is that the number of summands $N(t) + 1$ is *not* independent of the summand contributions themselves. For example, recall that in the Poisson discussion of Section 3, the total life β_t, the last summand involved in $S_{N(t)+1}$, has a mean that approached

twice the unconditional mean for t large. For this reason, it is *not* correct, in particular, that $E[S_{N(t)}]$ can be evaluated as the product of $E[X_1]$ and $E[N(t)]$. In view of these cautionary comments, identity (4.12) is more intriguing and remarkable. (It is actually a special case of what is known as the Wald identity; in this connection see Chapter 6 on martingales.)

To derive (4.12) we will use a renewal argument to establish a renewal equation for

$$A(t) = E[S_{N(t)+1}].$$

As usual, we condition on the time of the first renewal $X_1 = x$, and distinguish two contingencies: the first where $x > t$ so that $N(t) = 0$ and $S_{N(t)+1} = x$, and the second where $x \leq t$. A direct interpretation of the quantities involved readily validates the equation

$$E[S_{N(t)+1}|X_1 = x] = \begin{cases} x, & \text{if } x > t, \\ x + A(t-x), & \text{if } x \leq t. \end{cases}$$

Next, invoking the law of total probability yields

$$\begin{aligned} A(t) &= E[S_{N(t)+1}] \\ &= \int_0^\infty E[S_{N(t)+1}|X_1 = x]\, dF(x) \\ &= \int_0^t [x + A(t-x)]\, dF(x) + \int_t^\infty x\, dF(x) \\ &= \int_0^\infty x\, dF(x) + \int_0^t A(t-x)\, dF(x) \\ &= E[X_1] + \int_0^t A(t-x)\, dF(x). \end{aligned}$$

Thus $A(t) = E[S_{N(t)+1}]$ satisfies a renewal equation for which $a(t) =$ the constant $E[X_1]$. Theorem 4.1 states that

$$\begin{aligned} A(t) &= a(t) + \int_0^t a(t-x)\, dM(x) \\ &= E[X_1] + \int_0^t E[X_1]\, dM(x) \\ &= E[X_1] \times [1 + M(t)], \end{aligned}$$

which completes the proof of (4.12).

Observe that the excess life $\gamma_t = S_{N(t)+1} - t$ has

$$E[\gamma_t] = E[X_1] \cdot [1 + M(t)] - t. \tag{4.13}$$

At several occasions in Section 3 the intuitive result $M(t)/t \to 1/\mu$ as $t \to \infty$, where $\mu = E[X_1]$ was applied. We are now in a position to prove this important fact, commonly referred to as the *elementary renewal* theorem.

Theorem 4.2. *Let $\{X_i\}$ be a renewal process with $\mu = E[X_1] < \infty$. Then*

$$\lim_{t\to\infty} \frac{1}{t} M(t) = \frac{1}{\mu}.$$

Proof. It is always the case that $t < S_{N(t)+1}$. Combined with Eq. (4.12), we have

$$t < E[S_{N(t)+1}] = \mu[1 + M(t)],$$

and therefore

$$\frac{1}{t} M(t) > \frac{1}{\mu} - \frac{1}{t}.$$

It follows that

$$\liminf_{t\to\infty} \frac{1}{t} M(t) \geq \frac{1}{\mu}. \tag{4.14}$$

To establish the opposite inequality, let $c > 0$ be arbitrary, and set

$$X_i^c = \begin{cases} X_i, & \text{if } X_i \leq c, \\ c, & \text{if } X_i > c, \end{cases}$$

and consider the renewal process having lifetimes $\{X_i^c\}$.

Let S_n^c and $N^c(t)$ denote the waiting times and counting process, respectively, for this truncated renewal process generated by $\{X_i^c\}$. Since the random variables X_i^c are uniformly bounded by c, it is clear that $t + c \geq S^c_{N^c(t)+1}$, and therefore

$$t + c \geq E[S^c_{N^c(t)+1}] = \mu^c[1 + M^c(t)],$$

where

$$\mu^c = E[X_i^c] = \int_0^c \{1 - F(x)\}\, dx,$$

and

$$M^c(t) = E[N^c(t)].$$

Obviously $X_i^c \leq X_i$ entails $N^c(t) \geq N(t)$, and therefore $M^c(t) \geq M(t)$. It follows that

$$t + c \geq \mu^c[1 + M(t)],$$

and by rearrangement

$$\frac{1}{t} M(t) \leq \frac{1}{\mu_c} + \frac{1}{t}\left(\frac{c}{\mu^c} - 1\right).$$

Hence

$$\limsup_{t\to\infty} \frac{1}{t} M(t) \leq \frac{1}{\mu^c}, \qquad \text{for any} \quad c > 0. \tag{4.15}$$

Since

$$\lim_{c\to\infty} \mu^c = \lim_{c\to\infty} \int_0^c [1 - F(x)]\, dx$$
$$= \int_0^\infty [1 - F(x)]\, dx = \mu,$$

while the left-hand side of (4.15) is fixed, we deduce

$$\limsup_{t\to\infty} \frac{1}{t} M(t) \leq \lim_{c\to\infty} \frac{1}{\mu^c} = \frac{1}{\mu}. \tag{4.16}$$

Inequalities (4.14) and (4.16) in conjunction imply

$$\lim_{t\to\infty} \frac{1}{t} M(t) = \frac{1}{\mu},$$

and the proof of the theorem is complete. ■

5: *The Renewal Theorem*

The subject of this section involves one of the most basic theorems in applied probability. The renewal theorem can be regarded as a refinement of the asymptotic relation $M(t) \sim t/\mu$, $t \to \infty$, established in Theorem 4.2.

It can be interpreted as a differentiated form of the limit formula $\lim_{t\to\infty} M(t)/t = 1/\mu$. More explicitly, subject to certain mild conditions on $F(x)$, the renewal theorem asserts that for any $h > 0$

$$M(t + h) - M(t) \to \frac{h}{\mu}, \qquad \text{as} \quad t \to \infty. \tag{5.1}$$

In words, the expected number of renewals in an interval of length h is approximately h/μ, provided the process has been in operation for a long duration. The statement of Theorem 5.1, appearing in a different formulation but equivalent to (5.1), provides the optimum setting for application of the renewal theorem. Another perspective on the renewal theorem emphasizes its value in ascertaining the asymptotic character of solutions of renewal equations.

The proof of the renewal theorem is lengthy and demanding. We will omit the details and refer to Feller [1] for comprehensive details. However, its statement will be given with care so that the student can understand its meaning and be able to apply it unhesitatingly without ambiguity. Throughout the later sections, numerous applications will be forthcoming, and the implications of the basic renewal theorem will accordingly be well recognized. For the precise statement, we need several preliminary definitions. (The reader concerned only with application can read the remainder of this section cursorily.)

Definition 5.1. A point α of a distribution function F is called *a point of increase* if for every positive ε

$$F(\alpha + \varepsilon) - F(\alpha - \varepsilon) > 0.$$

A distribution function is said to be *arithmetic* if there exists a positive number λ such that F exhibits points of increase exclusively among the points $0, \pm\lambda, \pm 2\lambda, \ldots$. The largest such λ is called the *span* of F.

A distribution function F that has a continuous part is not arithmetic. The distribution function of a discrete random variable having possible values $0, 1, 2, \ldots$ is arithmetic with span 1.

Definition 5.2. Let g be a function defined on $[0, \infty)$. For every positive δ and $n = 1, 2, \ldots$, let

$$\underline{m}_n = \min\{g(t) : (n-1)\delta \leq t \leq n\delta\},$$
$$\bar{m}_n = \max\{g(t) : (n-1)\delta \leq t \leq n\delta\},$$
$$\underline{\sigma}(\delta) = \delta \sum_{n=1}^{\infty} \underline{m}_n, \qquad \text{and} \qquad \bar{\sigma}(\delta) = \delta \sum_{n=1}^{\infty} \bar{m}_n.$$

Then g is said to be *directly Riemann integrable* if both series $\underline{\sigma}(\delta)$ and $\bar{\sigma}(\delta)$ converge absolutely for every positive δ, and the difference $\bar{\sigma}(\delta) - \underline{\sigma}(\delta)$ goes to 0 as $\delta \to 0$.

Every monotonic function g which is absolutely integrable in the sense that

$$\int_0^{\infty} |g(t)|\, dt < \infty \tag{5.2}$$

is directly Riemann integrable, and this is the most important case for our purposes. Manifestly, all finite linear combinations of monotone functions satisfying (5.2) are also directly Riemann integrable.

Theorem 5.1. (The Basic Renewal Theorem). *Let F be the distribution function of a positive random variable with mean μ. Suppose that a is directly Riemann integrable and that A is the solution of the renewal equation*

$$A(t) = a(t) + \int_0^t A(t-x)\, dF(x). \tag{5.3}$$

(i) *If F is not arithmetic, then*

$$\lim_{t\to\infty} A(t) = \begin{cases} \dfrac{1}{\mu}\displaystyle\int_0^\infty a(x)\, dx, & \text{if } \mu < \infty, \\ 0, & \text{if } \mu = \infty. \end{cases}$$

(ii) *If F is arithmetic with span λ, then for all $c > 0$,*

$$\lim_{n\to\infty} A(c+n\lambda) = \begin{cases} \dfrac{\lambda}{\mu}\displaystyle\sum_{n=0}^{\infty} a(c+n\lambda), & \text{if } \mu < \infty, \\ 0, & \text{if } \mu = \infty. \end{cases}$$

There is a second form of the theorem, equivalent to that just given, but expressed more directly in terms of the renewal function. Let $h > 0$ be given, and examine the special prescription of

$$a(y) = \begin{cases} 1, & \text{if } 0 \le y < h, \\ 0, & \text{if } h \le y, \end{cases}$$

inserted in (5.3). In this example, for $t > h$, because of (4.10), we have

$$\begin{aligned} A(t) &= a(t) + \int_0^t a(t-x)\, dM(x) \\ &= \int_{t-h}^t dM(x) \\ &= M(t) - M(t-h), \end{aligned}$$

and $\mu^{-1}\int_0^\infty a(x)\, dx = h/\mu$. If F is not arithmetic, we may conclude on the basis of the renewal theorem that

$$\lim_{t\to\infty}[M(t) - M(t-h)] = h/\mu, \tag{5.4}$$

with the convention $h/\mu = 0$ when $\mu = \infty$.

The following converse prevails. Theorem 5.1 can be deduced from the fact of (5.4) by approximating a directly Riemann integrable function with step functions. The formal statement of the second form of the renewal theorem follows.

Theorem 5.2. *Let F be the distribution function of a positive random variable with mean μ. Let $M(t) = \sum_{k=1}^{\infty} F_k(t)$ be the renewal function associated with F. Let $h > 0$ be fixed.*

(i) *If F is not arithmetic, then*

$$\lim_{t\to\infty} [M(t+h) - M(t)] = h/\mu.$$

(ii) *If F is arithmetic, the same limit holds, provided h is a multiple of the span λ.*

We conclude this section by recovering Theorem 4.2, the elementary renewal theorem, namely,

$$\lim_{t\to\infty} \frac{1}{t} M(t) = \frac{1}{\mu}, \tag{5.5}$$

as a corollary of Theorem 5.2. To this end, set $b_n = M(n+1) - M(n)$. Then stipulating F to be not arithmetic, Theorem 5.2 tells us that $b_n \to 1/\mu$ as $n \to \infty$, so that also the average of b_n converges to the same limit. Thus

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} b_k = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} [M(k+1) - M(k)] = \lim_{n\to\infty} \frac{1}{n} M(n) = \frac{1}{\mu}.$$

Now for an arbitrary $t > 0$, let $[t]$ denote the largest integer not exceeding t. Cognizance of the monotone nature of $M(t)$ allows the facts of

$$\frac{[t]}{t} \frac{M([t])}{[t]} \leq \frac{M(t)}{t} \leq \frac{[t]+1}{t} \frac{M([t]+1)}{[t]+1}.$$

Since $t^{-1}M(t)$ is trapped by functions converging to μ^{-1}, (5.5) ipso facto follows. If F is arithmetic with span λ, we set $b_n = M[(n+1)\lambda] - M(n\lambda)$ and use an entirely parallel argument.

6: Applications of the Renewal Theorem

(a) *Limiting Distribution of the Excess Life*

Let $\gamma_t = S_{N(t)+1} - t$ be the excess life at time t and for a fixed $z > 0$, set

$$A_z(t) = \Pr\{\gamma_t > z\}.$$

We employ the renewal argument to establish a renewal equation for A_z in the usual way by conditioning on the time $X_1 = x$ of the first renewal. We obtain (we encourage the student to draw a picture)

$$\Pr\{\gamma_t > z | X_1 = x\} = \begin{cases} 1, & \text{if } x > t+z, \\ 0, & \text{if } t+z \geq x > t, \\ A_z(t-x), & \text{if } t \geq x > 0. \end{cases}$$

Then by the law of total probability,

$$\begin{aligned} A_z(t) &= \int_0^\infty \Pr\{\gamma_t > z | X_1 = x\}\, dF(x) \\ &= 1 - F(t+z) + \int_0^t A_z(t-x)\, dF(x). \end{aligned} \tag{6.1}$$

Theorem 4.1 yields

$$A_z(t) = 1 - F(t+z) + \int_0^t \{1 - F(t+z-x)\}\, dM(x).$$

To obtain a limiting distribution, we assume

$$\mu = E[X_1] = \int_0^\infty \{1 - F(x)\}\, dx < \infty.$$

Then

$$\int_0^\infty \{1 - F(t+z)\}\, dt = \int_z^\infty \{1 - F(y)\}\, dy < \infty,$$

and $\{1 - F(t+z)\}$ being monotonic, is directly Riemann integrable as a function of t with z fixed. Applying the renewal theorem yields

$$\lim_{t\to\infty} \Pr\{\gamma_t > z\} = \lim_{t\to\infty} A_z(t) = \mu^{-1} \int_z^\infty \{1 - F(y)\}\, dy, \qquad z > 0, \tag{6.2}$$

which displays the asymptotic distribution of the excess life.

Limiting distributions for the current life δ_t and the total life β_t can be deduced from the result of (6.2). Observe, with the aid of Fig. 6, we can directly corroborate the equivalence of the sets of events

$$\{\gamma_t \geq x \text{ and } \delta_t \geq y\} \qquad \text{if and only if} \quad \{\gamma_{t-y} \geq x + y\}. \tag{6.3}$$

It follows that

$$\begin{aligned} \lim_{t\to\infty} \Pr\{\delta_t \geq y,\, \gamma_t \geq x\} &= \lim_{t\to\infty} \Pr\{\gamma_{t-y} \geq x + y\} \\ &= \mu^{-1} \int_{x+y}^\infty \{1 - F(z)\}\, dz, \end{aligned} \tag{6.4}$$

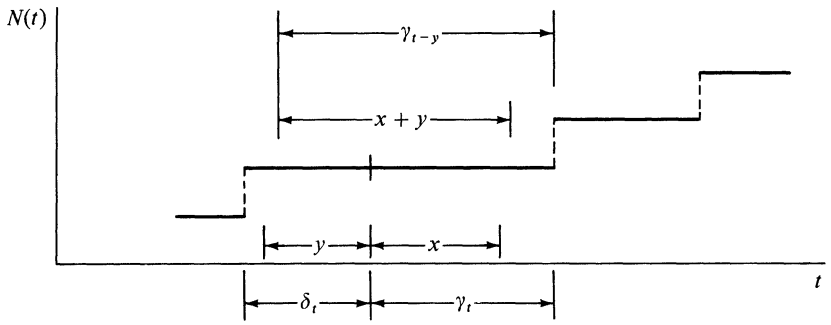


FIG. 6. Shows $\{\delta_t \geq y \text{ and } \gamma_t \geq x\}$ if and only if $\{\gamma_{t-y} \geq x+y\}$.

exhibiting the limiting joint distribution of $\{\delta_t, \gamma_t\}$. In particular,

$$\lim_{t\to\infty} \Pr\{\delta_t \geq y\} = \lim_{t\to\infty} \Pr\{\delta_t \geq y, \gamma_t \geq 0\}$$
$$= \mu^{-1} \int_y^\infty \{1 - F(z)\}\, dz. \tag{6.5}$$

The limiting distribution of $\beta_t = \delta_t + \gamma_t$ can be extracted from the limit formula (6.4). However it is equally quick to proceed via the renewal argument. We will be brief. Define

$$K_x(t) = \Pr\{\beta_t > x\}.$$

Conditioning on the time of the first renewal event, we have

$$\Pr\{\beta_t > x | X_1 = y\} = \begin{cases} 1, & \text{if } y > \max(x, t), \\ K_x(t-y), & \text{if } y \leq t, \\ 0, & \text{otherwise.} \end{cases}$$

The law of total probabilities produces the renewal equation

$$K_x(t) = 1 - F(\max(x, t)) + \int_0^t K_x(t-y)\, dF(y).$$

Application of the renewal theorem furnishes the limit law

$$\lim_{t\to\infty} \Pr\{\beta_t > x\} = \lim_{t\to\infty} K_x(t) = \frac{1}{\mu} \int_0^\infty [1 - F(\max(x, \tau))]\, d\tau$$
$$= \frac{1}{\mu} \int_x^\infty \xi\, dF(\xi), \tag{6.6}$$

where the last equality emanates from easy manipulations of the first integral together with integration by parts. Accordingly, we have established that the limiting distribution of the total life is

$$\lim_{t\to\infty} \Pr\{\beta_t \leq x\} = \frac{1}{\mu}\int_0^x \xi\, dF(\xi) = G(x).$$

When F has a density f, then the density of G is obviously $xf(x)/\mu$.

It is of interest to relate the mean of $G(x)$ with that of $F(x)$. Consider

$$\int_0^\infty x\, dG(x) = \frac{1}{\mu}\int_0^\infty x^2\, dF(x), \tag{6.7}$$

and let us compare this quantity with μ, the mean length of an arbitrary renewal interval. Note that the Schwarz inequality (see Chapter 1) implies

$$\int_0^\infty x^2\, dF(x) \geq \mu^2 = \left(\int_0^\infty x\, dF(x)\right)^2, \tag{6.8}$$

with strict inequality prevailing unless F is a degenerate distribution. The relation of (6.8) tells us that the mean limiting total life of the object in current operation strictly exceeds an ordinary mean lifetime. This fact is, of course, consistent with the analogous calculations done for the Poisson case (Section 3), which involved a mean total life exceeding the mean lifetime by a factor of 2. The inequality of (6.8) affirms the innate bias associated with sampling the lifetime interval that contains a prescribed point.

(b) *Asymptotic Expansion of the Renewal Function*

Suppose F is a nonarithmetic distribution with a finite variance σ^2. Under these assumptions we will determine the second term in the asymptotic expansion of $M(t)$ by proving

$$\lim_{t\to\infty} \{M(t) - \mu^{-1}t\} = \frac{\sigma^2 - \mu^2}{2\mu^2}.$$

Additional embellishments of the asymptotic behavior of $M(t)$ for t large can be ascertained employing parallel methods, where the existence of higher moments of F are stipulated.

Define

$$\begin{aligned} H(t) &= M(t) + 1 - \mu^{-1}t \\ &= E[N(t) + 1] - \mu^{-1}t \\ &= \mu^{-1}\{E[S_{N(t)+1}] - t\} \qquad \text{(by (4.12))} \\ &= \mu^{-1}E[\gamma_t] \qquad \text{(by (4.13)).} \end{aligned}$$

Once again, appeal to the renewal argument, conditioning on the time $X_1 = x$ of the first renewal, will provide a renewal equation for $H(t)$. Direct enumeration of cases leads to

$$E[\gamma_t | X_1 = x] = \begin{cases} x - t, & \text{if } x \geq t, \\ \mu H(t - x), & \text{if } x < t. \end{cases}$$

Invoking the law of total probability yields

$$\begin{aligned} \mu H(t) &= \int_0^\infty E[\gamma_t | X_1 = x] \, dF(x) \\ &= \int_t^\infty (x - t) \, dF(x) + \mu \int_0^t H(t - x) \, dF(x). \end{aligned}$$

Now

$$\begin{aligned} \int_t^\infty (x - t) \, dF(x) &= \int_0^\infty y \, dF(t + y) \\ &= \int_0^\infty \{1 - F(t + y)\} \, dy \end{aligned}$$

is a monotonic function of t, and expressing $1 - F(t + y) = \int_{t+y}^\infty dF(z)$ and interchanging the orders of integration leads to

$$\begin{aligned} &\int_0^\infty \left\{ \int_0^\infty [1 - F(t + y)] \, dy \right\} dt \\ &\quad = \int_0^\infty \int_0^\infty \int_{t+y}^\infty dF(z) \, dy \, dt \\ &\quad = \int_0^\infty \int_t^\infty \left\{ \int_0^{z-t} dy \right\} dF(z) \, dt \\ &\quad = \int_0^\infty \int_t^\infty (z - t) \, dF(z) \, dt \\ &\quad = \int_0^\infty \int_0^z (z - t) \, dt \, dF(z) \\ &\quad = \frac{1}{2} \int_0^\infty z^2 \, dF(z) = \frac{1}{2} (\sigma^2 + \mu^2) < \infty. \end{aligned}$$

Thus the renewal theorem implies

$$\lim_{t\to\infty} \mu H(t) = \mu^{-1}\tfrac{1}{2}(\sigma^2 + \mu^2),$$

or

$$\begin{aligned}\lim_{t\to\infty}\{M(t) - \mu^{-1}t\} &= \lim_{t\to\infty}\{H(t) - 1\} \\ &= \frac{\sigma^2 + \mu^2}{2\mu^2} - 1 \\ &= \frac{\sigma^2 - \mu^2}{2\mu^2},\end{aligned}$$

as was to be shown.

7: *Generalizations and Variations on Renewal Processes*

A. DELAYED RENEWAL PROCESSES

We continue to assume that $\{X_k\}$ are all independent positive random variables, but only $X_2, X_3, \ldots$ (*from the second on*) are identically distributed with distribution function F, while X_1 has possibly a different distribution function G. Such a process is called a *delayed renewal process.* We have all the ingredients for an ordinary renewal process except that the initial time to the first renewal has a distribution different from that of the other interoccurence times.

One way in which a delayed renewal process arises is when the component in operation at time $t = 0$ is not new. For example, suppose that the time origin is taken y time units after the start of an ordinary renewal process. Then the time to the first renewal after the origin in the delayed process will have the distribution of the excess life at time y of an ordinary renewal process.

As previously, let $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$, and let $N(t)$ count the number of renewals up to time t. But now it is essential to distinguish between the mean number of renewals in the delayed process

$$M_D(t) = E[N(t)],$$

and the renewal function associated with the distribution F,

$$M(t) = \sum_{k=1}^{\infty} F_k(t).$$

Conditioning on the time of the first renewal, noting

$$E[N(t)|X_1 = x] = \begin{cases} 0, & \text{if } x > t, \\ 1 + M(t-x), & \text{if } x \le t. \end{cases}$$

and following with implementation of the law of total probability gives

$$\begin{aligned} M_D(t) &= \int_0^\infty E[N(t)|X_1 = x]\, dG(x) \\ &= \int_0^t \{1 + M(t-x)\}\, dG(x) \\ &= G(t) + \int_0^t M(t-x)\, dG(x) \\ &= G(t) + \int_0^t G(t-x)\, dM(x). \end{aligned} \tag{7.1}$$

Manifestly, Eq. (7.1) displays $M_D(t)$ as the solution of the renewal equation [compare with (4.9) and (4.10)]

$$M_D(t) = G(t) + \int_0^t M_D(t-x)\, dF(x). \tag{7.2}$$

We will show that $M_D(t)$ obeys the renewal theorem, assuming that F is a nonarithmetic distribution. (An analogous approach works in the arithmetic case.) From (7.1) we recall, for any $t > 0$,

$$M_D(t) = G(t) + \int_0^t M(t-x)\, dG(x),$$

and in particular, for $t > h$,

$$M_D(t-h) = G(t-h) + \int_0^{t-h} M(t-h-x)\, dG(x).$$

Agreeing that $M(x) = 0$ for $x < 0$, the difference of these equations becomes

$$M_D(t) - M_D(t-h) = G(t) - G(t-h) + \int_0^t \{M(t-x) - M(t-h-x)\}\, dG(x).$$

It is convenient to decompose the integral into two sections, viz.,

$$\int_0^{t/2} \{M(t-x) - M(t-h-x)\}\, dG(x) + \int_{t/2}^t \{M(t-x) - M(t-h-x)\}\, dG(x).$$

Since $\lim_{t\to\infty} \{M(t-x) - M(t-h-x)\} = h/\mu$, the first integral converges to h/μ, while the second converges to zero, since $\{M(t-x) - M(t-h-x)\}$,

being convergent, is a bounded function of x. Of course, $G(t) - G(t/2) \to 0$ as $t \to \infty$, so that, in summary,

$$\lim_{t\to\infty} [M_D(t) - M_D(t-h)] = h/\mu.$$

B. STATIONARY RENEWAL PROCESSES

A delayed renewal process for which the first life has the distribution function

$$G(x) = \mu^{-1} \int_0^x \{1 - F(y)\}\, dy$$

is called a stationary renewal process. We are attempting to model a renewal process that began indefinitely far in the past, so that the remaining life of the item in service at the origin has the limiting distribution of the excess life in an ordinary renewal process. We recognize G as this limiting distribution.

It is anticipated that such a process exhibits a number of stationary or time-invariant properties. We will content ourselves with showing that, for a stationary renewal process,

$$M_D(t) = E[N(t)] \equiv t/\mu, \tag{7.3}$$

and

$$\Pr\{\gamma_t^D \le x\} = G(x),$$

for all t. Thus, what is in general only an asymptotic renewal relation becomes an identity, holding for all t, in a stationary renewal process.

Recall from Eq. (7.2) that $M_D(t)$ satisfies the renewal equation

$$M_D(t) = G(t) + \int_0^t M_D(t-x)\, dF(x), \tag{7.4}$$

and since the solution to such an equation is unique (modulo suitable boundedness restrictions, see Theorem 4.1), we need merely check that $M_D(t) \equiv t/\mu$ satisfies (7.4). We have

$$\begin{aligned} G(t) + \int_0^t M_D(t-x)\, dF(x) &= \mu^{-1} \int_0^t \{1 - F(x)\}\, dx + \mu^{-1} \int_0^t (t-x)\, dF(x) \\ &= \mu^{-1} t + \mu^{-1} \left\{ \int_0^t (t-x)\, dF(x) - \int_0^t F(y)\, dy \right\} \\ &= \mu^{-1} t \end{aligned}$$

since the part in braces vanishes as can be checked by performing an integration by parts. The identity $M_D(t) \equiv t/\mu$ is hereby confirmed.

We will follow the same procedure to validate the equation $\Pr\{\gamma_t^{\mathrm{D}} \leq x\} = G(x)$ for all x, where γ_t^{D} is the excess life in the delayed (stationary) renewal process. Let

$$A_x^{\mathrm{D}}(t) = \Pr\{\gamma_t^{\mathrm{D}} > x\}, \qquad \text{and} \qquad A_x(t) = \Pr\{\gamma_t > x\},$$

where γ_t is the excess life in an ordinary renewal process. The standard renewal argument leads to

$$A_x^{\mathrm{D}}(t) = 1 - G(t+x) + \int_0^t A_x(t-y)\, dG(y),$$

or

$$A_x^{\mathrm{D}}(t) = 1 - G(t+x) + G * A_x(t). \tag{7.5}$$

The renewal equation

$$A_x(t) = 1 - F(t+x) + F * A_x(t)$$

appeared earlier, in our deliberations of Section 6. By virtue of Theorem 4.1, the solution can be represented in the form

$$A_x(t) = a_x(t) + M * a_x(t), \tag{7.6}$$

where

$$a_x(t) = 1 - F(t+x).$$

Inserting (7.6) into (7.5) and citing the formula $M_{\mathrm{D}}(t) = G(t) + G * M(t)$ (this comes out directly from the definitions involved), we obtain

$$\begin{aligned} A_x^{\mathrm{D}}(t) &= 1 - G(t+x) + G * a_x(t) + G * M * a_x(t) \\ &= 1 - G(t+x) + M_{\mathrm{D}} * a_x(t) \\ &= 1 - G(t+x) + \int_0^t a_x(t-y)\, dM_{\mathrm{D}}(y). \end{aligned}$$

Now $a_x(t-y) = 1 - F(t+x-y)$ and $M_{\mathrm{D}}(y) \equiv y/\mu$, so that $dM_{\mathrm{D}}(y) = \mu^{-1}\, dy$. Then

$$\begin{aligned} A_x^{\mathrm{D}}(t) &= 1 - G(t+x) + \mu^{-1} \int_0^t \{1 - F(t+x-y)\}\, dy \\ &= 1 - G(t+x) + \mu^{-1} \int_x^{t+x} \{1 - F(u)\}\, du \\ &= 1 - G(t+x) + G(t+x) - G(x) \\ &= 1 - G(x), \end{aligned}$$

as was to be shown.

C. CUMULATIVE AND RELATED PROCESSES

Suppose associated with the ith unit or lifetime interval is a second random variable Y_i ($\{Y_i\}$ identically distributed) in addition to the lifetime X_i. We allow X_i and Y_i to be dependent, but assume that the pairs (X_1, Y_1), (X_2, Y_2), ... are independent. We use the notation $F(x) = \Pr\{X_i \leq x\}$, $G(y) = \Pr(Y_i \leq y\}$, $\mu = E[X_i]$, and $\nu = E[Y_i]$.

A number of problems of practical and theoretical interest have a natural formulation in these terms.

I. *Renewal Processes Involving Two Components to Each Renewal Interval*

Suppose that

$$Y_i \text{ represents a portion of the duration } X_i.$$

Figure 7 illustrates the model. In Fig. 7 we have depicted the Y portion occurring at the beginning of the interval, but this is not essential for the results that follow.

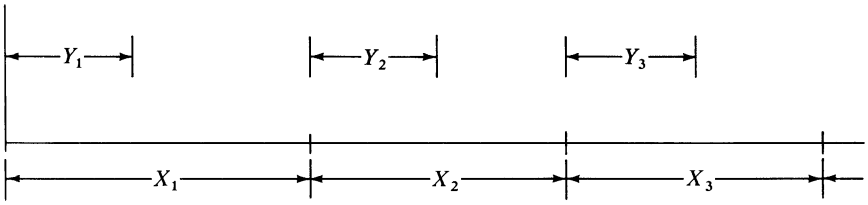


FIG. 7. A renewal process in which an associated random variable Y_i represents a portion of the ith renewal interval.

Let $p(t)$ be the probability that t falls in a Y portion of some renewal interval. By conditioning on the length of the first interval $X_1 = x$ and distinguishing the two possibilities $x < t$ and $x \geq t$, we arrive (by what is now a routine methodology) at the renewal equation

$$p(t) = \Pr\{t \text{ is covered by } Y_1\} + \int_0^t p(t - \xi)\, dF(\xi).$$

Let

$$I_{Y_1}(t) = \begin{cases} 1, & \text{if } Y_1 \text{ covers } t, \\ 0, & \text{if } Y_1 \text{ does not cover } t. \end{cases}$$

Then

$$\Pr\{t \text{ is covered by } Y_1\} = E[I_{Y_1}(t)],$$

and

$$\begin{aligned}\int_0^\infty \Pr\{t \text{ is covered by } Y_1\}\, dt &= \int_0^\infty E[I_{Y_1}(t)]\, dt \\ &= E\left[\int_0^\infty I_{Y_1}(t)\, dt\right] \\ &= E[Y_1] = \nu,\end{aligned}$$

since the totality of points covered by Y_1 is Y_1. Now applying the renewal theorem, we conclude that if F is nonarithmetic and $\Pr\{t$ is covered by $Y_1\}$ is directly Riemann integrable, then

$$\begin{aligned}\lim_{t\to\infty} p(t) &= \mu^{-1} \int_0^\infty \Pr\{t \text{ is covered by } Y_1\}\, dt \\ &= \nu/\mu. \end{aligned} \tag{7.7}$$

Here are some concrete examples.

(a) *A Replacement Model*

Consider a replacement model in which replacement is not instantaneous. Let Y_i be the operating time and Z_i the lag period preceding installment of the $(i+1)$st operating unit. (The delay in replacement can be conceived as a period of repair of the service unit.) We assume that the sequence of times between successive replacements $X_k = Y_k + Z_k$, $k = 1, 2, \ldots$, constitutes a renewal process. Then $p(t)$, the probability that the system is in operation at time t, converges to $E[Y_1]/E[X_1]$, provided the distribution of X_k is nonarithmetic.

(b) *A Queuing Model*

If arrivals to a queue follow a Poisson process, then the successive times X_k from the commencement of the kth busy period to the start of the next busy period form a renewal process. (A busy period is an uninterrupted duration when the queue is not empty.) Each X_k is composed of a busy portion Z_k and an idle portion Y_k. Then $p(t)$, the probability that the queue is empty at time t, converges to $E[Y_1]/E[X_1]$.

(c) *A Counter Problem*

Let X_k, $k = 1, 2, \ldots$, denote the sequence of times between successive recorded particles in a counter and let Y_k represent the dead (blocked) time during the X_k renewal period. Then $p(t)$, the probability that the counter is blocked at time t, converges to $E[Y_1]/E[X_1]$.

II. *Cumulative Processes*

Interpret Y_i as a cost or value, etc., associated with the ith renewal cycle. A class of problems with natural setting in this general context of pairs (X_i, Y_i), where X_i generates a renewal process, will now be considered. Interest here focuses on the so-called *cumulative process*

$$W(t) = \sum_{k=1}^{N(t)+1} Y_k,$$

the accumulated costs or what-have-you up to time t (assuming transactions are made at the beginning of a renewal cycle). By conditioning on the time $X_1 = x$ until the first renewal, and examining the two possibilities $x > t$ and $x \leq t$, we secure for $A(t) = E[W(t)]$ the renewal equation

$$A(t) = E[Y_1] + \int_0^t A(t-x)\, dF(x).$$

An appeal to Theorem 4.1 yields the formula

$$\begin{aligned} A(t) &= E[Y_1] + \int_0^t E[Y_1]\, dM(x) \\ &= E[Y_1][1 + M(t)]. \end{aligned}$$

It follows immediately that, where F is nonarithmetic and $h > 0$, then

$$\lim_{t\to\infty} [A(t) - A(t-h)] = E[Y_1]h/\mu,$$

and in any case,

$$\lim_{t\to\infty} \frac{1}{t} A(t) = E[Y_1]/\mu.$$

This justifies the interpretation of $E[Y_1]/\mu$ as a long-run mean cost, value, etc., per unit time, an interpretation that was used repeatedly in the examples of Section 3.

Here are some examples of cumulative processes.

(a) *Replacement Models*

Suppose Y_i is the cost of the ith replacement. Let us suppose that under an age-replacement strategy (see Example B, Section 3) a planned replacement at age T costs c_1 dollars, while a failure replaced at time $< T$ costs c_2 dollars. If Y_k is the cost incurred at the kth replacement cycle, then

$$Y_k = \begin{cases} c_1 & \text{with probability } 1 - F(T), \\ c_2 & \text{with probability } F(T), \end{cases}$$

and $E[Y_k] = c_1[1 - F(T)] + c_2 F(T)$. Since the expected length of a replacement cycle is

$$E[\min\{X_k, T\}] = \int_0^T [1 - F(x)]\, dx,$$

we have that the long-run cost per unit time is

$$\frac{c_1[1 - F(T)] + c_2 F(T)}{\int_0^T [1 - F(x)]\, dx},$$

and in any particular situation a routine calculus exercise or recourse to numerical computation produces the value of T that minimizes the long-run cost per unit time.

Under a block replacement policy, there is one planned replacement every T units of time and, on the average, $M(T)$ failure replacements, so the expected cost is $E[Y_k] = c_1 + c_2 M(T)$, and the long-run mean cost per unit time is $\{c_1 + c_2 M(T)\}/T$.

(b) *Counter Models*

In a counter model (see Example C, Section 3), let Y_k be the number of unregistered signals that arise during the period X_k between the $(k-1)$st and kth recorded signals. Then the long-run mean number of uncounted particles per unit time is $E[Y_1]/E[X_1]$.

(c) *Risk Theory*

Suppose claims arrive at an insurance company according to a renewal process with interoccurrence times $X_1, X_2, \ldots$. Let Y_k be the magnitude of the kth claim. Then $W(t) = \sum_{k=0}^{M(t)+1} Y_k$ represents the cumulative amount claimed up to time t, and the long-run mean claim rate is

$$\lim_{t\to\infty} \frac{1}{t} E[W(t)] = E[Y_1]/E[X_1].$$

D. TERMINATING RENEWAL PROCESSES

Suppose we allow the possibility of infinite interoccurrence times in a renewal process. Such a process is called a *terminating* renewal process, since the renewals cease at the first infinite interoccurrence time. The situation is diagrammed in Fig. 8.

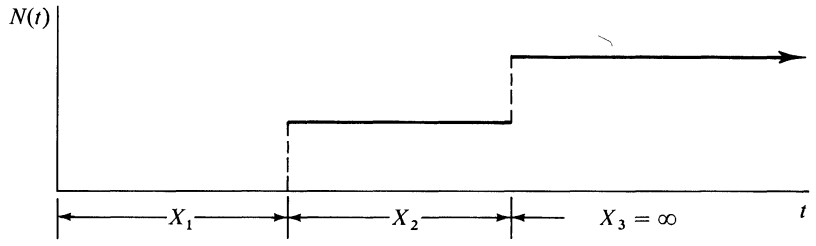


FIG. 8.

Let $L = F(\infty) = \Pr\{X_k < \infty\} < 1$ and $1 - L = \Pr\{X_k = \infty\} > 0$. Then the total number of renewals in all time, denoted by $N(\infty)$, is a finite-valued random variable and follows the geometric probability law

$$\Pr\{N(\infty) \geq k\} = L^k, \qquad k = 0, 1, 2, \ldots,$$

with

$$E[N(\infty)] = \sum_{k=1}^{\infty} \Pr\{N(\infty) \geq k\}$$
$$= L/(1 - L).$$

The realizations of the termination process still have the property $N(t) \geq k$ if and only if $S_k \leq t$, so that

$$\Pr\{N(t) \geq k\} = \Pr\{S_k \leq t\} = F_k(t),$$

and

$$M(t) = E[N(t)] = \sum_{k=1}^{\infty} F_k(t) < \sum_{k=1}^{\infty} L^k = L/(1 - L).$$

Moreover, the renewal argument continues to work, entailing the equation

$$M(t) = F(t) + \int_0^t M(t - x)\, dF(x).$$

However, the renewal theorem is *not* automatically applicable owing to the fact that F is not a proper probability distribution function. Fortunately, there is often a way to overcome this lacuna. Suppose that

$$g(s) = \int_0^{\infty} e^{sx}\, dF(x)$$

is a finite function of s for $s \geq 0$. Then g will be continuous and $g(0) = L < 1$ and $\lim_{s \to \infty} g(s) = \infty$, implying the existence of a unique positive value $s_0 = \lambda > 0$, for which

$$g(\lambda) = \int_0^{\infty} e^{\lambda x}\, dF(x) = 1.$$

Define $\hat{F}(t) = \int_0^t e^{\lambda x}\, dF(x)$. Then $\hat{F}(t)$ is nondecreasing and $\lim_{t\to\infty} \hat{F}(t) = 1$, showing that $\hat{F}$ is a proper distribution function. Now consider a renewal equation of the form

$$A(t) = a(t) + \int_0^t A(t-x)\, dF(x).$$

Set $\hat{A}(t) = e^{\lambda t}A(t)$, $\hat{a}(t) = e^{\lambda t}a(t)$, and verify

$$\begin{aligned}\hat{A}(t) &= e^{\lambda t}A(t)\\ &= e^{\lambda t}a(t) + \int_0^t e^{\lambda(t-x)}A(t-x)e^{\lambda x}\, dF(x)\\ &= \hat{a}(t) + \int_0^t \hat{A}(t-x)\, d\hat{F}(x),\end{aligned}$$

indicating that $\hat{A}$ satisfies a renewal equation now involving a proper distribution function, to which the renewal theorem can be applied. As a specific example, consider $A(t) = M(\infty) - M(t) = L/(1-L) - M(t)$. Equivalently, $A(t) = E[N(\infty) - N(t)]$ is the mean number of indices n for which $t < S_n < \infty$. We now develop a renewal equation satisfied by $A(t)$. In fact, we have

$$E[N(\infty) - N(t)|X_1 = x] = \begin{cases} 1 + L/(1-L), & \text{if } x > t,\\ A(t-x), & \text{if } 0 < x \le t.\end{cases}$$

It follows, using the law of total probabilities, that

$$\begin{aligned}A(t) &= \int_0^\infty E[N(\infty) - N(t)|X_1 = x]\, dF(x)\\ &= \{L - F(t)\}/(1-L) + \int_0^t A(t-x)\, dF(x).\end{aligned}$$

Next check that

$$\hat{a}(t) = e^{\lambda t}\{L - F(t)\}/(1-L)$$

is directly Riemann integrable. Note also

$$\begin{aligned}\int_0^\infty \hat{a}(t)\, dt &= \frac{1}{1-L}\int_0^\infty e^{\lambda t}\{L - F(t)\}\, dt\\ &= \frac{1}{1-L}\int_0^\infty e^{\lambda t}\int_t^\infty dF(x)\, dt\end{aligned}$$

$$= \frac{1}{1-L} \int_0^\infty \int_0^x e^{\lambda t}\, dt\, dF(x)$$

$$= \frac{1}{1-L} \int_0^\infty \frac{(e^{\lambda x}-1)}{\lambda}\, dF(x)$$

$$= \frac{1}{1-L} \frac{(1-L)}{\lambda} = \frac{1}{\lambda}.$$

Thus

$$\lim_{t\to\infty} \hat{A}(t) = \lim_{t\to\infty} e^{\lambda t}[M(\infty) - M(t)]$$

$$= \left\{\lambda \int_0^\infty x e^{\lambda x}\, dF(x)\right\}^{-1}$$

We conclude that $M(t)$ approaches $M(\infty) = L/(1-L)$ exponentially fast at rate λ.

E. ALTERNATING AND MARKOV RENEWAL PROCESSES

An *alternating renewal process* is a sequence $Y_1, Y_2, \ldots$ of independent random variables, where

$$\begin{array}{ll} Y_1, Y_{r+1}, Y_{2r+1}, \ldots & \text{have distribution function } F_1, \\ Y_2, Y_{r+2}, Y_{2r+2}, \ldots & \text{have distribution function } F_2, \\ \vdots & \\ Y_r, Y_{2r}, Y_{3r}, \ldots & \text{have distribution function } F_r. \end{array}$$

We think of a system passing successively through states 1, 2, ..., r, 1, 2, ..., r, 1, 2, ..., and sojourning a random time period during each visit to each state.

Let $p_i(t)$ be the probability that the system is in state i at time t. From relation (7.7) of Part C, we infer

$$\lim_{t\to\infty} p_i(t) = \mu_i/(\mu_1 + \cdots + \mu_r),$$

where

$$\mu_i = E[Y_i] < \infty, \qquad i = 1, \ldots, r,$$

provided the distribution $F = F_1 * F_2 \cdots * F_r$ is nonarithmetic.

A *Markov renewal* or *semi-Markov* process passes through states 1, ..., r according to a Markov chain having transition probability matrix $P = \|P_{ij}\|_{i,j=1}^r$. The time spent in state i, given that the next state is j, has distribution function F_{ij}, and, conditioned on the sequence of states,

all sojourn times are assumed independent. The unconditional distribution function of the sojourn time in a state i is $F_i(t) = \sum_{j=1}^{r} P_{ij} F_{ij}(t)$, which is postulated to have a finite mean μ_i. Assume that the Markov chain is irreducible and recurrent, with stationary distribution given by $\pi_j = \sum_i \pi_i P_{ij}$.

Suppose the process starts in a fixed state i, and let a state k be prescribed. Call the duration between one visit to state i and the next an *i-cycle.* The sequence of times between these successive visits to state i forms a renewal process. From relation (7.7) and assuming at least one F_i is not arithmetic, the probability $p_k(t)$ of being in state k at time t converges to the mean time spent in state k during an i-cycle divided by the mean duration of an i-cycle. By the law of total probability, the mean time in state k during an i-cycle is the product of μ_k times the mean number of visits to k in the intervening time between successive visits to state i. The second factor depends only on the discrete-time Markov chain of state visits and therefore is necessarily proportional to π_k. It follows that

$$\lim_{t\to\infty} p_k(t) = c\pi_k \mu_k,$$

when c is a constant of proportionality. Since these probabilities necessarily sum to 1, $c = 1/(\pi_1\mu_1 + \cdots + \pi_r \mu_r)$.

F. CENTRAL LIMIT THEOREM FOR RENEWALS

Theorem 7.1. *Let $\{X_n\}$ be a renewal process for which $\mu = E[X_1] < \infty$ and $\sigma^2 = E[(X_1 - \mu)^2] < \infty$. Then*

$$\lim_{t\to\infty} \Pr\left\{\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} < x\right\} = \Phi(x),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\tfrac{1}{2}u^2)\, du$$

is the normal integral.

Proof. The proof rests on the central limit theorem for $S_n = X_1 + \cdots + X_n$ and the basic identity of realizations of the process in the form $\{N(t) < n\}$ if and only if $\{S_n > t\}$.

Let x be fixed and let $n \to \infty$ and $t \to \infty$ in such a way that

$$\lim_{\substack{t\to\infty \\ n\to\infty}} \frac{t - n\mu}{\sigma\sqrt{n}} = -x.$$

Then, by the usual central limit theorem,

$$\lim_{\substack{t\to\infty\\ n\to\infty}} \Pr\{S_n > t\} = \lim_{\substack{t\to\infty\\ n\to\infty}} \Pr\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} > -x\right\} = 1 - \Phi(-x) = \Phi(x).$$

But then

$$\begin{aligned}\Phi(x) &= \lim_{\substack{t\to\infty\\ n\to\infty}} \Pr\{S_n > t\} \\ &= \lim_{\substack{t\to\infty\\ n\to\infty}} \Pr\{N(t) < n\} \\ &= \lim_{\substack{t\to\infty\\ n\to\infty}} \Pr\left\{\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} < \frac{n - t/\mu}{\sqrt{t\sigma^2/\mu^3}}\right\} \\ &= \lim_{t\to\infty} \Pr\left\{\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} < x\right\},\end{aligned}$$

since $(n - t/\mu)/\sqrt{t\sigma^2/\mu^3} \to x$ as $t \to \infty$, $n \to \infty$ in such a manner that $(t - n\mu)/\sqrt{n\sigma^2} \to -x$. ■

The preceding analysis was conducted in a formal manner and needs tightening. The student may try to supply the epsilonics.

G. RUIN IN RISK THEORY

Let $N(t)$ be the number of claims incurred by an insurance company over the time interval $(0, t]$. Assume $N(t)$ is a Poisson process with parameter λ. Assume, moreover, that the magnitudes of the successive claims $Y_1, Y_2, Y_3, \ldots$ are independent identically distributed random variables having distribution function $G(x)$. Let the inflow of cash (premiums, investments, etc.) be c dollars per unit time and suppose the initial capital of the company is z. Then at time t, the cash balance is

$$\Gamma(t) = z + ct - \sum_{i=1}^{N(t)} Y_i,$$

where Y_i is the magnitude of the ith successive claim. It is of interest to ascertain the probability of continual solvency as a function of z. That is, we wish to determine

$$R(z) = \Pr\left\{z + ct - \sum_{i=1}^{N(t)} Y_i > 0, \text{ for all } t\right\} \tag{7.8}$$

$$= \text{probability of no ruin with initial capital } z.$$

We apply the renewal argument conditioning on the time T_1 of the first Poisson event. Together with the law of total probabilities, we obtain

$$R(z) = \int_0^\infty \Pr\left\{z + ct - \sum_{i=1}^{N(t)} Y_i > 0 \text{ for all } t \mid T_1 = \tau\right\} \lambda e^{-\lambda\tau}\, d\tau. \tag{7.9}$$

But another conditioning on the value of Y_1 entails

$$\Pr\left\{z + ct - \sum_{i=1}^{N(t)} Y_i > 0 \text{ for all } t \mid T_1 = \tau\right\}$$
$$= \int_0^\infty \Pr\left\{z + ct - \sum_{i=1}^{N(t)} Y_i > 0 \text{ for all } t \mid Y_1 = y,\ T_1 = \tau\right\} dG(y). \tag{7.10}$$

The process $\Gamma(t)$ renews itself immediately after time τ holding the new initial capital $z + c\tau - y$, given $T_1 = \tau$, $Y_1 = y$. Therefore,

$$\Pr\left\{z + ct - \sum_{i=1}^{N(t)} Y_i > 0 \text{ for all } t \mid Y_1 = y,\ T_1 = \tau\right\} = R(z + c\tau - y). \tag{7.11}$$

Of course, $R(u) = 0$ for $u < 0$.

The facts of (7.10) and (7.11) implemented into (7.9) produce the integral equation

$$R(z) = \int_0^\infty \left(\int_0^{z+c\tau} R(z + c\tau - y)\, dG(y)\right) \lambda e^{-\lambda\tau}\, d\tau.$$

A change of variables $t = z + c\tau$ in the outer integral and rearrangement gives

$$R(z)e^{-\lambda z/c} = \frac{\lambda}{c} \int_z^\infty \left(\int_0^t R(t - y)\, dG(y)\right) e^{-\lambda t/c}\, dt.$$

The representation assures that $R(z)$ is differentiable, and differentiation yields

$$e^{-\lambda z/c}\left[R'(z) - \frac{\lambda}{c} R(z)\right] = -\frac{\lambda}{c} e^{-\lambda z/c} \int_0^z R(z - y)\, dG(y),$$

or, equivalently,

$$R'(z) = \frac{\lambda}{c} R(z) - \frac{\lambda}{c} \int_0^z R(z - y)\, dG(y).$$

Integrating both sides with respect to z gives

$$R(w) - R(0) = \frac{\lambda}{c}\int_0^w R(z)\,dz - \frac{\lambda}{c}\int_0^w \left(\int_0^z R(z-y)\,dG(y)\right) dz.$$

Interchanging the orders of integration and then a change of variable $\xi = z - y$ leads to

$$R(w) = R(0) + \frac{\lambda}{c}\int_0^w R(z)\,dz - \frac{\lambda}{c}\int_0^w \left(\int_0^{w-y} R(\xi)\,d\xi\right) dG(y).$$

Define $S(x) = \int_0^x R(\xi)\,d\xi$. Next perform an integration by parts to obtain

$$R(w) - R(0) = \frac{\lambda}{c}S(w) - \frac{\lambda}{c}\left\{S(w) - \int_0^w R(w-y)[1 - G(y)]\,dy\right\},$$

or

$$R(w) = R(0) + \frac{\lambda}{c}\int_0^w R(w-y)[1 - G(y)]\,dy.$$

These manipulations have produced a renewal equation with an improper density $(\lambda/c)[1 - G(y)]$, since

$$\int_0^\infty \frac{\lambda}{c}[1 - G(y)]\,dy = \frac{\lambda}{c}E[Y_1] = \frac{\lambda}{c}\mu.$$

If $\lambda\mu/c > 1$, it is certain that $R(z) = 0$ (why?). (Note that $\lambda\mu$ is the expected outflow per unit time servicing claims, while c is the rate of income.) Assume henceforth the case $\lambda\mu/c < 1$. With $a(w) = R(0)$, Theorem 4.1 continues to apply in this degenerate case to inform us

$$R(w) = a(w) + \int_0^w a(w-y)\,dM(y) = R(0)[1 + M(w)],$$

and since $M(w)$ corresponds to a terminating renewal process

$$\lim_{w\to\infty} M(w) = L/(1 - L) = \frac{\lambda\mu/c}{1 - \lambda\mu/c}$$

whence

$$\lim_{w\to\infty} R(w) = \frac{R(0)}{1 - (\lambda\mu/c)}.$$

But $R(\infty) = 1$ (why?), and we obtain

$$R(0) = 1 - \frac{\lambda\mu}{c}.$$

More precise asymptotic relations can be achieved by refining the analysis.

8: More Elaborate Applications of Renewal Theory

A. A GENETIC MODEL WITH MUTATION

Consider a finite population of constant size N and label the individuals by $j = 1, \ldots, N$. This comprises the first generation in an evolutionary process subject to certain natural selection effects and mutation pressures. We now delimit the nature and order of the forces governing the process.

Each individual of the population is endowed with a characteristic called "fitness." Loosely speaking, fitness is a measure of the individual's innate relative advantage in contributing offspring to the succeeding generation. Let w_k^1 denote the fitness of the kth individual of the first generation. Determine

$$u_k = \frac{w_k^1}{\sum_{j=1}^{N} w_j^1}, \qquad k = 1, 2, \ldots, N, \tag{8.1}$$

which connotes the relative fitness value of the kth individual. The next generation of progeny is formed by performing N independent random samplings following a multinomial distribution with probability vector (8.1). Thus an offspring carries the fitness value w_k^1 of his parental type and will be selected with probability u_k. Manifestly, individuals of high fitness value compared to the others have concordantly larger relative fitness values and manifestly have greater chance of propagating their own kind.

This multinomial reproduction procedure bears a population of offspring carrying fitness values

$$\tilde{w}^2 = (\tilde{w}_1^2, \tilde{w}_2^2, \ldots, \tilde{w}_N^2). \tag{8.2}$$

(Each of the values $\tilde{w}_i^2$ is, of course, one of the $\{w_k^1\}$. The type of an individual will be identified with his fitness.)

The vector $\tilde{w}^2$ does not yet comprise the mature population of the second generation. We will introduce the possibilities of mutation, so that an offspring can undergo a spontaneous change in fitness value. The precise assumption concerning the effects of the mutation changes is as

follows: We suppose that $\{V_i^j;\ i=1,\ \ldots,\ N,\ j=2,\ \ldots\}$ is a rectangular array of independent, identically distributed positive random variables, and then let

$$\begin{aligned} w_1^2 &= \tilde{w}_1^2 V_1^2, \\ w_2^2 &= \tilde{w}_2^2 V_2^2, \\ &\vdots \\ w_N^2 &= \tilde{w}_N^2 V_N^2. \end{aligned}$$

The vector $(w_1^2, \ldots, w_N^2)$ represents the fitnesses of mature individuals in the second generation. The above procedure is repeated, sequentially producing successive N-dimensional vectors that depict the evolution of the population through the changes of the fitness vector $w^k = (w_1^k, \ldots, w_N^k)$, and the relative fitness vector $u^k = (u_1^k, \ldots, u_N^k)$ [the superscript indicates the generation number counting from the initial specified population of (8.1)]. The probability law governing the determination of w^{k+1} from w^k and u^k in line with the formation of w^2 from w^1 goes as follows: Sample N independent values from among $w_1^k, w_2^k, \ldots, w_N^k$ with probabilities u_i^k of choosing w_i^k, $i=1, 2, \ldots, N$. Denote the resulting vector by $\tilde{w}^{k+1} = (\tilde{w}_1^{k+1}, \ldots, \tilde{w}_N^{k+1})$. Mutation changes then transform $\tilde{w}^{k+1}$ to w^{k+1} through multiplication by the positive random variables V_i^{k+1} in the explicit manner

$$w_i^{k+1} = \tilde{w}_i^{k+1} V_i^{k+1}, \qquad i=1, 2, \ldots, N.$$

Finally, determine the relative fitness vector u^{k+1} by the rule

$$u_i^{k+1} = \frac{w_i^{k+1}}{\sum_{j=1}^{N} w_j^{k+1}}, \qquad i=1, 2, \ldots, N.$$

The evolutionary process can be realized by the path of the point w^k, $k=1, 2, \ldots$, traversed in N-dimensional space.

The relative fitness u^k is the projection of the random vector w^k onto the N-dimensional simplex

$$\Delta_N = \{x = (x_1, \ldots, x_N) : x_i \geq 0 \text{ and } x_1 + \cdots + x_N = 1\}.$$

Figure 9 illustrates the projection when $N=3$. As generations pass, u^k describes the relative fitness point moving about the simplex Δ_N. It is natural to inquire concerning the long-run statistical behavior of u^k.

Define $T(0)$ as the elapsed number of generations (i.e., the smallest $k \quad 1$) until all components of $\tilde{w}^k$ coincide. Such a generation is called a *generation of equal components.*

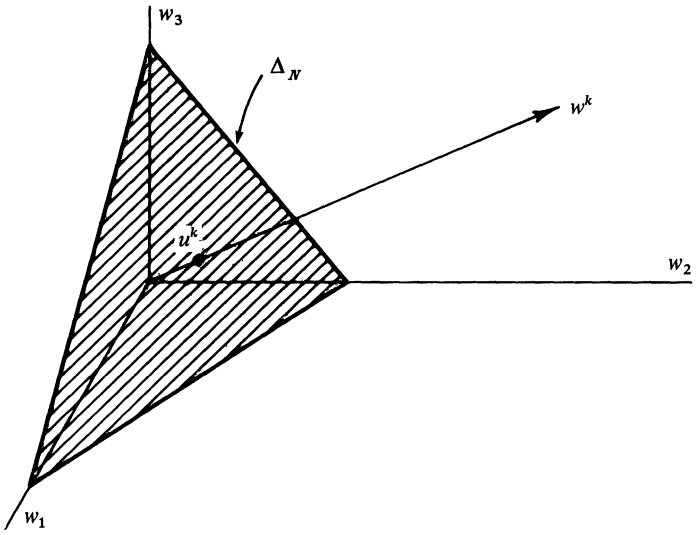


FIG. 9. u^k is the projection of w^k onto the simplex Δ_N.

Lemma 8.1. $\Pr\{T(0) < \infty\} = 1$. *In fact,* $E[T(0)] \leq N^N$.

Proof. Let $w^1 = (w_1^1, \ldots, w_N^1)$ be the fitnesses in the first generation. One way in which $T(0) = 1$ can occur is if the progeny population has

$$\tilde{w}_k^2 = w_v^1, \qquad \text{for all } k,$$

where

$$w_v^1 = \max\{w_1^1, \ldots, w_N^1\}.$$

This event clearly occurs with probability at least $\alpha = (1/N)^N$. Thus $\Pr\{T(0) = 1\} \geq \alpha$ and $\Pr\{T(0) > 1\} \leq 1 - \alpha$.

The same estimate applies for the transitions from the second to the third generation. Consequently, we have

$$\Pr\{T(0) > 2 \mid T(0) > 1\} \leq 1 - \alpha,$$

and

$$\Pr\{T(0) > 2\} \leq (1 - \alpha)^2.$$

By a direct induction, we deduce $\Pr\{T(0) > k\} \leq (1 - \alpha)^k$, so that $E[T(0)] \leq 1/\alpha = N^N$. ∎

Let $T(0) + T(1)$ be the first generation time exceeding $T(0)$ that has equal components, and define $T(0) + \cdots + T(k)$ to be the first generation of equal components after $T(0) + \cdots + T(k-1)$. The key observation

is that *the process of* $\{u^k\}$ *starts afresh at the generation times* $T(0)$, $T(0)+T(1)$, $\cdots$, and therefore this sequence of positive integer-valued random variables forms a renewal process. To see why this is so, let $\tilde{w}$ be the common fitness value at generation $T(k)$. Then all components of $\tilde{w}^{T(k)+1}$ have value $\tilde{w}$, and the jth component of $w^{T(k)+1}$ is $\tilde{w}V_j^{T(k)+1}$. The influence and relevance of $\tilde{w}$ disappears at the next step, because the multinomial probabilities are, in accordance with (8.1),

$$u_j = \frac{\tilde{w}V_j^{T(k)+1}}{\sum_{i=1}^{N} \tilde{w}V_i^{T(k)+1}} = \frac{V_j^{T(k)+1}}{\sum_{i=1}^{N} V_i^{T(k)+1}}, \qquad j=1, 2, \ldots, N,$$

independent of $\tilde{w}$. Hence, following every generation of equal components, the process is determined solely by the independent identically distributed set of mutation multipliers of random variables. It follows that the sequence $T(0)$, $T(1)$, ..., induces a delayed renewal process.

Let $N(t)$ denote the renewal counting process induced by $\{T(k)\}$ and let $S_k = T(0)+\cdots+T(k-1)$ with $S_0=0$. Choose a generation time m. Then $S_{N(m)}$ is the last generation time prior to m where all the premutation fitnesses are the same. What is the distribution of the relative fitness $u^m=(u_1^m, \ldots, u_N^m)$?

At time $S_{N(m)}$ the relative fitnesses have equal components $1/N$. Consider

$$\delta_m = m - S_{N(m)},$$

the number of generations elapsed between $S_{N(m)}$ and m. The effect of mutation followed by the multinomial selection process over one generation can be summarized by a transition distribution function Γ:

$$\Gamma(z_1, \ldots, z_N; \eta_1, \ldots, \eta_N) = \Pr\{\text{a relative fitness of } (\eta_1, \ldots, \eta_N) \text{ changes to } (\eta_1', \ldots, \eta_N') \text{ where } \eta_k' \le z_k,\ k=1, \ldots, N\}.$$

Hence

$$\Pr\{u_k^m \le z_k,\ k=1, \ldots, N \mid \delta_m = 1\} = \Gamma\left(z_1, \ldots, z_N; \frac{1}{N}, \ldots, \frac{1}{N}\right).$$

The result of k generations of the reproduction process with mutation and sampling selection effects starting from a generation of equal components can be expressed formally as

$$\Pr\{u_j^m \le z_j,\ j=1, \ldots, N \mid \delta_m = k\} = \Gamma^{(k)}\left(z_1, \ldots, z_N; \frac{1}{N}, \ldots, \frac{1}{N}\right),$$

where $\Gamma^{(k)}$ stands for the iterated k-fold transition distribution induced from the transformation Γ.

The law of total probabilities provides us with the representation

$$\Pr\{u_j^m \leq z_j, j = 1, \ldots, N\} = \sum_{k=1}^{m} \Pr\{\delta_m = k\}\Gamma^{(k)}\left(z_1, \ldots, z_N; \frac{1}{N}, \ldots, \frac{1}{N}\right) + \Pr\{T(0) \geq m\}\Gamma^{(m+1)}(z_1, \ldots, z_N; u_1^1, \ldots, u_N^1). \tag{8.3}$$

Now the limiting distribution of current life associated with the renewal process $\{T(k)\}_{k=1}^{\infty}$ asserts that

$$\lim_{n\to\infty} \Pr\{\delta_n = k\} = \frac{1 - \Pr\{T(1) \leq k\}}{E[T(1)]} \qquad \text{[a special case of (6.5)]}$$

Applying this fact in (8.2) leads to the result

$$\lim_{m\to\infty} \Pr\{u_j^m \leq z_j, j = 1, \ldots, n\} = \sum_{k=1}^{\infty} \left\{\frac{1 - \Pr\{T(1) \leq k\}}{E[T(1)]}\right\}\Gamma^{(k)}\left(z_1, \ldots, z_N; \frac{1}{N}, \ldots, \frac{1}{N}\right).$$

(The student should justify interchange of limit with sum; it is easy.)

This final relation describes the limiting behavior of relative fitness in the evolving population and, in fact, provides an explicit formula for the stationary distribution.

B. A BRANCHING PROCESS

Suppose at time $t = 0$ there is a single organism that lives a random time T_0, taken from a distribution F. At the end of its life, it produces j new organisms with probability $p_j, j = 1, 2, \ldots$. Each new organism lives and produces independently of the other members of the population and with the corresponding identical distributions governing all actions. Let $m = \sum jp_j$ be the mean of the offspring distribution. Let $M(t)$ denote the mean number of organisms living at time t. Using the law of total probability plus a renewal argument conditioning on the outcome after the death of the initial parent (see Chapter 8 for more details), we obtain

$$\begin{aligned} M(t) &= 1 - F(t) + \sum_{j=1}^{\infty} p_j \int_0^t jM(t - x)\, dF(x) \\ &= 1 - F(t) + m \int_0^t M(t - x)\, dF(x). \end{aligned} \tag{8.4}$$

Except for the factor m, Eq. (8.4) presents a renewal equation. When $m > 1$, we can transform (8.4) into a proper renewal equation. Let β be such that $\int_0^\infty e^{-\beta x}\, dF(x) = 1/m$. There exists a unique such $\beta > 0$, since $\int_0^\infty e^{-\lambda x}\, dF(x)$ is a strictly decreasing continuous function of λ, taking the value 1 for $\lambda = 0$ and approaching zero as $\lambda \to \infty$.

Define

$$\hat{F}(t) = m \int_0^t e^{-\beta x}\, dF(x),$$

$$\hat{M}(t) = e^{-\beta t} M(t),$$

and

$$g(t) = e^{-\beta t}[1 - F(t)].$$

Multiply Eq. (8.4) by $e^{-\beta t}$ to get

$$e^{-\beta t} M(t) = e^{-\beta t}[1 - F(t)] + \int_0^t e^{-\beta(t-x)} M(t-x) m\, e^{-\beta x}\, dF(x),$$

which written in the new notation has the form

$$\hat{M}(t) = g(t) + \int_0^t \hat{M}(t-x)\, d\hat{F}(x).$$

Straightforward verifications guarantee that $g(t)$ is directly Riemann integrable, so that, if F is nonarithmetic, by the renewal theorem,

$$\lim_{t\to\infty} \hat{M}(t) = \frac{\int_0^\infty g(x)\, dx}{\int_0^\infty x\, d\hat{F}(x)}.$$

Now $\int_0^\infty g(t)\, dt = (m-1)/\beta m$, so

$$\lim_{t\to\infty} e^{-\beta t} M(t) = \frac{m-1}{\beta m^2 \int_0^\infty x e^{-\beta x}\, dF(x)},$$

and, asymptotically, $M(t)$ increases exponentially at the rate $e^{\beta t}$. The parameter β is known as the Malthusian rate of growth of the population.

C. INVENTORY THEORY

A shopkeeper keeps a certain quantity of stock on hand. When the stock runs low, he places an order to replenish his supplies. The inventory policy in operation is assumed to be of (s, S) type. (This is in common practice.) Specifically, two levels $s < S$ are prescribed. Suppose the stock is originally at level S. A period length is also specified, and the stock level at the end of each period is checked. If at the close of a period, the stock level falls below s, a requisition (or order) is placed to return the level of stock up to S ready for dispensation at the start of the next period.

Let X_i be the quantity of demand accumulated during the ith period. We assume that $X_1, X_2, \ldots,$ are independent identically distributed positive random variables with distribution function F. Let $N(t)$ be the corresponding renewal counting process. Clearly $N(S-s)+1$ is the number of demand periods elapsed until the first order for refill is placed, at which time the stock level is again S. A little reflection reveals that we are dealing with two renewal processes; the first is the demand process and the second is the refill process.

Let the number of demand periods between the $(i-1)$st and the ith stock refill be θ_i. Then $\{\theta_i\}$ is a discrete (integer-valued) renewal process with mean $E(\theta_i) = 1 + M(S-s)$, and

$$\Pr\{\theta_i = k\} = F_{k-1}(S-s) - F_k(S-s). \tag{8.5}$$

Let W_n be the stock level at the end of the nth demand period. Define G_n to be the conditional distribution

$$G_n(x) = \Pr\{S - x \leq W_n | s \leq W_n\}.$$

This is the distribution of the stock level at the close of the nth period, knowing that the level has not fallen below s. This distribution is calculated by conditioning on δ_n, the number of demand periods since the last stock refill where the stock level was S. (Recall that by assumption at time 0 the stock quantity is S.) We get

$$\begin{aligned} G_n(x) &= \sum_{j=1}^{\infty} \Pr\{\delta_n = j\} \Pr\{X_1 + \cdots + X_j \leq x | X_1 + \cdots + X_j \leq S - s\} \\ &= \sum_{j=1}^{\infty} \Pr\{\delta_n = j\} \left(\frac{F_j(x)}{F_j(S-s)} \right). \end{aligned} \tag{8.6}$$

But from (8.5)

$$\Pr\{\theta_1 \leq j\} = 1 - F_j(S-s). \tag{8.7}$$

Hence, by appeal to the limit theorem for the excess random variable of the renewal process $\{\theta_i\}$ [see (6.2)] and by virtue of (8.7), we deduce

$$\lim_{n\to\infty} \Pr\{\delta_n = j\} = \frac{1}{E[\theta_1]} \Pr\{\theta_1 > j\} = \frac{F_j(S-s)}{E[\theta_1]}$$
$$= \frac{F_j(S-s)}{1+M(S-s)}. \tag{8.8}$$

Using the above result in (8.6), we may conclude

$$\lim_{n\to\infty} \Pr\{S - x \leq W_n | s \leq W_n\} = \lim_{n\to\infty} \sum_{j=1}^{\infty} \Pr\{\delta_n = j\} \frac{F_j(x)}{F_j(S-s)}$$
$$= \frac{M(x)}{1+M(S-s)},$$

which gives the limiting distribution of stock level in periods in which a requisition order is not pending.

D. CHARACTERIZATIONS OF THE POISSON PROCESS

The Poisson process is a very special renewal process. This section offers some characterizations of the Poisson process as a special process within the class of renewal processes. For this objective we will exploit several of the limit theorems of Section 6.

Let $\{X_k\}$ be a renewal process with $E[X_k] = \mu < \infty$, and $F(x) = \Pr\{X_k \leq x\}$. Assume $F(0) = 0$. Define

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

[compare to (3.5)]. Of course, $F_t(x)$ is the distribution function for $\min\{X_k, t\}$.

Theorem 8.1. (a) *If there exists a sequence* $\{t_j\}$, *where* $t_j \to \infty$ *as* $j \to \infty$, *and for which the current life* δ_t *satisfies*

$$F_{t_j}(x) = \Pr\{\delta_{t_j} \leq x\}, \qquad \text{for all } x,$$

then F *is an exponential distribution.*

(b) *If there exists a sequence* $\{t_j\}$, *where* $t_j \to \infty$ *as* $j \to \infty$, *and for which*

$$F(x) = \Pr\{\gamma_{t_j} \leq x\}, \qquad \text{for all } x,$$

[compare with (3.3)] *then* F *is exponential.*

Proof. We will demonstrate only (a) since (b) is quite similar. By the result of (6.5), the limiting distribution of the current life δ_t is

$$\lim_{t\to\infty} \Pr\{\delta_t > y\} = \mu^{-1} \int_y^\infty \{1 - F(z)\}\, dz.$$

Letting t increase along t_j with due account of the hypothesis of the theorem, we derive the functional equation

$$1 - F(y) = \mu^{-1} \int_y^\infty \{1 - F(z)\}\, dz.$$

The right-hand side is clearly differentiable in y, yielding the elementary first-order differential equation

$$\frac{d}{dy}\{1 - F(y)\} = -\frac{1}{\mu}\{1 - F(y)\},$$

whose solution, subject to $F(0) = 0$, is

$$1 - F(y) = e^{-\lambda y}, \qquad \lambda = 1/\mu,$$

The proof is complete. ■

Theorem 8.2. *Suppose, for some* $t_0 > 0$,

$$\Pr\{\delta_t \le x\} = F_t(x), \qquad 0 \le t < t_0,\ x \ge 0.$$

Then, for some $\lambda > 0$, $F(x) = 1 - e^{-\lambda x}$ *for* $0 \le x < t_0$.

Proof. For $0 \le x \le t$, we have

$$\begin{aligned}
\Pr\{\delta_t \le x\} &= \sum_{j=1}^{\infty} \Pr\{\delta_t \le x \quad \text{and} \quad N(t) = j\} \\
&= \sum_{j=1}^{\infty} \Pr\{t - x < S_j \le t \quad \text{and} \quad S_{j+1} > t\} \\
&= \sum_{j=1}^{\infty} \int_{t-x}^{t} [1 - F(t-y)]\, dF_j(y) \\
&= \int_{t-x}^{t} [1 - F(t-y)]\, dM(y).
\end{aligned}$$

Thus, by hypothesis, for $0 \le x \le t < t_0$,

$$F(x) = \int_{t-x}^{t} [1 - F(t-y)]\, dM(y), \tag{8.9}$$

and

$$\frac{1}{x} F(x) = \frac{1}{x} \int_{t-x}^{t} [1 - F(t-y)]\, dM(y). \tag{8.10}$$

The function $M(t)$ is finite, right continuous, and nondecreasing and thus possesses a finite derivative $M'(t)$ for infinitely many t dense in any $(0, t_0]$ interval. Choose such a $t = \tau$ for wich $M'(\tau) < \infty$. Then (8.10) has a limit as x decreases to zero at $t = \tau$, and

$$F'(0) = [1 - F(0)]M'(\tau) = M'(\tau).$$

But the left-hand side of (8.10) is independent of t. Thus the limit on the right must exist for all t, and

$$F'(0) = M'(t), \qquad \text{for all} \quad t < t_0.$$

Let $\lambda = F'(0)$, so that $M(t) = \lambda t$. We substitute this into (8.9) to obtain

$$F(x) = \int_{t-x}^{t} [1 - F(t-y)]\lambda \, dy.$$

We may now differentiate in x to obtain

$$dF(x)/dx = -\lambda[1 - F(x)], \qquad 0 \leq x < t_0,$$

or $F(x) = 1 - e^{-\lambda x}$, $0 \leq x < t_0$, as claimed. ■

9: Superposition of Renewal Processes

In this section we will establish, under certain conditions, that the superposition of indefinitely many uniformly sparse renewal processes tends to a Poisson process. The following Theorem 9.1 can serve to give a meaningful rationale for the Poisson assumption in a variety of circumstances, just as the central limit theorem provides justification for the widespread postulate of the normal distribution in representing certain random variables.

Several other results pertaining to the superposition of renewal processes are also highlighted, lending a glimpse into a currently popular area of research.

For each integer $n = 1, 2, \ldots$, and for each $i = 1, \ldots, k_n$, where $k_n \to \infty$ as $n \to \infty$, let $N_{ni}(t)$ be a renewal counting process with interoccurrence time distribution $F_{ni}(t)$. The collection $\{N_{ni}(t); n = 1, 2, \ldots, i = 1, \ldots, k_n\}$ constitutes a triangular array of stochastic processes. *For every n, we assume the processes $\{N_{n1}(t)\}, \ldots, \{N_{nk_n}(t)\}$ are independent.*

By the *superposition process* $N_n(t)$, we mean the aggregate counting process

$$N_n(t) = \sum_{i=1}^{k_n} N_{ni}(t), \qquad t \geq 0.$$

The superposition process is *not*, in general, a *renewal* counting process because the interoccurrence times are not independent and certainly not identically distributed. In fact, the distribution of the intervals between "events" is complicated and in general intractable.

Definition 9.1. The triangular array $\{N_{ni}(t)\}$ is called infinitesimal if for every $t \geq 0$

$$\lim_{n\to\infty} \max_{1\leq i\leq k_n} F_{ni}(t) = 0. \tag{9.1}$$

Prior to the main theorem we give a preliminary lemma.

Lemma 9.1. *Let $\{N_{ni}(t)\}$ be an infinitesimal array with interoccurrence distribution $\{F_{ni}(t)\}$. Let $F_{ni}^{*j}(t)$ denote the j-fold convolution of $F_{ni}(t)$. Suppose for some finite nondecreasing function $c(t)$ that*

$$\limsup_{n\to\infty} \sum_{i=1}^{k_n} F_{ni}(t) \leq c(t), \qquad \text{for all} \quad t. \tag{9.2}$$

Then

$$\text{(a)} \quad \lim_{n\to\infty} \sum_{i=1}^{k_n} \sum_{j=2}^{\infty} [F_{ni}(t)]^j = 0,$$

$$\text{(b)} \quad \lim_{n\to\infty} \sum_{i=1}^{k_n} \sum_{j=2}^{\infty} F_{ni}^{*j}(t) = 0,$$

and

$$\text{(c)} \quad \lim_{n\to\infty} \sum_{i=1}^{k_n} F_{ni}^{*j}(t) = 0, \qquad \text{uniformly in} \quad j = 2, 3, \ldots.$$

Proof. We prove only (a); (b) and (c) follow easily by virtue of the inequality $F_{ni}^{*j}(t) \leq [F_{ni}(t)]^j$ [cf. (4.2)]. Let

$$A_n(t) = \sum_{i=1}^{k_n} \sum_{j=2}^{k_n} [F_{ni}(t)]^j.$$

Let $\varepsilon > 0$ and $t > 0$ be given. Since $\{F_{ni}(t)\}$ are infinitesimal, there exists n_0 such that

$$F_{ni}(t) \leq \varepsilon, \qquad \text{for} \quad i = 1, \ldots, k_n,$$

whenever $n \geq n_0$. Then for $n \geq n_0$ and increasing along the limit superior in (9.2), we have

$$A_n(t) \leq \sum_{i=1}^{k_n} \sum_{j=2}^{\infty} \varepsilon^{j-1} F_{ni}(t)$$

$$= \frac{\varepsilon}{1-\varepsilon} \sum_{i=1}^{k_n} F_{ni}(t)$$

$$\leq 2 \frac{\varepsilon c(t)}{1-\varepsilon}.$$

Since $\varepsilon > 0$ is arbitrary, we conclude

$$\limsup_{n\to\infty} A_n(t) = 0,$$

as was to be shown. ■

Theorem 9.1. *Let $\{N_{ni}(t)\}$ be an infinitesimal array of renewal processes with superposition $N_n(t)$. Then*

$$\lim_{n\to\infty} \Pr\{N_n(t) = j\} = \frac{e^{-\lambda t}(\lambda t)^j}{j!}, \qquad j = 0, 1, 2, \ldots, \tag{9.3}$$

if and only if

$$\lim_{n\to\infty} \sum_{i=1}^{k_n} F_{ni}(t) = \lambda t. \tag{9.4}$$

Proof. (1) *Necessity.* Suppose (9.3) is true. For $j = 0$ we obtain

$$\lim_{n\to\infty} \Pr\{N_n(t) = 0\} = e^{-\lambda t},$$

or, equivalently,

$$\lim_{n\to\infty} [-\log \Pr\{N_n(t) = 0\}] = \lambda t.$$

Recalling that $N_{n,1}(t), \ldots, N_{n,k_n}(t)$ are nonnegative independent random variables by assumption, it follows that

$$-\log \Pr\{N_n(t) = 0\} = -\log \prod_{i=1}^{k_n} \Pr\{N_{ni}(t) = 0\}$$

$$= -\sum_{i=1}^{k_n} \log[1 - F_{ni}(t)].$$

Expanding in the Taylor series

$$-\log[1 - F_{ni}(t)] = \sum_{j=1}^{\infty} \frac{1}{j} [F_{ni}(t)]^j$$

gives

$$\lambda t = \lim_{n\to\infty} [-\log \Pr\{N_n(t) = 0\}]$$

$$= \lim_{n\to\infty} \left\{ \sum_{i=1}^{k_n} F_{ni}(t) + \sum_{i=1}^{k_n} \sum_{j=2}^{\infty} \frac{1}{j} [F_{ni}(t)]^j \right\}.$$

The second sum is nonnegative, so that

$$\limsup_{n\to\infty} \sum_{i=1}^{k_n} F_{ni}(t) \leq \lambda t,$$

and we may appeal to Lemma 9.1 to conclude that the second sum vanishes in the limit. Thus

$$\lambda t = \lim_{n\to\infty} \sum_{i=1}^{k_n} F_{ni}(t),$$

which is (9.4) as was desired to be shown.

(2) *Sufficiency.* Suppose (9.4) holds. The proof will consist of an induction on m to show

$$\lim_{n\to\infty} \Pr\{N_n(t) = m\} = e^{-\lambda t}\frac{(\lambda t)^m}{m!}, \qquad m = 0, 1, \ldots.$$

Step 1: $m = 0$. Using the same Taylor series argument as above

$$\Pr\{N_n(t) = 0\} = \prod_{i=1}^{k_n} [1 - F_{ni}(t)]$$

$$= \exp\left[-\sum_{i=1}^{k_n} F_{ni}(t) - \sum_{i=1}^{k_n}\sum_{j=2}^{\infty} \frac{[F_{ni}(t)]^j}{j}\right].$$

By virtue of assumption (9.4) and Lemma 9.1, the limit of the exponent is $-\lambda t$. Thus the limit relation

$$\lim_{n\to\infty} \Pr\{N_n(t) = 0\} = \exp\{-\lambda t\},$$

is confirmed.

Step 2: The induction step. Let m be given and suppose we have shown

$$\lim_{n\to\infty} \Pr\{N_n(t) = m - 1\} = \frac{e^{-\lambda t}(\lambda t)^{m-1}}{(m-1)!}.$$

We need an extension. Let $s_1, \ldots, s_r$ be a finite set of indices. Then, since the array is infinitesimal, it is correct that

$$\lim_{n\to\infty} \prod_{\substack{j=1 \\ j\neq s_1,\ldots,s_r}}^{k_n} [1 - F_{nj}(t)] = e^{-\lambda t}, \tag{9.5}$$

and, for each fixed r, the limit is uniform over the indices $s_1, \ldots, s_r$. Now

$$\Pr\{N_n(t) = m\} = P_n(t) + Q_n(t),$$

where

$$P_n(t) = \Pr\{N_n(t) = m \text{ and } N_{ni}(t) \leq 1,\ i = 1, \ldots, k_n\},$$

and

$$Q_n(t) = \Pr\{N_n(t) = m \text{ and } N_{ni}(t) \geq 2, \text{ for some } i\}.$$

We claim $Q_n(t) \to 0$ as $n \to \infty$. In fact, for any $\varepsilon > 0$

$$\begin{aligned} Q_n(t) &\leq \sum_{i=1}^{k_n} \Pr\{N_{ni}(t) \geq 2\} \\ &\leq \sum_{i=1}^{k_n} [F_{ni}(t)]^2 \\ &\leq \varepsilon \sum_{i=1}^{k_n} F_{ni}(t), \end{aligned}$$

if n is sufficiently large so that $F_{ni}(t) \leq \varepsilon$. Thus, using the hypothesis we have

$$\limsup_{n \to \infty} Q_n(t) \leq \varepsilon \lambda t,$$

and since ε is arbitrary, we have shown $Q_n(t) \to 0$ as $n \to \infty$. It remains to evaluate the limit of $P_n(t)$. Let $I(m)$ be all possible combinations $(i_1, \ldots, i_m)$ with $1 \leq i_j \leq k_n$. Then

$$\begin{aligned} P_n(t) &= \Pr\{N_n(t) = m \text{ and } N_{ni}(t) \leq 1,\ i = 1, \ldots, k_n\} \\ &= \sum_{I(m)} \Pr\{N_{ni_1}(t) = 1, \ldots, N_{ni_m}(t) = 1 \\ &\qquad \text{and } N_{ni}(t) = 0, \text{ for } i \notin I(m)\} \\ &= \frac{1}{m} \sum_{i=1}^{k_n} \Pr\{N_{ni}(t) = 1,\ N_n(t) - N_{ni}(t) = m - 1, \\ &\qquad \text{and } N_{nj}(t) \leq 1,\ j = 1, \ldots, k_n\} \\ &= \frac{1}{m} \sum_{i=1}^{k_n} \{F_{ni}(t) - F_{ni}^{*2}(t)\} R_{ni}(t) \end{aligned}$$

(the independence assumption on $N_{ni}(t)$ comes into play here) where

$$R_{ni}(t) = \Pr\{N_n(t) - N_{ni}(t) = m - 1,\ N_{nj}(t) \leq 1,\ j = 1, \ldots, k_n,\ j \neq i\}.$$

But by the induction step and (9.5), $R_{ni}(t) \to e^{-\lambda t}(\lambda t)^{m-1}/(m-1)!$ uniformly in i as $n \to \infty$. Thus

$$\begin{aligned}\lim_{n\to\infty} P_n(t) &= \lim_{n\to\infty} \frac{1}{m} \sum_{i=1}^{k_n} \{F_{ni}(t) - F_{ni}^{*2}(t)\} R_{ni}(t) \\ &= \lim_{n\to\infty} \frac{1}{m} \sum_{i=1}^{k_n} F_{ni}(t) e^{-\lambda t}(\lambda t)^{m-1} - /(m\,1)! \\ &= e^{-\lambda t}(\lambda t)^m/m!\end{aligned}$$

as was to be shown. ■

Example 1. Suppose $F(t)$ is a distribution function for which $F(0) = 0$ and $F'(0) = \lambda > 0$. Let

$$F_{ni}(t) = F(t/n), \qquad i = 1, \ldots, n,$$

and, for all n, let $N_{ni}(t)$, $i = 1, \ldots, n$, be independent renewal counting processes with interoccurrence distribution F_{ni}. Then $N_{ni}(t)$ is a triangular array. Furthermore, since

$$\lim_{n\to\infty} \max_{1\le i\le n} F_{ni}(t) = \lim_{n\to\infty} F(t/n) = 0,$$

the array is infinitesimal. To verify (9.4) we compute

$$\begin{aligned}\lim_{n\to\infty} \sum_{i=1}^{n} F_{ni}(t) &= \lim_{n\to\infty} nF(t/n) \\ &= t \lim_{n\to\infty} \frac{F(t/n)}{t/n} \\ &= \lambda t.\end{aligned}$$

Hence, the distribution of the superposition $N_n(t)$ converges to the Poisson process.

We end this section with two characterizations of the Poisson process involving composition of renewal processes. A sum of two independent Poisson processes persists as a Poisson process with the rate parameters merely adding. We will show that in essence only the Poisson process, among renewal processes, possesses this property.

Theorem 9.2. *Let $N_1(t)$ and $N_2(t)$ be two independent renewal processes with the same interoccurrence distribution F having mean μ. Let $N(t) = N_1(t) + N_2(t)$. If $N(t)$ is also a renewal process, then $N_1(t)$, $N_2(t)$, and $N(t)$ are all Poisson.*

Proof. Let H be the interoccurrence distribution for $N(t)$. Then

$$\begin{aligned} 1 - H(x) &= \Pr\{N(x) = 0\} \\ &= \Pr\{N_1(x) = 0,\ N_2(x) = 0\} \\ &= [1 - F(x)]^2. \end{aligned}$$

Let $\gamma_1(t)$, $\gamma_2(t)$, and $\gamma(t)$ be the excess life at time t for the processes N_1, N_2, and N, respectively. Then, because the processes N_1 and N_2 are composed, we necessarily have

$$\gamma(t) = \min\{\gamma_1(t),\ \gamma_2(t)\},$$

and

$$\Pr\{\gamma(t) > x\} = [\Pr\{\gamma_1(t) > x\}]^2.$$

Letting $t \to \infty$ and using the asymptotic distribution of excess life given in (6.5), we obtain

$$\frac{1}{\nu}\int_x^\infty [1 - H(y)]\,dy = \frac{1}{\mu^2}\left\{\int_x^\infty [1 - F(y)]\,dy\right\}^2, \tag{9.6}$$

where $\nu = \int_0^\infty [1 - H(y)]\,dy$. Both sides are differentiable† with respect to x, and earlier we noted $1 - H(x) = [1 - F(x)]^2$. Differentiating (9.6) gives

$$\frac{1}{\nu}[1 - F(x)]^2 = \frac{1}{\nu}[1 - H(x)] = \frac{2}{\mu^2}\left\{\int_x^\infty [1 - F(y)]\,dy\right\}[1 - F(x)],$$

or

$$1 - F(x) = \frac{2\nu}{\mu^2}\int_x^\infty [1 - F(y)]\,dy.$$

Letting $G(x) = 1 - F(x)$ this becomes, after differentiation,

$$\frac{dG(x)}{dx} = -\frac{2\nu}{\mu^2}G(x),$$

whose solution subject to $G(0) = 1$ is

$$G(x) = 1 - F(x) = e^{-\lambda x},$$

where $\lambda = 2\nu/\mu^2$, as desired. Thus, both $N_1(t)$ and $N_2(t)$ are Poisson, and, of course, then so must be their sum $N(t)$. ■

Theorem 9.3. *Let $N_1(t)$ be a Poisson process with parameter μ. Let $N_2(t)$ be a renewal process having a finite mean interoccurrence time and suppose N_1*

† This is immediate if F is continuous. The general case requires further argument.

and N_2 are independent. If $N(t) = N_1(t) + N_2(t)$ defines a renewal process, then $N_2(t)$ must also be Poisson.

Proof. We use the same technique as in Theorem 9.2. Suppose N_1, N_2, and N have interoccurrence distributions $1 - e^{-t/\mu}$, $G(t)$, and $H(t)$, respectively. Then

$$1 - H(t) = \{1 - G(t)\}e^{-t/\mu}, \tag{9.7}$$

and

$$\frac{1}{\mu_H}\int_x^\infty [1 - H(y)]\,dy = \frac{e^{-x/\mu}}{\mu_G}\int_x^\infty [1 - G(y)]\,dy,$$

where μ_H and μ_G are the means of H and G, respectively.

Differentiation leads to

$$\frac{\mu_G}{\mu_H}[1 - H(x)] = e^{-x/\mu}[1 - G(x)] + \frac{1}{\mu}e^{-x/\mu}\int_x^\infty [1 - G(y)]\,dy,$$

which, with (9.7), gives

$$[1 - G(x)]\left[\frac{\mu_G}{\mu_H} - 1\right] = \frac{1}{\mu}\int_x^\infty [1 - G(y)]\,dy. \tag{9.8}$$

Let

$$\lambda = \mu\left[\frac{\mu_G}{\mu_H} - 1\right], \qquad \text{and} \qquad F(x) = 1 - G(x).$$

Then differentiation of (9.8) gives

$$-\lambda\, dF(x)/dx = F(x),$$

or

$$F(x) = e^{-x/\lambda}, \qquad x \geq 0,$$

and

$$G(x) = 1 - e^{-x/\lambda}, \qquad x \geq 0,$$

as was to be shown. ■

Elementary Problems

1. If $\Pr\{X_i = 1\} = \frac{1}{3}$, $\Pr\{X_i = 2\} = \frac{2}{3}$, compute

$$\Pr\{N(1) = k\}, \qquad \Pr\{N(2) = k\}, \qquad \Pr\{N(3) = k\}.$$

2. A patient arrives at a doctor's office. With probability 1/5 he receives service immediately, while with probability 4/5 his service is deferred an hour. After an hour's wait again with probability 1/5 his needs are serviced instantly or another delay of an hour is imposed and so on.

(a) What is the waiting time distribution of the first arrival?

(b) What is the distribution of the number of patients who receive service over an 8-hr period assuming the same procedure is followed for every arrival and the arrival pattern is that of a Poisson process with parameter 1.

3. The weather in a certain locale A consists of rainy spells alternating with spells when the sun shines. Suppose that the number of days of each rainy spell is Poisson distributed with parameter 2 and a sunny spell is distributed according to a geometric distribution with mean 7 days. Assume that the successive random durations of rainy and sunny spells are statistically independent variables. In the long run, what is the probability on a given day that it will be raining?

4. The random lifetime of an item has distribution function $F(x)$. What is the mean remaining life of an item of age x?

Solution:

$$e(x) = E[X - x | X > x] = \frac{\int_x^\infty \{1 - F(t)\}\, dt}{1 - F(x)}.$$

5. If $f(x)$ is a probability density function associated with a lifetime distribution function $F(x)$, the *hazard rate* is the function $r(x) = f(x)/[1 - F(x)]$. Show that the replacement age T^* that minimizes the long run mean cost per unit time

$$\theta(T) = \frac{c_1[1 - F(T)] + c_2 F(T)}{\int_0^T [1 - F(x)]\, dx}$$

must satisfy

$$r(T^*) \times \int_0^{T^*} [1 - F(x)]\, dx - F(T^*) = \frac{c_1}{c_1 - c_2}.$$

6. Cars arrive at a gate. Each car is of random length L having distribution function $F(\xi)$. The first car arrives and parks against the gate. Each succeeding car parks behind the previous one at a distance that is random according to a uniform distribution $[0, 1]$. Consider the number of cars N_x that are lined up within a total distance x of the gate. Determine

$$\lim_{x \to \infty} E[N_x]/x,$$

for $F(\xi)$ a degenerate distribution of length c, and also for the case $F(\xi) = 1 - e^{-\xi}$.

Solution: $2/(1 + 2c)$ and $2/3$.

7. At the beginning of each day customers arrive at a taxi stand at times of a renewal process with distribution law $F(x)$. Assume an unlimited supply of cabs, as at an airport. Suppose each customer pays a random fee at the station following the distribution law $G(x)$, $x > 0$.

(i) Write an expression for the sum of money collected by the station by time t of the day.

(ii) Determine the limit expectation

$$\lim_{t\to\infty} E[\text{the money collected over an initial interval of time } t]/t.$$

8. Consider a counter of Type II, where the locking time with each pulse arrival is of fixed length of τ units. Assume pulses arrive according to a Poisson process with parameter λ. Determine the probability, $p(t)$, that the counter is free at time t.

Solution:

$$p(t) = \begin{cases} e^{-\lambda t}, & t < \tau, \\ e^{-\lambda\tau}, & t \geq \tau. \end{cases}$$

Problems

1. Find $\Pr\{N(t) \geq k\}$ in a renewal process having lifetime density

$$f(x) = \begin{cases} \rho e^{-\rho(x-\delta)}, & \text{for } x > \delta, \\ 0, & \text{for } x \leq \delta, \end{cases}$$

where $\delta > 0$ is fixed.

2. Throughout its lifetime, itself a random variable having distribution function $F(x)$, an organism produces offspring according to a nonhomogenous Poisson process with intensity function $\lambda(u)$. Independently, each offspring follows the same probabilistic pattern, and thus a population evolves. Assuming

$$1 < \int_0^\infty \{1 - F(u)\}\lambda(u)\, du < \infty,$$

show that the mean population size $m(t)$ asymptotically grows exponentially at rate $r > 0$, where r uniquely solves

$$1 = \int_0^\infty e^{-ru}\{1 - F(u)\}\lambda(u)\, du.$$

Hint: Develop a renewal equation for $B(t)$, the mean number of individuals born up to time t, and from this infer that $B(t)$ grows exponentially at rate r. Then express $m(t)$ in terms of $B(u)$ for $u \leq t$.

3. Show that $\lim_{t\to\infty} V(t)/t = \sigma^2/\mu^3$, where $V(t)$ is the variance of a renewal process $N(t)$ and μ and $\sigma^2 < \infty$ are the mean and variance, respectively, of the interarrival distribution.

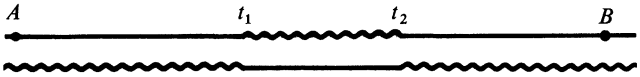
4. For a renewal process with distribution $F(x)$ compute

$$p(t) = \Pr\{\text{number of renewals in } (0, t] \text{ is odd}\}.$$

Obtain this explicitly for a Poisson process with parameter λ and also explicitly when $F(t) = \int_0^t xe^{-x}\,dx$.

5. Breaks and recombinations occur along the length of a pair of chromosomes according to a Poisson process with parameter λ. To illustrate suppose breaks occur at points t_1 and t_2

then the recombined chromosomes have the form



Determine the probability that a point B whose original distance from the location A is l will remain connected with A on the same chromosome after recombination.

6. Show that the renewal function corresponding to the lifetime density

$$f(x) = \lambda^2 xe^{-\lambda x}, \qquad x \geq 0,$$

is

$$M(t) = \tfrac{1}{2}\lambda t - \tfrac{1}{4}(1 - e^{-2\lambda t}).$$

7. Let c_1 be the planned replacement cost and c_2 the failure cost in a block replacement model. Using the long-run mean cost per unit time formula $[c_1 + c_2 M(T)]/T$, show that the cost minimizing block replacement time T^* satisfies

$$e^{-2\lambda T^*}(1 + 2\lambda T^*) = 1 - (4c_1/c_2),$$

where $c_2 > 4c_1$, and the lifetime density is that of Problem 6.

8. Let $X_1, X_2, \ldots,$ be i.i.d. uniformly distributed on (0, 1). Define $N_k =$ the index n satisfying $X_1^k + X_2^k + \cdots + X_n^k \leq 1 < X_1^k + \cdots + X_{n+1}^k$ (the kth powers). Determine

$$\lim_{k\to\infty} E(N_k)/k.$$

Hint: Establish and use the identity $E(S_{N_k+1}) = E(N_k + 1)E(X_1^k)$, where $S_r = X_1^k + \cdots + X_r^k$, $r = 1, 2, \ldots$.

9. Determine the distribution of the total life β_t of the Poisson process.

Answer: $\Pr\{\beta_t \leq x\} = 1 - [1 + \lambda \min\{t, x\}]e^{-\lambda x}$,

10. Show that the age $\{\delta_t; t \geq 0\}$ in a renewal process, considered as a stochastic process, is a Markov process, and derive its transition distribution function

$$F(y; t, x) = \Pr\{\delta_{s+t} \leq y | \delta_s = x\}.$$

11. Suppose $A(t)$ solves the renewal equation $A(t) = a(t) + \int_0^t A(t - y)\, dF(y)$, where $a(t)$ is a bounded nondecreasing function with $a(0) = 0$. Establish that $\lim_{t\to\infty} A(t)/t = a^*/\mu$, where $a^* = \lim_{t\to\infty} a(t)$ and $\mu < \infty$ is the mean of $F(x)$.

12. Consider a system that can be in one of two states: "on" or "off." At time zero it is "on." It then serves before breakdown for a random time T_{on} with distribution function $1 - e^{-t\lambda}$. It is then off before being repaired for a random time T_{off} with the same distribution function $1 - e^{-t\lambda}$. It then repeats a statistically independent and identically distributed similar cycle, and so on. Determine the mean of $W(t)$, the random variable measuring the total time the system is operating during the interval $(0, t)$.

13. Successive independent observations are taken from a distribution with density function

$$f(x) = \begin{cases} xe^{-x}, & x \geq 0, \\ 0, & x \leq 0, \end{cases}$$

until the sum of the observations exceeds the number t. Let $N + 1$ be the number of observations required. Prove that

$$\Pr\{N = n\} = \frac{t^{2n+1}e^{-t}}{\Gamma(2n + 2)} + \frac{t^{2n}e^{-t}}{\Gamma(2n + 1)}.$$

14. A renewal process is an integer-valued stochastic process that registers the number of points in $(0, t]$, when the interarrival times of the points are independent, identically distributed random variables with common distribution function $F(x)$ for $x \geq 0$ and zero elsewhere, and F is continuous at $x = 0$. A modified renewal process is one where the common distribution function $F(x)$ of the interarrival times has a jump q at zero. Show that a modified renewal process is equivalent to an ordinary renewal process, where the numbers of points registered at each arrival are independent identically distributed random variables, R_0, R_1, R_2, ..., with distribution

$$\Pr\{R_i = n\} = pq^n, \qquad n = 0, 1, 2, \ldots,$$

for all $i = 0, 1, 2, \ldots$, where $p = 1 - q$.

15. Consider a renewal process with underlying distribution function $F(x)$. Let W be the time when the interval duration from the preceding renewal event first exceeds $\xi > 0$ (a fixed constant). Determine an integral equation satisfied by

$$V(t) = \Pr\{W \leq t\}.$$

Calculate $E[W]$. (Assume an event occurs at time $t = 0$.)

16. Consider a renewal process $N(t)$ with associated distribution function $F(x)$. Define $m_k(t) = E[N(t)^k]$. Show that $m_k(t)$ satisfies the renewal equation

$$m_k(t) = z_k(t) + \int_0^t m_k(t-\tau)\, dF(\tau), \qquad k = 1, 2, \ldots,$$

where

$$z_k(t) = \int_0^t \sum_{j=0}^{k-1} \binom{k}{j} m_j(t-\tau)\, dF(\tau).$$

Hint: Use the renewal argument.

17. (Continuation of Problem 16). By induction show that

$$z_k(t) = (-1)^{k-1}[F(t) - \binom{k}{1}m_1(t) + \cdots + (-1)^k\binom{k}{k-1}m_{k-1}(t)].$$

18. Consider a stochastic process $X(t)$, $t \geq 0$, which alternates in 2 states A and B. Denote by $\xi_1, \eta_1, \xi_2, \eta_2, \ldots$, the successive sojourn times spent in states A and B, respectively, and suppose $X(0)$ is in A. Assume $\xi_1, \xi_2, \ldots$, are i.i.d.r.v.'s with distribution function $F(\xi)$ and $\eta_1, \eta_2, \ldots$, are i.i.d.r.v.'s with distribution function $G(\eta)$. Denote by $Z(t)$ and $W(t)$ the total sojourn time spent in states A and B during the time interval $(0, t)$. Clearly $Z(t)$ and $W(t)$ are random variables and $Z(t) + W(t) = t$. Let $N(t)$ be the renewal process generated by $\xi_1, \xi_2, \ldots$. Define

$$\theta(t) = \eta_1 + \eta_2 + \cdots + \eta_{N(t)}.$$

Prove

$$P\{W(t) \leq x\} = P\{\theta(t-x) \leq x\},$$

and express this in terms of the distributions F and G.

Answer:

$$\Pr\{W(t) \leq x\} = \sum_{n=1}^{\infty} G_n(t-x)[F_n(x) - F_{n+1}(x)],$$

where G_n and F_n are the usual convolutions.

19. Consider a renewal process with distribution $F(x)$. Suppose each event is erased with probability $1-q$. Expand the time scale by a factor $1/q$. Show that the resulting sequence of events constitutes a renewal process where the distribution function of the time between events is

$$\sum_{n=1}^{\infty} (1-q)^{n-1} q F_n(x/q) = F(x; q),$$

where F_n as usual denotes the n-fold convolution of F.

20. (Continuation of Problem 19). In the preceding problem let $\phi(s)$ be the Laplace transform of $F(x)$. Determine the Laplace transform of $F(x; q)$.

Answer:

$$\phi(s; q) = \frac{q\phi(sq)}{1-(1-q)\phi(sq)}.$$

21. (Continuation of Problem 20). If F has two moments, prove that

$$\phi(s; q) \to \frac{\lambda}{\lambda + s}, \qquad \text{as} \quad q \to 0+, \quad \text{for all} \quad s, \quad \text{Re } s \geq 0,$$

where $\lambda^{-1} = \int_0^{\infty} x\, dF(x)$.

22. (Continuation of Problem 21). Appealing to the convergence theorem, Chapter 1, p. 11, prove that

$$F(x; q) \to 1 - e^{-\lambda x}, \qquad \text{as} \quad q \to 0+.$$

23. Consider a renewal process with interarrival distribution $G_0(x)$. Suppose each event is kept with probability q and deleted with probability $1-q$, and then the time scale is expanded by a factor $1/q$ (see Problem 19). Show that the mean interarrival time is the same for the original and the new process. Repeat the above operation of deletion and scale expansion to obtain a sequence of renewal processes with interarrival distribution given by $G_{(n)}(x)$ after n such transformations of the process. In all these operations q is held fixed. Show that if $0 < q < 1$, then

$$\lim_{n \to \infty} G_{(n)}(x) = 1 - e^{-x\mu},$$

where $\mu = \int_0^{\infty} \{1 - G_{(0)}(\xi)\}\, d\xi$.

Answer: Set

$$\phi_0(s) = \int_0^{\infty} e^{-s\xi}\, dG_0(\xi), \qquad \phi_i(s) = \int_0^{\infty} e^{-s\xi}\, dG_{(i)}(\xi).$$

Establish by induction that

$$\phi_n(s) = \frac{q^n \phi_0(sq^n)}{1-(1-q^n)\phi_0(sq^n)}.$$

Now letting $n \to \infty$ leads to the same result as in Problems 20–22.

24. Consider a triangular array of identically distributed renewal processes $N_{ni}(t), 1 \leq i \leq n$, where the interarrival times have a distribution $F(t)$ with mean μ. Consider the nth row of the array. In each process of this row, retain an event with probability $1/n$ and discard the event with probability $1 - (1/n)$. This operation is applied independently to all events. Denote the new array of renewal processes obtained by this deletion operation by $N^*_{ni}(t)$. Next form the superposition of composed processes,

$$N^*_n(t) = \sum_{j=1}^{n} N^*_{nj}(t), \qquad 1 \leq n < \infty.$$

Show that

$$\lim_{n \to \infty} \Pr[N^*_n(t) = j] = \frac{e^{-t/\mu}}{j!}(t/\mu)^j,$$

if and only if $F(t) = 1 - e^{-t/\mu}$. In other words, the superpositions converge to a Poisson process if and only if all original renewal component processes were Poisson.

Answer: Verify first that the modified array $N^*_{nj}(t)$ is infinitesimal, i.e., show that

$$\lim_{n \to \infty} \left[\sup_{1 \leq j \leq n} F^*_{nj}(t) \right] = 0,$$

where $F^*_{ni}(t)$ is the interarrival distribution for the transformed process $N^*_{ni}(t)$. Indeed, paraphrasing the argument of Problem 19 gives,

$$\begin{aligned} F^*_{ni}(t) &= \sum_{j=1}^{\infty} \left(1 - \frac{1}{n}\right)^{j-1} \frac{1}{n} F_j(t) \\ &\leq \frac{1}{n} \sum_{j=1}^{\infty} F_j(t) \leq \frac{1}{n} \sum_{j=1}^{\infty} [F(t)]^j \\ &\leq \frac{1}{n} \frac{F(t)}{1 - F(t)}. \end{aligned}$$

Where $F(t) < 1$, then manifestly $\lim_{n \to \infty} [\sup_{1 \leq i \leq n} F^*_{ni}(t)] = 0$, while if $F(t) = 1$, we can determine an appropriate j such that $F_j(t) < 1$, and a similar estimate can be made.

Next apply the superposition Theorem 9.1,

$$\begin{aligned} \lim_{n \to \infty} \sum_{i=1}^{n} F^*_{ni}(t) &= \lim_{n \to \infty} \sum_{j=1}^{\infty} \left(1 - \frac{1}{n}\right)^{j-1} F_j(t) \\ &= \sum_{j=1}^{\infty} F_j(t) \\ &= M(t) \quad \text{the renewal function} \\ &= t/\mu, \end{aligned}$$

which is equivalent to $F(t) = 1 - e^{-t/\mu}$.

25. Given a renewal process with finite mean, suppose the excess life γ_t and current life δ_t are independent random variables for all t. Establish that the process is Poisson.

Hint: Use the limit theorem of Section 5 on the identity

$$\Pr\{\delta_t > x, \gamma_t > y\} = \Pr\{\delta_t > x\}\Pr\{\gamma_t > y\},$$

to derive a functional equation for

$$v(x) = \frac{1}{\mu}\int_x^\infty [1 - F(\xi)]\, d\xi,$$

and deduce thereby that $1 - F(\xi) = e^{-\xi/\mu}$.

26. (a) Assume orders for goods arrive at a central office according to a Poisson process with parameter λ. Suppose to fill each order takes a random length of time following a distribution $F(\xi)$. The number of workers available is infinite, so that all orders are handled without delay. Let $W(t)$ represent the number of orders requested but not yet filled by time t. Find

$$\lim_{t\to\infty} \Pr\{W(t) \leq k\}.$$

(b) Let $V(t)$ be the length of time required to fulfill all current orders given that at time 0 there are no unfilled orders. Determine the probability distribution of $V(t)$, i.e., find

$$\Pr\{V(t) < y\} = F(y, t).$$

Hint: Write out a recursion relation for $F(y, t)$ by conditioning on the time of the arrival of the first order.

27. The Laplace transform $g^*(\theta)$, $\theta > 0$, of a continuous function $g(x)$, $x \geq 0$, is defined by $g^*(\theta) = \int_0^\infty e^{-\theta x} g(x)\, dx$. Establish the formula

$$m^*(\theta) = \frac{f^*(\theta)}{1 - f^*(\theta)},$$

for a renewal process having lifetime density $f(x)$, where $m(t) = dM(t)/dt$ is the derivative of the renewal function. Compute $m^*(\theta)$ when

(i) $f(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$

(ii) $f(x) = xe^{-x}, \quad x \geq 0.$

28. Show that the limiting distribution as $t \to \infty$ of age δ_t in a renewal process has mean $(\sigma^2 + \mu^2)/2\mu$, where σ^2 and μ are the variance and mean, respectively, of the interoccurrence distribution.

29. Let δ_t be the age or current life in a renewal process in which the mean and variance of the interoccurrence distribution are μ and σ^2, respectively. Prove

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \delta_\tau\, d\tau = (\sigma^2 + \mu^2)/2\mu.$$

30. Let $X_1, X_2, \ldots,$ be the interoccurrence times in a renewal process. Suppose $\Pr\{X_k = 1\} = p$ and $\Pr\{X_k = 2\} = q = 1 - p$. Verify that

$$E[N_n] = \frac{n}{1+q} - \frac{q^2}{(1+q)^2} + \frac{q^{n+2}}{(1+q)^2}, \qquad n = 2, 4, \ldots,$$

where N_n is the mean number of renewals up to (discrete time) n.

REFERENCE

1. W. Feller, "An Introduction to Probability Theory and Its Applications," Vol. II. Wiley, New York, 1966.

Chapter 6

MARTINGALES

Stochastic processes are characterized by the dependence relationships among their variables. The martingale property expresses a relation that occurs in numerous contexts and has become a basic tool in both theoretical and applied probability. It is used for calculating absorption probabilities, analyzing the path structure of continuous time processes, deriving inequalities for stochastic processes, analyzing sequential decision and control models, and for a multitude of other purposes.

Martingale theory requires extensive use of conditional expectations. We suggest the reader review the properties of conditional expectation listed in Chapter 1 before continuing.

1: Preliminary Definitions and Examples

We initiate the formulation of the martingale concept with undoubtedly its earliest version, which although dated bears historical interest.

Definition 1.1. A stochastic process $\{X_n;\ n = 0, 1, \ldots\}$ is a martingale if, for $n = 0, 1, \ldots,$

$$\text{(i)} \quad E[|X_n|] < \infty,$$

and

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

Let X_n be a player's fortune at stage n of a game. The martingale property captures one notion of a game being fair in that the player's fortune on the next play is, on the average, his current fortune and is not otherwise affected by the previous history. In fact, the name "martingale" derives from a French acronym for the gambling strategy of doubling ones bets until a win is secured. At the present time, martingale theory has such broad scope and diverse domains of applications in general probability theory and mathematical analysis that to think of it purely in terms of gambling would be unduly restrictive and misleading.

A more general and pertinent definition follows. (In Section 7 we will elaborate the most up-to-date formulation.) Unless stated explicitly to the contrary, all random variables encountered are assumed to be real valued.

Definition 1.2. Let $\{X_n; n = 0, 1, \ldots\}$ and $\{Y_n; n = 0, 1, \ldots\}$ be stochastic processes. We say $\{X_n\}$ is a martingale with respect to $\{Y_n\}$ if, for $n = 0, 1, \ldots,$

$$\text{(i)} \quad E[|X_n|] < \infty, \tag{1.1}$$

and

$$\text{(ii)} \quad E[X_{n+1}|Y_0, \ldots, Y_n] = X_n. \tag{1.2}$$

It is useful to think of $(Y_0, \ldots, Y_n)$ as the information or history up to stage n. Thus, in the gambling context, this history could include more information than merely the sequence of past fortunes $(X_1, \ldots, X_n)$ as, for example, the outcomes on plays in which the player did not bet. Actually, there is no desire to restrict Y_k to being a real random variable. In general, it well may be a finite- or even infinite-dimensional vector. Whenever the particular sequence $\{Y_n\}$ is not vital or is evident from the context, we will suppress reference to it and say only "$\{X_n\}$ is a martingale".

The history determines X_n in the sense that X_n is a function of $Y_0, \ldots, Y_n$, i.e., knowledge of the values of $Y_0, Y_1, \ldots, Y_n$ determines the values of X_n. Note from (1.2) that X_n is the particular function

$$X_n = E[X_{n+1}|Y_0, \ldots, Y_n]$$

of $Y_0, \ldots, Y_n$. From the property of conditional expectation, viz.,

$$E[g(Y_0, \ldots, Y_n)|Y_0, \ldots, Y_n] = g(Y_0, \ldots, Y_n), \tag{1.3}$$

we infer that

$$E[X_n|Y_0, \ldots, Y_n] = X_n,$$

and invoking the law of total probability, now yields

$$\begin{aligned} E[X_{n+1}] &= E\{E[X_{n+1}|Y_0, \ldots, Y_n]\} \\ &= E[X_n], \end{aligned}$$

so that, by induction,

$$E[X_n] = E[X_0], \qquad \text{for all} \quad n.$$

SOME EXAMPLES

These examples were selected to demonstrate the immense variety and relevance of martingale processes. We start with some important concrete cases and later add more general constructions.

(a) *Sums of Independent Random Variables*

Let $Y_0 = 0$ and $Y_1, Y_2, \ldots$ be independent random variables with $E[|Y_n|] < \infty$ and $E[Y_n] = 0$ for all n. If $X_0 = 0$ and $X_n = Y_1 + \cdots + Y_n$ for $n \geq 1$, then $\{X_n\}$ is a martingale with respect to $\{Y_n\}$. We check (1.1) from

$$E[|X_n|] \leq E[|Y_1|] + \cdots + E[|Y_n|] < \infty,$$

and verify (1.2) from

$$\begin{aligned} E[X_{n+1}|Y_0, \ldots, Y_n] &= E[X_n + Y_{n+1}|Y_0, \ldots, Y_n] \\ &= E[X_n|Y_0, \ldots, Y_n] + E[Y_{n+1}|Y_0, \ldots, Y_n] \\ &= X_n + E[Y_{n+1}] \qquad \text{(because of the independence assumption on } \{Y_i\}) \\ &= X_n \qquad \text{(since } E[Y_m] = 0 \text{ by stipulation.)} \end{aligned}$$

(b) *More General Sums*

Suppose $Z_i = g_i(Y_0, \ldots, Y_i)$ for some arbitrary sequences of random variables Y_i and functions g_i. Let f be a function for which

$$E[|f(Z_k)|] < \infty, \qquad \text{for} \quad k = 0, 1, \ldots.$$

Let a_k be a bounded function of k real variables. Then

$$X_n = \sum_{k=0}^{n} \{f(Z_k) - E[f(Z_k)|Y_0, \ldots, Y_{k-1}]\} a_k(Y_0, \ldots, Y_{k-1})$$

defines a martingale with respect to $\{Y_n\}$. (By convention, $E[f(Z_k)|Y_0, \ldots, Y_{k-1}] = E[f(Z_k)]$ when $k = 0$.) Since a_k is bounded, say,

$$|a_k(y_0, \ldots, y_{k-1})| \leq A_k, \qquad \text{for all} \quad y_0, \ldots, y_{k-1},$$

we have

$$E[|X_n|] \leq 2 \sum_{k=0}^{n} A_k E[|f(Z_k)|] < \infty.$$

Let $B_k = \{f(Z_k) - E[f(Z_k)|Y_0, \ldots, Y_{k-1}]\} a_k(Y_0, \ldots, Y_{k-1})$. Then citing (1.3) we see that

$$E[B_k|Y_0, \ldots, Y_{k-1}] = 0.$$

Thus,

$$\begin{aligned} E[X_n|Y_0, \ldots, Y_{n-1}] &= E[X_{n-1}|Y_0, \ldots, Y_{n-1}] + E[B_n|Y_0, \ldots, Y_{n-1}] \\ &= X_{n-1}, \end{aligned}$$

which establishes the martingale property.

(c) *The Variance of a Sum as a Martingale*

Let $Y_0 = 0$ and $Y_1, Y_2, \ldots$, be independent identically distributed random variables with $E[Y_k] = 0$ and $E[Y_k^2] = \sigma^2$, $k = 1, 2, \ldots$. Let $X_0 = 0$ and

$$X_n = \left(\sum_{k=1}^{n} Y_k\right)^2 - n\sigma^2.$$

Then $E[|X_n|] \leq 2n\sigma^2 < \infty$, and

$$\begin{aligned} E[X_{n+1}|Y_0, \ldots, Y_n] &= E\left[\left(Y_{n+1} + \sum_{k=1}^{n} Y_k\right)^2 - (n+1)\sigma^2 | Y_0, \ldots, Y_n\right] \\ &= E\left[Y_{n+1}^2 + 2Y_{n+1}\sum_{k=1}^{n} Y_k + \left(\sum_{k=1}^{n} Y_k\right)^2 \right. \\ &\quad \left. - (n+1)\sigma^2 | Y_0, \ldots, Y_n\right] \\ &= X_n + E[Y_{n+1}^2 | Y_0, \ldots, Y_n] \\ &\quad + 2E[Y_{n+1}|Y_0, \ldots, Y_n]\left(\sum_{k=1}^{n} Y_k\right) - \sigma^2 \\ &= X_n. \end{aligned}$$

Thus $\{X_n\}$ is a martingale with respect to $\{Y_n\}$.

(d) *Right Regular Sequences and Induced Martingales for Markov Chains*

There is a routine and highly productive way of discovering martingales in association with Markov processes. Let $Y_0, Y_1, \ldots$, represent a Markov chain process governed by the transition probability matrix $P = \|P_{ij}\|$. Let f be a bounded right regular sequence for P, that is, $f(i)$ is non-negative and satisfies

$$f(i) = \sum_j P_{ij} f(j). \tag{1.4}$$

(See also Chapter 11, Volume II.) Set $X_n = f(Y_n)$. Then $E[|X_n|] < \infty$ since f is bounded, and

$$\begin{aligned} E[X_{n+1}|Y_0, \ldots, Y_n] &= E[f(Y_{n+1})|Y_0, \ldots, Y_n] \\ &= E[f(Y_{n+1})|Y_n] \qquad \text{(by the Markov property)} \\ &= \sum_j P_{Y_n, j} f(j) \\ &\quad \text{(since } E[f(Y_{n+1})|Y_n = i] = \sum_j P_{ij} f(j)\text{)} \\ &= f(Y_n) \qquad \text{[in accordance with (1.4)]} \\ &= X_n. \end{aligned}$$

Many martingales that at first glance seem unrelated actually arise in this manner, or in the generalization that is our next example.

(e) *Martingales Induced by Eigenvectors of the Transition Matrix*

Probably the most widespread method of forming martingales is covered by this example (see Elementary Problems 8, 15, 18, 19, 21, and 23). Let $Y_0, Y_1, \ldots$ be a Markov chain having transition probability matrix $P = \|P_{ij}\|$. A vector f is a right *eigenvector* of P if for some λ, called the *eigenvalue*,

$$\lambda f(i) = \sum_j P_{ij} f(j), \qquad \text{for all} \quad i.$$

If f is a right eigenvector of P for which $E[|f(Y_n)|] < \infty$ for all n, then

$$X_n = \lambda^{-n} f(Y_n)$$

is a martingale, since

$$\begin{aligned} E[X_{n+1}|Y_0, \ldots, Y_n] &= E[\lambda^{-n-1} f(Y_{n+1})|Y_0, \ldots, Y_n] \\ &= \lambda^{-n}\lambda^{-1} E[f(Y_{n+1})|Y_n] = \lambda^{-n}\lambda^{-1} \sum_j P_{Y_n,j} f(j) \\ &= \lambda^{-n} f(Y_n) = X_n. \end{aligned}$$

More generally, suppose $Y_0, Y_1, \ldots$ is a discrete-time Markov process governed by the transition distribution function

$$F(y|z) = \Pr\{Y_{n+1} \leq y | Y_n = z\}. \tag{1.5}$$

If

$$E[|f(Y_n)|] < \infty, \qquad \text{for all} \quad n,$$

and

$$\lambda f(y) = \int f(z)\, dF(z|y), \qquad \text{for all} \quad y,$$

then $X_n = \lambda^{-n} f(Y_n)$ is a martingale.

The subsequent examples amply demonstrate the power and versatility of this technique for producing martingales.

(f) *A Branching Process*

Let $\{Y_n\}$ specify a branching process (Chapter 2, Section 2, Example F) and suppose that the mean of the progeny distribution is $m < \infty$. Then $X_n = m^{-n} Y_n$ is a martingale. In order to validate this claim we designate by $Z^{(n)}(j)$ the number of progeny produced by the jth existing parent in the nth generation. Then

$$Y_{n+1} = Z^{(n)}(1) + \cdots + Z^{(n)}(Y_n),$$

where $Z^{(n)}(i)$, $i = 1, 2, \ldots, Y_n$ are independent and identically distributed.

Manifestly

$$E[Y_{n+1}|Y_n] = Y_n E[Z^{(n)}(1)] = mY_n,$$

so that m is an eigenvalue for the function $f(y) = y$. It follows that $X_n = m^{-n}Y_n$ is a martingale as asserted.

(g) *Wald's Martingale*

Let $Y_0 = 0$ and suppose $Y_1, Y_2, \ldots$ are independent identically distributed random variables having a finite *moment generating function* $\phi(\lambda) = E[\exp\{\lambda Y_k\}]$ existing for some $\lambda \neq 0$. Then $X_0 = 1$ and $X_n = \phi(\lambda)^{-n} \exp\{\lambda(Y_1 + \cdots + Y_n)\}$ determines a martingale with respect to $\{Y_n\}$, because the function $f(y) = e^{\lambda y}$ is an eigenfunction for the Markov process of partial sums $S_n = Y_1 + \cdots + Y_n$, and the associated eigenvalue is $\phi(\lambda)$. Indeed, in line with (1.5), the transition distribution function in the case at hand is

$$P\{S_{n+1} \leq y | S_n = x\} = G(y - x),$$

where G is the common distribution function of Y_k.

Now we can calculate $E[f(S_{n+1})|S_n = x]$, executing an obvious change of variable, to obtain

$$\int e^{\lambda y}\, d_y G(y - x) = e^{\lambda x} \int e^{\lambda \xi}\, dG(\xi) = e^{\lambda x}\phi(\lambda),$$

and this identity clearly validates the claim made before.

As an illustration suppose $Y_1, Y_2 \ldots$ are independent and normally distributed with mean zero and variance σ^2. Then

$$\phi(\lambda) = E[\exp\{\lambda Y_1\}] = \exp\{\tfrac{1}{2}\lambda^2\sigma^2\}$$

and

$$X_n = \exp\left\{\lambda(Y_1 + \cdots + Y_n) - \frac{n}{2}\lambda^2\sigma^2\right\}.$$

For the choice $\lambda = \mu/\sigma^2$, where μ is an arbitrary constant, we get

$$X_n = \exp\left\{\frac{\mu}{\sigma^2}(Y_1 + \cdots + Y_n) - \frac{n\mu^2}{2\sigma^2}\right\}.$$

This martingale appears again in Example (j).

(h) *Generalization of the Eigenvector Argument*

Let $Y_0, Y_1, \ldots$ be arbitrary random variables but having finite absolute means $E[|Y_n|] < \infty$. Suppose, for $n = 0, 1, 2, \ldots$,

$$E[Y_{n+1}|Y_0, \ldots, Y_n] = a_n + b_n Y_n, \qquad b_n \neq 0. \tag{1.6}$$

Let $l_{n+1}(z)$ be the linear function $l_{n+1}(z) = a_n + b_n z$, whose inverse is $l_{n+1}^{-1}(y) = (y - a_n)/b_n$, and let $L_n(y) = l_1^{-1}(l_2^{-1}(\cdots (l_n^{-1}(y)\cdots)))$. Then $X_n = kL_n(Y_n)$ is a martingale, for any constant k, because

$$\begin{aligned}\frac{1}{k}E[X_{n+1}|Y_0, \ldots, Y_n] &= E[L_{n+1}(Y_{n+1})|Y_0, \ldots, Y_n] \\ &= L_{n+1}\{E[Y_{n+1}|Y_0, \ldots, Y_n]\} \qquad \text{(since } L_{n+1} \text{ is a linear function of its argument)} \\ &= L_{n+1}(l_{n+1}(Y_n)), \qquad \text{by (1.6)} \\ &= L_n(Y_n) = \frac{1}{k}X_n.\end{aligned}$$

To illustrate concretely, let Y_0 be uniformly distributed over [0, 1], and given Y_n, suppose Y_{n+1} is uniformly distributed on $[Y_n, 1]$. Then $X_n = 2^n[1 - Y_n]$ is a martingale. We check directly:

$$\begin{aligned}E[X_{n+1}|Y_0, \ldots, Y_n] &= 2^{n+1}[1 - E[Y_{n+1}|Y_n]] \\ &= 2^{n+1}[1 - \tfrac{1}{2}(1 + Y_n)] \\ &= 2^n(1 - Y_n) = X_n.\end{aligned}$$

(i) *An Urn Scheme*

Here is another example of the generalized eigenvector argument. The model has application in the study of growth processes.

Consider an urn that at stage 0 contains one red and one green ball. A ball is drawn at random from the urn and it and one more of the same color are then returned. The experiment is repeated indefinitely. Let X_n be the fraction of red balls at stage n, and let $Y_n = (n + 2)X_n$ be the number of red balls. Then $\{X_n\}$ is a martingale with respect to $\{Y_n\}$. We have that, given $Y_n = k$,

$$Y_{n+1} = \begin{cases} k+1, & \text{with probability } k/(n+2), \\ k, & \text{with probability } 1 - k/(n+2). \end{cases}$$

Hence

$$E[Y_{n+1}|Y_n = k] = \frac{(k+1)k + k(n+2-k)}{n+2} = k(n+3)/(n+2).$$

That is,

$$E[Y_{n+1}|Y_n] = b_n Y_n,$$

where $b_n = (n+3)/(n+2)$. Then, using the notation of (h), $l_n(z) = b_{n-1}z$, $l_n^{-1}(y) = z/b_{n-1}$, and

$$\begin{aligned} L_n(y) &= \frac{y}{b_0 b_1 \cdots b_{n-1}} \\ &= \frac{2}{3} \cdot \frac{3}{4} \cdots \frac{n+1}{n+2} y \\ &= \frac{2}{n+2} y. \end{aligned}$$

Thus

$$X_n = \frac{1}{2} L_n(Y_n) = \frac{1}{n+2} Y_n$$

is a martingale.

(j) *Likelihood Ratios*

Let $Y_0, Y_1 \ldots$ be independent, identically distributed random variables, and let f_0 and f_1 be probability density functions. A stochastic process of fundamental importance in the theory of testing statistical hypotheses is the sequence of likelihood ratios

$$X_n = \frac{f_1(Y_0) f_1(Y_1) \cdots f_1(Y_n)}{f_0(Y_0) f_0(Y_1) \cdots f_0(Y_n)}, \qquad n = 0, 1, \ldots.$$

To assure the definition, assume $f_0(y) > 0$ for all y. Since $Y_0, Y_1, \ldots$ are independent,

$$\begin{aligned} E[X_{n+1} | Y_0, \ldots, Y_n] &= E\left[X_n \frac{f_1(Y_{n+1})}{f_0(Y_{n+1})} \,\middle|\, Y_0, \ldots, Y_n \right] \\ &= X_n E\left[\frac{f_1(Y_{n+1})}{f_0(Y_{n+1})} \right]. \end{aligned}$$

When the common distribution of the Y_k's has f_0 as its probability density function, $\{X_n\}$ is a martingale with respect to $\{Y_n\}$. To confirm this claim, we need only verify

$$E\left[\frac{f_1(Y_{n+1})}{f_0(Y_{n+1})} \right] = 1.$$

But

$$\begin{aligned} E\left[\frac{f_1(Y_{n+1})}{f_0(Y_{n+1})} \right] &= \int \left\{ \frac{f_1(y)}{f_0(y)} \right\} f_0(y)\, dy \\ &= \int f_1(y)\, dy = 1, \end{aligned}$$

as desired.

As an example, suppose f_0 is the normal density with mean zero and variance σ^2, and f_1 is normal with mean μ and variance σ^2. Then

$$\frac{f_1(y)}{f_0(y)} = \exp\left\{\frac{2\mu y - \mu^2}{2\sigma^2}\right\},$$

and

$$X_n = \exp\left\{\frac{\mu}{\sigma^2}(Y_1 + \cdots + Y_n) - \frac{n\mu^2}{2\sigma^2}\right\}.$$

This martingale occurred earlier in Example (g).

Martingales constructed from likelihood ratios have many uses in evaluating the properties of sequential procedures for hypothesis testing.

(k) *Doob's Martingale Process*

Let $Y_0, Y_1, \ldots$ be an arbitrary sequence of random variables and suppose X is a random variable satisfying $E[|X|] < \infty$. Then

$$X_n = E[X|Y_0, \ldots, Y_n]$$

forms a martingale with respect to $\{Y_n\}$, called *Doob's process*. First

$$\begin{aligned} E[|X_n|] &= E\{|E[X|Y_0, \ldots, Y_n]|\} \\ &\leq E\{E[|X|\,|\,Y_0, \ldots, Y_n]\} \\ &= E[|X|] < \infty. \end{aligned}$$

Second and last, by the law of total probability for conditional expectations,†

$$\begin{aligned} E[X_{n+1}|Y_0, \ldots, Y_n] &= E\{E[X|Y_0, \ldots, Y_{n+1}]|Y_0, \ldots, Y_n\} \\ &= E[X|Y_0, \ldots, Y_n] = X_n. \end{aligned}$$

(l) *Radon–Nikodym Derivatives*

Suppose Z is a uniformly distributed random variable on [0, 1), and define the random variables Y_n by setting

$$Y_n = k/2^n,$$

for the unique k (depending on n and Z) that satisfies

$$\frac{k}{2^n} \leq Z < \frac{k+1}{2^n}.$$

† The law of total probability for conditional expectation extends the usual law by introducing further conditioning on a random variable Z. The law states $E[X|Z] = E\{E[X|Y, Z]|Z\}$, valid whenever $E[|X|] < \infty$. The student should supply a proof.

Notice how Y_n provides increasingly more information about Z as n increases. Indeed, $Y_n \leq Z < Y_n + (\frac{1}{2})^n$ so that Y_n determines the first n bits in Z's terminating binary expansion.

Let f be a bounded function on [0, 1] and form the difference quotient

$$X_n = 2^n\{f(Y_n + 2^{-n}) - f(Y_n)\}.$$

We claim that $\{X_n\}$ is a martingale with respect to $\{Y_n\}$. First observe that Z, conditional on $Y_0, \ldots, Y_n$, has a uniform distribution on $[Y_n, Y_n + 2^{-n})$, and thus Y_{n+1} is equally likely to be Y_n or $Y_n + 2^{-(n+1)}$ Thus

$$\begin{aligned} E[X_{n+1}|Y_0, \ldots, Y_n] &= 2^{n+1}E[f(Y_{n+1} + 2^{-(n+1)}) - f(Y_{n+1})|Y_0, \ldots, Y_n] \\ &= 2^{n+1}\{\tfrac{1}{2}[f(Y_n + 2^{-(n+1)}) - f(Y_n)] \\ &\quad + \tfrac{1}{2}[f(Y_n + 2^{-n}) - f(Y_n + 2^{-(n+1)})]\} \\ &= 2^n\{f(Y_n + 2^{-n}) - f(Y_n)\} = X_n. \end{aligned}$$

Note that

$$X_n = \frac{f(Y_n + 2^{-n}) - f(Y_n)}{2^{-n}}$$

is approximately the derivative of f at Z. In fact, under quite general conditions it can be shown that the sequence $\{X_n\}$ of approximate derivatives converges with probability one to a random variable $X_\infty = X_\infty(Z)$, called the Radon–Nykodym derivative of f evaluated at Z, and that $X_n = E[X_\infty|Y_0, \ldots, Y_n]$ (see the close of Section 7). Thus martingale properties find alliance and relevance in the theory of differentiation of functions and indeed in numerous other facets of mathematical analysis.

PREVIEW OF RESULTS

The next section treats generalizations where the martingale equality is replaced by an inequality. Following that, we deal with the two major results of martingale theory, the *optional sampling theorem* and the *martingale convergence theorem*, including a diversity of applications of these theorems.

The optional sampling theorem tells us that, under quite general circumstances, whenever X_n is a martingale, then $X_{T_n} = Z_n$ also constitutes a martingale for a collection of randomly selected times $\{T_n\}$, which form an increasing sequence of "Markov times." A Markov time T has the property that the event $\{T = n\}$ is determined only by the history $(Y_0, \ldots, Y_n)$ up to stage n. The optional sampling (or stopping) theorem finds frequent application in sequential decision problems and in deriving

inequalities and estimates of probabilities of various events associated with certain stochastic processes.

Martingale convergence theorems provide general conditions under which a martingale X_n will converge to a limit random variable X_∞ as n increases. These theorems are of value for analyzing the path structure of processes and in determining the asymptotic distribution of a variety of functionals of quite general stochastic processes.

2: *Supermartingales and Submartingales*

For many purposes it is desirable to have available a more general concept, built around an inequality.

Definition 2.1. Let $\{X_n, n = 0, 1, \ldots\}$ and $\{Y_n, n = 0, 1, \ldots\}$ be stochastic processes. Then $\{X_n\}$ is called a supermartingale with respect to $\{Y_n\}$ if, for all n,

$$\begin{aligned} &\text{(i)} && E[X_n^-] > -\infty, \qquad \text{where} \quad x^- = \min\{x, 0\}, \\ &\text{(ii)} && E[X_{n+1} | Y_0, \ldots, Y_n] \leq X_n, \\ &\text{(iii)} && X_n \quad \text{is a function of} \quad (Y_0, \ldots, Y_n). \end{aligned} \tag{2.1}$$

We call $\{X_n\}$ a submartingale with respect to $\{Y_n\}$ if, for all n,

$$\begin{aligned} &\text{(i)} && E[X_n^+] < \infty, \qquad \text{where} \quad x^+ = \max\{0, x\}, \\ &\text{(ii)} && E[X_{n+1} | Y_0, \ldots, Y_n] \geq X_n, \\ &\text{(iii)} && X_n \quad \text{is a function of} \quad (Y_0, \ldots, Y_n). \end{aligned} \tag{2.2}$$

As we did with martingales, we will omit mention of $\{Y_n\}$ when it is either not important or else evident from the context which particular sequence $\{Y_n\}$ is involved.

The third stipulation in each definition states that X_n must be determined by the history up to time n, or equivalently, the information available to time n includes the value of X_n. As noted earlier, the determination is automatically satisfied in the martingale case with the requisite expression of X_n as a function of $\{Y_i\}_{i=0}^n$ being

$$X_n = E[X_{n+1} | Y_0, \ldots, Y_n]. \tag{2.3}$$

In the super- and submartingale cases, the martingale equality is replaced by an inequality. Therefore the requirement that X_n be determined by $(Y_0, \ldots, Y_n)$ must be explicitly imposed. We will sometimes indicate this functional relation by writing

$$X_n = X_n(Y_0, \ldots, Y_n).$$

Note that $\{X_n\}$ *is a supermartingale with respect to* $\{Y_n\}$ *if and only if* $\{-X_n\}$ *is a submartingale. Similarly,* $\{X_n\}$ *is a martingale with respect to* $\{Y_n\}$ *if and only if* $\{X_n\}$ *is both a submartingale and a supermartingale.* By this means, statements about supermartingales can be transcribed into equivalent statements concerning both submartingales and martingales. This will save us substantial writing, since often a proof in only one of the three cases need be given.

Example. Let $\{Y_n\}$ be a Markov chain having the transition probability matrix $P = \|P_{ij}\|$. If f is a right superregular sequence for P (i.e., a nonnegative sequence satisfying $\sum_j P_{ij} f(j) \leq f(i)$ for all i), then $X_n = f(Y_n)$ defines a supermartingale with respect to Y_n. The proof of this assertion paraphrases the analysis of Example (d), Section 1.

There is, of course, a parallel correspondence between submartingales and subregular sequences [nonnegative sequences $f(i)$ for which $f(i) \leq \sum_j P_{ij} f(j)$], provided we assume $E[f(Y_n)] < \infty$.

Jensen's Inequality. A function ϕ defined on an interval I is said to be *convex* if for every $x_1, x_2 \in I$ and $0 < \alpha < 1$, we have

$$\alpha\phi(x_1) + (1 - \alpha)\phi(x_2) \geq \phi(\alpha x_1 + (1 - \alpha)x_2). \tag{2.4}$$

A straightforward induction commencing from (2.4) proves

$$\sum_{i=1}^{m} \alpha_i \phi(x_i) \geq \phi\left(\sum_{i=1}^{m} \alpha_i x_i\right), \tag{2.5}$$

valid for all $x_1, x_2, \ldots, x_m \in I$ and $\alpha_i \geq 0$, $\sum_{i=1}^{m} \alpha_i = 1$. If ϕ is twice differentiable, then ϕ is convex if and only if $d^2\phi/dx^2 \geq 0$ for all x. Thus, convexity is often easy to verify. If X is a random variable that takes the value x_i with probability α_i $(i = 1, 2, \ldots, m)$, then Eq. (2.5) can be succinctly written in the form

$$E[\phi(X)] \geq \phi(E[X]). \tag{2.6}$$

Jensen's inequality states that (2.6) prevails for all real random variables X whenever ϕ is convex on $(-\infty, \infty)$. Inequality (2.6) can be viewed as a continuous integrated version of (2.5). The same is true for conditional expectations: Thus, if ϕ is convex, we have

$$E[\phi(X)|\, Y_0, \ldots, Y_n] \geq \phi(E[X|\, Y_0, \ldots, Y_n]). \tag{2.7}$$

With these facts in hand we provide some ways of constructing submartingales from martingales.

Lemma 2.1. *Let $\{X_n\}$ be a martingale with respect to $\{Y_n\}$. If ϕ is a convex function for which $E[\phi(X_n)^+] < \infty$ for all n, then $\{\phi(X_n)\}$ is a submartingale with respect to $\{Y_n\}$. In particular, $\{|X_n|\}$ is always a submartingale and $\{|X_n|^2\}$ is a submartingale whenever $E[X_n^2] < \infty$ for all n.*

Proof. We need only show the submartingale inequality, the other properties being rather easily demonstrated. Using Jensen's inequality, we have

$$\begin{aligned} E[\phi(X_{n+1})|\, Y_0, \ldots, Y_n] &\geq \phi(E[X_{n+1}|\, Y_0, \ldots, Y_n]) \\ &= \phi(X_n). \quad \blacksquare \end{aligned}$$

Here is a similar result whose proof is omitted.

Lemma 2.2. *Let $\{X_n\}$ be a submartingale with respect to $\{Y_n\}$. If ϕ is a convex and increasing function, then $\phi(X_n)$ is a submartingale, provided $E[\phi(X_n)^+] < \infty$.*

(Note that less is demanded of $\{X_n\}$, merely a submartingale, but more of ϕ in that ϕ is increasing besides convex.)

Thus, for example, if $\{X_n\}$ is a submartingale and

$$\tilde{X}_n = \begin{cases} X_n, & \text{if } X_n > -c, \\ -c, & \text{if } X_n \leq -c, \end{cases} \tag{2.8}$$

where c is fixed, then $\{\tilde{X}_n\}$ is a submartingale for which $E[|\tilde{X}_n|] < \infty$ for all n, and as a special case, $\{X_n^+\}$ is a submartingale whenever $\{X_n\}$ is.

ELEMENTARY PROPERTIES

We include both the supermartingale and the martingale results in a single statement, the hypothesis and conclusion for the supermartingale being enclosed in parentheses. (The corresponding results for submartingales are derived by passing from $\{X_n\}$ to $\{-X_n\}$.)

(a) If $\{X_n\}$ is a (super) martingale with respect to $\{Y_n\}$, then

$$E[X_{n+k}|\, Y_0, \ldots, Y_n](\leq) = X_n, \qquad \text{for every} \quad k \geq 0. \tag{2.9}$$

Proof. We proceed by induction. By definition, (2.9) is correct for $k = 1$. Suppose (2.9) holds for k. Then

$$\begin{aligned} E[X_{n+k+1}|\, Y_0, \ldots, Y_n] &= E\{E[X_{n+k+1}|\, Y_0, \ldots, Y_n, \ldots, Y_{n+k}]|\, Y_0, \ldots, Y_n\} \\ &(\leq) = E\{X_{n+k}|\, Y_0, \ldots, Y_n\} \\ &(\leq) = X_n. \end{aligned}$$

(b) If $\{X_n\}$ is a (super) martingale, then for $0 \leq k \leq n$

$$E[X_n](\leq) = E[X_k](\leq) = E[X_0]. \tag{2.10}$$

Proof. Using (2.9) we take expectations in

$$E[X_n | Y_0, \ldots, Y_k](\leq) = X_k,$$

to conclude

$$E[X_n] = E\{E[X_n | Y_0, \ldots, Y_k]\}(\leq) = E[X_k].$$

The case $E[X_k](\leq) = E[X_0]$ uses the same argument. ■

(c) Suppose $\{X_n\}$ is a (super) martingale with respect to $\{Y_n\}$ and that g is a (nonnegative) function of $Y_0, \ldots, Y_n$ for which the expectations that follow exist. Then

$$E[g(Y_0, \ldots, Y_n)X_{n+k} | Y_0, \ldots, Y_n](\leq) = g(Y_0, \ldots, Y_n)X_n. \tag{2.11}$$

Proof. Since $g(Y_0, \ldots, Y_n)$ is determined by $(Y_0, \ldots, Y_n)$, using a basic property of conditional expectation we have

$$\begin{aligned} E[g(Y_0, \ldots, Y_n)X_{n+k} | Y_0, \ldots, Y_n] &= g(Y_0, \ldots, Y_n)E[X_{n+k} | Y_0, \ldots, Y_n] \\ (\leq) &= g(Y_0, \ldots, Y_n)X_n. \end{aligned}$$

(For the supermartingale case $(\leq)$ we need $g \geq 0$.) ■

A SEQUENTIAL DECISION MODEL

Consider a system with a finite number S of states, labeled by the integers 1, 2, ..., S. Periodically, say once a day, we observe the current state of the system, and then choose an action from a set containing a finite number A of possible actions, labeled 1, 2, ..., A. As a joint result of the current state s and the chosen action a, two things happen: (i) we receive an immediate income $i(s, a)$; and (ii) the system moves to a new state, where the probability of a particular state s' being attained is determined by a known function $q = q(s'|s, a)$. Our problem is to ascertain the policy for choosing actions that maximize the total expected income over an N period horizon.

Let $S_0, A_0, S_1, \ldots, A_{N-2}, S_{N-1}, A_{N-1}$ describe the sequence of alternating states and acts. A policy π is a set of functions $\pi_0, \ldots, \pi_{N-1}$, where π_n prescribes the act A_n as a function of the observed history $S_0, A_0, \ldots, A_{n-1}, S_n$. That is, if policy π is used, then

$$A_n = \pi_n(S_0, A_0, \ldots, A_{n-1}, S_n).$$

The expected reward under policy π as a function of the initial state $S_0 = s$ is

$$I(\pi, s) = E\left[\sum_{k=0}^{N-1} i(S_k, A_k)\right]$$

We want to choose π so as to maximize $I(\pi, s)$.

Define the functions $f_0, \ldots, f_N$ recursively backwards according to

$$f_N(s) = 0, \qquad \text{for all} \quad s, \tag{2.12}$$

and, for $n = 1, 2, \ldots, N$,

$$f_{n-1}(s) = \max_a \left\{ i(s, a) + \sum_{s'} f_n(s') q(s'|s, a) \right\}. \tag{2.13}$$

Then

$$X_n = \sum_{k=1}^{n} \{f_k(S_k) - E[f_k(S_k)|S_0, A_0, \ldots, S_{k-1}, A_{k-1}]\}$$

defines a martingale with respect to $\{Y_n\} = \{(S_n, A_n)\}$ by Example (b) of Section 1. Thus

$$E[X_n] = 0. \tag{2.14}$$

Now (2.13) implies

$$f_{n-1}(s) \geq i(s, a) + \sum_{s'} f_n(s') q(s'|s, a), \qquad \text{for all} \quad s \text{ and } a,$$

so that, in particular,

$$\begin{aligned} f_{k-1}(S_{k-1}) &\geq i(S_{k-1}, A_{k-1}) + \sum_{s'} f_k(s') q(s'|S_{k-1}, A_{k-1}) \\ &= i(S_{k-1}, A_{k-1}) + E[f_k(S_k)|S_0, A_0, \ldots, S_{k-1}, A_{k-1}], \end{aligned} \tag{2.15}$$

and this holds no matter what policy is used. We substitute into (2.14) to conclude

$$\begin{aligned} 0 = E[X_N] &= E\left[\sum_{k=1}^{N} \{f_k(S_k) - E[f_k(S_k)|S_0, A_0, \ldots, S_{k-1}, A_{k-1}]\}\right] \\ &\geq E\left[\sum_{k=1}^{N} \{f_k(S_k) + i(S_{k-1}, A_{k-1}) - f_{k-1}(S_{k-1})\}\right] \\ &= E\left[\sum_{k=0}^{N-1} i(S_k, A_k) + f_N(S_N) - f_0(S_0)\right]. \end{aligned}$$

That is,

$$E[f_0(S_0)] \geq E\left[\sum_{k=0}^{N-1} i(S_k, A_k)\right].$$

If $S_0 = s$, then this says, for any policy π,

$$f_0(s) \geq I(\pi, s).$$

In words, no policy can achieve an expected reward that exceeds $f_0(s)$. Thus, if we exhibit a policy π^* satisfying

$$f_0(s) = I(\pi^*, s),$$

then this policy is manifestly optimal. For each s, let

$$\pi^*_{n-1}(S_0, A_0, \ldots, A_{n-2}, s)$$

be the action that maximizes the right-hand side of (2.13). That is, if $A^*_{k-1} = \pi^*_{k-1}(S_0, A^*_0, \ldots, S_{k-1})$, then (2.15) reduces to the equality

$$f_{k-1}(S_{k-1}) = i(S_{k-1}, A^*_{k-1}) + E[f_k(S_k)|S_0, A^*_0, \ldots, S_{k-1}, A^*_{k-1}].$$

Continuing the same argument as before, we further obtain the equality

$$E[f_0(S_0)] = E\left[\sum_{k=0}^{N-1} i(S_k, A^*_k)\right]$$

or

$$f_0(s) = I(\pi^*, s).$$

Thus π^*, the policy that in state $S_{n-1} = s$ at stage $n-1$ selects the action that maximizes the right-hand side of (2.13), is optimal.

3: The Optional Sampling Theorem

Consider a fair game in which on each play a dollar is won or lost with equal probability. We let $Y_1, Y_2, \ldots$ be independent identically distributed random variables with $\Pr\{Y_k = +1\} = \Pr\{Y_k = -1\} = \frac{1}{2}$. Let $X_n = Y_1 + \cdots + Y_n$ be the player's net gain at stage n. We know $E[X_n] = 0$, the mean net gain is zero. But the player need not play forever, nor need he predetermine a particular time n for stopping. Rather, he might let the choice of when to stop be determined depending on how the game evolves. For example he might try to stop when ahead.

Let T be the time the player ends his play and let X_T be his net gain then. We know $E[X_n] = 0$ for all n, but is it necessarily true that $E[X_T] = 0$? Can the player stop when ahead? The answer is "yes," but there are a number of qualifications.

First, we must outlaw clairvoyance and require that the choice of when to stop depends only on the information observed to date. That is,

we require that the event that the player stops on the nth turn depends only on $Y_0, \ldots, Y_n$. A random variable T that satisfies this requirement for every n is called a *Markov time*, or sometimes a *stopping time* or *random variable independent of the future* (with respect to $\{Y_n\}$). We will give a crisper definition shortly.

Even with this restriction, we can achieve $E[X_T] > 0$! For example, $T = \min\{n : X_n = 1\}$ is a Markov time, since $\{T = n\}$ if and only if $Y_1 + \cdots + Y_k < 1$ for $k < n$, and $Y_1 + \cdots + Y_n = 1$. Since the random walk is recurrent, $T < \infty$, and $X_T \equiv 1$, so that $E[X_T] = 1 > 0$.

It is the purpose of this section to examine this matter more closely and in more generality. We will show that the Markov time T defined above has a number of adverse properties and, in a very real sense, is not physically realizable. For example, the mean of T is infinite, and indefinitely large losses occur, on the average, before stopping, so that a player would need an infinite fortune to adopt this strategy successfully.

On the other hand, under quite general conditions $E[X_T] = E[X_0]$ for a martingale, whenever T is a Markov time with finite expectation, and this conclusion has applications far beyond the gambling setup.

MARKOV TIMES

Markov times occur in many contexts. Here is an example of their use in Markov chains. Suppose i is a recurrent state in a Markov chain $\{Y_n\}$. We want to show

$$\Pr\{Y_n \text{ returns to } i \text{ at least twice}\} = 1.$$

Since i is recurrent we know

$$\Pr\{Y_n \text{ returns to } i \text{ at least once}\} = \Pr\{T_i < \infty\} = 1,$$

where $T_i = \min\{n \geq 1 : Y_n = i\}$ is the time of first return to i. We have, using the Markov property,

$$\begin{aligned}\Pr\{\text{return to } i \text{ at least twice}\} &= \Pr\{T_i < \infty\} \Pr\{\text{return after } T_i \mid T_i < \infty\} \\ &= \Pr\{T_i < \infty\} \Pr\{T_i < \infty \mid Y_0 = i\} \\ &= 1 \times 1 = 1.\end{aligned}$$

There is something that needs more discussion here. Why is the probability of returning to i after time T_i the same as the probability of ever returning to i? Here is a more detailed proof that shows where the Markov property is used and what attributes of the random time T_i make this possible. We note

$$\begin{aligned}
&\Pr\{\text{return after } T_i \mid T_i = k\} \\
&\quad = \Pr\{Y_{k+n} = i, \text{ for some } n = 1, 2, \ldots \mid Y_j \neq i, j = 1, \ldots, k-1; Y_k = i\} \\
&\quad = \Pr\{Y_{k+n} = i, \text{ for some } n = 1, 2, \ldots \mid Y_k = i\} \\
&\qquad \text{(By the Markov property)} \\
&\quad = \Pr\{Y_n = i, \text{ for some } n = 1, 2, \ldots \mid Y_0 = i\} \\
&\qquad \text{(Using the fact of stationary transition probabilities)} \\
&\quad = \Pr\{T_i < \infty \mid Y_0 = i\} = 1.
\end{aligned}$$

Thus

$$\begin{aligned}
&\Pr\{T_i < \infty \text{ and return after } T_i\} \\
&\quad = \sum_{k=1}^{\infty} \Pr\{\text{return after } T_i \mid T_i = k\} \Pr\{T_i = k\} \\
&\quad = \sum_{k=1}^{\infty} 1 \times \Pr\{T_i = k\} = 1.
\end{aligned}$$

The key is that the event $\{T_i = k\}$ is the same as the event $\{Y_j \neq i$, for $j = 1, \ldots, k-1$; $Y_k = i\}$ and, in particular, depends only on $(Y_0, \ldots, Y_k)$.

Definition 3.1. *A random variable T is called a Markov time with respect to $\{Y_n\}$ if T takes values in $\{0, 1, \ldots, \infty\}$ and if, for every $n = 0, 1, \ldots$, the event $\{T = n\}$ is determined by $(Y_0, \ldots, Y_n)$. By "determined" we mean the indicator function of the event $\{T = n\}$ can be written as a function of $Y_0, \ldots, Y_n$,* i.e., we can decide whether $T = n$ or $T \neq n$ from knowledge of the values of the process $Y_0, Y_1, \ldots, Y_n$ only up to time n. *We signify this by writing*

$$\begin{aligned}
I_{\{T=n\}} &= I_{\{T=n\}}(Y_0, \ldots, Y_n) \\
&= \begin{cases} 1, & \text{if } T = n, \\ 0, & \text{if } T \neq n. \end{cases}
\end{aligned}$$

We often omit mention of $\{Y_n\}$ and say only "T is a Markov time." If T is a Markov time, then for every n the events $\{T \leq n\}$, $\{T > n\}$, $\{T \geq n\}$, and $\{T < n\}$ are also determined by $(Y_0, \ldots, Y_n)$. In fact, we have

$$I_{\{T \leq n\}} = \sum_{k=0}^{n} I_{\{T=k\}}(Y_0, \ldots, Y_k),$$

$$I_{\{T > n\}} = 1 - I_{\{T \leq n\}}(Y_0, \ldots, Y_n),$$

and so on.

Conversely, if for every n, the event $\{T \leq n\}$ is determined by $\{Y_0, \ldots, Y_n\}$, then T is a Markov time. Or, if for every n, the event $\{T > n\}$ is determined by $(Y_0, \ldots, Y_n)$, then T is a Markov time. (But see Problem 20.)

If $\{X_n\}$ is a martingale with respect to $\{Y_n\}$, then for every n, X_n is determined by $(Y_0, \ldots, Y_n)$. It follows that every Markov time with respect to $\{X_n\}$ is also a Markov time with respect to $\{Y_n\}$. The same statement holds for supermartingales and submartingales, of course.

Some Examples of Markov Times

(a) The fixed (that is, constant) time $T \equiv k$ is a Markov time. For all $Y_0, Y_1, \ldots$, we have

$$I_{\{T=n\}}(Y_0, \ldots, Y_n) = \begin{cases} 0, & \text{if } n \neq k, \\ 1, & \text{if } n = k. \end{cases}$$

(b) The first time the process $Y_0, Y_1, \ldots$ reaches a subset A of the state space is a Markov time. That is, for

$$T(A) = \min\{n : Y_n \in A\},$$

we have

$$I_{\{T(A)=n\}}\{Y_0, \ldots, Y_n\} = \begin{cases} 1, & \text{if } Y_j \notin A, \text{ for } j = 0, \ldots, n-1, \ Y_n \in A, \\ 0, & \text{otherwise.} \end{cases}$$

(c) More generally, for any fixed k, the kth time the process visits a set A is a Markov time. However, the *last* time a process visits a set is *not* a Markov time. To determine whether or not a particular visit is the last, the entire future must be known.

ELEMENTARY PROPERTIES

(a) If S and T are Markov times, then so is $S + T$. We have

$$I_{\{S+T=n\}} = \sum_{k=0}^{n} I_{\{S=k\}} I_{\{T=n-k\}}.$$

(b) The smaller of two Markov times S, T, denoted

$$S \wedge T = \min\{S, T\},$$

is also a Markov time. This is clear because of the relation

$$I_{\{S \wedge T > n\}} = I_{\{S>n\}} I_{\{T>n\}}.$$

Thus, if T is a Markov time, then so is $T \wedge n = \min\{n, T\}$, for any fixed $n = 0, 1, \ldots$.

(c) If S and T are Markov times, then so is the larger $S \vee T = \max\{S, T\}$, since

$$I_{\{S \vee T \leq n\}} = I_{\{S \leq n\}} I_{\{T \leq n\}}.$$

OPTIONAL SAMPLING THEOREM*

Suppose $\{X_n\}$ is a martingale and T is a Markov time with respect to $\{Y_n\}$. We will establish later

$$E[X_0] = E[X_{T\wedge n}] = \lim_{n\to\infty} E[X_{T\wedge n}].$$

If $T < \infty$, then $\lim_{n\to\infty} X_{T\wedge n} = X_T$; actually $X_{T\wedge n} = X_T$ whenever $n > T$. *Thus, whenever we can justify the interchange of limit $n \to \infty$ and expectation*, we can deduce the important identity

$$E[X_0] = \lim_{n\to\infty} E[X_{T\wedge n}] \stackrel{?}{=} E[\lim_{n\to\infty} X_{T\wedge n}] = E[X_T]. \tag{3.1}$$

We will later offer several conditions where, indeed, this interchange is legitimate.

Lemma 3.1. *Let $\{X_n\}$ be a (super) martingale and T a Markov time with respect to $\{Y_n\}$. Then for all $n \geq k$,*

$$E[X_n I_{\{T=k\}}](\leq) = E[X_k I_{\{T=k\}}]. \tag{3.2}$$

Proof. By the law of total probability and (2.9),

$$\begin{aligned} E[X_n I_{\{T=k\}}] &= E\{E[X_n I_{\{T=k\}}(Y_0, \ldots, Y_k)|\, Y_0, \ldots, Y_k]\} \\ &= E\{I_{\{T=k\}} E[X_n|\, Y_0, \ldots, Y_k]\} \\ (\leq) &= E\{I_{\{T=k\}} X_k\}. \quad \blacksquare \end{aligned}$$

Lemma 3.2. *If $\{X_n\}$ is a (super) martingale and T a Markov time, then for all $n = 1, 2, \ldots$*

$$E[X_0](\geq) = E[X_{T\wedge n}](\geq) = E[X_n]. \tag{3.3}$$

Proof. Using Lemma 3.1,

$$\begin{aligned} E[X_{T\wedge n}] &= \sum_{k=0}^{n-1} E[X_T I_{\{T=k\}}] + E[X_n I_{\{T\geq n\}}] \\ &= \sum_{k=0}^{n-1} E[X_k I_{\{T=k\}}] + E[X_n I_{\{T\geq n\}}] \qquad [X_T = X_k \text{ when } T = k] \\ (\geq) &= \sum_{k=0}^{n-1} E[X_n I_{\{T=k\}}] + E[X_n I_{\{T\geq n\}}] \qquad [\text{on the basis of (3.2)}] \\ &= E[X_n]. \end{aligned}$$

* Referred to also as the *optional stopping theorem*.

For a martingale, $E[X_n] = E[X_0]$, which completes the proof in this case. For a supermartingale, we have shown $E[X_{T\wedge n}] \geq E[X_n]$ and have yet to establish $E[X_0] \geq E[X_{T\wedge n}]$. We will do this assuming $E[|X_n|] < \infty$, for all n. The general case may be obtained by truncation, along the lines suggested in the remark following Lemma 2.2.

The sequence defined by

$$\tilde{X}_n = \sum_{k=1}^{n} \{X_k - E[X_k | Y_0, \ldots, Y_{k-1}]\}$$

is a martingale ($\tilde{X}_0 = 0$) [cf. Example (b) of Section 1]. Thus,

$$\begin{aligned} 0 &= E[\tilde{X}_{T\wedge n}] \\ &= E\left[\sum_{k=1}^{T\wedge n} \{X_k - E[X_k | Y_0, \ldots, Y_{k-1}]\}\right] \\ &\geq E\left[\sum_{k=1}^{T\wedge n} \{X_k - X_{k-1}\}\right] = E[X_{T\wedge n}] - E[X_0], \end{aligned}$$

and consequently

$$E[X_0] \geq E[X_{T\wedge n}],$$

which completes the proof. ■

We return to the matter of justifying an interchange of limit and expectation as $n \to \infty$ in (3.1). The most general conditions that guarantee this operation are that the random variables $\{X_{T\wedge n}\}$ be *uniformly integrable* in the sense that

$$\lim_{a\to\infty} \sup_{n\geq 0} E[|X^a_{T\wedge n}|] = 0,$$

where

$$X^a_{T\wedge n} = \begin{cases} 0, & \text{if } |X_{T\wedge n}| < a, \\ X_{T\wedge n}, & \text{if } |X_{T\wedge n}| \geq a, \end{cases}$$

and that $T < \infty$. We skip to conditions not quite so general, but still covering a number of cases of importance.

Lemma 3.3. *Let W be an arbitrary random variable satisfying $E[|W|] < \infty$, and let T be a Markov time for which $\Pr\{T < \infty\} = 1$. Then*

$$\lim_{n\to\infty} E[WI_{\{T>n\}}] = 0, \tag{3.4}$$

and

$$\lim_{n\to\infty} E[WI_{\{T\leq n\}}] = E[W]. \tag{3.5}$$

Proof. We reduce the problem to one involving elementary convergence properties of series having nonnegative terms. First,

$$\begin{aligned} E[|W|] &\geq E[|W| I_{\{T \leq n\}}] \\ &= \sum_{k=0}^{n} E[|W| \,\Big|\, T = k] \Pr\{T = k\} \qquad \text{(law of total probabilities)} \\ &\xrightarrow[n \to \infty]{} \sum_{k=0}^{\infty} E[|W| \,\Big|\, T = k] \Pr\{T = k\} \\ &= E[|W|]. \end{aligned}$$

Hence

$$\lim_{n \to \infty} E[|W| I_{\{T \leq n\}}] = E[|W|],$$

and

$$\lim_{n \to \infty} E[|W| I_{\{T > n\}}] = 0.$$

Next, observe

$$\begin{aligned} 0 &\leq |E[W] - E[W I_{\{T \leq n\}}]| \\ &= |E[W I_{\{T > n\}}]| \\ &\leq E[|W| I_{\{T > n\}}] \to 0, \end{aligned}$$

which completes the proof. ∎

The following theorem, the optional sampling theorem for dominated martingales, is a direct consequence.

Theorem 3.1. *Suppose* $\{X_n\}$ *is a martingale and* T *is a Markov time. If* $\Pr\{T < \infty\} = 1$ *and* $E[\sup_{n \geq 0} |X_{T \wedge n}|] < \infty$, *then*

$$E[X_T] = E[X_0].$$

Proof. Set $W = \sup_{n \geq 0} |X_{T \wedge n}|$. Starting with the decomposition

$$X_T = \sum_{k=0}^{\infty} X_k \, I_{\{T = k\}} = \sum_{k=0}^{\infty} X_{T \wedge k} \, I_{\{T = k\}},$$

valid by virtue of the assumption $\Pr\{T < \infty\} = 1$, we find that $|X_T| \leq W$, and therefore $E[|X_T|] \leq E[W] < \infty$, so that the expectation of X_T is defined. We need only show $\lim_{n \to \infty} E[X_{T \wedge n}] = E[X_T]$. We have

$$\begin{aligned} |E[X_{T \wedge n}] - E[X_T]| &\leq E[|(X_{T \wedge n} - X_T)| I_{\{T > n\}}] \\ &\leq 2E[W I_{\{T > n\}}]. \end{aligned}$$

But $\lim_{n\to\infty} E[WI_{\{T>n\}}] = 0$ by Lemma 3.3. With reference to (3.3) the proof is complete. ■

Corollary 3.1. *Suppose $\{X_n\}$ is a martingale and T is a Markov time with respect to $\{Y_n\}$. If*

(i) $$E[T] < \infty,$$

and there exists a constant $K < \infty$, for which

(ii) $$E[|X_{n+1} - X_n| \,|\, Y_0, \ldots, Y_n] \leq K, \qquad \text{for} \quad n < T,$$

then $E[X_T] = E[X_0]$.

Proof. Define $Z_0 = |X_0|$ and $Z_n = |X_n - X_{n-1}|$, $n = 1, 2, \ldots$, and set

$$W = Z_0 + \cdots + Z_T.$$

Then $W \geq |X_T|$, and

$$\begin{aligned} E[W] &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} E[Z_k I_{\{T=n\}}] \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} E[Z_k I_{\{T=n\}}] \\ &= \sum_{k=0}^{\infty} E[Z_k I_{\{T\geq k\}}]. \end{aligned}$$

Observe that $I_{\{T\geq k\}} = 1 - I_{\{T\leq k-1\}}$ is a function only of $\{Y_0, \ldots, Y_{k-1}\}$, and from (ii) the inequalities $E[Z_k | Y_0, \ldots, Y_{k-1}] \leq K$ hold if $k \leq T$. Hence

$$\begin{aligned} \sum_{k=0}^{\infty} E[Z_k I_{\{T\geq k\}}] &= \sum_{k=0}^{\infty} E\{E[Z_k I_{\{T\geq k\}} | Y_0, \ldots, Y_{k-1}]\} \\ &= \sum_{k=0}^{\infty} E\{I_{\{T\geq k\}} E[Z_k | Y_0, \ldots, Y_{k-1}]\} \\ &\leq K \sum_{k=0}^{\infty} \Pr\{T \geq k\} \\ &\leq K(1 + E[T]) < \infty. \end{aligned}$$

Thus $E[W] < \infty$. Since $|X_{T\wedge n}| \leq W$ for all n by the definition of W, the result follows from Theorem 3.1. ■

Theroem 3.2. *(Optional Stopping Theorem). Let* $\{X_n\}$ *be a martingale and* T *a Markov time. If*

$$\text{(i)} \qquad \Pr\{T < \infty\} = 1,$$

$$\text{(ii)} \qquad E[|X_T|] < \infty,$$

$$\text{(iii)} \qquad \lim_{n\to\infty} E[X_n I_{\{T>n\}}] = 0,$$

then $E[X_T] = E[X_0]$.

Proof. We emphasize that (ii) must be assumed and is *not* a consequence of $E[|X_n|] < \infty$ for all n.

We write, for all n,

$$\begin{aligned} E[X_T] &= E[X_T I_{\{T\le n\}}] + E[X_T I_{\{T>n\}}] \\ &= E[X_{T\wedge n}] - E[X_n I_{\{T>n\}}] + E[X_T I_{\{T>n\}}]. \end{aligned}$$

Now $E[X_{T\wedge n}] = E[X_0]$ by Lemma 3.1, and $\lim_{n\to\infty} E[X_n I_{\{T>n\}}] = 0$ by assumption (iii). Lastly, we use Lemma 3.3 with $W = X_T$ and assumption (ii) to infer $\lim_{n\to\infty} E[X_T I_{\{T>n\}}] = 0$. Thus

$$E[X_T] = \lim_{n\to\infty} E[X_{T\wedge n}] = E[X_0],$$

as was to be shown. ■

Here are some sample consequences of this fundamental theorem.

Corollary 3.2. *Suppose* $\{X_n\}$ *is a martingale and* T *is a Markov time. If*

$$\text{(i)} \quad \Pr\{T < \infty\} = 1,$$

and for some $K < \infty$,

$$\text{(ii)} \quad E[X_{T\wedge n}^2] \le K, \quad \textit{for all } n,$$

then $E[X_T] = E[X_0]$.

Proof. Since $X_{T\wedge n}^2 \ge 0$, condition (ii) implies

$$\begin{aligned} K &\ge E[X_{T\wedge n}^2 I_{\{T\le n\}}] \\ &= \sum_{k=0}^{n} E[X_T^2 | T = k] \Pr\{T = k\} \\ &\xrightarrow[n\to\infty]{} \sum_{k=0}^{\infty} E[X_T^2 | T = k] \Pr\{T = k\} \\ &= E[X_T^2]. \end{aligned}$$

It follows from Schwarz' inequality that $E[|X_T|] \leq (E[X_T^2])^{1/2} < \infty$, which verifies condition (ii) of Theorem 3.2. For (iii) we use Schwarz' inequality again to conclude that

$$\begin{aligned}\{E[X_n I_{\{T>n\}}]\}^2 &= \{E[X_{T\wedge n} I_{\{T>n\}}]\}^2 \\ &\leq E[X_{T\wedge n}^2]E[I_{\{T>n\}}^2] \\ &\leq \mathrm{K}\ \Pr\{T>n\} \to 0, \qquad \text{as} \quad n \to \infty. \quad \blacksquare\end{aligned}$$

Corollary 3.3. *Suppose $Y_0 = 0$ and $Y_1, Y_2, \ldots$ are independent identically distributed random variables for which $E[Y_k] = \mu$ and $\mathrm{Var}[Y_k] = \sigma^2 < \infty$. Set $X_n = S_n - n\mu$, where $S_n = Y_0 + \cdots + Y_n$. If T is a Markov time for which $E[T] < \infty$, then $E[|X_T|] < \infty$ and $E[X_T] = E[S_T] - \mu E[T] = 0$.*

Proof. We apply the theorem to the martingale $\{S_n - n\mu\}$. Let $Y_0' = 0$ and $Y_k' = Y_k - \mu$, $k = 1, 2, \ldots$. To show $E[|X_T|] < \infty$, we have

$$\begin{aligned}E[|X_T|] &\leq E\left[\sum_{k=1}^{T} |Y_k'|\right] \\ &= E\left[\sum_{n=1}^{\infty}\sum_{k=1}^{n} |Y_k'| I_{\{T=n\}}\right] \\ &= E\left[\sum_{k=1}^{\infty} |Y_k'| I_{\{T\geq k\}}\right].\end{aligned}$$

Now $I_{\{T\geq k\}} = I_{\{T>k-1\}}$ depends only on $\{Y_0, \ldots, Y_{k-1}\}$ and thus is independent of Y_k'. Hence

$$\begin{aligned}E\left[\sum_{k=1}^{\infty} |Y_k'| I_{\{T\geq k\}}\right] &= E[|Y_1'|]\sum_{k=1}^{\infty} \Pr\{T \geq k\} \\ &= E[|Y_1'|]E[T] < \infty.\end{aligned}$$

To confirm condition (iii) of Theorem 3.2, we have, using Schwarz' inequality,

$$\begin{aligned}(E[X_n I_{\{T>n\}}])^2 &\leq E[X_n^2]E[I_{\{T>n\}}] \\ &\leq n\sigma^2 \Pr\{T \geq n\}.\end{aligned}$$

But $\infty > E[T] = \sum_{k=0}^{\infty} k \Pr\{T = k\}$, so that

$$\begin{aligned}0 &= \lim_{n\to\infty} \sum_{k\geq n} k \Pr\{T = k\} \\ &\geq \lim_{n\to\infty} n \Pr\{T \geq n\} \geq 0. \quad \blacksquare\end{aligned}$$

4: *Some Applications of the Optional Sampling Theorem*

As we shall see, the optional sampling theorem finds ready use in computing and bounding certain probabilities connected with stochastic processes. More applications relevant to Brownian motion appear in Chapter 7, Section 5.

(a) *Random Walks*

The optional sampling theorem quickly yields a number of important results in connection with random walks. First let us examine one of the examples that opened this chapter. Let $Y_0 = 0$ and, for $i = 1, 2, \ldots$, let Y_i be independent identically distributed random variables with $\Pr\{Y_i = 1\} = p$ and $\Pr\{Y_i = -1\} = q = 1 - p$. Let $S_0 = 0$ and $S_n = Y_1 + \cdots + Y_n$, $n \geq 1$.

Suppose first that $p = q = \frac{1}{2}$. Then $\{S_n\}$ is a martingale. If $T = \min\{n : S_n = 1\}$, then $\Pr\{T < \infty\} = 1$, since S_n is recurrent. But $S_T \equiv 1$, so $E[S_T] \neq E[S_0] = 0$, which contradicts the result of Corollary 3.3. Thus the hypothesis of this corollary cannot hold, and in particular $E[T] = \infty$.

Continue assuming $p = q = \frac{1}{2}$, but now let

$$T = \min\{n : S_n = -a \quad \text{or} \quad S_n = b\} \qquad (a, b \text{ positive integers}).$$

Let v_a be the probability that S_n reaches $-a$ before reaching b. Then, from Theorem 3.1,

$$0 = E[S_T] = v_a(-a) + (1 - v_a)b$$

or

$$v_a = \frac{b}{a + b},$$

which was determined by other means in Chapter 3.

$Z_n = S_n^2 - n$ also defines a martingale, and

$$E[Z_T] = 0 = [v_a a^2 + (1 - v_a)b^2] - E[T],$$

which reduces to give

$$E[T] = ab.$$

Now suppose $p > q$, and set $\mu = E[Y_k] = p - q > 0$. Then

$$X_n = S_n - n\mu, \tag{4.1}$$

and

$$X_n' = (q/p)^{S_n}, \tag{4.2}$$

are martingales. From (4.1) and Corollary 3.1, we extract the identity

$$E[S_T] = \mu E[T],$$

applicable for any Markov time T satisfying $E[T] < \infty$.

To use (4.2) let

$$T = \min\{n : S_n = -a \quad \text{or} \quad S_n = b\}.$$

Then

$$1 = E[X_T'] = v_a\left(\frac{q}{p}\right)^{-a} + (1 - v_a)\left(\frac{q}{p}\right)^{b},$$

or

$$v_a = \frac{1 - (q/p)^b}{(q/p)^{-a} - (q/p)^b},$$

where, as before, v_a is the probability that S_n reaches $-a$ before b. Again, this agrees with a formula derived in Chapter 3 through other means.

(b) *Wald's Identity*

Let $Y_0 = 0$ and $Y_1, Y_2, \ldots$ be nondegenerate independent identically distributed random variables having the *moment generating function*

$$\phi(\theta) = E[\exp\{\theta Y_1\}],$$

defined and finite for θ in some open interval containing the origin. Set $S_0 = 0$ and $S_n = Y_1 + \cdots + Y_n$. Finally, fix values $-a < 0$ and $b > 0$ and set

$$T = \min\{n : S_n \leq -a \quad \text{or} \quad S_n \geq b\}.$$

The fundamental identity of Wald is

$$E[\phi(\theta)^{-T} \exp\{\theta S_T\}] = 1, \tag{4.3}$$

valid for any θ satisfying $\phi(\theta) \geq 1$. This identity bears numerous applications throughout applied probability and statistics.

We will use Corollary 3.1 to establish (4.3). Recall from Example (g) of Section 1 that $X_0 = 1$, and

$$X_n = \phi(\theta)^{-n} \exp\{\theta S_n\}, \qquad n \geq 1,$$

defines a martingale with respect to $\{Y_n\}$. Then

$$E[|X_{n+1} - X_n| \,|\, Y_0, \ldots, Y_n] = X_n E[|\phi(\theta)^{-1} \exp(\theta Y_{n+1}) - 1|].$$

By assumption, $\phi(\theta)^{-n} \leq 1$, and, for $n < T$, $\exp\{\theta S_n\} \leq e^b$. Thus $X_n \leq e^b$ for $n < T$. In addition,

$$E[|\phi(\theta)^{-1} \exp(\theta Y_{n+1}) - 1|] \leq \phi(\theta)^{-1}\{E[\exp \theta Y_{n+1}] + \phi(\theta)\} = 2.$$

Thus

$$E[|X_{n+1} - X_n| \, | \, Y_0, \ldots, Y_n] \leq 2e^b, \qquad \text{for} \quad n < T,$$

and we need only verify that $E[T] < \infty$ in order to apply Corollary 3.1. Let $c = a + b$. Since Y_k is stipulated nondegenerate, there exists an integer N and a $\delta > 0$ such that $\Pr\{|S_N| > c\} > \delta$. Define $S_1' = S_N$, $S_2' = S_{2N} - S_N$, ..., and $S_k' = S_{kN} - S_{\{k-1\}N}$. Then

$$\begin{aligned} \Pr\{T \geq kN\} &\leq \Pr\{|S_1'| \leq c\} \cdots \Pr\{|S_k'| \leq c\} \\ &\leq (1 - \delta)^k, \end{aligned}$$

and also bringing in the fact that $\Pr\{T \geq n\}$ decreases in n, we secure

$$\begin{aligned} E[T] &= \sum_{n=1}^{\infty} \Pr\{T \geq n\} \\ &\leq N \sum_{k=0}^{\infty} \Pr\{T \geq kN\} \leq N/\delta < \infty. \end{aligned}$$

To see how Wald's identity is commonly applied, let us suppose there exists a value $\theta_0 \neq 0$ for which $\phi(\theta_0) = 1$. Then (4.3) becomes

$$E[\exp\{\theta_0 S_T\}] = 1.$$

Setting

$$E_a = E[\exp\{\theta_0 S_T\} | S_T \leq -a],$$

and

$$E_b = E[\exp\{\theta_0 S_T\} | S_T \geq b],$$

we conclude that

$$\begin{aligned} 1 &= E_a \Pr\{S_T \leq -a\} + E_b \Pr\{S_T \geq b\} \\ &= E_a + (E_b - E_a) \Pr\{S_T \geq b\}, \end{aligned}$$

or

$$\Pr\{S_T \geq b\} = \frac{1 - E_a}{E_b - E_a}.$$

One might expect $E_a \cong \exp\{-\theta_0 a\}$ and $E_b \cong \exp\{\theta_0 b\}$, provided that when S_n leaves the interval $[-a, b]$, it does so without jumping too far from the boundary. This is the intuition underlying Wald's approximation

$$\Pr\{S_T \geq b\} \cong \frac{1 - \exp\{-\theta_0 a\}}{\exp\{\theta_0 b\} - \exp\{-\theta_0 a\}}.$$

Return to identity (4.3) and formally differentiate it with respect to θ to obtain

$$0 = \frac{d}{d\theta} E[\phi(\theta)^{-T} \exp\{\theta S_T\}]$$

$$= E[(-T\phi(\theta)^{-T-1}\phi'(\theta) + \phi(\theta)^{-T} S_T) \exp\{\theta S_T\}]$$

$$= -\phi'(\theta) E[T\phi(\theta)^{-T-1} \exp\{\theta S_T\}] + E[\phi(\theta)^{-T} S_T \exp\{\theta S_T\}].$$

Set $\theta = 0$, using $\phi(0) = 1$ and $\phi'(0) = E[Y_1]$. Then

$$0 = -E[Y_1]E[T] + E[S_T],$$

or

$$E[S_T] = E[Y_1]E[T].$$

OPTIONAL STOPPING FOR SUPERMARTINGALES

Let $\{X_n\}$ be a supermartingale with respect to $\{Y_n\}$. According to Lemma 3.2, $E[X_0] \geq E[X_{T \wedge n}]$ for any Markov time T, and we may infer $E[X_0] \geq E[X_T]$ provided we can justify the interchange of limit and expectation as $n \to \infty$, just as in the case with martingales. The following two theorems are important cases.

Theorem 4.1. *Let $\{X_n\}$ be a supermartingale and T a Markov time. If $\Pr\{T < \infty\} = 1$ and there exists a random variable $W \geq 0$ for which $E[W] < \infty$ and $X_{T \wedge n} > -W$ for all n, then*

$$E[X_0] \geq E[X_T].$$

Proof. Let $c > 0$ be fixed and define $X_n^c = \min\{c, X_n\}$, for $n = 0, 1, \ldots$. Then $\{X_n^c\}$ is also a supermartingale with respect to $\{Y_n\}$ (in this connection consult p. 250), so that

$$E[X_0^c] \geq E[X_{T \wedge n}^c],$$

and since

$$|X_{T \wedge n}^c| \leq \max\{c, W\},$$

for all n, we may interchange the limit as $n \to \infty$ and expectation as in Theorem 3.1, and deduce that

$$E[X_0^c] \geq E[X_T^c].$$

But clearly $E[X_0] \geq E[X_0^c]$, while

$$\lim_{c \to \infty} E[X_T^c] = \lim_{c \to \infty} \int_{-\infty}^{c} x \, d \Pr\{X_T \leq x\} = E[X_T].$$

Thus $E[X_0] \geq E[X_T]$ is achieved as forecast. ■

Theorem 4.2. *Let $\{X_n\}$ be a supermartingale and T a Markov time. If $X_n \geq 0$ for all n, then*

$$E[X_0] \geq E[X_T I_{\{T<\infty\}}].$$

Proof. As usual,

$$E[X_0] \geq E[X_{T\wedge n}].$$

Since $X_n \geq 0$, $X_{T\wedge n} = X_T I_{\{T\leq n\}} + X_n I_{\{T>n\}} \geq X_T I_{\{T\leq n\}}$, so that $E[X_0] \geq E[X_T I_{\{T\leq n\}}]$, and

$$\begin{aligned} E[X_0] &\geq \lim_{n\to\infty} E[X_T I_{\{T\leq n\}}] \\ &= \lim_{n\to\infty} \sum_{k=0}^{n} E[X_T | T=k] \Pr\{T=k\} \\ &= \sum_{k=0}^{\infty} E[X_T | T=k] \Pr\{T=k\} \\ &= E[X_T I_{\{T<\infty\}}]. \quad \blacksquare \end{aligned}$$

BOUNDS ON THE VALUE OF AN OPTION

Let W_n be the price on day n of an asset, say a share of stock, that is traded in a public market. Let Y_n be the ratio of the price on day n to the price on day $n-1$, so that $W_n = w \times Y_1 \times \cdots \times Y_n$, where $W_0 = w$ is today's price. In this context, the historically famous and controversial "random walk hypothesis" asserts that $Y_1, Y_2, \ldots$ are independent and identically distributed positive random variables. It can be shown that certain assumptions characterizing a "perfect market" lead to this behavior. Recently interest has centered in replacing the random walk assumption by a weaker assumption phrased in terms of martingales. Let us recognize a long-term growth and inflation rate $\alpha \geq 0$ and assume that $e^{-\alpha n} W_n$ is a martingale with respect to Y_n. Then $E[e^{-\alpha n} W_n] = E[W_0] = w$ or $E[W_n] = we^{\alpha n}$, so that α represents the mean growth rate of the market price of the asset.

Consider now an *option contract* that entitles the holder to purchase the asset at any time he pleases at a fixed stated price, regardless of what the market price might be. In the stock market such options are called "warrants" and "calls." By changing our scale of values, if necessary, we may suppose that the fixed stated price is one. Then if $W_n > 1$, the option holder may "exercise" his option, purchase at the stated price of

one, and resell in the market at W_n for a profit of $W_n - 1$. If $W_n \leq 1$, no such profit is possible. Thus the potential profit to the option holder is

$$r(W_n) = (W_n - 1)^+ = \max\{W_n - 1, 0\}.$$

An alternative to holding the option is to hold the asset itself, and the rate of return here, on the average, is α. Since this alternative is always available, one could justify holding the option only if it provided a greater rate of return $\beta > \alpha$. Equivalently, we discount the potential return $r(W_n)$ on day n by the factor $e^{-\beta n}$. Let T be the time the option is exercised. In this general martingale model we seek a bound on the mean discounted profit $E[e^{-\beta T} r(W_T)]$ as a function of the current price w, under the moment condition that there exists a $\theta > 1$ for which

$$E[Y_n^\theta | Y_1, \ldots, Y_{n-1}] \leq e^\beta, \qquad \text{for} \quad n = 1, 2, \ldots. \tag{4.4}$$

In fact, (4.4) is the sole assumption needed for what follows. Our results do not depend on either the random walk model or the martingale model for market prices but are compatible with both these models. Note, we interpret $e^{-\beta T} r(W_T) = 0$ if $T = \infty$ so that no profit is made if the option is never exercised.

Example. Suppose $Y_1, Y_2, \ldots$ are independent identically distributed log normal random variables, i.e., $V_k = \ln Y_k$ is normally distributed with mean μ and variance σ^2. Then

$$E[Y_n | Y_1, \ldots, Y_{n-1}] = E[Y_n] = E[\exp\{V_n\}] = \exp\{\mu + \tfrac{1}{2}\sigma^2\},$$

so that $\alpha = \mu + \frac{1}{2}\sigma^2$, while

$$E[Y_n^\theta | Y_1, \ldots, Y_{n-1}] = E[\exp\{\theta V_n\}] = \exp\{\theta\mu + \tfrac{1}{2}\theta^2\sigma^2\}.$$

We solve $\theta\mu + \frac{1}{2}\theta^2\sigma^2 = \beta$ to see that (4.4) holds as an equality for

$$\theta = \frac{(\mu^2 + 2\sigma^2\beta)^{1/2} - \mu}{\sigma^2}.$$

Since $\beta > \alpha = \mu + \frac{1}{2}\sigma^2$,

$$\theta > \frac{(\mu^2 + 2\sigma^2\alpha)^{1/2} - \mu}{\sigma^2} = 1,$$

as required.

To return to the general case, we will show that $E[e^{-\beta T} r(W_T)] \leq f(w)$ for all $w \geq 0$, where

$$f(w) = \begin{cases} w^\theta(\theta - 1)^{\theta - 1}/\theta^\theta, & \text{if} \quad w \leq \theta/(\theta - 1), \\ w - 1, & \text{if} \quad w > \theta/(\theta - 1). \end{cases}$$

Notice that $f(w) \leq w^\theta(\theta-1)^{\theta-1}/\theta^\theta$ for all $w > 0$.
This bound holds no matter what strategy the option owner uses in deciding when, if ever, to exercise his option. Thus the bound might be used by sellers of options to limit their mean loss. Note also that if today's price $W_0 = w$ exceeds $\theta/(\theta-1)$, then the value of the option is at most $f(w) = w - 1$, the amount that could be obtained by exercising immediately, and thus the option should be exercised if the market price is this high. This conclusion requires only the moment assumption (4.4) and subject to this, holds regardless of the form of the probability distribution for daily price changes!

To verify the bound, set $X_n = e^{-\beta n} f(W_n)$. We claim $\{X_n\}$ is a nonnegative supermartingale with respect to $\{Y_n\}$. It suffices to show

$$f(w) \geq e^{-\beta} E[f(w \times Y)], \qquad \text{for all} \quad w \geq 0, \tag{4.5}$$

whenever

$$E[Y^\theta] \leq e^\beta, \tag{4.6}$$

since once (4.5) is established then using (4.4) we obtain

$$\begin{aligned} E[X_n | Y_0, \ldots, Y_{n-1}] &= e^{-\beta n} E[f(W_{n-1} \times Y_n) | Y_0, \ldots, Y_{n-1}] \\ &\leq e^{-\beta(n-1)} f(W_{n-1}) = X_{n-1}. \end{aligned}$$

It remains to prove (4.5). For a fixed $t > 1$, define

$$v(t, w) = w^t(t-1)^{t-1}/t^t, \qquad w \geq 0.$$

Then

$$v(t, w) \geq f(w), \quad 1 < t \leq \theta, \quad w \geq 0. \tag{4.7}$$

We leave to the reader the exercise of verifying (4.7). [With $g_t(w) = v(t, w) - (w-1)$ and $w_0 = t/(t-1) > \theta/(\theta-1)$ he should first check that both $g_t(w_0) = 0$ and $g_t'(w_0) = 0$, where prime denotes the derivative with respect to w. Then, after verifying the positive second derivative $g_t''(w) > 0$ he knows $g_t(w) \geq 0$ or $v(t, w) \geq w - 1$, for all w. Second, he should compute the derivative of $\log v(t, w)$ with respect to t and simplify it to get

$$\frac{d \log v(t, w)}{dt} = \log\left\{\frac{w}{t/(t-1)}\right\}$$

which is negative when $w < \theta/(\theta-1) < t/(t-1)$. That is $v(t, w)$ increases as t decreases from θ for $w \leq \theta/(\theta-1)$. Therefore $v(t, w) > v(\theta, w) = f(w)$ for w in this region.]

To continue we next consider two cases. Suppose first that $w \leq \theta/(\theta - 1)$. Then

$$\begin{aligned} f(w) &= v(\theta, w) \\ &\geq e^{-\beta} v(\theta, w) E[Y^\theta] && \text{[because of (4.6)]} \\ &= e^{-\beta} E[v(\theta, w \times Y)] && \text{[by the definition of } v(\theta, w)\text{]} \qquad (*) \\ &\geq e^{-\beta} E[f(w \times Y)] && \text{[by (4.7)].} \end{aligned}$$

The second case is $w > \theta/(\theta - 1)$. Now $E[Y^a]$ is a convex function of $a \in [0, \theta]$, and $E[Y^0] = 1 < e^\beta$. It follows using Jensen's inequality and the hypothesis (4.6) that $E[Y^{w/(w-1)}] \leq e^\beta$ for $w/(w-1) < \theta$. Now consider

$$\begin{aligned} f(w) &= w - 1 \\ &= v(w/(w-1), w) \\ &\geq e^{-\beta} E[v(w/(w-1), w \times Y)] && \text{[since } E(Y^{w/w-1}) \\ & && \leq e^\beta\text{]} \qquad (**) \\ &\geq e^{-\beta} E[f(w \times Y)] && \text{[by (4.7)].} \end{aligned}$$

The deliberations of (*) and (**) verify (4.5). Thus $X_n = e^{-\beta n} f(W_n)$ forms a nonnegative supermartingale, and so

$$X_0 = f(w) \geq E[e^{-\beta T} f(W_T)], \qquad (4.8)$$

for any Markov time T. Lastly, we check that $f(w) \geq r(w) = (w-1)^+$, and then (4.8) will imply

$$f(w) \geq E[e^{-\beta T} r(W_T)],$$

as asserted earlier. Manifestly, $f(w) \geq 0$ for all w, and $f(w) = w - 1$ for $w \geq \theta/(\theta - 1)$. Furthermore,

$$df/dw = [(\theta - 1)/\theta]^{\theta - 1} w^{\theta - 1} < 1, \qquad \text{for} \quad w < \theta/(\theta - 1).$$

It follows by integration that

$$f\left(\frac{\theta}{\theta - 1}\right) - f(w) < \frac{\theta}{\theta - 1} - w,$$

and therefore $f(w) > w - 1$ for $w < \theta/(\theta - 1)$. This completes the verification of the inequality $f(w) \geq (w-1)^+$.

BOUNDS IN A RESERVOIR MODEL

Let Z_t be the water level at time t in a dam of finite capacity b. Let I_t be the random inflow in the time interval $(t, t+1]$ and O_t the outflow. The balance equation

$$Z_{t+1} = \min\{(Z_t + I_t - O_t)^+, b\}$$

expresses the fact that the water level cannot be negative nor can it exceed the capacity b. (We use "x^+" to denote "$\max\{x, 0\}$".)

Let us suppose that it is desirable for navigation, recreation, or emergency supply purposes always to maintain a water level above some minimal acceptable level a. Then

$$T = \min\{t : Z_t \leq a\}$$

is the first time that this requirement is not met. The demands $\{O_t\}$ and the capacity b are controllable or design parameters. For various values of these parameters the performance of the dam might be compared using the mean time $E[T]$ as a criterion, the longer mean times being the better.

A common approach is to make exact assumptions concerning the inflows and outflows and then to compute $E[T]$ exactly. However, it is often difficult to justify exact assumptions for water reservoir systems because of seasonal effects and upstream surface water storage, which may affect inflows for several successive time periods. It is also true that little information is generally available concerning the distribution of inflows. Moreover, where past information does exist, it often bears little relevance to present conditions because of the topographical changes that are constantly occurring in any watershed system.

In this example, we will obtain a bound on the mean time $E[T]$ under the rather weak general assumption that the conditional distribution of the net inflow

$$Y_{t+1} = I_{t+1} - O_{t+1}$$

given the past satisfies

$$E[Y_{t+1} | Y_1, \ldots, Y_t] \leq m, \tag{4.9}$$

and

$$E[\exp\{-\lambda Y_{t+1}\} | Y_1, \ldots, Y_t] \leq 1 \tag{4.10}$$

where m and λ are known positive constants. We then show $E[T] \geq f(z)$, where

$$f(z) = \frac{1}{m}\left\{e^{\lambda(b-a)} \frac{1 - e^{-\lambda(z-a)}}{\lambda} - (z-a)\right\}, \qquad \text{for} \quad a \leq z \leq b,$$

and $Z_0 = z$ is the initial dam content. A conservative designer might plan with $f(z)$ as his criterion, since this represents the worst case or earliest mean time in which the critical level a could be reached.

We claim

$$X_n = f(Z_n) + n$$

forms a submartingale with respect to $\{Y_n\}$. Extend the domain of f by setting $f(z) = 0$ for $z < a$, and $f(z) = f(b)$ for $z > b$. Then $f(Z_{n+1}) = f(Z_n + Y_{n+1})$. Set

$$g(z) = \frac{1}{m}\left\{e^{\lambda(b-a)}\frac{1-e^{-\lambda(z-a)}}{\lambda} - (z-a)\right\}, \qquad \text{for all} \quad z.$$

Since $g(z)$ is increasing up to $z = b$ and decreasing thereafter, we find that $f(z) \geq g(z)$ for all z, and

$$\begin{aligned} E[X_{n+1}|Y_0, \ldots, Y_n] &= E[f(Z_{n+1})|Y_0, \ldots, Y_n)] + (n+1) \\ &= E[f(Z_n + Y_{n+1})|Y_0, \ldots, Y_n] + (n+1) \\ &\geq E[g(Z_n + Y_{n+1})|Y_0, \ldots, Y_n] + (n+1). \end{aligned}$$

Now, if U is a random variable for which $E[U] \leq m$ and $E[e^{-\lambda U}] \leq 1$, then

$$\begin{aligned} E[g(z+U)] &= \frac{1}{m}\left\{e^{\lambda(b-a)}\frac{1-E[e^{-\lambda(z+U-a)}]}{\lambda} - (z+E[U]-a)\right\} \\ &\geq f(z) - 1, \qquad \text{for} \quad a \leq z \leq b. \end{aligned}$$

By virtue of the foregoing analyses and in view of the stipulations of (4.9) and (4.10), we infer that

$$E[X_{n+1}|Y_0, \ldots, Y_n] \geq f(Z_n) + n = X_n,$$

establishing that X_n is a submartingale as claimed. Now let $T = \min\{n : Z_n \leq a\}$ be the first time the water level ever reaches the critical height a. From the optional stopping theorem for submartingales, we have

$$E[f(Z_{T\wedge n}) + (T \wedge n)] \geq f(z),$$

where $z = Z_0$ is the initial dam content. Since f is a bounded function, we may appeal to Lemma 3.3, validating the results

$$\lim_{n\to\infty} E[f(Z_{T\wedge n})] = E[f(Z_T)] = 0, \qquad (\text{since } Z_T \leq a)$$

and

$$\begin{aligned} \lim_{n\to\infty} E[T \wedge n] &= \lim_{n\to\infty} \sum_{k=1}^{n} \Pr\{T \geq k\} \\ &= \sum_{k=1}^{\infty} \Pr\{T \geq k\} = E[T]. \end{aligned}$$

Hence

$$\begin{aligned} f(z) &\leq \lim_{n\to\infty} \{E[f(Z_{T\wedge n})] + E[T \wedge n]\} \\ &= E[T], \end{aligned}$$

as desired to be shown.

The conditions (4.9) and (4.10) are certainly satisfied where $Y_1, Y_2, \ldots, Y_i, \ldots$ are independent normal random variables having means $\mu_i \leq m$ and variances $\sigma_i^2 \leq 2\mu_i/\lambda$.

THE CROSSINGS INEQUALITY

Given a submartingale $\{X_n\}$ with respect to a sequence $\{Y_n\}$, real numbers $a < b$, and a positive integer N, define $V_{a,b}$ to be the number of pairs (i, j) with $0 \leq i < j \leq N$, for which $X_i \leq a$, $a < X_k < b$, for $i < k < j$ and $X_j \geq b$. Then $V_{a,b}$ counts the number of times X_n upcrosses the interval (a, b) for $n = 0, 1, \ldots, N$, i.e., the number of crosses from a level below a to a level above b. We will prove the fundamental *upcrossings inequality*

$$E[V_{a,b}] \leq \frac{E[(X_N - a)^+] - E[(X_0 - a)^+]}{b - a}. \tag{4.11}$$

The upcrossings inequality indicates limits on the oscillations permissible for a submartingale and suggests that the paths or sample trajectories behave rather regularly. The inequality, its extensions and variations are used widely in probability analysis to prove convergence theorems and to investigate the growth and continuity properties of sample paths for continuous parameter stochastic processes.

We will need the following extension of Lemma 3.2, which covers two Markov times simultaneously in the submartingale case. This is a special circumstance of a more general optional sampling theorem that compares several, or even denumerably many, Markov times.

Lemma 4.1. *Let $\{X_n\}$ be a submartingale and S, T Markov times with respect to $\{Y_n\}$. Suppose $0 \leq S \leq T \leq N$, where N is a fixed positive integer. Then*

$$E[X_S] \leq E[X_T].$$

Proof. Let k be fixed. For $k \leq n \leq N$, since $I_{\{T>n\}}$ depends only on $Y_0, \ldots, Y_n$, using obvious properties of conditional probabilities we have

$$\begin{aligned} &E[X_{n+1} I_{\{T>n\}} I_{\{S=k\}}] \\ &\quad = E[E\{X_{n+1} | Y_0, \ldots, Y_n\} I_{\{T>n\}} I_{\{S=k\}}] \\ &\quad \geq E[X_n I_{\{T>n\}} I_{\{S=k\}}]. \end{aligned}$$

Thus

$$\begin{aligned} E[X_{T\wedge n} I_{\{S=k\}}] &= E[X_T I_{\{T\leq n\}} I_{\{S=k\}}] + E[X_n I_{\{T>n\}} I_{\{S=k\}}] \\ &\leq E[X_T I_{\{T\leq n\}} I_{\{S=k\}}] + E[X_{n+1} I_{\{T>n\}} I_{\{S=k\}}] \\ &= E[X_{T\wedge(n+1)} I_{\{S=k\}}] \end{aligned}$$

is a monotonic increasing function of n. Setting $n = k$ and then $n = N$, using the hypothesis $S \leq T \leq N$ leads to the relation

$$E[X_k I_{\{S=k\}}] \leq E[X_T I_{\{S=k\}}].$$

Now

$$\begin{aligned} E[X_S] &= \sum_{k=0}^{N} E[X_S I_{\{S=k\}}] \\ &= \sum_{k=0}^{N} E[X_k I_{\{S=k\}}] \\ &\leq \sum_{k=0}^{N} E[X_T I_{\{S=k\}}] \\ &= E[X_T], \end{aligned}$$

as required to be shown. ■

We apply the lemma to obtain the upcrossings inequality. Define

$$\hat{X}_n = (X_n - a)^+.$$

Since $g(x) = (x-a)^+ = \max\{(x-a), 0\}$ is a convex increasing function of x, Lemma 2.2 tells us that $\{\hat{X}_n\}$ is also a submartingale with respect to $\{Y_n\}$. Define $T_1 \equiv 0$, and, for $k = 1, \ldots, N$, set

$$T_k = \begin{cases} N, & \text{if } X_j \neq 0, \text{ for } j > T_{k-1}, \\ \min\{j : j > T_{k-1},\ X_j = 0\}, & \text{otherwise,} \end{cases}$$

if k is even, and

$$T_k = \begin{cases} N, & \text{if } X_j < b - a, \text{ for all } j > T_{k-1}, \\ \min\{j : j > T_{k-1},\ X_j \geq b - a\}, & \text{otherwise,} \end{cases}$$

if k is odd. Set $T_{N+1} = N$. Then each T_k is a Markov time (validate this rigorously) and $T_k \leq T_{k+1}$, so that, using the lemma just obtained,

$$E[\hat{X}_{T_k}] \leq E[\hat{X}_{T_{k+1}}]. \tag{4.12}$$

Now

$$\begin{aligned} \hat{X}_N - \hat{X}_0 &= \sum_{k=1}^{N} (\hat{X}_{T_{k+1}} - \hat{X}_{T_k}) \\ &= \sum_{k=2,4,\ldots} (\hat{X}_{T_{k+1}} - \hat{X}_{T_k}) + \sum_{k=1,3,\ldots} (\hat{X}_{T_{k+1}} - \hat{X}_{T_k}). \end{aligned}$$

Now, if k is even, $\hat{X}_{T_{k+1}} - \hat{X}_{T_k}$ is nonzero only if an upcrossing occurs and then is at least $(b-a)$, while the expected value of the second sum is nonnegative by (4.12). Thus

$$E[\hat{X}_N - \hat{X}_0] \geq (b-a)E[V_{a,b}],$$

or

$$E[V_{a,b}] \leq \frac{E[(X_N - a)^+] - E[(X_0 - a)^+]}{b-a},$$

as was to be shown. ■

AN INEQUALITY FOR PARTIAL SUMS

Let $X_1, X_2, \ldots$ be random variables having finite conditional moments

$$M_k = E[X_k | X_1, \ldots, X_{k-1}],$$

and

$$V_k = E[(X_k - M_k)^2 | X_1, \ldots, X_{k-1}].$$

Let $S_n = X_1 + \cdots + X_n$ ($S_0 = 0$). Under the condition

$$M_k \leq -\alpha V_k, \qquad \text{for all} \quad k, \tag{4.13}$$

for some *fixed positive* α, we will show

$$\Pr\left\{\sup_{n \geq 0}[x + S_n] > l\right\} \leq \frac{1}{1 + \alpha(l-x)}, \qquad \text{for} \quad x < l. \tag{4.14}$$

For example, if the summands are independent and identically distributed with mean $\mu < 0$ and variance σ^2, then, by the law of large numbers, $S_n \to -\infty$ as $n \to \infty$ and $M = \max_{n \geq 0} S_n < \infty$ is defined (see Chapter 17). The inequality (4.14) with the choice $\alpha = |\mu|/\sigma^2$ then says

$$\Pr\{M > l\} \leq \sigma^2/[\sigma^2 + |\mu| l].$$

We return to the general case in which independence of the summands is not assumed. Define

$$f(z) = \begin{cases} \dfrac{1}{1 + \alpha(l-z)}, & \text{for} \quad z < l, \\ 1, & \text{for} \quad z \geq l. \end{cases}$$

Our program is to show that, subject to condition (4.13), $\{f(x + S_n)\}$ determines a nonnegative supermartingale. Assuming this fact for the

moment, the application of Theorem 4.2 yields

$$f(x) \geq E[f(x+S_T)I_{\{T<\infty\}}]$$
$$= \Pr\left\{\sup_{n\geq 0}[x+S_n] > l\right\},$$

when T is the first n, if any, for which $x+S_n > l$.

For later use, record the derivatives

$$f'(y) = \frac{\alpha}{[1+\alpha(l-y)]^2} = \alpha[f(y)]^2, \qquad \text{for} \quad y < l, \tag{4.15}$$

and

$$f''(y) = 2\alpha f(y)f'(y), \qquad \text{for} \quad y < l, \tag{4.16}$$

which implies, because $0 \leq f(y) < 1$,

$$f''(y) < 2\alpha f'(y), \qquad \text{for} \quad y < l. \tag{4.17}$$

Fix an arbitrary point $z < l$. Let $g(y)$ be a quadratic in y, tangent to $f(y)$ at $y = z$. Specifically, $g(y)$ has the form

$$g(y) = f(z) + f'(z)(y-z) + a(y-z)^2,$$

where a is a suitable constant. We also want $g(l) = f(l) = 1$, and accordingly select

$$a = \alpha f'(z)$$

This achieves

$$\begin{aligned} g(l) &= f(z) + f'(z)(l-z) + \alpha f'(z)(l-z)^2 \\ &= f(z)\{1+\alpha(l-z)[1+\alpha(l-z)]f(z)\} \qquad \text{[by (4.15)]} \\ &= f(z)\{1+\alpha(l-z)\} \\ &= 1, \end{aligned}$$

as desired.

Specified thus, we get

$$g(y) \geq f(y), \qquad \text{for all} \quad y. \tag{4.18}$$

The key to seeing this is contained in (4.17), whence

$$g''(z) = 2a = 2\alpha f'(z) > f''(z)$$

Since $g(z) = f(z)$ and $g'(z) = f'(z)$, it follows that (4.18) prevails in some neighborhood of z. Now (4.18) holds if and only if

$$h(y) = \frac{g(y)}{f(y)} - 1 \geq 0, \qquad \text{for all} \quad y \leq l$$

But $h(y)$ is a cubic and therefore admits at most three real roots, two of which are located at the point of tangency z, and the third at l. Since (4.18) holds in a neighborhood of z, the only possibility is $g(y) \geq f(y)$ for all $y \leq l$. It is easy to check that $g(y) \geq f(y) = 1$ for $y \geq l$. Thus (4.18) is established. With these preliminaries complete, we turn to the task of proving that $f(x + S_n)$ generates a supermartingale. Let X be an arbitrary random variable having a mean m and a variance v^2, which together satisfy

$$m \leq -\alpha v^2. \tag{4.19}$$

Let $\xi \leq l$ be an arbitrary point and apply the preceding analysis with $z = \xi + m \leq l$ [by (4.19) m is negative]. Using (4.18) for the first inequality,

$$\begin{aligned} E[f(\xi + X)] &= E[f(z + X - m)] \\ &\leq E[g(z + X - m)] \\ &= f(z) + \alpha f'(z)v^2 \\ &\leq f(z) - mf'(z) \qquad \text{[since } f'(z) > 0 \text{ by (4.15),} \\ &\qquad\qquad\qquad\quad \text{and then use (4.19)]} \\ &\leq f(z - m) \qquad \text{[since } f''(y) \geq 0 \text{ for } y < l\text{]} \\ &= f(\xi). \end{aligned}$$

Thus, $f(\xi) \geq E[f(\xi + X)]$ for any $\xi \leq l$ [it trivially holds for $\xi > l$ since $f(z) \leq 1$ everywhere], provided X is a random variable satisfying (4.19). With this fact and recalling (4.13), we obtain immediately

$$\begin{aligned} E[f(x + S_{n+1})|X_1, \ldots, X_n] &= E[f(x + S_n + X_{n+1})|X_1, \ldots, X_n] \\ &\leq f(x + S_n), \end{aligned}$$

which verifies the supermartingale inequality. The validation of (4.14) is now complete.

Here is an alternative form or equivalent result, from which a number of important inequalities can be obtained.

Let $X_1, X_2, \ldots$ be jointly distributed random variables having finite conditional moments

$$M_k = E[X_k'|X_1', \ldots, X_{k-1}'],$$

and

$$V_k = E[(X_k' - M_k)^2|X_1', \ldots, X_{k-1}'].$$

Then

$$\Pr\{X_1' + \cdots + X_n' - (M_1 + \cdots + M_n) \geq a(V_1 + \cdots + V_n) + b, \text{ for some } n \geq 1\} \leq \frac{1}{1 + ab}, \qquad a, b \geq 0.$$

The desired conclusion is equivalent to

$$\Pr\{X_1 + \cdots + X_n \geq b, \text{ for some } n \geq 1\} \leq \frac{1}{1+ab},$$

with

$$X_k = X_n' - M_k - aV_k,$$

and the conditional mean of X_k is $-aV_k$, while the conditional variance remains V_k. Thus the new inequality is immediate from the old.

5: Martingale Convergence Theorems

Under quite general conditions, a martingale X_n will converge to a limit random variable X as n increases. Precise statements of these results form some of the most far reaching and powerful theorems in probability theory. We highlight immediately the *basic martingale convergence theorem*. (Chapter 1 reviewed several notions for the convergence of a sequence of random variables.)

Theorem 5.1. (a) *Let* $\{X_n\}$ *be a submartingale satisfying*

$$\sup_{n \geq 0} E[|X_n|] < \infty. \tag{5.1}$$

Then there exists a random variable X_∞ *to which* $\{X_n\}$ *converges with probability one,*

$$\Pr\left\{\lim_{n \to \infty} X_n = X_\infty\right\} = 1. \tag{5.2}$$

(b) *If* $\{X_n\}$ *is a martingale and is uniformly integrable* (*see later Remark* 5.3) *then, in addition to* (5.2), $\{X_n\}$ *converges in the mean, that is,*

$$\lim_{n \to \infty} E[|X_n - X_\infty|] = 0, \tag{5.3}$$

and

$$E[X_\infty] = E[X_n], \qquad \textit{for all} \quad n.$$

Remark 5.1. If $E[|X_0|] < \infty$ for a submartingale $\{X_n\}$, then the condition

$$\sup_{n \geq 1} E[|X_n|] < \infty \qquad \text{holds if and only if} \quad \sup_{n \geq 1} E[X_n^+] < \infty$$

also is maintained, where $X^+ = \max\{X, 0\}$. This equivalence emanates from the elementary inequality $X_n^+ \leq |X_n|$ and the relation $|X_n| = 2X_n^+ - X_n$, yielding thereby

$$E[|X_n|] = 2E[X_n^+] - E[X_n] \leq 2E[X_n^+] - E[X_0].$$

The theorem informs us that, in particular, *every nonpositive submartingale converges with probability one, and so does a nonnegative supermartingale, or a martingale that is uniformly bounded from above or from below.*

Remark 5.2. Convergence with probability one, (5.2), does not entail convergence in the mean, (5.3), nor vice versa. However, both these modes of convergence do imply convergence in probability,

$$\lim_{n\to\infty} \Pr\{|X_n - X_\infty| > \varepsilon\} = 0, \qquad \text{for every} \quad \varepsilon > 0.$$

Indeed, Chebyshev's inequality (consult Chapter 1, Section 1), in the form

$$\Pr\{|X_n - X_\infty| \geq \varepsilon\} \leq \frac{E[|X_n - X_\infty|]}{\varepsilon}, \qquad \varepsilon > 0,$$

shows that convergence in the mean implies convergence in probability.

Remark 5.3. Recall from Section 3 that, by definition, a sequence X_n is *uniformly integrable if*

$$\lim_{c\to\infty} \sup_{n\geq 0} E[|X_n| I\{|X_n| > c\}] = 0. \tag{5.4}$$

(The notation $I\{|X_n| > c\}$ stands for the indicator function of the event where $|X_n| > c$). Specifically,

$$I\{|X_n| > c\} = \begin{cases} 1, & \text{if} \quad |X_n| > c, \\ 0, & \text{if} \quad |X_n| \leq c. \end{cases}$$

Henceforth, generally I of a relation means the indicator function corresponding to the relation. We write also (see Section 2) $I_{\{|X_n|>c\}}$ to signify $I\{|X_n| > c\}$.

Stipulation (5.4) is implied by either of the following conditions:

$$\text{(i)} \quad |X_n| \leq W, \qquad \text{for all} \quad n, \tag{5.5}$$

where W is a random variable satisfying $E[W] < \infty$;

$$\text{(ii)} \quad E[|X_n|^{1+\rho}] \leq K < \infty, \qquad \text{for all} \quad n \tag{5.6}$$

where K and ρ are constants with $\rho > 0$. (The student should prove these statements.)