

Stochastic Processes

Models, Methods, and Explorations

Randall J. Swift

California State Polytechnic University, Pomona

Stochastic Processes
Models, Methods, and Explorations
R. J. Swift
Department of Mathematics & Statistics
Cal Poly Pomona

Copyright ©2027 R.J. Swift
Exclusive Publication and Distribution Rights: Textbook Media Press

Formats/ISBNs

Black & white paperback: 979-8-953597-14-2
Black & white loose-leaf: 979-8-953597-15-9
e-Book access card for Bookstores: 979-8-953597-18-0
eBook Basic: 979-8-953597-17-3
eBook Plus: 979-8-953597-16-6

ALL RIGHTS RESERVED. No part of this work covered by the copyright hereon may be reproduced or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, Web distribution or information storage and retrieval systems—without the written permission of the publisher.

For permission to use material from this text or product, submit a request online at:
info@textbookmedia.com

Textbook Media Press
Suite A
848 Chippewa Avenue
Saint Paul, MN 55107

Textbook Media Press is a Minnesota-based educational publisher. We deliver textbooks and supplements with the quality instructors expect, while providing students with unique media options at uniquely affordable prices.

Also, from R.J. Swift @ Textbook Media Press:

Problems and Explorations in Probability
A Problem-Centered Introduction
And
Essential Combinatorics
A Working Course in Counting Methods
And
THE FIELD GUIDE SERIES FOR MATHEMATICS



The Field Guide Series for Mathematics is a coordinated collection of concise, visual resources designed to help readers learn mathematical structures, recognize them in context, practice using them, and understand the connections among related ideas.

Each topic may be explored through a Field Guide, Field Cards, a Field Workbook, or an integrated Field Companion, allowing readers and instructors to choose the format that best suits how they want to learn.

The Field Guide Series has four coordinated volumes for each topic:

Field Guide — **Learn.** *Essential ideas, maps, methods, examples, warnings, and connections.*

Field Cards — **Recognize.** *Rapid identification, formulas, decision cues, and portable reference.*

Field Workbook — **Practice.** *Tailored calculation, investigation, explanation, and exploration.*

Field Companion — **Integrate.** *Learn, Recognize, and Practice combined by topic.*

The Field Guide Series for Mathematics covers the following 10 topics:

1. Trigonometry
2. Integration Techniques and Infinite Series
3. Vectors
4. Div, Grad, Curl and the Integral Theorems
5. Ordinary Differential Equations
6. Combinatorics
7. Probability Distributions
8. Mathematical Statistics
9. Real Analysis
10. Complex Analysis

For more information, availability and prices, please visit:

www.textbookmedia.com

*To Professor Alan Krinik,
mentor, colleague, coauthor,
trusted ally and lifelong friend*

Preface

Stochastic processes have been a central part of my mathematical career. I met Alan Krinik in 1983, and the following year I took my first course in stochastic processes from him. At his encouragement, I pursued graduate study in mathematics. In 1985, I went to the University of California, Santa Barbara. There I met Joe Gani, whose beautiful lectures on applied stochastic processes still influence me today. Much later, Joe and I collaborated and authored many articles.

In 1988, I attended the University of California, Riverside, where I studied under M. M. Rao and completed my PhD. M. M. showed me the depth that mathematical abstraction can bring to the broad solution of problems. Both Joe Gani and M. M. Rao shaped my mathematical life in profound ways; Alan launched it, and throughout the years he has continued to believe in me.

Stochastic processes provide a language for systems that evolve under uncertainty. Probability sits at the meeting point of several parts of mathematics, from analysis and algebra through combinatorics, statistics, and mathematical modeling. A deep study of stochastic processes eventually requires measure theory, functional analysis, and more advanced probability than we develop in this undergraduate-level text. The aim here is to be self-contained, to develop essential ideas carefully, and to show how mathematics grows from questions arising in applications.

The book also moves in directions that are unusual for an undergraduate treatment. These directions express my view of the topics while reflecting the direct influence of Alan Krinik, Joe Gani, and M. M. Rao.

The text grew from lectures and courses given at Western Kentucky University, Cal Poly Pomona, The Australian National University, and the Claremont Colleges. The book assumes a first calculus-based course in probability and some comfort with elementary matrix calculations, but it does not assume that the reader already knows the language of filtrations, stopping times, generators, or stationary processes. Those ideas are introduced when they are needed, with reminders and supporting material in the appendices. The result can be used for a one-semester course, a reading course, or independent study. The exercises begin with computation and move toward proof, interpretation, and model construction; the appendix provides an answer, a hint, or a solution for every exercise.

The dependency map on the following page shows the main route through the book. Solid arrows indicate material that is used directly in a later chapter; dashed arrows indicate a useful enrichment rather than a prerequisite. The map is a guide for instructors choosing a shorter route as well as for readers working through the full text.

There is a personal history behind the book as well. Alan and I have been friends and colleagues for more than four decades, and our paths have been intertwined in ways that extend well beyond that first course. We share the same PhD advisor, M. M. Rao; we have published papers together and edited volumes of research together. We have also supported one another through departmental storms, changes in universities, and the ordinary uncertainties of academic life. I am grateful for the mathematics we have shared, and even more grateful for the deep and enduring friendship.

Randall Swift
Pomona, California

A dependency map

The chapters are arranged in a principal sequence, but stochastic-process theory is not a single chain of prerequisites. Chapter 1 supplies the common probability language. Finite chains and first-step methods then lead to random walks and branching processes, while event-time models lead to continuous-time Markov chains and population processes. Reliability, risk, and epidemic models draw on several of these strands at once. The second-order and signal chapters take a different analytical route, beginning with mean-square geometry rather than a state transition rule.

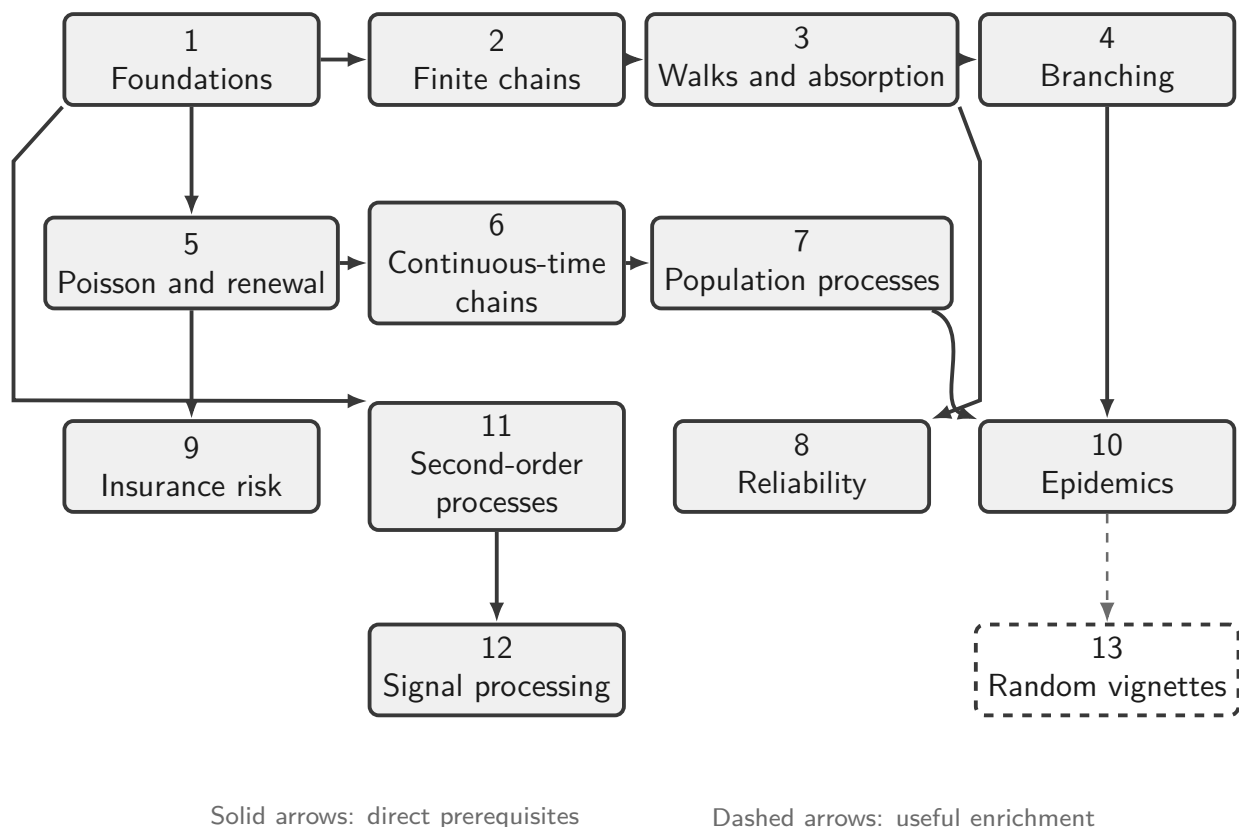


Figure 1: The chapter dependencies and the principal threads of the course.

The map should be read as a guide to emphasis rather than as a rigid requirement. For example, Chapter 8 can be read after Chapters 3 and 6, while Chapter 11 can be entered directly from Chapter 1 if the reader is comfortable with inner products and finite second moments. Chapter 13 is intentionally optional: its vignettes revisit ideas from several earlier chapters and point toward further study.

Contents

Preface	v
A dependency map	vi
I Discrete-Time Models	1
1 From Random Variables to Random Processes	2
1.1 Probability review for process models	2
1.2 Processes, time, and state	7
1.3 State models, dependence, and prediction	10
1.3.1 Mean, covariance, and dependence across time	10
1.3.2 Conditional prediction, martingales, and stopping	11
1.3.3 A deeper look at conditional prediction	12
1.4 Stationarity, increments, and standard process types	15
1.4.1 Increments and standard process types	15
1.5 Three threads through the course	16
Exercises	20
2 Finite Markov Chains: Structure and Long-Run Behavior	22
2.1 Motion, matrices, and path representations	23
2.1.1 Finite Markov Chains and Matrix Methods	23
Exercises	30
2.1.2 Lattice Paths and Finite-Chain Counting	31
Exercises	34
2.1.3 Random allocation: an occupancy process	35
2.1.4 The profile chain and its generating functions	36
2.1.5 Dependence, Poisson asymptotics, and the occupancy peak	37
2.1.6 Thinning, infection, and a continuous-time embedding	38
2.2 Hitting Times and the Structure of a Chain	40
Exercises	44
2.3 Stationary Distributions and Long-Run Behavior	45
Exercises	54

2.3.1	Further structure: reversibility, trees, and stationary flow	56
2.3.2	Ergodic averages and time reversal	57
2.3.3	Case study: a random ladder, shuffles, permutations, and mixing	58
2.3.4	The permutation chain	58
2.3.5	The one-marcher chain	59
3	Random Walks, Networks, and Absorption	64
3.1	Random walks and boundary problems	64
3.1.1	Birth–death chains and random walks	64
	Exercises	68
3.1.2	A carried path: reflection, queues, and workload	69
3.1.3	Catastrophes, storage, and first passage	71
3.1.4	Recurrence, transience, and scaling	72
3.2	Absorption, occupation, and accumulated reward	75
3.2.1	Finite absorbing chains	75
	Exercises	78
3.2.2	Case study: the cover-up game	79
3.2.3	Green functions, occupation, and potentials	81
3.3	Networks, circuits, and areas	83
3.3.1	Random walks, electrical networks, and areas	83
3.3.2	Deterministic circuit models: the same balances without randomness	86
	Exercises	93
3.4	Markov-chain applications	94
3.4.1	Red and gray squirrels in Great Britain	94
3.4.2	Harvesting Scots pines	96
4	Branching Processes	100
4.1	Population trees and the Galton–Watson construction	100
4.2	Generating functions, extinction, and moments	102
4.2.1	Generating functions and extinction	102
4.2.2	Moments and martingales	104
4.3	Total progeny, exploration paths, and lattice geometry	105
4.3.1	Total progeny and the branching property	105
4.3.2	Genealogical exploration as a lattice path	106
4.4	Genetics, geometry, and epidemic approximations	108

4.4.1	Percolation and the early phase of an epidemic	109
4.4.2	A first multitype calculation	110
	Synthesis exercises	111
II	Continuous-Time and Counting Models	113
5	Poisson and Renewal Processes	114
5.1	The Poisson Process	114
	Exercises	119
5.2	Poisson extensions and combinatorial structure	120
5.2.1	Factorial moments, Stirling numbers, and Bell numbers	121
5.2.2	Extensions of the Poisson Process	122
	Exercises	126
5.3	Renewal Processes and Long-Run Averages	128
5.3.1	Why the renewal–reward limit holds	135
	Exercises	135
5.4	Applications: gaps, queues, and controlled traffic	136
5.4.1	Crossing a Busy Street	136
5.4.2	Traffic as a Stochastic System	138
5.4.3	Extended data project: arrivals at a traffic light	142
6	Continuous-Time Markov Chains	148
6.1	From Exponential Clocks to Generators	148
	Exercises	150
6.2	Transition Semigroups and Kolmogorov Equations	150
6.3	Hitting, Occupation, Recurrence, and Stationarity	152
6.3.1	Return times and recurrence	154
6.3.2	Stationarity, detailed balance, and convergence	154
6.4	Birth–Death Chains, Queues, and Applied Models	155
6.4.1	The $M/M/1$ queue	156
6.4.2	Finite capacity, many servers, and infinite servers	156
6.4.3	Little’s law and performance measures	157
6.4.4	State augmentation: vacations, catastrophes, and inventories	158
6.5	Uniformization, Lattice Paths, and Phase-Type Extensions	158
6.5.1	Phase-type holding times	159

6.5.2	Simulation as an application	160
7	Population Processes	162
7.1	Deterministic population models and stochastic translation	162
7.1.1	Growth, crowding, and thresholds	162
7.1.2	The generator as the bridge between curves and events	164
7.2	Birth–death processes and extinction	164
7.3	Branching populations and type structure	166
7.4	Immigration, catastrophes, and stationary populations	167
7.5	Density dependence, interactions, and delayed events	169
7.6	Factorial moments, scaling, and model comparison	170
7.6.1	Large populations and diffusion scale	171
III	System and Risk Models	174
8	Reliability of Components and Systems	175
8.1	Component reliability	175
8.1.1	Cumulative hazard and remaining life	177
8.2	System architecture: series, parallel, and bridge networks	179
8.3	Repairable systems, availability, and operating state	181
8.4	Alternating renewal availability	184
8.5	Maintenance, replacement, and renewal	187
8.6	Dependence and reliability bounds	189
8.6.1	Alternating reliability policies	190
	Exercises	191
9	Renewal Models for Insurance Risk	194
9.1	Aggregate loss: frequency, severity, and renewal structure	194
9.2	General renewal arrivals and discrete-time risk	197
9.3	Aggregate-loss approximations and premium principles	199
9.4	Surplus, reserves, and boundary crossing	200
9.5	Finite-horizon ruin	203
9.6	Renewal equations for ruin	204
9.7	Reinsurance, diversification, and common shocks	206
9.8	Extensions, prediction, and connections	208
	Exercises	208

10 Epidemic Modeling	212
10.1 From Ross to compartment models	212
10.2 Biology, contact, and the law of mass action	213
10.2.1 The Kermack–McKendrick framework and the classical threshold	214
10.3 Deterministic SI and SIR models	215
10.4 The deterministic final-size theorem	216
10.5 Stochastic SI, SIR, and chain-binomial models	217
10.6 Early branching, major outbreaks, and extinction	220
10.7 Epidemic Histories as Weighted Lattice Paths	221
10.8 Two Detailed Examples	222
10.8.1 The common cold: a susceptible-count approximation	223
10.8.2 Kermack–McKendrick removals and the exponential approximation	224
10.9 Epidemics with allocation, carriers, and treatment	225
10.10 Latency, vital dynamics, and intervention	226
10.11 Environmental and ecological epidemics: the cholera model	227
Exercises	230
11 Second-Order Stationary Processes	232
11.1 The mean-square viewpoint	232
11.2 Means, covariances, and second-order stationarity	233
11.3 Covariance existence and the spectral theorem	234
11.4 Ergodicity and time averages	236
11.5 Gaussian processes and the Wiener process	237
11.6 Mean-square continuity, integration, and differentiation	239
11.7 Linear stochastic differential equations	240
11.8 Linear prediction and estimation	242
11.9 Linear filters and covariance in the frequency domain	243
11.10 Spectral distributions and harmonizable processes	244
Exercises	247
12 Elements of Signal Processing	249
12.1 Signals, systems, and convolution	249
12.2 Fourier series and the frequency viewpoint	251
12.3 Random signals and spectral density	252
12.4 Sampling, reconstruction, and aliasing	253

12.5 Finite records and the periodogram	254
Exercises	257
13 Random Vignettes	258
13.1 Random breakage of a rod	258
13.2 Kelly growth and sequential risk	260
13.3 Finite-state scoring and clinical decisions	260
13.4 A birth process driven by a dying enzyme population	262
13.5 Urns and stochastic allocation	263
13.6 Collisions, networks, and reversible chains	264
13.7 Brownian motion and Hermite martingales	265
13.8 Beyond Markov chains: increments and infinite divisibility	266
13.9 Independence Without Coins: Digits and Continued Fractions	268
13.10 The Great Filter and unlikely events	269
Exercises	271
A Foundations Review	272
A.1 Conditional probability	272
A.2 Expectation and conditioning	272
A.3 Covariance and the L^2 viewpoint	273
A.4 Martingales and stopping times	273
A.5 Common distributions	274
A.6 Generating functions and transforms	274
A.7 Matrices	275
A.7.1 Tail formulas	276
A.8 Ordinary differential equations used in the text	276
A.8.1 First-order autonomous equations	277
A.8.2 The method of characteristics	279
A.8.3 Higher-order linear equations with constant coefficients	282
B Notation and Conventions	286
Answers and Solutions	289
Index	350

Part I

Discrete-Time Models

From Random Variables to Random Processes

A stochastic process is a family of random variables that can be viewed from two directions: fix a time to obtain a random variable, or fix an outcome to obtain a sample path. This chapter develops those viewpoints and shows how a sequence of conditional probabilities becomes a Markov model.

1.1 Probability review for process models

The later chapters use a small collection of probability ideas repeatedly. We begin here so that the reader has a comfortable place to return whenever a conditioning or moment calculation needs to be refreshed. Throughout, probabilities are assigned to events in a sample space Ω , and random variables are functions on that space.

We will use indicator functions in our work, specifically if A is a set or an event, its indicator function is

$$\mathbf{1}_A = \mathbf{1}_A = \begin{cases} 1, & \text{if the relevant outcome belongs to } A, \\ 0, & \text{otherwise.} \end{cases}$$

Thus an indicator converts a yes–no condition into a random variable. For example, $Y\mathbf{1}_A$ agrees with Y on A and is zero outside A .

An event is said to occur *almost surely* if it has probability one. In other words, the outcomes for which the event fails form a set of probability zero. “Almost surely” does not mean that there are no exceptional outcomes; it means that those exceptions are invisible to probability calculations. For example, saying that a chain reaches a state almost surely allows, in principle, a collection of paths that never reach it, provided that collection has probability zero.

For events A and B with $\mathbb{P}(B) > 0$, the conditional probability is

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

The multiplication rule and the law of total probability are the two most useful rearrangements:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B), \quad \mathbb{P}(A) = \sum_j \mathbb{P}(A \mid B_j)\mathbb{P}(B_j),$$

when $B_1, B_2, \dots$ partition the sample space. Bayes’ theorem reverses the conditioning:

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_k \mathbb{P}(A \mid B_k)\mathbb{P}(B_k)}.$$

Example 1.1 (A positive test is not a diagnosis). Suppose 1% of a population has a condition, a test is positive with probability .95 when the condition is present, and is positive with probability .05 when it is absent. Then

$$\mathbb{P}(\text{condition} \mid \text{positive}) = \frac{(.95)(.01)}{(.95)(.01) + (.05)(.99)} \approx .161.$$

The calculation is a warning against replacing a conditional probability by its reverse. The same distinction appears when interpreting transition rates, observed states, and posterior model probabilities.

Two events are independent when

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

or, when $\mathbb{P}(B) > 0$, when $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Independence is stronger than having zero covariance and is a modeling assumption about the joint law, not a consequence of two variables having the same marginal distribution. Conditional independence is weaker and more local: events can be independent after a state or latent variable is specified while being dependent after that variable is averaged out.

Independence has consequences for expectations, for a discrete random variable Y , the expectation and variance are defined as

$$\mathbb{E}(Y) = \sum_y y\mathbb{P}(Y = y), \quad \text{Var}(Y) = \mathbb{E}((Y - \mathbb{E}(Y))^2).$$

Linearity of expectation follows from the linearity of summation (or integration); however, linearity of expectation does not require independence:

$$\mathbb{E}\left(\sum_{j=1}^n Y_j\right) = \sum_{j=1}^n \mathbb{E}(Y_j).$$

Independence is needed for the familiar variance sum $\text{Var}(\sum_j Y_j) = \sum_j \text{Var}(Y_j)$; in general,

$$\text{Var}\left(\sum_{j=1}^n Y_j\right) = \sum_j \text{Var}(Y_j) + 2 \sum_{j < k} \text{Cov}(Y_j, Y_k).$$

The covariance term is where temporal dependence enters many stochastic process calculations.

A random variable Y is called *integrable* when its absolute value has finite expectation, $\mathbb{E}(|Y|) < \infty$. For a discrete variable this means $\sum_y |y| \mathbb{P}(Y = y) < \infty$. We will also say that a sequence or series is *summable* when the corresponding sum of its terms is finite; for example, $\sum_k |a_k| < \infty$ means that (a_k) is absolutely summable. Integrability is the condition that makes the expectation and the conditional mean below finite and therefore available for prediction.

Proposition 1.2 (Conditional expectation and total variance). *For an integrable Y and a finite partition $B_1, \dots, B_m$ with positive probabilities,*

$$\mathbb{E}(Y) = \sum_{j=1}^m \mathbb{E}(Y \mid B_j) \mathbb{P}(B_j),$$

and, when Y has finite variance,

$$\text{Var}(Y) = \mathbb{E}(\text{Var}(Y \mid B)) + \text{Var}(\mathbb{E}(Y \mid B)).$$

Proof. The first identity follows by writing $Y = Y \sum_j \mathbf{1}_{B_j}$ and taking expectations. For the second, let $m(B) = \mathbb{E}(Y | B)$. Add and subtract $m(B)$ inside $(Y - \mathbb{E}(Y))^2$. The cross term has conditional expectation zero, so averaging leaves the conditional variance plus the variance of the conditional mean. $\square$

Example 1.3 (A mixture of two workloads). Suppose a randomly selected day is light with probability .7 and heavy with probability .3. The number of calls has conditional means 10 and 30, and conditional variances 10 and 25. Then

$$\mathbb{E}(Y) = .7(10) + .3(30) = 16,$$

while

$$\text{Var}(Y) = .7(10) + .3(25) + .7(10 - 16)^2 + .3(30 - 16)^2 = 98.5.$$

The variance is larger than the within-regime average because the regime means are themselves variable.

In our work, several standard probability distributions will recur. A Bernoulli variable records one success; the binomial variable counts successes in a fixed number of independent trials; the geometric variable counts trials until the first success; and the Poisson variable counts events in a time interval. Their means and variances are displayed below. The field guide by author provides a compact reference for these and other standard distributions, including their properties and connections [88].

Y	$\mathbb{E}(Y)$	$\text{Var}(Y)$
Bernoulli(p)	p	$p(1 - p)$
Binomial(n, p)	np	$np(1 - p)$
Geometric(p)	$1/p$	$(1 - p)/p^2$
Poisson(λ)	λ	λ

The geometric and exponential distributions are *memoryless*:

$$\mathbb{P}(Y > s + t | Y > s) = \mathbb{P}(Y > t).$$

This single identity explains why geometric waiting times generate discrete Markov chains and exponential holding times generate continuous-time Markov chains.

Probability generating functions are another useful tool when working with nonnegative integer-valued variables:

$$G_Y(s) = \mathbb{E}(s^Y), \quad 0 \leq s \leq 1.$$

Differentiating gives $G'_Y(1) = \mathbb{E}(Y)$ and $G''_Y(1) = \mathbb{E}(Y(Y - 1))$, when the derivatives exist. Independent sums multiply generating functions; this will be a useful tool when we study branching, Poisson, and population models.

Extended exercise: Probability review

- Derive Bayes' theorem from the multiplication rule and the law of total probability.
- A disease has prevalence .02, sensitivity .90, and false-positive rate .04. Compute the probability of disease after a positive test and after a negative test.
- Let Y be 5 with probability .4 and 15 with probability .6. Find its mean and variance

directly and using the total-variance identity.

- (d) For dependent variables with covariance c , derive the variance of $Y_1 + Y_2$ and explain the effect of positive versus negative dependence.
- (e) Verify the means and variances in the table for Bernoulli, binomial, geometric, and Poisson variables.
- (f) Prove the memoryless property of the geometric distribution and explain why it supports a Markov state description.
- (g) Find the probability generating function of a binomial variable and use it to recover its mean and variance.
- (h) A population is light with probability .6 and heavy with probability .4. Conditional on the regime, daily births are Poisson with means 8 and 20. Find the unconditional mean and variance, and explain the extra variance.

The preceding calculations form the working probability language for the book. One optional example shows that even independence can be constrained by the arithmetic of a finite sample space.

Optional exploration: when is independence possible.

Independence is often introduced through product experiments: toss two coins, choose two cards, or run two copies of a random mechanism, but a finite sample space can make that intuition misleading. Sometimes the arithmetic of the probabilities rules out every pair of nontrivial independent events. This is useful preparation for stochastic processes, because a process is determined not by its one-time distributions but by the joint probabilities that connect different times.

Let $X_n = \{1, 2, \dots, n\}$ have the uniform probability

$$\mathbb{P}(A) = \frac{|A|}{n}.$$

We call this a *dependent probability space* when no two nonempty proper subsets are independent. The terminology concerns events in one probability space; it does not mean that every pair of random variables built on the space must be dependent.

Example 1.4 (A prime-sized space has no independent pair). Take $X_5 = \{1, 2, 3, 4, 5\}$. If nonempty proper subsets A and B were independent, then

$$\frac{|A \cap B|}{5} = \frac{|A|}{5} \frac{|B|}{5}, \quad \text{so} \quad 5|A \cap B| = |A||B|.$$

Because 5 is prime and $1 \leq |A|, |B| \leq 4$, the right-hand side is not divisible by 5. Thus the uniform space with five points is dependent, even though it contains many pairs of events with positive and negative association.

This dependence structure disappears as soon as the number of points factors.

Example 1.5 (A product hidden in six equally likely outcomes). In X_6 , let $A = \{1, 2\}$ and

$B = \{1, 2, 3\}$. Then

$$\begin{aligned}\mathbb{P}(A) &= \frac{1}{3}, & \mathbb{P}(B) &= \frac{1}{2}, \\ \mathbb{P}(A \cap B) &= \frac{1}{6} = \mathbb{P}(A)\mathbb{P}(B).\end{aligned}$$

so the two events are independent. One can picture the six outcomes as the cells of a 2×3 array: A selects one column and B selects one row. The same construction works whenever $n = rs$ with $r, s > 1$: identify the outcomes with pairs in $\{1, \dots, r\} \times \{1, \dots, s\}$ and use a row event and a column event.

Theorem 1.6 (Uniform finite spaces and primality). *The uniform probability space X_n contains a pair of nonempty proper independent events if and only if n is composite.*

Proof. If $n = rs$ with $r, s > 1$, arrange the n outcomes in an $r \times s$ array. Let A be one column and B one row. Then $|A| = r$, $|B| = s$, and $|A \cap B| = 1$, so

$$\mathbb{P}(A \cap B) = \frac{1}{rs} = \frac{r}{rs} \frac{s}{rs} = \mathbb{P}(A)\mathbb{P}(B).$$

Conversely, suppose n is prime and A, B are nonempty proper subsets. The independence equation gives

$$n|A \cap B| = |A||B|.$$

Since n divides the product $|A||B|$, primality forces n to divide $|A|$ or $|B|$. Neither cardinality lies between 1 and $n - 1$, so this is impossible. $\square$

This theorem is elementary, but it has an important modeling lesson, a finite space may have the same marginal probabilities as another space while allowing different joint events. In a process, the corresponding question is whether the prescribed state space and transition rule leave room for the conditional independences we want. If a state has hidden information, enlarging the state can reveal conditional independence even when the observed process remains dependent. This is the same state-design principle that will lead to Markov chains in the next section.

The finite uniform theorem is adapted from Edwards, Shiflett, and Shultz [22]. Their article studies more general rational weights and shows how arithmetic structure can force dependence even outside the uniform case. We use the finite theorem here as a self-contained example; the general classification is a useful optional direction for further reading.

Extended exercise: Dependent probability spaces

1. Show directly that if A and B are independent, then each of the pairs (A^c, B) , (A, B^c) , and (A^c, B^c) is independent.
2. Find two independent nonempty proper events in the uniform space X_{12} , and display them in a rectangular array of outcomes.
3. Prove that X_7 is dependent without invoking the theorem; use only cardinalities and the independence equation.
4. In X_{10} , construct independent events with probabilities $1/2$ and $1/5$. Explain why their intersection has probability $1/10$.

5. Give an example of two dependent events with the same marginal probabilities as an independent pair. Compute the three relevant probabilities and identify the source of dependence.
6. Let H be a fair bit. Given $H = 0$, let U, V be independent fair bits; given $H = 1$, let $U = V$ be fair. Find $\mathbb{P}(U = V)$ and $\mathbb{P}(U = 1, V = 1)$. Are U and V independent? Are they conditionally independent given H ?
7. **Advanced.** Let $n = rs$ and identify X_n with an $r \times s$ array. Describe all events formed by choosing a fixed collection of rows and a fixed collection of columns, and determine when a row event and a column event are independent.

1.2 Processes, time, and state

In elementary probability, a random variable is a function $X : \Omega \rightarrow \mathbb{R}$. It assigns a numerical value to each outcome in a sample space. A stochastic process adds an index—usually time—and therefore allows the random quantity itself to evolve.

Definition 1.7 (Stochastic process). Let T be an index set and S a state space. A *stochastic process* is a collection

$$\{X_t : t \in T\},$$

where every $X_t : \Omega \rightarrow S$ is a random variable on the same probability space. Equivalently, the process is a map $X : T \times \Omega \rightarrow S$.

The notation also separates two ideas that will be used repeatedly. A sequence is *independent* when observations from different times are independent; it is *identically distributed* when every observation has the same distribution. The abbreviation *i.i.d.* means both properties. Independence describes how times fit together, whereas identical distribution describes what is seen at each time.

The words *random process* and *stochastic process* are interchangeable. The notation contains two complementary objects:

- For fixed t , the map $\omega \mapsto X_t(\omega)$ is a random variable and has a probability distribution.
- For fixed ω , the map $t \mapsto X_t(\omega)$ is a deterministic function called a *sample path* or *realization*.

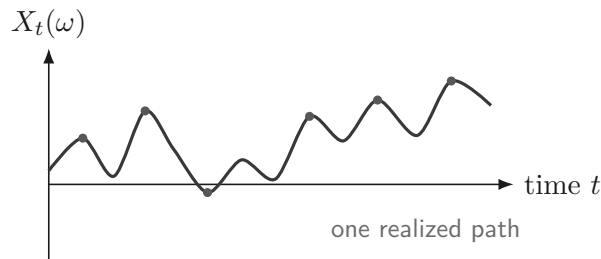


Figure 1.1: A sample path turns the random process into a function of time.

Time and state spaces.

Both the time set T and the state space S may be discrete or continuous. This gives four broad settings.

Table 1.1: Four basic types of stochastic process.

Time	State space	Typical notation	Example
Discrete	Discrete	$X_0, X_1, \dots$	repeated coin tosses; population by generation
Discrete	Continuous	$X_0, X_1, \dots$	rainfall measured each day
Continuous	Discrete	$X(t), t \geq 0$	arrivals by time t ; queue length
Continuous	Continuous	$X(t), t \geq 0$	a noisy signal or asset price

This classification describes the coordinates, not the dependence. Two processes can have the same one-time distributions and very different sample paths. What distinguishes a process is the way its random variables fit together.

Definition 1.8 (Finite-dimensional distributions). For times $t_1 < \dots < t_k$, the joint law of

$$(X_{t_1}, \dots, X_{t_k})$$

is a finite-dimensional distribution of the process.

The one-dimensional distribution of X_t tells us what may happen at time t . The joint distribution tells us how values at different times depend on one another. That dependence is at the heart of stochastic-process theory.

Example 1.9 (The same marginals, different processes). Let $Y \sim \text{Bernoulli}(p)$. The sequence $X_1, X_2, \dots$ may be i.i.d. $\text{Bernoulli}(p)$, or we may set $Z_n = Y$ for every n . In both cases each single observation has the same Bernoulli distribution. In the first case the observations give new information at every time; in the second, one observation reveals the entire path. A stochastic process therefore requires joint distributions, not only one-time distributions.

A first process: Bernoulli trials.

Suppose a fair coin is tossed three times and define

$$X_k = \begin{cases} 1, & \text{the } k\text{th toss is heads,} \\ 0, & \text{the } k\text{th toss is tails.} \end{cases}$$

Then $\{X_1, X_2, X_3\}$ is a discrete-time, discrete-state process. The familiar binomial random variable is recovered by aggregation:

$$N_3 = X_1 + X_2 + X_3.$$

More generally, $N_n = \sum_{k=1}^n X_k$ counts successes by time n . The sequence $\{X_k\}$ records local outcomes; $\{N_n\}$ records their accumulated effect. These are different processes constructed from the same experiment.

Key idea

A process is more than a list of marginal distributions. It includes the temporal relationship among the variables. Independence is one relationship; the Markov property will be another.

From a probability tree diagram to a state model.

Consider families in which a parent is classified as college educated C or not college educated N . Suppose

$$\mathbb{P}(C | C) = 0.85, \quad \mathbb{P}(N | N) = 0.65,$$

where the condition refers to the parent's classification and the event to the child's classification. If 30% of the parents are college educated, then

$$\mathbb{P}(C \text{ for child}) = (0.30)(0.85) + (0.70)(0.35) = 0.50.$$

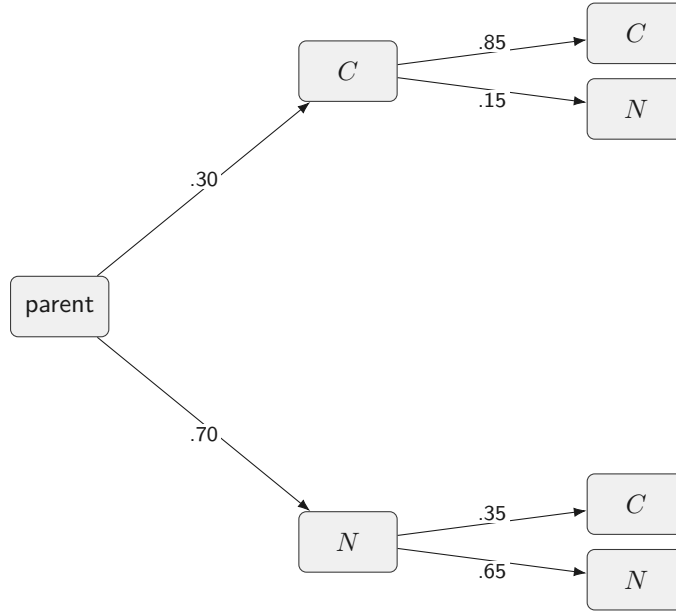


Figure 1.2: The one-generation calculation as a probability tree.

If these conditional probabilities are assumed to remain the same from one generation to the next, a single calculation becomes a process. Let $X_n \in \{C, N\}$ denote the classification in generation n . The essential modeling assumption is that, once X_n is known, earlier generations add no further information about X_{n+1} .

Definition 1.10 (Markov property). A discrete-time process $\{X_n : n \geq 0\}$ on a countable state space is a *Markov chain* if

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i)$$

whenever the conditional probabilities are defined.

The property is often summarized by saying that the future depends on the present state, not on the full past. This is not a claim that the real system has no memory. It says that the chosen *state* contains all the information the model uses to predict one step ahead.

If the conditional probabilities in the definition do not depend on n , the chain is called *time-homogeneous*. Time homogeneity is an additional assumption, not part of the Markov property itself. It allows one transition matrix to describe every step and is the convention used in the finite-chain chapters.

Example 1.11 (A two-step memory model). Suppose a device records whether it is quiet or busy, but the chance of tomorrow's state depends on today's state and yesterday's state. The pair process $Y_n = (X_{n-1}, X_n)$ has four states. A transition from (a, b) to (b, c) is possible, while a transition to (d, c) with $d \neq b$ is not. The enlarged state remembers exactly the information needed for one-step prediction. This simple construction is called a *state augmentation*; it is often the first repair when a proposed state is not Markov.

Modeling note

When the Markov property appears implausible, first ask whether the state has been defined richly enough. A process that is not Markov in X_n may become Markov in an expanded state such as (X_n, X_{n-1}) .

1.3 State models, dependence, and prediction

For a time-homogeneous Markov chain, define

$$p_{ij} = \mathbb{P}(X_{n+1} = i \mid X_n = j).$$

The value is independent of n . We use the *column convention*: the arrow leaves state j , enters state i , and the matrix entry is p_{ij} . Thus columns sum to one.

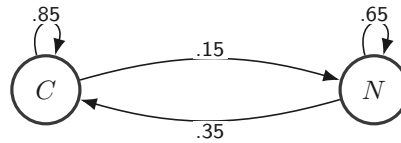


Figure 1.3: A redrawn state diagram for the intergenerational model.

With state order C, N , the transition matrix and initial distribution are

$$P = \begin{bmatrix} .85 & .35 \\ .15 & .65 \end{bmatrix}, \quad p_0 = \begin{bmatrix} .30 \\ .70 \end{bmatrix}.$$

Then $p_1 = Pp_0 = (.50, .50)^T$. The next chapter develops this matrix viewpoint systematically.

1.3.1 Mean, covariance, and dependence across time

The distribution of each X_t answers a one-time question, but a process also carries information about how observations at different times move together. When the required moments exist, three functions give a useful first description. For random variables U and V ,

$$\text{Cov}(U, V) = \mathbb{E}((U - \mathbb{E}(U))(V - \mathbb{E}(V))).$$

Thus

$$m(t) = \mathbb{E}(X_t), \quad r(s, t) = \text{Cov}(X_s, X_t), \quad \rho(s, t) = \frac{r(s, t)}{\sqrt{r(s, s)r(t, t)}},$$

when the two variances are positive. They are called the *mean function*, *covariance function*, and *correlation function*. These functions do not usually determine the full law of a process, but they summarize level, variability, and temporal dependence.

Example 1.12 (The Bernoulli counting process). Let $X_1, X_2, \dots$ be independent Bernoulli variables with success probability p , and set $N_n = \sum_{k=1}^n X_k$. Then

$$\mathbb{E}(N_n) = np, \quad \text{Var}(N_n) = np(1-p).$$

If $m \leq n$, write $N_n = N_m + (N_n - N_m)$. The two terms on the right are independent, so

$$\text{Cov}(N_m, N_n) = \text{Var}(N_m) = mp(1-p).$$

Consequently $\text{Cov}(N_m, N_n) = p(1-p) \min(m, n)$. Counts at distant times remain dependent because they share all successes recorded up to the earlier time.

1.3.2 Conditional prediction, martingales, and stopping

The Markov property is a statement about conditional probabilities. A second basic language for stochastic processes is conditional expectation. Conditional expectation lets us express the best prediction based on the information available at a given time. We now describe that information record before using the notation in the tower rule below.

An information record must satisfy one simple principle: if an event can be decided from the record, then its complement can also be decided, and a countable collection of decisions can be combined into one decision. This is the reason probability theory does not assign probabilities to arbitrary collections of subsets of the sample space. It works with a collection that is stable under the operations needed to describe events and to take limits.

Definition 1.13 (Sigma-field). Let Ω be a sample space. A collection $\mathcal{F}$ of subsets of Ω is a *sigma-field* (or σ -field) if

- (i) $\Omega \in \mathcal{F}$;
- (ii) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$; and
- (iii) if $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

The members of $\mathcal{F}$ are the events whose probabilities and other conditional questions are available in the model.

Example 1.14 (An information record from successive tosses). Let $\Omega = \{HH, HT, TH, TT\}$ describe two tosses of a coin. Write X_0 for the first toss and X_1 for the second. Before either toss, the available information is

$$\mathcal{F}_0 = \{\emptyset, \Omega\}.$$

After the first toss, the information has grown to

$$\mathcal{F}_1 = \{\emptyset, \Omega, \{HH, HT\}, \{TH, TT\}\} = \sigma(X_0),$$

because the first toss distinguishes the outcomes beginning with H from those beginning with T , but it does not distinguish the second tosses. After both tosses every outcome is distinguishable, and hence

$$\mathcal{F}_2 = 2^\Omega = \sigma(X_0, X_1).$$

Thus $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2$: each stage reveals more information without revealing the future in advance. The event $\{X_0 = H\}$ belongs to $\mathcal{F}_1$, whereas $\{X_1 = H\}$ does not. This finite example is the basic pattern for a filtration: it records what can be known at each stage and therefore determines what conditional predictions are legitimate.

The word “sigma” records closure under countably many operations. From the definition, $\emptyset \in \mathcal{F}$, countable intersections belong to $\mathcal{F}$, and differences such as $A \setminus B = A \cap B^c$ belong to $\mathcal{F}$. On a finite sample space one may take the full power set 2^Ω , but in an infinite model the chosen sigma-field expresses which events are observable or measurable. At the other extreme, $\{\emptyset, \Omega\}$ contains no information beyond certainty and impossibility.

For a random variable X , the notation $\sigma(X)$ means the smallest sigma-field that makes X observable: it contains events such as $\{X \in B\}$ for the relevant sets B , together with every event forced by the three closure rules. Thus $\sigma(X_0, \dots, X_n)$ records exactly the information supplied by the observations through time n . This is the bridge from a list of random variables to an information-based description of a process.

Here $\mathcal{F}_n$ is best understood as the information record available after the first n steps have been observed. More formally, a filtration is a sequence of sigma-fields

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots,$$

where $\mathcal{F}_n$ contains the events whose occurrence can be decided from the information available through time n . The natural filtration of a process is $\mathcal{F}_n^X = \sigma(X_0, \dots, X_n)$. In applications the filtration may also contain external observations. The record grows with time because a new observation adds information without erasing the old information.

The inclusion $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ is more than notation. It says that observing one more step may reveal new events, but it cannot make an event that was already knowable become unknowable. Conditional expectation with respect to $\mathcal{F}_n$ is therefore a prediction based on precisely the information available at that stage. This is the interpretation that will make the martingale condition in Definition 1.14 a statement about fair prediction rather than just an equation involving symbols.



Figure 1.4: A filtration records the information available as observations accumulate.

Thus conditioning on $\mathcal{F}_n$ means making the best prediction possible from everything known at time n .

$$\mathbb{E}(\mathbb{E}(Z \mid \mathcal{F}_n) \mid \mathcal{F}_m) = \mathbb{E}(Z \mid \mathcal{F}_m), \quad m \leq n.$$

Forecasting in two stages is therefore the same as forecasting directly from the coarser information.

$$\mathbb{E}(Z) = \mathbb{E}(\mathbb{E}(Z \mid \mathcal{F}_n)), \quad \text{Var}(Z) = \mathbb{E}(\text{Var}(Z \mid \mathcal{F}_n)) + \text{Var}(\mathbb{E}(Z \mid \mathcal{F}_n)).$$

These identities separate randomness within a conditional case from variation between conditional means. They will explain why compound totals can be more variable than their conditional means suggest.

1.3.3 A deeper look at conditional prediction

Conditional expectation is more than a computational device. It is the best mean-square prediction based on the information available. If $\mathcal{G}$ contains the information currently known, then for every square-integrable random variable Y and every square-integrable, $\mathcal{G}$ -measurable Z ,

$$\mathbb{E}[(Y - \mathbb{E}(Y \mid \mathcal{G}))Z] = 0. \quad (1.1)$$

Indeed, the pull-out property and the defining property of conditional expectation give

$$\mathbb{E}[YZ] = \mathbb{E}[\mathbb{E}(YZ \mid \mathcal{G})] = \mathbb{E}[Z\mathbb{E}(Y \mid \mathcal{G})].$$

Taking $Z = 1$ gives the mean-preserving identity $\mathbb{E}(\mathbb{E}(Y \mid \mathcal{G})) = \mathbb{E}(Y)$, and taking $Z = \mathbb{E}(Y \mid \mathcal{G}) - \mathbb{E}(Y)$ gives the variance decomposition

$$\text{Var}(Y) = \mathbb{E}(\text{Var}(Y \mid \mathcal{G})) + \text{Var}(\mathbb{E}(Y \mid \mathcal{G})).$$

Thus the conditional mean removes exactly the predictable component and leaves a residual orthogonal to every prediction formed from the available data.

Example 1.15 (A two-stage population forecast). Let G record the environmental condition before the recruits are observed. We use two possible conditions:

condition	$\mathbb{P}(\text{condition})$	$\mathbb{E}(Y \mid \text{condition})$	$\text{Var}(Y \mid \text{condition})$
favorable habitat	1/3	18	12
unfavorable habitat	2/3	6	8

Here Y is the number of recruits in the next period. The four numbers in the last two columns are model inputs: 18 and 6 are the two conditional forecast means, while 12 and 8 describe the remaining within-environment variation around those means. They are not calculated from an earlier equation; they represent information supplied by observation or by a chosen model. The probability 2/3 is the complement of the stated favorable-habitat probability.

The unconditional mean is obtained by averaging the two conditional means with their environmental probabilities:

$$\mathbb{E}(Y) = \frac{1}{3}(18) + \frac{2}{3}(6) = 10.$$

For the variance, the conditional-variance identity separates two sources of uncertainty:

$$\text{Var}(Y) = \underbrace{\frac{1}{3}(12) + \frac{2}{3}(8)}_{\text{variation within environments}} + \underbrace{\frac{1}{3}(18 - 10)^2 + \frac{2}{3}(6 - 10)^2}_{\text{variation between environments}} = \frac{124}{3} \approx 41.33.$$

The first contribution is the average of the two conditional variances. The second contribution is the uncertainty created by not knowing which environment will occur: the conditional means 18 and 6 lie on opposite sides of the overall mean 10. If we replaced the two environments by their average before observing G , this between-environment contribution would be lost.

Extended exercise: Conditional prediction

1. Prove (1.1) for bounded Z , then extend the result to square-integrable Z by approximation.
2. In the population example, calculate the mean-square error of the unconditional predictor and of the conditional predictor.
3. Explain why the martingale difference $X_{n+1} - \mathbb{E}(X_{n+1} \mid \mathcal{F}_n)$ is orthogonal to every $\mathcal{F}_n$ -measurable square-integrable variable.

Definition 1.16 (Martingale). Let $\mathcal{F}_n$ denote everything known from observing the process through time n , with the records increasing as time advances. An integrable process M_n is a *martingale* if

$$\mathbb{E}(M_{n+1} \mid \mathcal{F}_n) = M_n$$

for every n . In words, after all information through time n has been used, the best prediction of the next value is the present value. A process is a *submartingale* if the left side is at least M_n , and a *supermartingale* if it is at most M_n .

The simple symmetric random walk $S_n = \sum_{k=1}^n X_k$, with independent steps $X_k \in \{-1, 1\}$ having mean zero, is a martingale. The next step has conditional mean zero, so $\mathbb{E}(S_{n+1} \mid \mathcal{F}_n) = S_n$. The compensated Poisson process $N(t) - \lambda t$ is the continuous-time analogue: subtracting the predictable mean removes the drift.

Definition 1.17 (Stopping time). Let $(\mathcal{F}_n)$ be the information available through time n . A nonnegative integer-valued random variable T is a *stopping time* if the event $\{T \leq n\}$ belongs to $\mathcal{F}_n$ for every n .

Thus, by time n , one can decide whether stopping has already occurred without looking into the future. First-passage and absorption times are the main examples.

For a process X_n , the stopped process is written $X_{n \wedge T}$, where the symbol $\wedge$ means “the smaller of”: $a \wedge b = \min(a, b)$. Thus $n \wedge T = \min(n, T)$. Stopping a process does not mean that the process has been observed for a deterministic number of steps. It means that the number of steps is itself determined by the observed path.

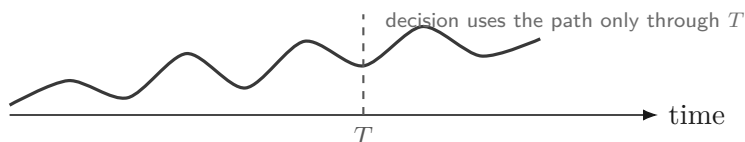


Figure 1.5: A stopping time is determined by information available up to the stopping time.

Theorem 1.18 (Bounded optional stopping). If $(M_n, \mathcal{F}_n)$ is a martingale and T is a bounded stopping time, then

$$\mathbb{E}(M_T) = \mathbb{E}(M_0).$$

The same conclusion holds for suitable unbounded stopping times when the stopped family is uniformly integrable; boundedness is the form used first here.

Martingales identify quantities whose expected value is preserved while information accumulates. If S_n is a fair walk and T is a stopping time satisfying a standard optional-stopping condition (for example, T is bounded), then

$$\mathbb{E}(S_T) = \mathbb{E}(S_0).$$

This optional-stopping identity is not automatic for every unbounded stopping time; boundedness, integrability, or a suitable uniform-integrability condition is part of the theorem.

Example 1.19 (A fair-game accounting identity). Suppose a fair walk starts at i and stops when it reaches either 0 or N . The hitting time is not bounded by a fixed deterministic number, but it is finite almost surely for this finite absorbing walk, and $S_{T \wedge n}$ is bounded between 0 and N .

Applying optional stopping to $T \wedge n$ and then using bounded convergence gives $\mathbb{E}(S_T) = \mathbb{E}(S_0) = i$. If p_i is the probability of reaching N , then

$$i = \mathbb{E}(S_T) = Np_i, \quad p_i = \frac{i}{N}.$$

The same hitting probability will later be obtained from a first-step difference equation. One method uses conditional expectation; the other uses local balance.

1.4 Stationarity, increments, and standard process types

A process is *strictly stationary* if shifting every observation time by the same amount leaves every finite-dimensional distribution unchanged. In discrete time,

$$(X_{t_1}, \dots, X_{t_k}) \stackrel{d}{=} (X_{t_1+h}, \dots, X_{t_k+h})$$

whenever all indices are valid. A process is *second-order stationary* if its mean is constant and its covariance depends only on the lag:

$$\mathbb{E}(X_t) = \mu, \quad \text{Cov}(X_s, X_t) = \gamma(t - s).$$

Strict stationarity implies second-order stationarity when second moments exist, but the converse need not hold.

A Markov chain is not automatically stationary. The Markov property describes conditional dependence; stationarity describes invariance under shifts in time. A time-homogeneous chain becomes stationary when its initial distribution is stationary. Starting from another distribution usually produces a nonstationary process even though the same transition matrix is used at every step.

A point of care

Time-homogeneous, stationary, and stationary-increment are three different descriptions. Time-homogeneous concerns the transition rule, stationary concerns the joint law of levels, and stationary increments concerns the law of changes.

1.4.1 Increments and standard process types

For $s < t$, the difference $X_t - X_s$ is the *increment* over $(s, t]$. A process has *independent increments* when increments over disjoint time intervals are independent. It has *stationary increments* when the distribution of $X_{s+h} - X_s$ depends on the length h , but not on the starting time s . These two properties describe changes, not levels, and neither should be confused with stationarity of the process itself.

A *counting process* $\{N(t) : t \geq 0\}$ has integer-valued, nondecreasing paths, starts at zero, and records how many events have occurred. A *partial-sum process* has the form

$$S_n = Y_1 + \dots + Y_n.$$

Random walks, cumulative demand, aggregate claims, and renewal arrival times all arise from this construction. The same underlying variables may therefore be viewed locally as increments or globally through their accumulated sum.

Example 1.20 (A daily count and its path). Let N_n be the number of service calls received by day n , and let $Y_n = N_n - N_{n-1}$. The process N_n is nondecreasing, while the increment path Y_n need not be. If the Y_n are independent with mean μ , then $\mathbb{E}(N_n) = n\mu$ and $\text{Var}(N_n) = n \text{Var}(Y_1)$. The level path displays accumulation; the increment path displays the local workload. This distinction is useful whenever a model records both a stock and the flow that changes it.

Modeling note

Before choosing a model, identify the observed object. Is it a state, a count, a waiting time, or a cumulative total? Then ask which features are plausible: Markov dependence, independent increments, stationary increments, or none of these. Similar-looking sample paths can arise from very different assumptions.

1.5 Three threads through the course

The later chapters are organized around three questions that recur in different settings. First, how does an object move through a state space? Second, what clock records the events? Third, what is accumulated before the observation ends? Keeping these questions visible makes it easier to recognize when two apparently different models use the same mathematics.

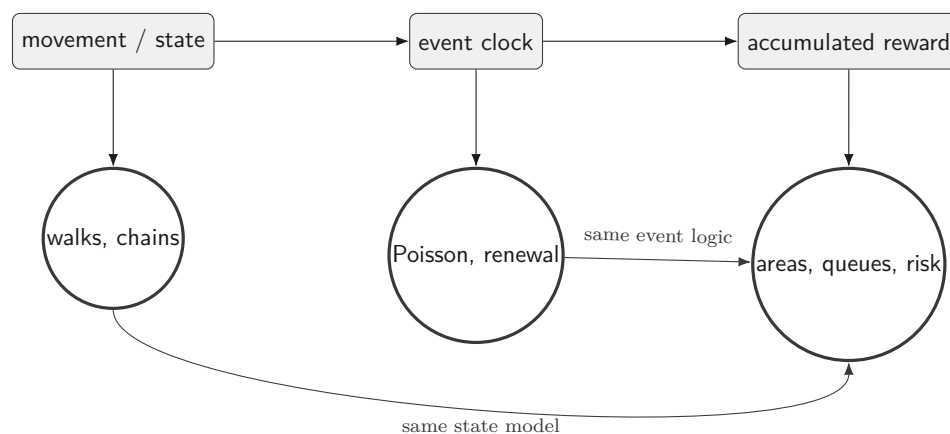


Figure 1.6: Three recurring questions organize the examples that follow.

The movement thread.

The intergenerational education chain in this chapter becomes a finite chain with a stationary distribution in Chapter 2. A random walk in Chapter 3 uses the same transition idea, but its geometry makes first-passage probabilities and electrical potentials visible. The squirrel, hermit-crab, clinical, and bonus-malus examples return to the same state-space question: what must be remembered for one-step prediction to be valid?

The clock thread.

The Bernoulli counting process records events one trial at a time. Chapter 5 replaces the deterministic trial clock by Poisson and renewal clocks. Chapter 6 then attaches exponentially distributed

holding times to state changes. Queues, population processes, epidemics, and reliability models differ in interpretation but share the mechanism “an event occurs, the state changes, and a new clock starts.”

The accumulation thread.

Partial sums lead to random-walk areas, occupation times, queueing costs, population exposure, and aggregate insurance losses. In each case the useful quantity is not only the state at the final time. It is a sum of rewards along the path. The reward vector in an absorbing chain, the area under a walk, and the compound-claim total are three versions of one construction.

Example 1.21 (One state, three interpretations). Let X_n be a nonnegative integer-valued state with transitions to nearby states. If X_n counts occupied territories, a downward move may mean a vacancy and an upward move a colonization. If it counts customers, the same moves are departures and arrivals. If it counts infected individuals, they are recoveries and new infections. The transition probabilities are not justified by the notation alone: each interpretation requires its own mechanism and time scale. Once the state and mechanism have been chosen, however, the same first-step, stationary, and reward calculations can be reused.

Two running examples

Two models will return in several forms. A repairable component begins as a two-state model of working and failed status, becomes a continuous-time chain when failure and repair receive rates, and becomes a reliability model when we distinguish mission survival from long-run availability. An infection count begins as a discrete state process, becomes a branching approximation when the population is small, becomes a finite epidemic chain when susceptibles are limited, and becomes a compartment model when deterministic population curves are useful. In each return, the state, clock, and accumulated quantity are made explicit again; the same notation is never allowed to hide a change in mechanism.

Extended exercise: Following a model through the book

- (a) Choose one of the squirrel, queue, or epidemic models and state exactly what its state records and what it deliberately forgets.
- (b) Identify its movement rule, event clock, and accumulated quantity.
- (c) Give one observation that would contradict the proposed Markov state.
- (d) Describe how the model would change if the observation ended at a first-passage or absorption time rather than at a fixed time.

Optional exploration: independent signs and random series.

In his wonderful classic text, Kac begins with a deliberately simple model: an infinite sequence of independent fair signs. The model is useful here not because we need a specialized theory of random series, but because one object can be read in three familiar ways. Its partial sums form a random walk, its weighted partial sums form a process with changing fluctuation size, and its successive signs provide a particularly transparent example of independence. We keep the discussion at that level and use it to preview two tools that will recur later.

Let $\varepsilon_1, \varepsilon_2, \dots$ be independent random variables with

$$\mathbb{P}(\varepsilon_k = 1) = \mathbb{P}(\varepsilon_k = -1) = \frac{1}{2}.$$

For a finite list of signs,

$$\mathbb{P}(\varepsilon_1 = e_1, \dots, \varepsilon_n = e_n) = 2^{-n}, \quad e_k \in \{-1, 1\}. \quad (1.2)$$

An event that fixes finitely many coordinates is called a *cylinder event*. Its probability is obtained by multiplying the probabilities of the specified signs. This finite calculation is the foundation for the infinite sign model; one does not need to manipulate an infinite product in order to calculate any particular finite-time probability.

Example 1.22 (One sequence, three readings). Set $S_n = \varepsilon_1 + \dots + \varepsilon_n$. Then (S_n) is the simple symmetric random walk. If deterministic coefficients a_k are inserted, the weighted partial sums

$$R_n = \sum_{k=1}^n a_k \varepsilon_k$$

are again a process whose next increment has conditional mean zero. The second moment is especially simple:

$$\mathbb{E}(R_n^2) = \sum_{k=1}^n a_k^2. \quad (1.3)$$

The cross terms vanish because distinct signs are independent and have mean zero. Thus the sum of squares, rather than the sum of absolute values, measures the accumulated random fluctuation.

The word *summable* has two related uses in this setting. Absolute summability, $\sum_k |a_k| < \infty$, guarantees convergence for every choice of signs. Square summability, $\sum_k a_k^2 < \infty$, is weaker, but it guarantees that the random partial sums are Cauchy in mean square, since

$$\mathbb{E}((R_n - R_m)^2) = \sum_{k=m+1}^n a_k^2 \longrightarrow 0.$$

For independent signs, a deeper but standard result says that square summability is also exactly the condition for almost-sure convergence:

$$\sum_{k=1}^{\infty} a_k \varepsilon_k \text{ converges almost surely} \iff \sum_{k=1}^{\infty} a_k^2 < \infty. \quad (1.4)$$

The forward direction reflects the fact that persistent quadratic fluctuation prevents the partial sums from settling; the reverse direction requires a maximal estimate for partial sums and is beyond what is needed for the present introduction. The point for this chapter is the distinction between a deterministic bound, which controls every sign pattern, and a probabilistic bound, which controls almost every sign pattern.

Two elementary probability principles help explain the almost-sure language. The *Borel–Cantelli lemma* says that if $A_1, A_2, \dots$ are events with

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

then, with probability one, only finitely many of the events occur. In other words, a summable sequence of failure probabilities cannot produce infinitely many failures. If the events are independent, the converse also holds: divergence of the probability sum implies that infinitely many occur with probability one. In more advanced arguments, the first direction is often used to turn a sequence of estimates into an almost-sure conclusion.

The *Paley–Zygmund inequality* gives a complementary lower bound. If $Y \geq 0$, $0 < \mathbb{E}(Y^2) < \infty$, and $0 < \theta < 1$, then

$$\mathbb{P}(Y > \theta \mathbb{E}(Y)) \geq (1 - \theta)^2 \frac{\mathbb{E}(Y)^2}{\mathbb{E}(Y^2)}.$$

It says that when the second moment is not too large compared with the square of the first moment, a nonnegative random quantity has a definite chance of being genuinely positive. Borel–Cantelli controls repeated small-probability events from above, while Paley–Zygmund prevents repeated fluctuations from becoming too rare. These ideas explain, at a conceptual level, why the two directions of (1.4) require different arguments.

Example 1.23 (Two contrasting random series). For $a_k = 1/k$, the series of absolute values diverges, but $\sum_k a_k^2 = \sum_k k^{-2} < \infty$. The random-sign series therefore converges almost surely, even though it does not converge absolutely for any choice of signs. For $a_k = 1/\sqrt{k}$, the terms tend to zero but $\sum_k a_k^2 = \sum_k k^{-1} = \infty$, and the criterion says that the weighted walk does not converge almost surely. Small individual steps do not by themselves imply that the accumulated fluctuation settles.

This is the modest lesson we retain from Kac’s example. Independence makes finite calculations transparent, while the second moment records the scale of the unresolved randomness. The same pattern reappears in martingales, counting processes, and signal-like sums later in the text [54].

Extended exercise: Independent signs and random series

1. Verify the cylinder probability in (1.2) for a specified pattern of five signs.
2. Derive (1.3) by expanding the square and evaluating the cross terms.
3. Show directly that absolute summability of (a_k) implies convergence of $\sum_k a_k \varepsilon_k$ for every sign sequence.
4. For $a_k = 1/k$, compare absolute convergence with almost-sure convergence of the random-sign series.
5. Use the Borel–Cantelli lemma to show that if $\mathbb{P}(A_n) \leq 2^{-n}$, then only finitely many A_n occur almost surely.
6. Apply Paley–Zygmund with a convenient value of θ to a nonnegative variable Y whose first two moments are known.
7. Show that $R_n = \sum_{k=1}^n a_k \varepsilon_k$ is a martingale with respect to the sign history.
8. Recover the binomial distribution of the simple symmetric random walk at time n from the cylinder probabilities.

Exercises

Computational practice

1. For each process, identify whether time and state space are discrete or continuous: daily temperature; the number of calls received by time t ; the result of each roll of a die; the distance traveled by a particle at time t .
2. Let $X_1, X_2, \dots$ be Bernoulli trials. Sketch two possible sample paths of $N_n = \sum_{k=1}^n X_k$. What geometric property must every path have?
3. Give two processes with $X_n \sim \text{Bernoulli}(1/2)$ for every n , one with independent variables and one with perfect dependence.
4. In the college model, compute the probabilities for the grandchild of a parent chosen from the initial population.
5. Show that the process $X_n = X_0$ for every n is a Markov chain, regardless of the distribution of X_0 .
6. A weather model has states dry and wet. Tomorrow is dry with probability 0.8 after a dry day and 0.4 after a wet day. Draw its state diagram and write its column-stochastic transition matrix.
7. Suppose tomorrow's weather depends on the weather during each of the preceding two days. Construct a four-state description that makes the process Markov.

Interpretation

8. Explain why the sentence "the future is independent of the past" is an incomplete description of the Markov property.
9. For the Bernoulli counting process, find $\text{Corr}(N_m, N_n)$ when $m \leq n$.
10. Give an example of a process with stationary increments whose level distribution changes with time.
11. Let $S_n = Y_1 + \dots + Y_n$, where the Y_i are independent with mean μ and variance σ^2 . Find $\mathbb{E}(S_n)$, $\text{Var}(S_n)$, and $\text{Cov}(S_m, S_n)$.
12. Decide which of the following are necessarily counting processes: cumulative customer arrivals, the number in a queue, cumulative machine failures, and daily rainfall totals. Explain each answer.
13. Give a time-homogeneous Markov chain that is not stationary, and state an initial distribution that makes it stationary.
14. Determine whether the Bernoulli counting process N_n is stationary, has stationary increments, or has independent increments.
15. Verify directly that a centered sum of independent mean-zero increments is a martingale.

Synthesis

16. Show that $N(t) - \lambda t$ has conditional mean-zero increments.
17. Use bounded optional stopping for a fair walk between 0 and N to derive the probability of hitting N first.
18. Explain why an optional-stopping argument requires a condition such as bounded stopping time or uniform integrability.
19. Build the four-state pair chain for a binary process whose next-state probability depends on the preceding two observations.
20. For independent increments with means m_k and variances v_k , derive the mean and covariance of the partial-sum process.
21. Give a process with stationary increments that is neither Markov nor stationary, and justify both claims from its finite-dimensional distributions.
22. Prove the total-variance identity by conditioning on a finite partition, then explain how it separates within-group variability from between-group variability in a population model.
23. Let $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$ for a simple random walk. Describe what events belong to $\mathcal{F}_2$ and explain why the event $\{T \leq 2\}$ belongs to $\mathcal{F}_2$ when T is the first time the walk reaches either -1 or 1 .
24. Decide whether each of the following is a stopping time for a simple random walk: the first time the walk reaches 3; the last time it visits 0 before time 10; the first time it reaches its running maximum plus 2. Justify each answer.

Chapter summary

A process may be read across outcomes at a fixed time or along time for a fixed outcome. Its joint distributions describe temporal dependence; its mean and covariance functions summarize part of that structure. Counts, increments, and partial sums provide standard process constructions. A Markov chain is obtained when the present state contains the information used to predict the next state. Throughout the discrete-time chapters, p_{ij} denotes motion from j to i , so transition matrices are column-stochastic.