

Introduction to Probability

Dimitri P. Bertsekas and John N. Tsitsiklis

Massachusetts Institute of Technology

WWW site for book information and orders

<http://www.athenasc.com>



Athena Scientific, Belmont, Massachusetts

**Athena Scientific
Post Office Box 391
Belmont, Mass. 02478-9998
U.S.A.**

**Email: info@athenasc.com
WWW: <http://www.athenasc.com>**

Cover Design: *Ann Gallager*

© 2002 Dimitri P. Bertsekas and John N. Tsitsiklis
All rights reserved. No part of this book may be reproduced in any form by any
electronic or mechanical means (including photocopying, recording, or informa-
tion storage and retrieval) without permission in writing from the publisher.

Publisher's Cataloging-in-Publication Data

Bertsekas, Dimitri P., Tsitsiklis, John N.
Introduction to Probability
Includes bibliographical references and index
1. Probabilities. 2. Stochastic Processes. I. Title.
QA273.B475 2002 519.2 – 21
Library of Congress Control Number: 2002092167

ISBN 1-886529-40-X

*To the memory of
Pantelis Bertsekas and Nikos Tsitsiklis*

Contents

1. Sample Space and Probability	p. 1
1.1. Sets	p. 3
1.2. Probabilistic Models	p. 6
1.3. Conditional Probability	p. 18
1.4. Total Probability Theorem and Bayes' Rule	p. 28
1.5. Independence	p. 34
1.6. Counting	p. 43
1.7. Summary and Discussion	p. 50
Problems	p. 52
2. Discrete Random Variables	p. 71
2.1. Basic Concepts	p. 72
2.2. Probability Mass Functions	p. 74
2.3. Functions of Random Variables	p. 80
2.4. Expectation, Mean, and Variance	p. 81
2.5. Joint PMFs of Multiple Random Variables	p. 92
2.6. Conditioning	p. 98
2.7. Independence	p. 110
2.8. Summary and Discussion	p. 116
Problems	p. 119
3. General Random Variables	p. 139
3.1. Continuous Random Variables and PDFs	p. 140
3.2. Cumulative Distribution Functions	p. 148
3.3. Normal Random Variables	p. 152
3.4. Conditioning on an Event	p. 158
3.5. Multiple Continuous Random Variables	p. 164
3.6. Derived Distributions	p. 179
3.7. Summary and Discussion	p. 190
Problems	p. 192

4. Further Topics on Random Variables	p. 209
4.1. Transforms	p. 210
4.2. Sums of Independent Random Variables - Convolution	p. 221
4.3. More on Conditional Expectation and Variance	p. 225
4.4. Sum of a Random Number of Independent Random Variables	p. 232
4.5. Covariance and Correlation	p. 236
4.6. Least Squares Estimation	p. 240
4.7. The Bivariate Normal Distribution	p. 247
4.8. Summary and Discussion	p. 255
Problems	p. 257
5. The Bernoulli and Poisson Processes	p. 271
5.1. The Bernoulli Process	p. 273
5.2. The Poisson Process	p. 285
5.3. Summary and Discussion	p. 299
Problems	p. 301
6. Markov Chains	p. 313
6.1. Discrete-Time Markov Chains	p. 314
6.2. Classification of States	p. 321
6.3. Steady-State Behavior	p. 326
6.4. Absorption Probabilities and Expected Time to Absorption	p. 337
6.5. Continuous-Time Markov Chains	p. 344
6.6. Summary and Discussion	p. 352
Problems	p. 354
7. Limit Theorems	p. 379
7.1. Markov and Chebyshev Inequalities	p. 381
7.2. The Weak Law of Large Numbers	p. 383
7.3. Convergence in Probability	p. 386
7.4. The Central Limit Theorem	p. 388
7.5. The Strong Law of Large Numbers	p. 395
7.6. Summary and Discussion	p. 397
Problems	p. 399
Index	p. 411

Preface

Probability is common sense reduced to calculation
Laplace

This book is an outgrowth of our involvement in teaching an introductory probability course (“Probabilistic Systems Analysis”) at the Massachusetts Institute of Technology.

The course is attended by a large number of students with diverse backgrounds, and a broad range of interests. They span the entire spectrum from freshmen to beginning graduate students, and from the engineering school to the school of management. Accordingly, we have tried to strike a balance between simplicity in exposition and sophistication in analytical reasoning. Our key aim has been to develop the ability to construct and analyze probabilistic models in a manner that combines intuitive understanding and mathematical precision.

In this spirit, some of the more mathematically rigorous analysis has been just sketched or intuitively explained in the text, so that complex proofs do not stand in the way of an otherwise simple exposition. At the same time, some of this analysis is developed (at the level of advanced calculus) in theoretical problems, that are included at the end of the corresponding chapter. Furthermore, some of the subtler mathematical issues are hinted at in footnotes addressed to the more attentive reader.

The book covers the fundamentals of probability theory (probabilistic models, discrete and continuous random variables, multiple random variables, and limit theorems), which are typically part of a first course on the subject. It also contains, in Chapters 4–6 a number of more advanced topics, from which an instructor can choose to match the goals of a particular course. In particular, in Chapter 4, we develop transforms, a more advanced view of conditioning, sums of random variables, least squares estimation, and the bivariate normal distribu-

tion. Furthermore, in Chapters 5 and 6, we provide a fairly detailed introduction to Bernoulli, Poisson, and Markov processes.

Our M.I.T. course covers all seven chapters in a single semester, with the exception of the material on the bivariate normal (Section 4.7), and on continuous-time Markov chains (Section 6.5). However, in an alternative course, the material on stochastic processes could be omitted, thereby allowing additional emphasis on foundational material, or coverage of other topics of the instructor's choice.

Our most notable omission in coverage is an introduction to statistics. While we develop all the basic elements of Bayesian statistics, in the form of Bayes' rule for discrete and continuous models, and least squares estimation, we do not enter the subjects of parameter estimation, or non-Bayesian hypothesis testing.

The problems that supplement the main text are divided in three categories:

- (a) *Theoretical problems*: The theoretical problems (marked by *) constitute an important component of the text, and ensure that the mathematically oriented reader will find here a smooth development without major gaps. Their solutions are given in the text, but an ambitious reader may be able to solve many of them, especially in earlier chapters, before looking at the solutions.
- (b) *Problems in the text*: Besides theoretical problems, the text contains several problems, of various levels of difficulty. These are representative of the problems that are usually covered in recitation and tutorial sessions at M.I.T., and are a primary mechanism through which many of our students learn the material. Our hope is that students elsewhere will attempt to solve these problems, and then refer to their solutions to calibrate and enhance their understanding of the material. The solutions are posted on the book's www site

<http://www.athenasc.com/probbook.html>

- (c) *Supplementary problems*: There is a large (and growing) collection of additional problems, which is not included in the book, but is made available at the book's www site. Many of these problems have been assigned as homework or exam problems at M.I.T., and we expect that instructors elsewhere will use them for a similar purpose. While the statements of these additional problems are publicly accessible, the solutions are made available from the authors only to course instructors.

We would like to acknowledge our debt to several people who contributed in various ways to the book. Our writing project began when we assumed responsibility for a popular probability class at M.I.T. that our colleague Al Drake had taught for several decades. We were thus fortunate to start with an organization of the subject that had stood the test of time, a lively presentation of the various topics in Al's classic textbook, and a rich set of material that had been used in recitation sessions and for homework. We are thus indebted to Al Drake

for providing a very favorable set of initial conditions.

We are thankful to the several colleagues who have either taught from the draft of the book at various universities or have read it, and have provided us with valuable feedback. In particular, we thank Ibrahim Abou Faycal, Gustavo de Veciana, Eugene Feinberg, Bob Gray, Muriel Médard, Jason Papastavrou, Ilya Pollak, David Tse, and Terry Wagner.

The teaching assistants for the M.I.T. class have been very helpful. They pointed out corrections to various drafts, they developed problems and solutions suitable for the class, and through their direct interaction with the student body, they provided a robust mechanism for calibrating the level of the material.

Reaching thousands of bright students at M.I.T. at an early stage in their studies was a great source of satisfaction for us. We thank them for their valuable feedback and for being patient while they were taught from a textbook-in-progress.

Last but not least, we are grateful to our families for their support throughout the course of this long project.

Dimitri P. Bertsekas, dimitrib@mit.edu

John N. Tsitsiklis, jnt@mit.edu

Cambridge, Mass., May 2002

ATHENA SCIENTIFIC BOOKS

1. Introduction to Probability, by Dimitri P. Bertsekas and John N. Tsitsiklis, 2002, ISBN 1-886529-40-X, 430 pages
2. Dynamic Programming and Optimal Control: Second Edition, Vols. I and II, by Dimitri P. Bertsekas, 2001, ISBN 1-886529-08-6, 704 pages
3. Nonlinear Programming, Second Edition, by Dimitri P. Bertsekas, 1999, ISBN 1-886529-00-0, 791 pages
4. Network Optimization: Continuous and Discrete Models by Dimitri P. Bertsekas, 1998, ISBN 1-886529-02-7, 608 pages
5. Network Flows and Monotropic Optimization by R. Tyrrell Rockafellar, 1998, ISBN 1-886529-06-X, 634 pages
6. Introduction to Linear Optimization by Dimitris Bertsimas and John N. Tsitsiklis, 1997, ISBN 1-886529-19-1, 608 pages
7. Parallel and Distributed Computation: Numerical Methods by Dimitri P. Bertsekas and John N. Tsitsiklis, 1997, ISBN 1-886529-01-9, 718 pages
8. Neuro-Dynamic Programming, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1996, ISBN 1-886529-10-8, 512 pages
9. Constrained Optimization and Lagrange Multiplier Methods, by Dimitri P. Bertsekas, 1996, ISBN 1-886529-04-3, 410 pages
10. Stochastic Optimal Control: The Discrete-Time Case by Dimitri P. Bertsekas and Steven E. Shreve, 1996, ISBN 1-886529-03-5, 330 pages

*Sample Space and
Probability*

Contents

1.1. Sets	p. 3
1.2. Probabilistic Models	p. 6
1.3. Conditional Probability	p. 18
1.4. Total Probability Theorem and Bayes' Rule	p. 28
1.5. Independence	p. 34
1.6. Counting	p. 43
1.7. Summary and Discussion	p. 50
Problems	p. 52

“Probability” is a very useful concept, but can be interpreted in a number of ways. As an illustration, consider the following.

A patient is admitted to the hospital and a potentially life-saving drug is administered. The following dialog takes place between the nurse and a concerned relative.

RELATIVE: Nurse, what is the probability that the drug will work?

NURSE: I hope it works, we'll know tomorrow.

RELATIVE: Yes, but what is the probability that it will?

NURSE: Each case is different, we have to wait.

RELATIVE: But let's see, out of a hundred patients that are treated under similar conditions, how many times would you expect it to work?

NURSE (somewhat annoyed): I told you, every person is different, for some it works, for some it doesn't.

RELATIVE (insisting): Then tell me, if you had to bet whether it will work or not, which side of the bet would you take?

NURSE (cheering up for a moment): I'd bet it will work.

RELATIVE (somewhat relieved): OK, now, would you be willing to lose two dollars if it doesn't work, and gain one dollar if it does?

NURSE (exasperated): What a sick thought! You are wasting my time!

In this conversation, the relative attempts to use the concept of probability to discuss an **uncertain** situation. The nurse's initial response indicates that the meaning of “probability” is not uniformly shared or understood, and the relative tries to make it more concrete. The first approach is to define probability in terms of **frequency of occurrence**, as a percentage of successes in a moderately large number of similar situations. Such an interpretation is often natural. For example, when we say that a perfectly manufactured coin lands on heads “with probability 50%,” we typically mean “roughly half of the time.” But the nurse may not be entirely wrong in refusing to discuss in such terms. What if this was an experimental drug that was administered for the very first time in this hospital or in the nurse's experience?

While there are many situations involving uncertainty in which the frequency interpretation is appropriate, there are other situations in which it is not. Consider, for example, a scholar who asserts that the Iliad and the Odyssey were composed by the same person, with probability 90%. Such an assertion conveys some information, but not in terms of frequencies, since the subject is a one-time event. Rather, it is an expression of the scholar's **subjective belief**. One might think that subjective beliefs are not interesting, at least from a mathematical or scientific point of view. On the other hand, people often have to make choices in the presence of uncertainty, and a systematic way of making use of their beliefs is a prerequisite for successful, or at least consistent, decision making.

In fact, the choices and actions of a rational person, can reveal a lot about the inner-held subjective probabilities, even if the person does not make conscious use of probabilistic reasoning. Indeed, the last part of the earlier dialog was an attempt to infer the nurse's beliefs in an indirect manner. Since the nurse was willing to accept a one-for-one bet that the drug would work, we may infer that the probability of success was judged to be at least 50%. And had the nurse accepted the last proposed bet (two-for-one), that would have indicated a success probability of at least 2/3.

Rather than dwelling further into philosophical issues about the appropriateness of probabilistic reasoning, we will simply take it as a given that the theory of probability is useful in a broad variety of contexts, including some where the assumed probabilities only reflect subjective beliefs. There is a large body of successful applications in science, engineering, medicine, management, etc., and on the basis of this empirical evidence, probability theory is an extremely useful tool.

Our main objective in this book is to develop the art of describing uncertainty in terms of probabilistic models, as well as the skill of probabilistic reasoning. The first step, which is the subject of this chapter, is to describe the generic structure of such models, and their basic properties. The models we consider assign probabilities to collections (sets) of possible outcomes. For this reason, we must begin with a short review of set theory.

1.1 SETS

Probability makes extensive use of set operations, so let us introduce at the outset the relevant notation and terminology.

A **set** is a collection of objects, which are the **elements** of the set. If S is a set and x is an element of S , we write $x \in S$. If x is not an element of S , we write $x \notin S$. A set can have no elements, in which case it is called the **empty set**, denoted by \emptyset .

Sets can be specified in a variety of ways. If S contains a finite number of elements, say x_1, x_2, \dots, x_n , we write it as a list of the elements, in braces:

$$S = \{x_1, x_2, \dots, x_n\}.$$

For example, the set of possible outcomes of a die roll is $\{1, 2, 3, 4, 5, 6\}$, and the set of possible outcomes of a coin toss is $\{H, T\}$, where H stands for “heads” and T stands for “tails.”

If S contains infinitely many elements x_1, x_2, \dots , which can be enumerated in a list (so that there are as many elements as there are positive integers) we write

$$S = \{x_1, x_2, \dots\},$$

and we say that S is **countably infinite**. For example, the set of even integers can be written as $\{0, 2, -2, 4, -4, \dots\}$, and is countably infinite.

Alternatively, we can consider the set of all x that have a certain property P , and denote it by

$$\{x \mid x \text{ satisfies } P\}.$$

(The symbol “ \mid ” is to be read as “such that.”) For example, the set of even integers can be written as $\{k \mid k/2 \text{ is integer}\}$. Similarly, the set of all scalars x in the interval $[0, 1]$ can be written as $\{x \mid 0 \leq x \leq 1\}$. Note that the elements x of the latter set take a continuous range of values, and cannot be written down in a list (a proof is sketched in the end-of-chapter problems); such a set is said to be **uncountable**.

If every element of a set S is also an element of a set T , we say that S is a **subset** of T , and we write $S \subset T$ or $T \supset S$. If $S \subset T$ and $T \subset S$, the two sets are **equal**, and we write $S = T$. It is also expedient to introduce a **universal set**, denoted by Ω , which contains all objects that could conceivably be of interest in a particular context. Having specified the context in terms of a universal set Ω , we only consider sets S that are subsets of Ω .

Set Operations

The **complement** of a set S , with respect to the universe Ω , is the set $\{x \in \Omega \mid x \notin S\}$ of all elements of Ω that do not belong to S , and is denoted by S^c . Note that $\Omega^c = \emptyset$.

The **union** of two sets S and T is the set of all elements that belong to S or T (or both), and is denoted by $S \cup T$. The **intersection** of two sets S and T is the set of all elements that belong to both S and T , and is denoted by $S \cap T$. Thus,

$$S \cup T = \{x \mid x \in S \text{ or } x \in T\},$$

and

$$S \cap T = \{x \mid x \in S \text{ and } x \in T\}.$$

In some cases, we will have to consider the union or the intersection of several, even infinitely many sets, defined in the obvious way. For example, if for every positive integer n , we are given a set S_n , then

$$\bigcup_{n=1}^{\infty} S_n = S_1 \cup S_2 \cup \dots = \{x \mid x \in S_n \text{ for some } n\},$$

and

$$\bigcap_{n=1}^{\infty} S_n = S_1 \cap S_2 \cap \dots = \{x \mid x \in S_n \text{ for all } n\}.$$

Two sets are said to be **disjoint** if their intersection is empty. More generally, several sets are said to be **disjoint** if no two of them have a common element. A collection of sets is said to be a **partition** of a set S if the sets in the collection are disjoint and their union is S .

If x and y are two objects, we use (x, y) to denote the **ordered pair** of x and y . The set of scalars (real numbers) is denoted by \mathfrak{R} ; the set of pairs (or triplets) of scalars, i.e., the two-dimensional plane (or three-dimensional space, respectively) is denoted by \mathfrak{R}^2 (or \mathfrak{R}^3 , respectively).

Sets and the associated operations are easy to visualize in terms of **Venn diagrams**, as illustrated in Fig. 1.1.

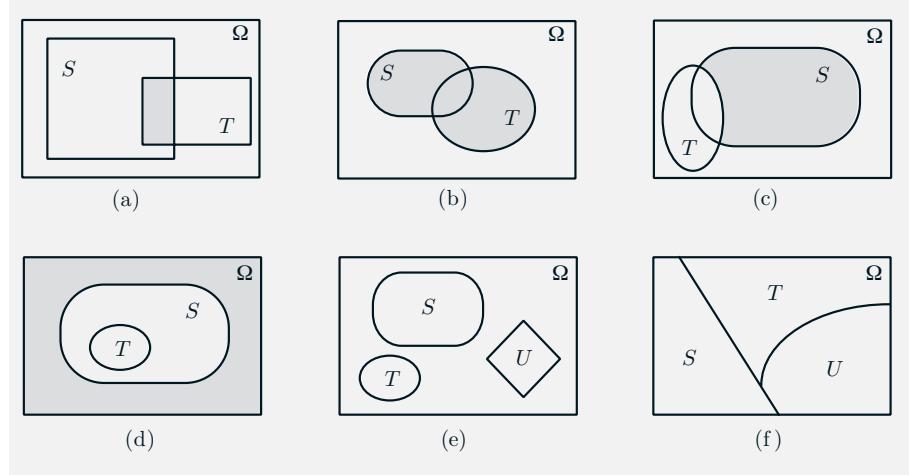


Figure 1.1: Examples of Venn diagrams. (a) The shaded region is $S \cap T$. (b) The shaded region is $S \cup T$. (c) The shaded region is $S \cap T^c$. (d) Here, $T \subset S$. The shaded region is the complement of S . (e) The sets S , T , and U are disjoint. (f) The sets S , T , and U form a partition of the set Ω .

The Algebra of Sets

Set operations have several properties, which are elementary consequences of the definitions. Some examples are:

$$\begin{array}{ll}
 S \cup T = T \cup S, & S \cup (T \cup U) = (S \cup T) \cup U, \\
 S \cap (T \cup U) = (S \cap T) \cup (S \cap U), & S \cup (T \cap U) = (S \cup T) \cap (S \cup U), \\
 (S^c)^c = S, & S \cap S^c = \emptyset, \\
 S \cup \Omega = \Omega, & S \cap \Omega = S.
 \end{array}$$

Two particularly useful properties are given by **De Morgan's laws** which state that

$$\left(\bigcup_n S_n \right)^c = \bigcap_n S_n^c, \quad \left(\bigcap_n S_n \right)^c = \bigcup_n S_n^c.$$

To establish the first law, suppose that $x \in (\bigcup_n S_n)^c$. Then, $x \notin \bigcup_n S_n$, which implies that for every n , we have $x \notin S_n$. Thus, x belongs to the complement

of every S_n , and $x_n \in \cap_n S_n^c$. This shows that $(\cup_n S_n)^c \subset \cap_n S_n^c$. The converse inclusion is established by reversing the above argument, and the first law follows. The argument for the second law is similar.

1.2 PROBABILISTIC MODELS

A probabilistic model is a mathematical description of an uncertain situation.

It must be in accordance with a fundamental framework that we discuss in this section. Its two main ingredients are listed below and are visualized in Fig. 1.2.

Elements of a Probabilistic Model

- The **sample space** Ω , which is the set of all possible **outcomes** of an experiment.
- The **probability law**, which assigns to a set A of possible outcomes (also called an **event**) a nonnegative number $\mathbf{P}(A)$ (called the **probability** of A) that encodes our knowledge or belief about the collective “likelihood” of the elements of A . The probability law must satisfy certain properties to be introduced shortly.

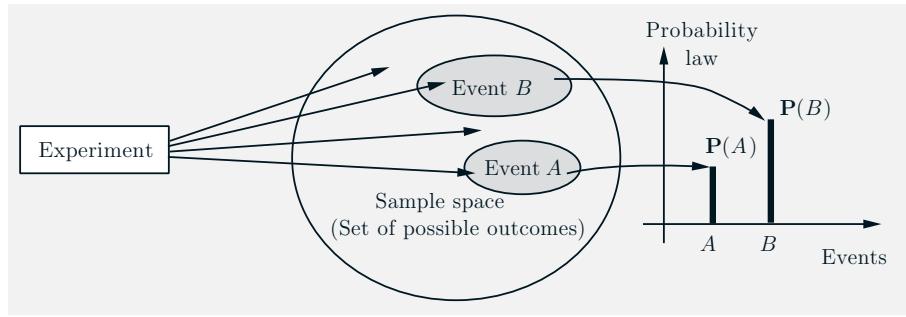


Figure 1.2: The main ingredients of a probabilistic model.

Sample Spaces and Events

Every probabilistic model involves an underlying process, called the **experiment**, that will produce exactly one out of several possible **outcomes**. The set of all possible outcomes is called the **sample space** of the experiment, and is denoted by Ω . A subset of the sample space, that is, a collection of possible

outcomes, is called an **event**.[†] There is no restriction on what constitutes an experiment. For example, it could be a single toss of a coin, or three tosses, or an infinite sequence of tosses. However, it is important to note that in our formulation of a probabilistic model, there is only one experiment. So, three tosses of a coin constitute a single experiment, rather than three experiments.

The sample space of an experiment may consist of a finite or an infinite number of possible outcomes. Finite sample spaces are conceptually and mathematically simpler. Still, sample spaces with an infinite number of elements are quite common. For an example, consider throwing a dart on a square target and viewing the point of impact as the outcome.

Choosing an Appropriate Sample Space

Regardless of their number, different elements of the sample space should be distinct and **mutually exclusive** so that when the experiment is carried out, there is a unique outcome. For example, the sample space associated with the roll of a die cannot contain “1 or 3” as a possible outcome and also “1 or 4” as another possible outcome, because we would not be able to assign a unique outcome when the roll is a 1.

A given physical situation may be modeled in several different ways, depending on the kind of questions that we are interested in. Generally, the sample space chosen for a probabilistic model must be **collectively exhaustive**, in the sense that no matter what happens in the experiment, we always obtain an outcome that has been included in the sample space. In addition, the sample space should have enough detail to distinguish between all outcomes of interest to the modeler, while avoiding irrelevant details.

Example 1.1. Consider two alternative games, both involving ten successive coin tosses:

Game 1: We receive \$1 each time a head comes up.

Game 2: We receive \$1 for every coin toss, up to and including the first time a head comes up. Then, we receive \$2 for every coin toss, up to the second time a head comes up. More generally, the dollar amount per toss is doubled each time a head comes up.

[†] Any collection of possible outcomes, including the entire sample space Ω and its complement, the empty set \emptyset , may qualify as an event. Strictly speaking, however, some sets have to be excluded. In particular, when dealing with probabilistic models involving an uncountably infinite sample space, there are certain unusual subsets for which one cannot associate meaningful probabilities. This is an intricate technical issue, involving the mathematics of measure theory. Fortunately, such pathological subsets do not arise in the problems considered in this text or in practice, and the issue can be safely ignored.

In game 1, it is only the total number of heads in the ten-toss sequence that matters, while in game 2, the order of heads and tails is also important. Thus, in a probabilistic model for game 1, we can work with a sample space consisting of eleven possible outcomes, namely, $0, 1, \dots, 10$. In game 2, a finer grain description of the experiment is called for, and it is more appropriate to let the sample space consist of every possible ten-long sequence of heads and tails.

Sequential Models

Many experiments have an inherently sequential character, such as for example tossing a coin three times, or observing the value of a stock on five successive days, or receiving eight successive digits at a communication receiver. It is then often useful to describe the experiment and the associated sample space by means of a **tree-based sequential description**, as in Fig. 1.3.

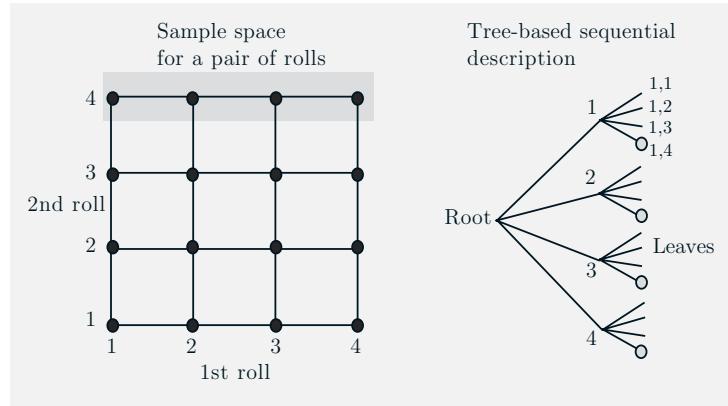


Figure 1.3: Two equivalent descriptions of the sample space of an experiment involving two rolls of a 4-sided die. The possible outcomes are all the ordered pairs of the form (i, j) , where i is the result of the first roll, and j is the result of the second. These outcomes can be arranged in a 2-dimensional grid as in the figure on the left, or they can be described by the tree on the right, which reflects the sequential character of the experiment. Here, each possible outcome corresponds to a leaf of the tree and is associated with the unique path from the root to that leaf. The shaded area on the left is the event $\{(1, 4), (2, 4), (3, 4), (4, 4)\}$ that the result of the second roll is 4. That same event can be described by the set of leaves highlighted on the right. Note also that every node of the tree can be identified with an event, namely, the set of all leaves downstream from that node. For example, the node labeled by a 1 can be identified with the event $\{(1, 1), (1, 2), (1, 3), (1, 4)\}$ that the result of the first roll is 1.

Probability Laws

Suppose we have settled on the sample space Ω associated with an experiment. Then, to complete the probabilistic model, we must introduce a **probability**

law. Intuitively, this specifies the “likelihood” of any outcome, or of any set of possible outcomes (an event, as we have called it earlier). More precisely, the probability law assigns to every event A , a number $\mathbf{P}(A)$, called the **probability** of A , satisfying the following axioms.

Probability Axioms

1. **(Nonnegativity)** $\mathbf{P}(A) \geq 0$, for every event A .
2. **(Additivity)** If A and B are two disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

More generally, if the sample space has an infinite number of elements and A_1, A_2, \dots is a sequence of disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A_1 \cup A_2 \cup \dots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots$$

3. **(Normalization)** The probability of the entire sample space Ω is equal to 1, that is, $\mathbf{P}(\Omega) = 1$.

In order to visualize a probability law, consider a unit of mass which is “spread” over the sample space. Then, $\mathbf{P}(A)$ is simply the total mass that was assigned collectively to the elements of A . In terms of this analogy, the additivity axiom becomes quite intuitive: the total mass in a sequence of disjoint events is the sum of their individual masses.

A more concrete interpretation of probabilities is in terms of relative frequencies: a statement such as $\mathbf{P}(A) = 2/3$ often represents a belief that event A will occur in about two thirds out of a large number of repetitions of the experiment. Such an interpretation, though not always appropriate, can sometimes facilitate our intuitive understanding. It will be revisited in Chapter 7, in our study of limit theorems.

There are many natural properties of a probability law, which have not been included in the above axioms for the simple reason that they can be **derived** from them. For example, note that the normalization and additivity axioms imply that

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(\Omega \cup \emptyset) = \mathbf{P}(\Omega) + \mathbf{P}(\emptyset) = 1 + \mathbf{P}(\emptyset),$$

and this shows that the probability of the empty event is 0:

$$\mathbf{P}(\emptyset) = 0.$$

As another example, consider three disjoint events A_1 , A_2 , and A_3 . We can use the additivity axiom for two disjoint events repeatedly, to obtain

$$\begin{aligned}\mathbf{P}(A_1 \cup A_2 \cup A_3) &= \mathbf{P}(A_1 \cup (A_2 \cup A_3)) \\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2 \cup A_3) \\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3).\end{aligned}$$

Proceeding similarly, we obtain that the probability of the union of finitely many disjoint events is always equal to the sum of the probabilities of these events. More such properties will be considered shortly.

Discrete Models

Here is an illustration of how to construct a probability law starting from some common sense assumptions about a model.

Example 1.2. Consider an experiment involving a single coin toss. There are two possible outcomes, heads (H) and tails (T). The sample space is $\Omega = \{H, T\}$, and the events are

$$\{H, T\}, \{H\}, \{T\}, \emptyset.$$

If the coin is fair, i.e., if we believe that heads and tails are “equally likely,” we should assign equal probabilities to the two possible outcomes and specify that $\mathbf{P}(\{H\}) = \mathbf{P}(\{T\}) = 0.5$. The additivity axiom implies that

$$\mathbf{P}(\{H, T\}) = \mathbf{P}(\{H\}) + \mathbf{P}(\{T\}) = 1,$$

which is consistent with the normalization axiom. Thus, the probability law is given by

$$\mathbf{P}(\{H, T\}) = 1, \quad \mathbf{P}(\{H\}) = 0.5, \quad \mathbf{P}(\{T\}) = 0.5, \quad \mathbf{P}(\emptyset) = 0,$$

and satisfies all three axioms.

Consider another experiment involving three coin tosses. The outcome will now be a 3-long string of heads or tails. The sample space is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

We assume that each possible outcome has the same probability of $1/8$. Let us construct a probability law that satisfies the three axioms. Consider, as an example, the event

$$A = \{\text{exactly 2 heads occur}\} = \{HHT, HTH, THH\}.$$

Using additivity, the probability of A is the sum of the probabilities of its elements:

$$\begin{aligned}\mathbf{P}(\{HHT, HTH, THH\}) &= \mathbf{P}(\{HHT\}) + \mathbf{P}(\{HTH\}) + \mathbf{P}(\{THH\}) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \quad \text{[Speech bubble icon]} \\ &= \frac{3}{8}.\end{aligned}$$

Similarly, the probability of any event is equal to $1/8$ times the number of possible outcomes contained in the event. This defines a probability law that satisfies the three axioms.

By using the additivity axiom and by generalizing the reasoning in the preceding example, we reach the following conclusion.

Discrete Probability Law

If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. In particular, the probability of any event $\{s_1, s_2, \dots, s_n\}$ is the sum of the probabilities of its elements:

$$\mathbf{P}(\{s_1, s_2, \dots, s_n\}) = \mathbf{P}(s_1) + \mathbf{P}(s_2) + \dots + \mathbf{P}(s_n).$$



Note that we are using here the simpler notation $\mathbf{P}(s_i)$ to denote the probability of the event $\{s_i\}$, instead of the more precise $\mathbf{P}(\{s_i\})$. This convention will be used throughout the remainder of the book.

In the special case where the probabilities $\mathbf{P}(s_1), \dots, \mathbf{P}(s_n)$ are all the same (by necessity equal to $1/n$, in view of the normalization axiom), we obtain the following.

Discrete Uniform Probability Law

If the sample space consists of n possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{n}.$$

Let us provide a few more examples of sample spaces and probability laws.

Example 1.3. Consider the experiment of rolling a pair of 4-sided dice (cf. Fig. 1.4). We assume the dice are fair, and we interpret this assumption to mean that each of the sixteen possible outcomes [pairs (i, j) , with $i, j = 1, 2, 3, 4$], has the same probability of $1/16$. To calculate the probability of an event, we must count the number of elements of the event and divide by 16 (the total number of possible

outcomes). Here are some event probabilities calculated in this way:

$$\mathbf{P}(\{\text{the sum of the rolls is even}\}) = 8/16 = 1/2,$$

$$\mathbf{P}(\{\text{the sum of the rolls is odd}\}) = 8/16 = 1/2,$$

$$\mathbf{P}(\{\text{the first roll is equal to the second}\}) = 4/16 = 1/4,$$

$$\mathbf{P}(\{\text{the first roll is larger than the second}\}) = 6/16 = 3/8,$$

$$\mathbf{P}(\{\text{at least one roll is equal to 4}\}) = 7/16.$$

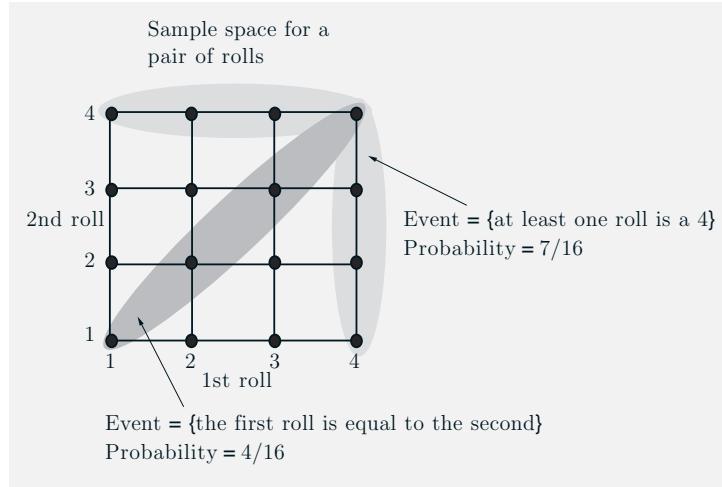


Figure 1.4: Various events in the experiment of rolling a pair of 4-sided dice, and their probabilities, calculated according to the discrete uniform law.

Continuous Models

Probabilistic models with continuous sample spaces differ from their discrete counterparts in that the probabilities of the single-element events may not be sufficient to characterize the probability law. This is illustrated in the following examples, which also indicate how to generalize the uniform probability law to the case of a continuous sample space.

Example 1.4. A wheel of fortune is continuously calibrated from 0 to 1, so the possible outcomes of an experiment consisting of a single spin are the numbers in the interval $\Omega = [0, 1]$. Assuming a fair wheel, it is appropriate to consider all outcomes equally likely, but what is the probability of the event consisting of a single element? It cannot be positive, because then, using the additivity axiom, it would follow that events with a sufficiently large number of elements would have

probability larger than 1. Therefore, the probability of any event that consists of a single element must be 0. 

In this example, it makes sense to assign probability $b - a$ to any subinterval $[a, b]$ of $[0, 1]$, and to calculate the probability of a more complicated set by evaluating its “length.”[†] This assignment satisfies the three probability axioms and qualifies as a legitimate probability law.

Example 1.5. Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour, with all pairs of delays being equally likely. The first to arrive will wait for 15 minutes and will leave if the other has not yet arrived. What is the probability that they will meet?

Let us use as sample space the unit square, whose elements are the possible pairs of delays for the two of them. Our interpretation of “equally likely” pairs of delays is to let the probability of a subset of Ω be equal to its area. This probability law satisfies the three probability axioms. The event that Romeo and Juliet will meet is the shaded region in Fig. 1.5, and its probability is calculated to be $7/16$.

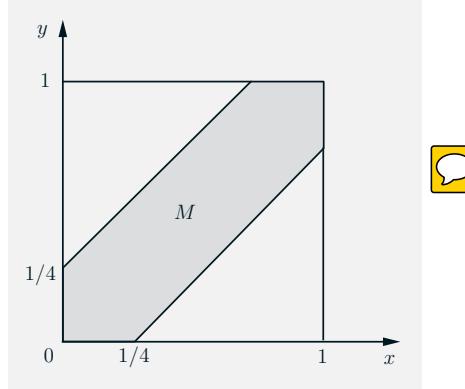


Figure 1.5: The event M that Romeo and Juliet will arrive within 15 minutes of each other (cf. Example 1.5) is

$$M = \{(x, y) \mid |x - y| \leq 1/4, 0 \leq x \leq 1, 0 \leq y \leq 1\},$$

and is shaded in the figure. The area of M is 1 minus the area of the two unshaded triangles, or $1 - (3/4) \cdot (3/4) = 7/16$. Thus, the probability of meeting is $7/16$.

[†] The “length” of a subset S of $[0, 1]$ is the integral $\int_S dt$, which is defined, for “nice” sets S , in the usual calculus sense. For unusual sets, this integral may not be well defined mathematically, but such issues belong to a more advanced treatment of the subject. Incidentally, the legitimacy of using length as a probability law hinges on the fact that the unit interval has an uncountably infinite number of elements. Indeed, if the unit interval had a countable number of elements, with each element having zero probability, the additivity axiom would imply that the whole interval has zero probability, which would contradict the normalization axiom.

Properties of Probability Laws

Probability laws have a number of properties, which can be deduced from the axioms. Some of them are summarized below.

Some Properties of Probability Laws

Consider a probability law, and let A , B , and C be events.

- (a) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.
- (b) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.
- (c) $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.
- (d) $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

These properties, and other similar ones, can be visualized and verified graphically using Venn diagrams, as in Fig. 1.6. Note that property (c) can be generalized as follows:

$$\mathbf{P}(A_1 \cup A_2 \cup \dots \cup A_n) \leq \sum_{i=1}^n \mathbf{P}(A_i). \quad \text{💡}$$

To see this, we apply property (c) to the sets A_1 and $A_2 \cup \dots \cup A_n$, to obtain

$$\mathbf{P}(A_1 \cup A_2 \cup \dots \cup A_n) \leq \mathbf{P}(A_1) + \mathbf{P}(A_2 \cup \dots \cup A_n).$$

We also apply property (c) to the sets A_2 and $A_3 \cup \dots \cup A_n$, to obtain

$$\mathbf{P}(A_2 \cup \dots \cup A_n) \leq \mathbf{P}(A_2) + \mathbf{P}(A_3 \cup \dots \cup A_n).$$

We continue similarly, and finally add.

Models and Reality

The framework of probability theory can be used to analyze uncertainty in a wide variety of physical contexts. Typically, this involves two distinct stages.

- (a) In the first stage, we construct a probabilistic model, by specifying a probability law on a suitably defined sample space. There are no hard rules to guide this step, other than the requirement that the probability law conform to the three axioms. Reasonable people may disagree on which model best represents reality. In many cases, one may even want to use a somewhat “incorrect” model, if it is simpler than the “correct” one or allows for tractable calculations. This is consistent with common practice in science

and engineering, where the choice of a model often involves a tradeoff between accuracy, simplicity, and tractability. Sometimes, a model is chosen on the basis of historical data or past outcomes of similar experiments, using methods from the field of **statistics**.

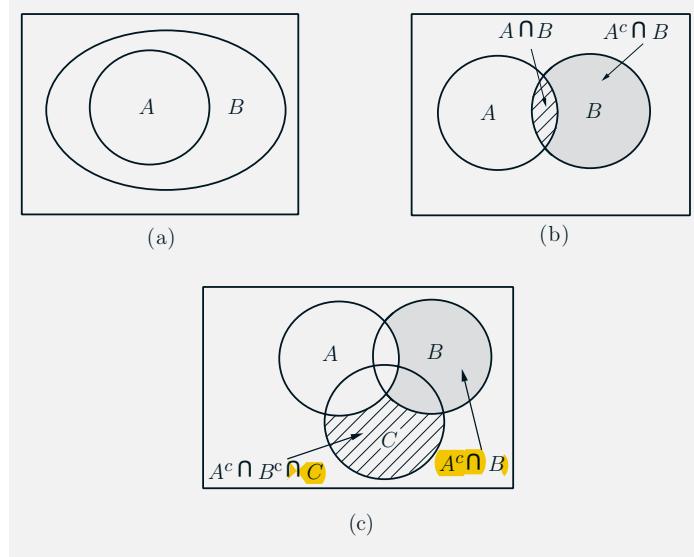


Figure 1.6: Visualization and verification of various properties of probability laws using Venn diagrams. If $A \subset B$, then B is the union of the two disjoint events A and $A^c \cap B$; see diagram (a). Therefore, by the additivity axiom, we have

$$\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) \geq \mathbf{P}(A),$$

where the inequality follows from the nonnegativity axiom, and verifies property (a).

From diagram (b), we can express the events $A \cup B$ and B as unions of disjoint events:

$$A \cup B = A \cup (A^c \cap B), \quad B = (A \cap B) \cup (A^c \cap B).$$

Using the additivity axiom, we have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B), \quad \mathbf{P}(B) = \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B).$$

Subtracting the second equality from the first and rearranging terms, we obtain $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$, verifying property (b). Using also the fact $\mathbf{P}(A \cap B) \geq 0$ (the nonnegativity axiom), we obtain $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$, verifying property (c).

From diagram (c), we see that the event $A \cup B \cup C$ can be expressed as a union of three disjoint events:

$$A \cup B \cup C = A \cup (A^c \cap B) \cup (A^c \cap B^c \cap C),$$

so property (d) follows as a consequence of the additivity axiom.

- (b) In the second stage, we work within a fully specified probabilistic model and derive the probabilities of certain events, or deduce some interesting properties. While the first stage entails the often open-ended task of connecting the real world with mathematics, the second one is tightly regulated by the rules of ordinary logic and the axioms of probability. Difficulties may arise in the latter if some required calculations are complex, or if a probability law is specified in an indirect fashion. Even so, there is no room for ambiguity: all conceivable questions have precise answers and it is only a matter of developing the skill to arrive at them.

Probability theory is full of “paradoxes” in which different calculation methods seem to give different answers to the same question. Invariably though, these apparent inconsistencies turn out to reflect poorly specified or ambiguous probabilistic models. An example, **Bertrand’s paradox**, is shown in Fig. 1.7.

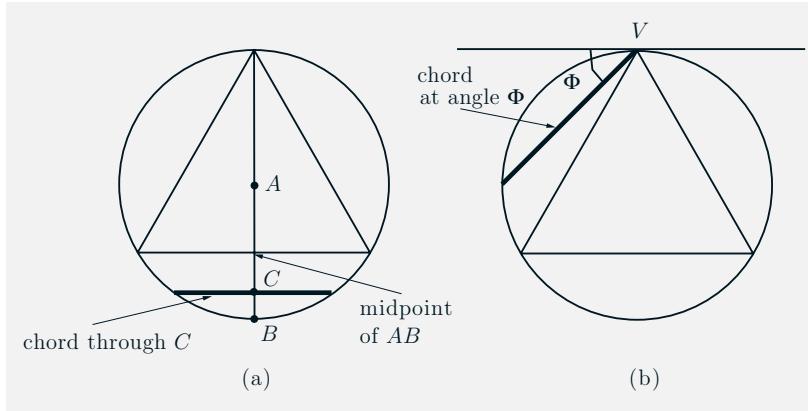


Figure 1.7: This example, presented by L. F. Bertrand in 1889, illustrates the need to specify unambiguously a probabilistic model. Consider a circle and an equilateral triangle inscribed in the circle. What is the probability that the length of a randomly chosen chord of the circle is greater than the side of the triangle? The answer here depends on the precise meaning of “randomly chosen.” The two methods illustrated in parts (a) and (b) of the figure lead to contradictory results.

In (a), we take a radius of the circle, such as AB , and we choose a point C on that radius, with all points being equally likely. We then draw the chord through C that is orthogonal to AB . From elementary geometry, AB intersects the triangle at the midpoint of AB , so the probability that the length of the chord is greater than the side is $1/2$.

In (b), we take a point on the circle, such as the vertex V , we draw the tangent to the circle through V , and we draw a line through V that forms a random angle Φ with the tangent, with all angles being equally likely. We consider the chord obtained by the intersection of this line with the circle. From elementary geometry, the length of the chord is greater than the side of the triangle if Φ is between $\pi/3$ and $2\pi/3$. Since Φ takes values between 0 and π , the probability that the length of the chord is greater than the side is $1/3$.

A Brief History of Probability

- **B.C. Games** of chance were popular in ancient Greece and Rome, but no scientific development of the subject took place, possibly because the number system used by the Greeks did not facilitate algebraic calculations. The development of probability based on sound scientific analysis had to await the development of the modern arithmetic system by the Hindus and the Arabs in the second half of the first millennium, as well as the flood of scientific ideas generated by the Renaissance.
- 16th century. Girolamo Cardano, a colorful and controversial Italian mathematician, publishes the first book describing correct methods for calculating probabilities in games of chance such as dice and cards.
- 17th century. A correspondence between Fermat and Pascal touches upon several interesting probability questions, and motivates further study in the field.
- 18th century. Jacob Bernoulli studies repeated coin tossing and introduces the first law of large numbers, which lays a foundation for linking theoretical probability concepts and empirical fact. Several mathematicians, such as Daniel Bernoulli, Leibnitz, Bayes, and Lagrange, make important contributions to probability theory and its use in analyzing real-world phenomena. De Moivre introduces the normal distribution and proves the first form of the central limit theorem.
- 19th century. Laplace publishes an influential book that establishes the importance of probability as a quantitative field and contains many original contributions, including a more general version of the central limit theorem. Legendre and Gauss apply probability to astronomical predictions, using the method of least squares, thus pointing the way to a vast range of applications. Poisson publishes an influential book with many original contributions, including the Poisson distribution. Chebyshev, and his students Markov and Lyapunov, study limit theorems and raise the standards of mathematical rigor in the field. Throughout this period, probability theory is largely viewed as a natural science, its primary goal being the explanation of physical phenomena. Consistently with this goal, probabilities are mainly interpreted as limits of relative frequencies in the context of repeatable experiments.
- 20th century. Relative frequency is abandoned as the conceptual foundation of probability theory in favor of the axiomatic system that is universally used now. Similar to other branches of mathematics, the development of probability theory from the axioms relies only on logical correctness, regardless of its relevance to physical phenomena. Nonetheless, probability theory is used pervasively in science and engineering because of its ability to describe and interpret most types of uncertain phenomena in the real world.

1.3 CONDITIONAL PROBABILITY

Conditional probability provides us with a way to reason about the outcome of an experiment, based on **partial information**. Here are some examples of situations we have in mind:

- (a) In an experiment involving two successive rolls of a die, you are told that the sum of the two rolls is 9. How likely is it that the first roll was a 6?
- (b) In a word guessing game, the first letter of the word is a “t”. What is the likelihood that the second letter is an “h”?
- (c) How likely is it that a person has a disease given that a medical test was negative?
- (d) A spot shows up on a radar screen. How likely is it that it corresponds to an aircraft?

In more precise terms, given an experiment, a corresponding sample space, and a probability law, suppose that we know that the outcome is within some given event B . We wish to quantify the likelihood that the outcome also belongs to some other given event A . We thus seek to construct a new probability law, which takes into account the available knowledge and which, for any event A , gives us the **conditional probability of A given B** , denoted by $\mathbf{P}(A|B)$.

We would like the conditional probabilities $\mathbf{P}(A|B)$ of different events A to constitute a legitimate probability law, that satisfies the probability axioms. The conditional probabilities should also be consistent with our intuition in important special cases, e.g., when all possible outcomes of the experiment are equally likely. For example, suppose that all six possible outcomes of a fair die roll are equally likely. If we are told that the outcome is even, we are left with only three possible outcomes, namely, 2, 4, and 6. These three outcomes were equally likely to start with, and so they should remain equally likely given the additional knowledge that the outcome was even. Thus, it is reasonable to let

$$\mathbf{P}(\text{the outcome is 6} | \text{the outcome is even}) = \frac{1}{3}.$$

This argument suggests that an appropriate definition of conditional probability when all outcomes are equally likely, is given by

$$\mathbf{P}(A|B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

Generalizing the argument, we introduce the following definition of conditional probability:

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

where we assume that $\mathbf{P}(B) > 0$; the conditional probability is undefined if the conditioning event has zero probability. In words, out of the total probability of the elements of B , $\mathbf{P}(A|B)$ is the fraction that is assigned to possible outcomes that also belong to A .

Conditional Probabilities Specify a Probability Law

For a fixed event B , it can be verified that the conditional probabilities $\mathbf{P}(A|B)$ form a legitimate probability law that satisfies the three axioms. Indeed, non-negativity is clear. Furthermore,

$$\mathbf{P}(\Omega|B) = \frac{\mathbf{P}(\Omega \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B)}{\mathbf{P}(B)} = 1,$$

and the normalization axiom is also satisfied. To verify the additivity axiom, we write for any two disjoint events A_1 and A_2 ,

$$\begin{aligned} \mathbf{P}(A_1 \cup A_2|B) &= \frac{\mathbf{P}((A_1 \cup A_2) \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}((A_1 \cap B) \cup (A_2 \cap B))}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1 \cap B) + \mathbf{P}(A_2 \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)} + \frac{\mathbf{P}(A_2 \cap B)}{\mathbf{P}(B)} \\ &= \mathbf{P}(A_1|B) + \mathbf{P}(A_2|B), \end{aligned}$$

where for the third equality, we used the fact that $A_1 \cap B$ and $A_2 \cap B$ are disjoint sets, and the additivity axiom for the (unconditional) probability law. The argument for a countable collection of disjoint sets is similar.

Since conditional probabilities constitute a legitimate probability law, all general properties of probability laws remain valid. For example, a fact such as $\mathbf{P}(A \cup C) \leq \mathbf{P}(A) + \mathbf{P}(C)$ translates to the new fact

$$\mathbf{P}(A \cup C|B) \leq \mathbf{P}(A|B) + \mathbf{P}(C|B).$$

Let us also note that since we have $\mathbf{P}(B|B) = \mathbf{P}(B)/\mathbf{P}(B) = 1$, all of the conditional probability is concentrated on B . Thus, we might as well discard all possible outcomes outside B and treat the conditional probabilities as a probability law defined on the new universe B .

Let us summarize the conclusions reached so far.

Properties of Conditional Probability

- The conditional probability of an event A , given an event B with $\mathbf{P}(B) > 0$, is defined by

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

and specifies a new (conditional) probability law on the same sample space Ω . In particular, all known properties of probability laws remain valid for conditional probability laws.

- Conditional probabilities can also be viewed as a probability law on a new universe B , because all of the conditional probability is concentrated on B .
- In the case where the possible outcomes are **finitely many and equally likely**, we have

$$\mathbf{P}(A | B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

Example 1.6. We toss a fair coin three successive times. We wish to find the conditional probability $\mathbf{P}(A | B)$ when A and B are the events

$$A = \{\text{more heads than tails come up}\}, \quad B = \{\text{1st toss is a head}\}.$$

The sample space consists of eight sequences,

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

which we assume to be equally likely. The event B consists of the four elements HHH, HHT, HTH, HTT , so its probability is

$$\mathbf{P}(B) = \frac{4}{8}.$$

The event $A \cap B$ consists of the three elements HHH, HHT, HTH , so its probability is

$$\mathbf{P}(A \cap B) = \frac{3}{8}.$$

Thus, the conditional probability $\mathbf{P}(A | B)$ is

$$\mathbf{P}(A | B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{3/8}{4/8} = \frac{3}{4}.$$

Because all possible outcomes are equally likely here, we can also compute $\mathbf{P}(A|B)$ using a shortcut. We can bypass the calculation of $\mathbf{P}(B)$ and $\mathbf{P}(A \cap B)$, and simply divide the number of elements shared by A and B (which is 3) with the number of elements of B (which is 4), to obtain the same result 3/4.

Example 1.7. A fair 4-sided die is rolled twice and we assume that all sixteen possible outcomes are equally likely. Let X and Y be the result of the 1st and the 2nd roll, respectively. We wish to determine the conditional probability $\mathbf{P}(A|B)$, where

$$A = \{\max(X, Y) = m\}, \quad B = \{\min(X, Y) = 2\},$$

and m takes each of the values 1, 2, 3, 4.

As in the preceding example, we can first determine the probabilities $\mathbf{P}(A \cap B)$ and $\mathbf{P}(B)$ by counting the number of elements of $A \cap B$ and B , respectively, and dividing by 16. Alternatively, we can directly divide the number of elements of $A \cap B$ with the number of elements of B ; see Fig. 1.8.

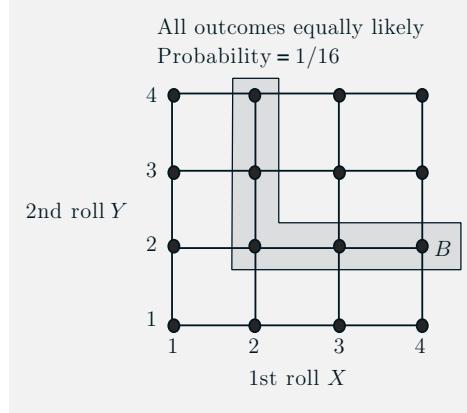


Figure 1.8: Sample space of an experiment involving two rolls of a 4-sided die. (cf. Example 1.7). The conditioning event $B = \{\min(X, Y) = 2\}$ consists of the 5-element shaded set. The set $A = \{\max(X, Y) = m\}$ shares with B two elements if $m = 3$ or $m = 4$, one element if $m = 2$, and no element if $m = 1$. Thus, we have

$$\mathbf{P}(\{\max(X, Y) = m\} | B) = \begin{cases} 2/5, & \text{if } m = 3 \text{ or } m = 4, \\ 1/5, & \text{if } m = 2, \\ 0, & \text{if } m = 1. \end{cases}$$

Example 1.8. A conservative design team, call it C, and an innovative design team, call it N, are asked to separately design a new product within a month. From past experience we know that:

- (a) The probability that team C is successful is 2/3.
- (b) The probability that team N is successful is 1/2.

(c) The probability that at least one team is successful is $3/4$.

Assuming that exactly one successful design is produced, what is the probability that it was designed by team N? 

There are four possible outcomes here, corresponding to the four combinations of success and failure of the two teams:

SS : both succeed,
 SF : C succeeds, N fails,

FF : both fail,
 FS : C fails, N succeeds.

We are given that the probabilities of these outcomes satisfy

$$\mathbf{P}(SS) + \mathbf{P}(SF) = \frac{2}{3}, \quad \mathbf{P}(SS) + \mathbf{P}(FS) = \frac{1}{2}, \quad \mathbf{P}(SS) + \mathbf{P}(SF) + \mathbf{P}(FS) = \frac{3}{4}.$$

From these relations, together with the normalization equation

$$\mathbf{P}(SS) + \mathbf{P}(SF) + \mathbf{P}(FS) + \mathbf{P}(FF) = 1,$$

we can obtain the probabilities of all the outcomes:

$$\mathbf{P}(SS) = \frac{5}{12}, \quad \mathbf{P}(SF) = \frac{1}{4}, \quad \mathbf{P}(FS) = \frac{1}{12}, \quad \mathbf{P}(FF) = \frac{1}{4}.$$

The desired conditional probability is

$$\mathbf{P}(FS \mid \{SF, FS\}) = \frac{\frac{1}{12}}{\frac{1}{4} + \frac{1}{12}} = \frac{1}{4}$$

Using Conditional Probability for Modeling

When constructing probabilistic models for experiments that have a sequential character, it is often natural and convenient to first specify conditional probabilities and then use them to determine unconditional probabilities. The rule $\mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A \mid B)$, which is a restatement of the definition of conditional probability, is often helpful in this process.

Example 1.9. Radar Detection. If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of false alarm (a false indication of aircraft presence), and the probability of missed detection (nothing registers, even though an aircraft is present)?

A sequential representation of the experiment is appropriate here, as shown in Fig. 1.9. Let A and B be the events

$$A = \{\text{an aircraft is present}\},$$

$$B = \{\text{the radar registers an aircraft presence}\},$$

and consider also their complements

$$A^c = \{\text{an aircraft is not present}\},$$

$$B^c = \{\text{the radar does not register an aircraft presence}\}.$$

The given probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.9. Each possible outcome corresponds to a leaf of the tree, and its probability is equal to the product of the probabilities associated with the branches in a path from the root to the corresponding leaf. The desired probabilities of false alarm and missed detection are

$$\mathbf{P}(\text{false alarm}) = \mathbf{P}(A^c \cap B) = \mathbf{P}(A^c)\mathbf{P}(B | A^c) = 0.95 \cdot 0.10 = 0.095,$$

$$\mathbf{P}(\text{missed detection}) = \mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c | A) = 0.05 \cdot 0.01 = 0.0005.$$

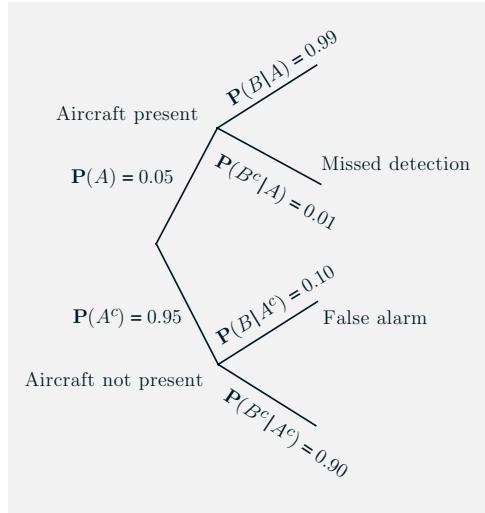


Figure 1.9: Sequential description of the experiment for the radar detection problem in Example 1.9.

Extending the preceding example, we have a general rule for calculating various probabilities in conjunction with a tree-based sequential description of an experiment. In particular:

- We set up the tree so that an event of interest is associated with a leaf. We view the occurrence of the event as a sequence of steps, namely, the traversals of the branches along the path from the root to the leaf.
- We record the conditional probabilities associated with the branches of the tree.
- We obtain the probability of a leaf by multiplying the probabilities recorded along the corresponding path of the tree.

In mathematical terms, we are dealing with an event A which occurs if and only if each one of several events A_1, \dots, A_n has occurred, i.e., $A = A_1 \cap A_2 \cap \dots \cap A_n$. The occurrence of A is viewed as an occurrence of A_1 , followed by the occurrence of A_2 , then of A_3 , etc., and it is visualized as a path with n branches, corresponding to the events A_1, \dots, A_n . The probability of A is given by the following rule (see also Fig. 1.10).

Multiplication Rule

Assuming that all of the conditioning events have positive probability, we have

$$\mathbf{P}(\cap_{i=1}^n A_i) = \mathbf{P}(A_1)\mathbf{P}(A_2 | A_1)\mathbf{P}(A_3 | A_1 \cap A_2) \cdots \mathbf{P}(A_n | \cap_{i=1}^{n-1} A_i).$$

The multiplication rule can be verified by writing

$$\mathbf{P}(\cap_{i=1}^n A_i) = \mathbf{P}(A_1) \cdot \frac{\mathbf{P}(A_1 \cap A_2)}{\mathbf{P}(A_1)} \cdot \frac{\mathbf{P}(A_1 \cap A_2 \cap A_3)}{\mathbf{P}(A_1 \cap A_2)} \cdots \frac{\mathbf{P}(\cap_{i=1}^n A_i)}{\mathbf{P}(\cap_{i=1}^{n-1} A_i)},$$

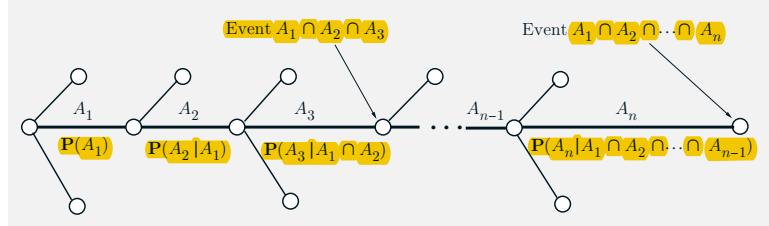


Figure 1.10: Visualization of the multiplication rule. The intersection event $A = A_1 \cap A_2 \cap \dots \cap A_n$ is associated with a particular path on a tree that describes the experiment. We associate the branches of this path with the events A_1, \dots, A_n , and we record next to the branches the corresponding conditional probabilities.

The final node of the path corresponds to the intersection event A , and its probability is obtained by multiplying the conditional probabilities recorded along the branches of the path

$$\mathbf{P}(A_1 \cap A_2 \cap \dots \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 | A_1) \cdots \mathbf{P}(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1}).$$

Note that any intermediate node along the path also corresponds to some intersection event and its probability is obtained by multiplying the corresponding conditional probabilities up to that node. For example, the event $A_1 \cap A_2 \cap A_3$ corresponds to the node shown in the figure, and its probability is

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 | A_1)\mathbf{P}(A_3 | A_1 \cap A_2).$$

and by using the definition of conditional probability to rewrite the right-hand side above as

$$\mathbf{P}(A_1)\mathbf{P}(A_2 | A_1)\mathbf{P}(A_3 | A_1 \cap A_2) \cdots \mathbf{P}(A_n | \cap_{i=1}^{n-1} A_i).$$

For the case of just two events, A_1 and A_2 , the multiplication rule is simply the definition of conditional probability.

Example 1.10. Three cards are drawn from an ordinary 52-card deck without replacement (drawn cards are not placed back in the deck). We wish to find the probability that none of the three cards is a heart. We assume that at each step, each one of the remaining cards is equally likely to be picked. By symmetry, this implies that every triplet of cards is equally likely to be drawn. A cumbersome approach, that we will not use, is to count the number of all card triplets that do not include a heart, and divide it with the number of all possible card triplets. Instead, we use a sequential description of the experiment in conjunction with the multiplication rule (cf. Fig. 1.11).

Define the events

$$A_i = \{\text{the } i\text{th card is not a heart}\}, \quad i = 1, 2, 3.$$

We will calculate $\mathbf{P}(A_1 \cap A_2 \cap A_3)$, the probability that none of the three cards is a heart, using the multiplication rule

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 | A_1)\mathbf{P}(A_3 | A_1 \cap A_2).$$

We have

$$\mathbf{P}(A_1) = \frac{39}{52},$$

since there are 39 cards that are not hearts in the 52-card deck. Given that the first card is not a heart, we are left with 51 cards, 38 of which are not hearts, and

$$\mathbf{P}(A_2 | A_1) = \frac{38}{51}.$$

Finally, given that the first two cards drawn are not hearts, there are 37 cards which are not hearts in the remaining 50-card deck, and

$$\mathbf{P}(A_3 | A_1 \cap A_2) = \frac{37}{50}.$$

These probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.11. The desired probability is now obtained by multiplying the probabilities recorded along the corresponding path of the tree:

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50}.$$

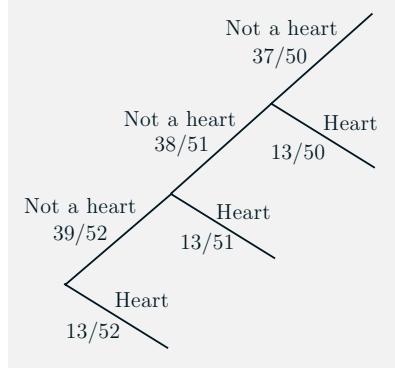


Figure 1.11: Sequential description of the experiment in the 3-card selection problem of Example 1.10.

Note that once the probabilities are recorded along the tree, the probability of several other events can be similarly calculated. For example,

$$\mathbf{P}(\text{1st is not a heart and 2nd is a heart}) = \frac{39}{52} \cdot \frac{13}{51},$$

$$\mathbf{P}(\text{1st two are not hearts and 3rd is a heart}) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{13}{50}.$$

Example 1.11. A class consisting of 4 graduate and 12 undergraduate students is randomly divided into 4 groups of 4. What is the probability that each group includes a graduate student? We interpret “randomly” to mean that given the assignment of some students to certain slots, any of the remaining students is equally likely to be assigned to any of the remaining slots. We then calculate the desired probability using the multiplication rule, based on the sequential description shown in Fig. 1.12. Let us denote the four graduate students by 1, 2, 3, 4, and consider the events

$$\begin{aligned} A_1 &= \{\text{students 1 and 2 are in different groups}\}, \\ A_2 &= \{\text{students 1, 2, and 3 are in different groups}\}, \\ A_3 &= \{\text{students 1, 2, 3, and 4 are in different groups}\}. \end{aligned}$$

We will calculate $\mathbf{P}(A_3)$ using the multiplication rule:

$$\mathbf{P}(A_3) = \mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 | A_1)\mathbf{P}(A_3 | A_1 \cap A_2).$$

We have

$$\mathbf{P}(A_1) = \frac{12}{15},$$

since there are 12 student slots in groups other than the one of student 1, and there are 15 student slots overall, excluding student 1. Similarly,

$$\mathbf{P}(A_2 | A_1) = \frac{8}{14},$$

since there are 8 student slots in groups other than those of students 1 and 2, and there are 14 student slots, excluding students 1 and 2. Also,

$$\mathbf{P}(A_3 | A_1 \cap A_2) = \frac{4}{13},$$

since there are 4 student slots in groups other than those of students 1, 2, and 3, and there are 13 student slots, excluding students 1, 2, and 3. Thus, the desired probability is

$$\frac{12}{15} \cdot \frac{8}{14} \cdot \frac{4}{13},$$

and is obtained by multiplying the conditional probabilities along the corresponding path of the tree of Fig. 1.12.

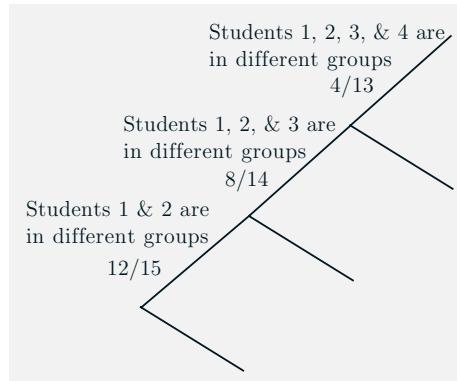


Figure 1.12: Sequential description of the experiment in the student problem of Example 1.11.

Example 1.12. The Monty Hall Problem. This is a much discussed puzzle, based on an old American game show. You are told that a prize is equally likely to be found behind any one of three closed doors in front of you. You point to one of the doors. A friend opens for you one of the remaining two doors, after making sure that the prize is not behind it. At this point, you can stick to your initial choice, or switch to the other unopened door. You win the prize if it lies behind your final choice of a door. Consider the following strategies:

- (a) Stick to your initial choice.
- (b) Switch to the other unopened door.
- (c) You first point to door 1. If door 2 is opened, you do not switch. If door 3 is opened, you switch.

Which is the best strategy? To answer the question, let us calculate the probability of winning under each of the three strategies.

Under the strategy of no switching, your initial choice will determine whether you win or not, and the probability of winning is 1/3. This is because the prize is equally likely to be behind each door.

Under the strategy of switching, if the prize is behind the initially chosen door (probability 1/3), you do not win. If it is not (probability 2/3), and given that

another door without a prize has been opened for you, you will get to the winning door once you switch. Thus, the probability of winning is now $2/3$, so (b) is a better strategy than (a).

Consider now strategy (c). Under this strategy, there is insufficient information for determining the probability of winning. The answer depends on the way that your friend chooses which door to open. Let us consider two possibilities.

Suppose that if the prize is behind door 1, your friend always chooses to open door 2. (If the prize is behind door 2 or 3, your friend has no choice.) If the prize is behind door 1, your friend opens door 2, you do not switch, and you win. If the prize is behind door 2, your friend opens door 3, you switch, and you win. If the prize is behind door 3, your friend opens door 2, you do not switch, and you lose. Thus, the probability of winning is $2/3$, so strategy (c) in this case is as good as strategy (b).

Suppose now that if the prize is behind door 1, your friend is equally likely to open either door 2 or 3. If the prize is behind door 1 (probability $1/3$), and if your friend opens door 2 (probability $1/2$), you do not switch and you win (probability $1/6$). But if your friend opens door 3, you switch and you lose. If the prize is behind door 2, your friend opens door 3, you switch, and you win (probability $1/3$). If the prize is behind door 3, your friend opens door 2, you do not switch and you lose. Thus, the probability of winning is $1/6 + 1/3 = 1/2$, so strategy (c) in this case is inferior to strategy (b).

1.4 TOTAL PROBABILITY THEOREM AND BAYES' RULE

In this section, we explore some applications of conditional probability. We start with the following theorem, which is often useful for computing the probabilities of various events, using a “divide-and-conquer” approach.

Total Probability Theorem

Let A_1, \dots, A_n be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events A_1, \dots, A_n) and assume that $\mathbf{P}(A_i) > 0$, for all i . Then, for any event B , we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1 \cap B) + \dots + \mathbf{P}(A_n \cap B) \\ &= \mathbf{P}(A_1)\mathbf{P}(B | A_1) + \dots + \mathbf{P}(A_n)\mathbf{P}(B | A_n).\end{aligned}$$

The theorem is visualized and proved in Fig. 1.13. Intuitively, we are partitioning the sample space into a number of scenarios (events) A_i . Then, the probability that B occurs is a weighted average of its conditional probability under each scenario, where each scenario is weighted according to its (unconditional) probability. One of the uses of the theorem is to compute the probability of various events B for which the conditional probabilities $\mathbf{P}(B | A_i)$ are known or

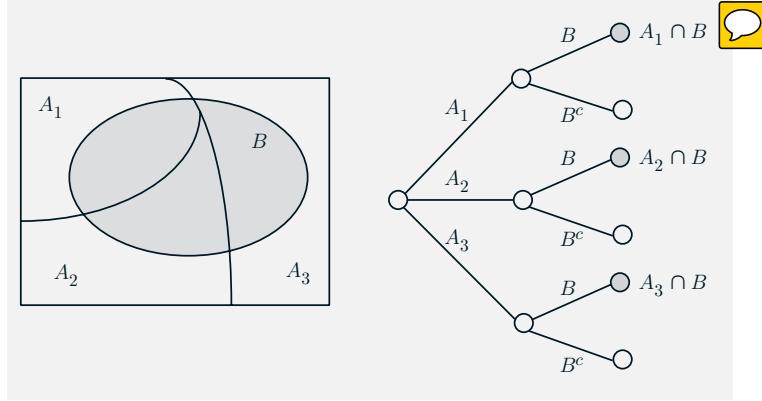


Figure 1.13: Visualization and verification of the total probability theorem. The events A_1, \dots, A_n form a partition of the sample space, so the event B can be decomposed into the disjoint union of its intersections $A_i \cap B$ with the sets A_i , i.e.,

$$B = (A_1 \cap B) \cup \dots \cup (A_n \cap B).$$

Using the additivity axiom, it follows that

$$\mathbf{P}(B) = \mathbf{P}(A_1 \cap B) + \dots + \mathbf{P}(A_n \cap B).$$

Since, by the definition of conditional probability, we have

$$\mathbf{P}(A_i \cap B) = \mathbf{P}(A_i) \mathbf{P}(B | A_i),$$

the preceding equality yields

$$\mathbf{P}(B) = \mathbf{P}(A_1) \mathbf{P}(B | A_1) + \dots + \mathbf{P}(A_n) \mathbf{P}(B | A_n).$$

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability of the leaf $A_i \cap B$ is the product $\mathbf{P}(A_i) \mathbf{P}(B | A_i)$ of the probabilities along the path leading to that leaf. The event B consists of the three highlighted leaves and $\mathbf{P}(B)$ is obtained by adding their probabilities.

easy to derive. The key is to choose appropriately the partition A_1, \dots, A_n , and this choice is often suggested by the problem structure. Here are some examples.

Example 1.13. You enter a chess tournament where your probability of winning a game is 0.3 against half the players (call them type 1), 0.4 against a quarter of the players (call them type 2), and 0.5 against the remaining quarter of the players (call them type 3). You play a game against a randomly chosen opponent. What is the probability of winning?

Let A_i be the event of playing with an opponent of type i . We have

$$\mathbf{P}(A_1) = 0.5, \quad \mathbf{P}(A_2) = 0.25, \quad \mathbf{P}(A_3) = 0.25.$$

Let also B be the event of winning. We have

$$\mathbf{P}(B | A_1) = 0.3, \quad \mathbf{P}(B | A_2) = 0.4, \quad \mathbf{P}(B | A_3) = 0.5.$$

Thus, by the total probability theorem, the probability of winning is

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1)\mathbf{P}(B|A_1) + \mathbf{P}(A_2)\mathbf{P}(B|A_2) + \mathbf{P}(A_3)\mathbf{P}(B|A_3) \\ &= 0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5 \\ &= 0.375.\end{aligned}$$

Example 1.14. You roll a fair four-sided die. If the result is 1 or 2, you roll once more but otherwise, you stop. What is the probability that the sum total of your rolls is at least 4?

Let A_i be the event that the result of first roll is i , and note that $\mathbf{P}(A_i) = 1/4$ for each i . Let B be the event that the sum total is at least 4. Given the event A_1 , the sum total will be at least 4 if the second roll results in 3 or 4, which happens with probability 1/2. Similarly, given the event A_2 , the sum total will be at least 4 if the second roll results in 2, 3, or 4, which happens with probability 3/4. Also, given the event A_3 , you stop and the sum total remains below 4. Therefore,

$$\mathbf{P}(B|A_1) = \frac{1}{2}, \quad \mathbf{P}(B|A_2) = \frac{3}{4}, \quad \mathbf{P}(B|A_3) = 0, \quad \mathbf{P}(B|A_4) = 1.$$

By the total probability theorem,

$$\mathbf{P}(B) = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 = \frac{9}{16}.$$

The total probability theorem can be applied repeatedly to calculate probabilities in experiments that have a sequential character, as shown in the following example.

Example 1.15. Alice is taking a probability class and at the end of each week she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.4 (or 0.6, respectively). Alice is (by default) up-to-date when she starts the class. What is the probability that she is up-to-date after three weeks?

Let U_i and B_i be the events that Alice is up-to-date or behind, respectively, after i weeks. According to the total probability theorem, the desired probability $\mathbf{P}(U_3)$ is given by

$$\mathbf{P}(U_3) = \mathbf{P}(U_2)\mathbf{P}(U_3|U_2) + \mathbf{P}(B_2)\mathbf{P}(U_3|B_2) = \mathbf{P}(U_2) \cdot 0.8 + \mathbf{P}(B_2) \cdot 0.4.$$

The probabilities $\mathbf{P}(U_2)$ and $\mathbf{P}(B_2)$ can also be calculated using the total probability theorem:

$$\mathbf{P}(U_2) = \mathbf{P}(U_1)\mathbf{P}(U_2|U_1) + \mathbf{P}(B_1)\mathbf{P}(U_2|B_1) = \mathbf{P}(U_1) \cdot 0.8 + \mathbf{P}(B_1) \cdot 0.4,$$

$$\mathbf{P}(B_2) = \mathbf{P}(U_1)\mathbf{P}(B_2 | U_1) + \mathbf{P}(B_1)\mathbf{P}(B_2 | B_1) = \mathbf{P}(U_1) \cdot 0.2 + \mathbf{P}(B_1) \cdot 0.6.$$

Finally, since Alice starts her class up-to-date, we have

$$\mathbf{P}(U_1) = 0.8, \quad \mathbf{P}(B_1) = 0.2.$$

We can now combine the preceding three equations to obtain

$$\mathbf{P}(U_2) = 0.8 \cdot 0.8 + 0.2 \cdot 0.4 = 0.72,$$

$$\mathbf{P}(B_2) = 0.8 \cdot 0.2 + 0.2 \cdot 0.6 = 0.28,$$

and by using the above probabilities in the formula for $\mathbf{P}(U_3)$:

$$\mathbf{P}(U_3) = 0.72 \cdot 0.8 + 0.28 \cdot 0.4 = 0.688.$$

Note that we could have calculated the desired probability $\mathbf{P}(U_3)$ by constructing a tree description of the experiment, by calculating the probability of every element of U_3 using the multiplication rule on the tree, and by adding. However, there are cases where the calculation based on the total probability theorem is more convenient. For example, suppose we are interested in the probability $\mathbf{P}(U_{20})$ that Alice is up-to-date after 20 weeks. Calculating this probability using the multiplication rule is very cumbersome, because the tree representing the experiment is 20-stages deep and has 2^{20} leaves. On the other hand, with a computer, a sequential calculation using the total probability formulas

$$\mathbf{P}(U_{i+1}) = \mathbf{P}(U_i) \cdot 0.8 + \mathbf{P}(B_i) \cdot 0.4,$$

$$\mathbf{P}(B_{i+1}) = \mathbf{P}(U_i) \cdot 0.2 + \mathbf{P}(B_i) \cdot 0.6,$$

and the initial conditions $\mathbf{P}(U_1) = 0.8$, $\mathbf{P}(B_1) = 0.2$, is very simple.

Inference and Bayes' Rule

The total probability theorem is often used in conjunction with the following celebrated theorem, which relates conditional probabilities of the form $\mathbf{P}(A | B)$ with conditional probabilities of the form $\mathbf{P}(B | A)$, in which the order of the conditioning is reversed.

Bayes' Rule

Let A_1, A_2, \dots, A_n be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$, for all i . Then, for any event B such that $\mathbf{P}(B) > 0$, we have

$$\begin{aligned} \mathbf{P}(A_i | B) &= \frac{\mathbf{P}(A_i)\mathbf{P}(B | A_i)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_i)\mathbf{P}(B | A_i)}{\mathbf{P}(A_1)\mathbf{P}(B | A_1) + \dots + \mathbf{P}(A_n)\mathbf{P}(B | A_n)}. \end{aligned}$$

To verify Bayes' rule, note that $\mathbf{P}(A_i)\mathbf{P}(B|A_i)$ and $\mathbf{P}(A_i|B)\mathbf{P}(B)$ are equal, because they are both equal to $\mathbf{P}(A_i \cap B)$. This yields the first equality. The second equality follows from the first by using the total probability theorem to rewrite $\mathbf{P}(B)$.

Bayes' rule is often used for **inference**. There are a number of "causes" that may result in a certain "effect." We observe the effect, and we wish to infer the cause. The events A_1, \dots, A_n are associated with the causes and the event B represents the effect. The probability $\mathbf{P}(B|A_i)$ that the effect will be observed when the cause A_i is present amounts to a probabilistic model of the cause-effect relation (cf. Fig. 1.14). Given that the effect B has been observed, we wish to evaluate the probability $\mathbf{P}(A_i|B)$ that the cause A_i is present.

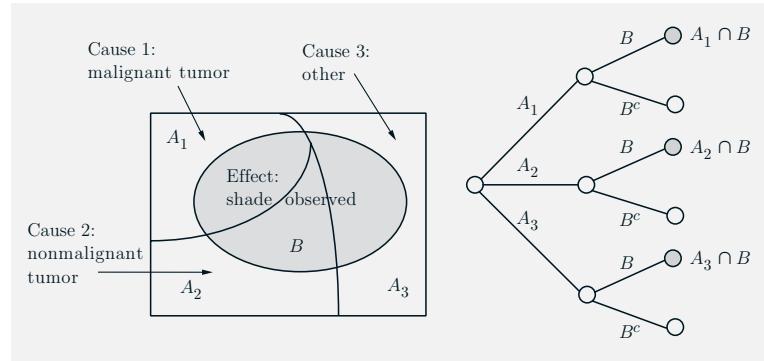


Figure 1.14: An example of the inference context that is implicit in Bayes' rule. We observe a shade in a person's X-ray (this is event B , the "effect") and we want to estimate the likelihood of three mutually exclusive and collectively exhaustive potential causes: cause 1 (event A_1) is that there is a malignant tumor, cause 2 (event A_2) is that there is a nonmalignant tumor, and cause 3 (event A_3) corresponds to reasons other than a tumor. We assume that we know the probabilities $\mathbf{P}(A_i)$ and $\mathbf{P}(B|A_i)$, $i = 1, 2, 3$. Given that we see a shade (event B occurs), Bayes' rule gives the conditional probabilities of the various causes as

$$\mathbf{P}(A_i|B) = \frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\mathbf{P}(A_1)\mathbf{P}(B|A_1) + \mathbf{P}(A_2)\mathbf{P}(B|A_2) + \mathbf{P}(A_3)\mathbf{P}(B|A_3)}, \quad i = 1, 2, 3.$$

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability $\mathbf{P}(A_1|B)$ of a malignant tumor is the probability of the first highlighted leaf, which is $\mathbf{P}(A_1 \cap B)$, divided by the total probability of the highlighted leaves, which is $\mathbf{P}(B)$.

Example 1.16. Let us return to the radar detection problem of Example 1.9 and Fig. 1.9. Let

$$\begin{aligned} A &= \{\text{an aircraft is present}\}, \\ B &= \{\text{the radar registers an aircraft presence}\}. \end{aligned}$$

We are given that

$$\mathbf{P}(A) = 0.05, \quad \mathbf{P}(B | A) = 0.99, \quad \mathbf{P}(B | A^c) = 0.1.$$

Applying Bayes' rule, with $A_1 = A$ and $A_2 = A^c$, we obtain

$$\begin{aligned} \mathbf{P}(\text{aircraft present} | \text{radar registers}) &= \mathbf{P}(A | B) \\ &= \frac{\mathbf{P}(A)\mathbf{P}(B | A)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A)\mathbf{P}(B | A)}{\mathbf{P}(A)\mathbf{P}(B | A) + \mathbf{P}(A^c)\mathbf{P}(B | A^c)} \\ &= \frac{0.05 \cdot 0.99}{0.05 \cdot 0.99 + 0.95 \cdot 0.1} \\ &\approx 0.3426. \end{aligned}$$

Example 1.17. Let us return to the chess problem of Example 1.13. Here, A_i is the event of getting an opponent of type i , and

$$\mathbf{P}(A_1) = 0.5, \quad \mathbf{P}(A_2) = 0.25, \quad \mathbf{P}(A_3) = 0.25.$$

Also, B is the event of winning, and

$$\mathbf{P}(B | A_1) = 0.3, \quad \mathbf{P}(B | A_2) = 0.4, \quad \mathbf{P}(B | A_3) = 0.5.$$

Suppose that you win. What is the probability $\mathbf{P}(A_1 | B)$ that you had an opponent of type 1?

Using Bayes' rule, we have

$$\begin{aligned} \mathbf{P}(A_1 | B) &= \frac{\mathbf{P}(A_1)\mathbf{P}(B | A_1)}{\mathbf{P}(A_1)\mathbf{P}(B | A_1) + \mathbf{P}(A_2)\mathbf{P}(B | A_2) + \mathbf{P}(A_3)\mathbf{P}(B | A_3)} \\ &= \frac{0.5 \cdot 0.3}{0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5} \\ &= 0.4. \end{aligned}$$

Example 1.18. The False-Positive Puzzle. A test for a certain rare disease is assumed to be correct 95% of the time: if a person has the disease, the test results are positive with probability 0.95, and if the person does not have the disease, the test results are negative with probability 0.95. A random person drawn from a certain population has probability 0.001 of having the disease. Given that the person just tested positive, what is the probability of having the disease?

If A is the event that the person has the disease, and B is the event that the test results are positive, the desired probability, $\mathbf{P}(A | B)$, is

$$\begin{aligned} \mathbf{P}(A | B) &= \frac{\mathbf{P}(A)\mathbf{P}(B | A)}{\mathbf{P}(A)\mathbf{P}(B | A) + \mathbf{P}(A^c)\mathbf{P}(B | A^c)} \\ &= \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05} \\ &= 0.0187. \end{aligned}$$

Note that even though the test was assumed to be fairly accurate, a person who has tested positive is still very unlikely (less than 2%) to have the disease. According to *The Economist* (February 20th, 1999), 80% of those questioned at a leading American hospital substantially missed the correct answer to a question of this type. Most of them said that the probability that the person has the disease is 0.95!

1.5 INDEPENDENCE

We have introduced the conditional probability $\mathbf{P}(A|B)$ to capture the partial information that event B provides about event A . An interesting and important special case arises when the occurrence of B provides no such information and does not alter the probability that A has occurred, i.e.,

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

When the above equality holds, we say that A is **independent** of B . Note that by the definition $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$, this is equivalent to

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

We adopt this latter relation as the definition of independence because it can be used even if $\mathbf{P}(B) = 0$, in which case $\mathbf{P}(A|B)$ is undefined. The symmetry of this relation also implies that independence is a symmetric property; that is, if A is independent of B , then B is independent of A , and we can unambiguously say that A and B are **independent events**.

Independence is often easy to grasp intuitively. For example, if the occurrence of two events is governed by distinct and noninteracting physical processes, such events will turn out to be independent. On the other hand, independence is not easily visualized in terms of the sample space. A common first thought is that two events are independent if they are disjoint, but in fact the opposite is true: two disjoint events A and B with $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$ are never independent, since their intersection $A \cap B$ is empty and has probability 0.

Example 1.19. Consider an experiment involving two successive rolls of a 4-sided die in which all 16 possible outcomes are equally likely and have probability 1/16.

(a) Are the events

$$A_i = \{\text{1st roll results in } i\}, \quad B_j = \{\text{2nd roll results in } j\},$$



independent? We have

$$\mathbf{P}(A_i \cap B_j) = \mathbf{P}(\text{the result of the two rolls is } (i, j)) = \frac{1}{16},$$

$$\mathbf{P}(A_i) = \frac{\text{number of elements of } A_i}{\text{total number of possible outcomes}} = \frac{4}{16},$$

$$\mathbf{P}(B_j) = \frac{\text{number of elements of } B_j}{\text{total number of possible outcomes}} = \frac{4}{16}.$$

We observe that $\mathbf{P}(A_i \cap B_j) = \mathbf{P}(A_i)\mathbf{P}(B_j)$, and the independence of A_i and B_j is verified. Thus, our choice of the discrete uniform probability law (which might have seemed arbitrary) models the independence of the two rolls.

- (b) Are the events

$$A = \{\text{1st roll is a 1}\}, \quad B = \{\text{sum of the two rolls is a 5}\},$$

independent? The answer here is not quite obvious. We have

$$\mathbf{P}(A \cap B) = \mathbf{P}(\text{the result of the two rolls is (1,4)}) = \frac{1}{16},$$

and also

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{\text{total number of possible outcomes}} = \frac{4}{16}.$$

The event B consists of the outcomes (1,4), (2,3), (3,2), and (4,1), and

$$\mathbf{P}(B) = \frac{\text{number of elements of } B}{\text{total number of possible outcomes}} = \frac{4}{16}.$$

Thus, we see that $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, and the events A and B are **independent**.

- (c) Are the events

$$A = \{\text{maximum of the two rolls is 2}\}, \quad B = \{\text{minimum of the two rolls is 2}\},$$

independent? Intuitively, the answer is “no” because the minimum of the two rolls conveys some information about the maximum. For example, if the minimum is 2, the maximum cannot be 1. More precisely, to verify that A and B are not independent, we calculate

$$\mathbf{P}(A \cap B) = \mathbf{P}(\text{the result of the two rolls is (2,2)}) = \frac{1}{16},$$

and also

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{\text{total number of possible outcomes}} = \frac{3}{16},$$

$$\mathbf{P}(B) = \frac{\text{number of elements of } B}{\text{total number of possible outcomes}} = \frac{5}{16}.$$

We have $\mathbf{P}(A)\mathbf{P}(B) = 15/(16)^2$, so that $\mathbf{P}(A \cap B) \neq \mathbf{P}(A)\mathbf{P}(B)$, and A and B are **not independent**.

Conditional Independence

We noted earlier that the conditional probabilities of events, conditioned on a particular event, form a legitimate probability law. We can thus talk about independence of various events with respect to this conditional law. In particular, given an event C , the events A and B are called **conditionally independent** if

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

To derive an alternative characterization of conditional independence, we use the definition of the conditional probability and the multiplication rule, to write

$$\begin{aligned}\mathbf{P}(A \cap B | C) &= \frac{\mathbf{P}(A \cap B \cap C)}{\mathbf{P}(C)} \\ &= \frac{\mathbf{P}(C)\mathbf{P}(B | C)\mathbf{P}(A | B \cap C)}{\mathbf{P}(C)} \\ &= \mathbf{P}(B | C)\mathbf{P}(A | B \cap C).\end{aligned}$$

We now compare the preceding two expressions, and after eliminating the common factor $\mathbf{P}(B | C)$, assumed nonzero, we see that conditional independence is the same as the condition

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

In words, this relation states that if C is known to have occurred, the additional knowledge that B also occurred does not change the probability of A .

Interestingly, independence of two events A and B with respect to the unconditional probability law, does not imply conditional independence, and vice versa, as illustrated by the next two examples.

Example 1.20. Consider two independent fair coin tosses, in which all four possible outcomes are equally likely. Let

$$\begin{aligned}H_1 &= \{\text{1st toss is a head}\}, \\ H_2 &= \{\text{2nd toss is a head}\}, \\ D &= \{\text{the two tosses have different results}\}.\end{aligned}$$

The events H_1 and H_2 are (unconditionally) independent. But

$$\mathbf{P}(H_1 | D) = \frac{1}{2}, \quad \mathbf{P}(H_2 | D) = \frac{1}{2}, \quad \mathbf{P}(H_1 \cap H_2 | D) = 0,$$

so that $\mathbf{P}(H_1 \cap H_2 | D) \neq \mathbf{P}(H_1 | D)\mathbf{P}(H_2 | D)$, and H_1, H_2 are not conditionally independent.

Example 1.21. There are two coins, a blue and a red one. We choose one of the two at random, each being chosen with probability $1/2$, and proceed with two independent tosses. The coins are biased: with the blue coin, the probability of heads in any given toss is 0.99 , whereas for the red coin it is 0.01 .

Let B be the event that the blue coin was selected. Let also H_i be the event that the i th toss resulted in heads. Given the choice of a coin, the events H_1 and H_2 are independent, because of our assumption of independent tosses. Thus,

$$\mathbf{P}(H_1 \cap H_2 | B) = \mathbf{P}(H_1 | B)\mathbf{P}(H_2 | B) = 0.99 \cdot 0.99.$$

On the other hand, the events H_1 and H_2 are not independent. Intuitively, if we are told that the first toss resulted in heads, this leads us to suspect that the blue coin was selected, in which case, we expect the second toss to also result in heads. Mathematically, we use the total probability theorem to obtain

$$\mathbf{P}(H_1) = \mathbf{P}(B)\mathbf{P}(H_1 | B) + \mathbf{P}(B^c)\mathbf{P}(H_1 | B^c) = \frac{1}{2} \cdot 0.99 + \frac{1}{2} \cdot 0.01 = \frac{1}{2},$$

as should be expected from symmetry considerations. Similarly, we have $\mathbf{P}(H_2) = 1/2$. Now notice that

$$\begin{aligned} \mathbf{P}(H_1 \cap H_2) &= \mathbf{P}(B)\mathbf{P}(H_1 \cap H_2 | B) + \mathbf{P}(B^c)\mathbf{P}(H_1 \cap H_2 | B^c) \\ &= \frac{1}{2} \cdot 0.99 \cdot 0.99 + \frac{1}{2} \cdot 0.01 \cdot 0.01 \approx \frac{1}{2}. \end{aligned}$$

Thus, $\mathbf{P}(H_1 \cap H_2) \neq \mathbf{P}(H_1)\mathbf{P}(H_2)$, and the events H_1 and H_2 are dependent, even though they are conditionally independent given B .

As mentioned earlier, if A and B are independent, the occurrence of B does not provide any new information on the probability of A occurring. It is then intuitive that the non-occurrence of B should also provide no information on the probability of A . Indeed, it can be verified that if A and B are independent, the same holds true for A and B^c (see the end-of-chapter problems).

We now summarize.

Independence

- Two events A and B are said to be **independent** if

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

If in addition, $\mathbf{P}(B) > 0$, independence is equivalent to the condition

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

- If A and B are independent, so are A and B^c .
- Two events A and B are said to be **conditionally independent**, given another event C with $\mathbf{P}(C) > 0$, if

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

If in addition, $\mathbf{P}(B \cap C) > 0$, conditional independence is equivalent to the condition

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

- Independence does not imply conditional independence, and vice versa.

Independence of a Collection of Events

The definition of independence can be extended to multiple events.

Definition of Independence of Several Events

We say that the events A_1, A_2, \dots, A_n are **independent** if

$$\mathbf{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i), \quad \text{for every subset } S \text{ of } \{1, 2, \dots, n\}.$$

For the case of three events, A_1 , A_2 , and A_3 , independence amounts to satisfying the four conditions

$$\begin{aligned} \mathbf{P}(A_1 \cap A_2) &= \mathbf{P}(A_1)\mathbf{P}(A_2), \\ \mathbf{P}(A_1 \cap A_3) &= \mathbf{P}(A_1)\mathbf{P}(A_3), \\ \mathbf{P}(A_2 \cap A_3) &= \mathbf{P}(A_2)\mathbf{P}(A_3), \\ \mathbf{P}(A_1 \cap A_2 \cap A_3) &= \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3). \end{aligned}$$

The first three conditions simply assert that any two events are independent, a property known as **pairwise independence**. But the fourth condition is also important and does not follow from the first three. Conversely, the fourth condition does not imply the first three; see the two examples that follow.

Example 1.22. Pairwise Independence does not Imply Independence.
Consider two independent fair coin tosses, and the following events:

$$\begin{aligned}
 H_1 &= \{1\text{st toss is a head}\}, \\
 H_2 &= \{2\text{nd toss is a head}\}, \\
 D &= \{\text{the two tosses have different results}\}.
 \end{aligned}$$

The events H_1 and H_2 are independent, by definition. To see that H_1 and D are independent, we note that

$$\mathbf{P}(D | H_1) = \frac{\mathbf{P}(H_1 \cap D)}{\mathbf{P}(H_1)} = \frac{1/4}{1/2} = \frac{1}{2} = \mathbf{P}(D).$$

Similarly, H_2 and D are independent. On the other hand, we have

$$\mathbf{P}(H_1 \cap H_2 \cap D) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathbf{P}(H_1)\mathbf{P}(H_2)\mathbf{P}(D),$$

and these three events are not independent.

Example 1.23. The Equality $\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3)$ is not Enough for Independence. Consider two independent rolls of a fair six-sided die, and the following events:

$$\begin{aligned}
 A &= \{1\text{st roll is 1, 2, or 3}\}, \\
 B &= \{1\text{st roll is 3, 4, or 5}\}, \\
 C &= \{\text{the sum of the two rolls is 9}\}.
 \end{aligned}$$

We have

$$\begin{aligned}
 \mathbf{P}(A \cap B) &= \frac{1}{6} \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbf{P}(A)\mathbf{P}(B), \\
 \mathbf{P}(A \cap C) &= \frac{1}{36} \neq \frac{1}{2} \cdot \frac{4}{36} = \mathbf{P}(A)\mathbf{P}(C), \\
 \mathbf{P}(B \cap C) &= \frac{1}{12} \neq \frac{1}{2} \cdot \frac{4}{36} = \mathbf{P}(B)\mathbf{P}(C).
 \end{aligned}$$

Thus the three events A , B , and C are not independent, and indeed no two of these events are independent. On the other hand, we have

$$\mathbf{P}(A \cap B \cap C) = \frac{1}{36} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{36} = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).$$

The intuition behind the independence of a collection of events is analogous to the case of two events. Independence means that the occurrence or non-occurrence of **any number** of the events from that collection carries no information on the remaining events or their complements. For example, if the events A_1, A_2, A_3, A_4 are independent, one obtains relations such as

$$\mathbf{P}(A_1 \cup A_2 | A_3 \cap A_4) = \mathbf{P}(A_1 \cup A_2)$$

or

$$\mathbf{P}(A_1 \cup A_2^c | A_3^c \cap A_4) = \mathbf{P}(A_1 \cup A_2^c);$$

see the end-of-chapter problems.

Reliability

In probabilistic models of complex systems involving several components, it is often convenient to assume that the behaviors of the components are uncoupled (independent). This typically simplifies the calculations and the analysis, as illustrated in the following example.

Example 1.24. Network Connectivity. A computer network connects two nodes A and B through intermediate nodes C, D, E, F, as shown in Fig. 1.15(a). For every pair of directly connected nodes, say i and j , there is a given probability p_{ij} that the link from i to j is up. We assume that link failures are independent of each other. What is the probability that there is a path connecting A and B in which all links are up?

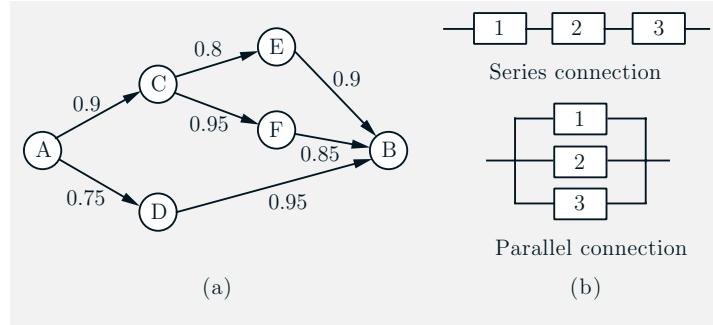


Figure 1.15: (a) Network for Example 1.24. The number next to each link indicates the probability that the link is up. (b) Series and parallel connections of three components in a reliability problem.

This is a typical problem of assessing the reliability of a system consisting of components that can fail independently. Such a system can often be divided into subsystems, where each subsystem consists in turn of several components that are connected either in **series** or in **parallel**; see Fig. 1.15(b).

Let a subsystem consist of components $1, 2, \dots, m$, and let p_i be the probability that component i is up (“succeeds”). Then, a series subsystem succeeds if **all** of its components are up, so its probability of success is the product of the probabilities of success of the corresponding components, i.e.,

$$\mathbf{P}(\text{series subsystem succeeds}) = p_1 p_2 \cdots p_m.$$

A parallel subsystem succeeds if **any one** of its components succeeds, so its probability of failure is the product of the probabilities of failure of the corresponding components, i.e.,

$$\begin{aligned} \mathbf{P}(\text{parallel subsystem succeeds}) &= 1 - \mathbf{P}(\text{parallel subsystem fails}) \\ &= 1 - (1 - p_1)(1 - p_2) \cdots (1 - p_m). \end{aligned}$$

Returning now to the network of Fig. 1.15(a), we can calculate the probability of success (a path from A to B is available) sequentially, using the preceding formulas, and starting from the end. Let us use the notation $X \rightarrow Y$ to denote the event that there is a (possibly indirect) connection from node X to node Y . Then,

$$\begin{aligned}\mathbf{P}(C \rightarrow B) &= 1 - (1 - \mathbf{P}(C \rightarrow E \text{ and } E \rightarrow B))(1 - \mathbf{P}(C \rightarrow F \text{ and } F \rightarrow B)) \\ &= 1 - (1 - p_{CE}p_{EB})(1 - p_{CF}p_{FB}) \\ &= 1 - (1 - 0.8 \cdot 0.9)(1 - 0.95 \cdot 0.85) \\ &= 0.946,\end{aligned}$$

$$\mathbf{P}(A \rightarrow C \text{ and } C \rightarrow B) = \mathbf{P}(A \rightarrow C)\mathbf{P}(C \rightarrow B) = 0.9 \cdot 0.946 = 0.851,$$

$$\mathbf{P}(A \rightarrow D \text{ and } D \rightarrow B) = \mathbf{P}(A \rightarrow D)\mathbf{P}(D \rightarrow B) = 0.75 \cdot 0.95 = 0.712,$$

and finally we obtain the desired probability

$$\begin{aligned}\mathbf{P}(A \rightarrow B) &= 1 - (1 - \mathbf{P}(A \rightarrow C \text{ and } C \rightarrow B))(1 - \mathbf{P}(A \rightarrow D \text{ and } D \rightarrow B)) \\ &= 1 - (1 - 0.851)(1 - 0.712) \\ &= 0.957.\end{aligned}$$

Independent Trials and the Binomial Probabilities

If an experiment involves a sequence of independent but identical stages, we say that we have a sequence of **independent trials**. In the special case where there are only two possible results at each stage, we say that we have a sequence of **Bernoulli trials**. The two possible results can be anything, e.g., “it rains” or “it doesn’t rain,” but we will often think in terms of coin tosses and refer to the two results as “heads” (H) and “tails” (T).

Consider an experiment that consists of n independent tosses of a coin, in which the probability of heads is p , where p is some number between 0 and 1. In this context, independence means that the events A_1, A_2, \dots, A_n are independent, where $A_i = \{i\text{th toss is a head}\}$.

We can visualize independent Bernoulli trials by means of a sequential description, as shown in Fig. 1.16 for the case where $n = 3$. The conditional probability of any toss being a head, conditioned on the results of any preceding tosses is p , because of independence. Thus, by multiplying the conditional probabilities along the corresponding path of the tree, we see that any particular outcome (3-long sequence of heads and tails) that involves k heads and $3 - k$ tails has probability $p^k(1 - p)^{3-k}$. This formula extends to the case of a general number n of tosses. We obtain that the probability of any particular n -long sequence that contains k heads and $n - k$ tails is $p^k(1 - p)^{n-k}$, for all k from 0 to n .

Let us now consider the probability

$$p(k) = \mathbf{P}(k \text{ heads come up in an } n\text{-toss sequence}),$$

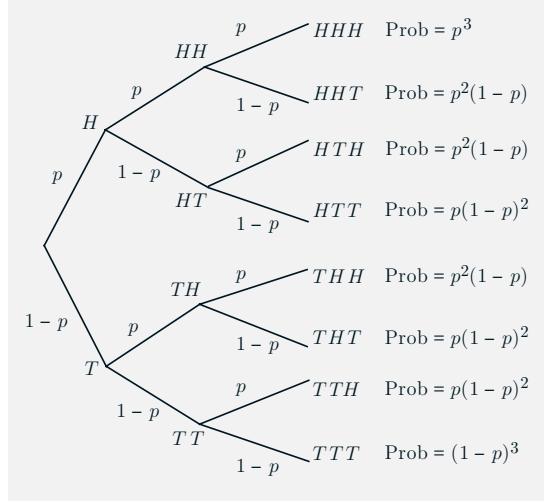


Figure 1.16: Sequential description of an experiment involving three independent tosses of a coin. Along the branches of the tree, we record the corresponding conditional probabilities, and by the multiplication rule, the probability of obtaining a particular 3-toss sequence is calculated by multiplying the probabilities recorded along the corresponding path of the tree.

which will play an important role later. We showed above that the probability of any given sequence that contains k heads is $p^k(1-p)^{n-k}$, so we have

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{💡}$$

where we use the notation

$$\binom{n}{k} = \text{number of distinct } n\text{-toss sequences that contain } k \text{ heads.}$$

The numbers $\binom{n}{k}$ (called “ n choose k ”) are known as the **binomial coefficients**, while the probabilities $p(k)$ are known as the **binomial probabilities**. Using a counting argument, to be given in Section 1.6, we can show that

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad k = 0, 1, \dots, n,$$

where for any positive integer i we have

$$i! = 1 \cdot 2 \cdots (i-1) \cdot i,$$

and, by convention, $0! = 1$. An alternative verification is sketched in the end-of-chapter problems. Note that the binomial probabilities $p(k)$ must add to 1, thus

showing the **binomial formula**

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Example 1.25. Grade of Service. An internet service provider has installed c modems to serve the needs of a population of n customers. It is estimated that at a given time, each customer will need a connection with probability p , independently of the others. What is the probability that there are more customers needing a connection than there are modems?

Here we are interested in the probability that more than c customers simultaneously need a connection. It is equal to

$$\sum_{k=c+1}^n p(k),$$

where

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

are the binomial probabilities. For instance, if $n = 100$, $p = 0.1$, and $c = 15$, the desired probability turns out to be 0.0399.

This example is typical of problems of sizing a facility to serve the needs of a homogeneous population, consisting of independently acting customers. The problem is to select the facility size to achieve a certain threshold probability (sometimes called **grade of service**) that no user is left unserved.

1.6 COUNTING

The calculation of probabilities often involves counting the number of outcomes in various events. We have already seen two contexts where such counting arises.

- (a) When the sample space Ω has a finite number of equally likely outcomes, so that the discrete uniform probability law applies. Then, the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{\text{number of elements of } \Omega},$$

and involves counting the elements of A and of Ω .

- (b) When we want to calculate the probability of an event A with a finite number of equally likely outcomes, **each of which has an already known probability p** . Then the probability of A is given by

$$\mathbf{P}(A) = p \cdot (\text{number of elements of } A),$$

and involves counting the number of elements of A . An example of this type is the calculation of the probability of k heads in n coin tosses (the binomial probabilities). We saw in the preceding section that the probability of each distinct sequence involving k heads is easily obtained, but the calculation of the number of all such sequences is somewhat intricate, as will be seen shortly.

While counting is in principle straightforward, it is frequently challenging; the art of counting constitutes a large portion of the field of **combinatorics**. In this section, we present the basic principle of counting and apply it to a number of situations that are often encountered in probabilistic models.

The Counting Principle

The counting principle is based on a divide-and-conquer approach, whereby the counting is broken down into stages through the use of a tree. For example, consider an experiment that consists of two consecutive stages. The possible results of the first stage are a_1, a_2, \dots, a_m ; the possible results of the second stage are b_1, b_2, \dots, b_n . Then, the possible results of the two-stage experiment are all possible **ordered** pairs (a_i, b_j) , $i = 1, \dots, m$, $j = 1, \dots, n$. Note that the number of such ordered pairs is equal to mn . This observation can be generalized as follows (see also Fig. 1.17).

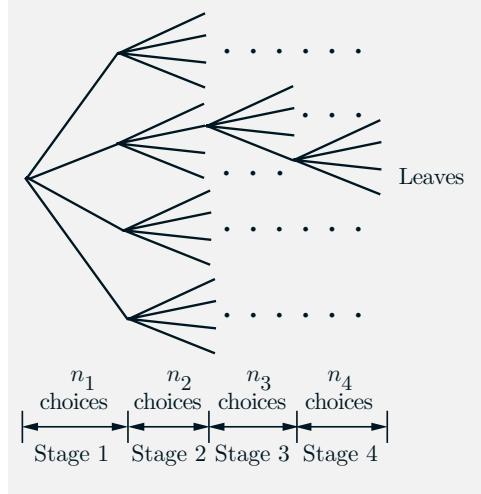


Figure 1.17: Illustration of the basic counting principle. The counting is carried out in r stages ($r = 4$ in the figure). The first stage has n_1 possible results. For every possible result of the first $i - 1$ stages, there are n_i possible results at the i th stage. The number of leaves is $n_1 n_2 \dots n_r$. This is the desired count.

The Counting Principle

Consider a process that consists of r stages. Suppose that:

- (a) There are n_1 possible results at the first stage.
- (b) For every possible result of the first stage, there are n_2 possible results at the second stage.
- (c) More generally, for any possible results of the first $i - 1$ stages, there are n_i possible results at the i th stage.

Then, the total number of possible results of the r -stage process is

$$n_1 n_2 \cdots n_r.$$

Example 1.26. The Number of Telephone Numbers. A telephone number is a 7-digit sequence, but the first digit has to be different from 0 or 1. How many distinct telephone numbers are there? We can visualize the choice of a sequence as a sequential process, where we select one digit at a time. We have a total of 7 stages, and a choice of one out of 10 elements at each stage, except for the first stage where we only have 8 choices. Therefore, the answer is

$$8 \cdot \underbrace{10 \cdot 10 \cdots 10}_{6 \text{ times}} = 8 \cdot 10^6.$$

Example 1.27. The Number of Subsets of an n -Element Set. Consider an n -element set $\{s_1, s_2, \dots, s_n\}$. How many subsets does it have (including itself and the empty set)? We can visualize the choice of a subset as a sequential process where we examine one element at a time and decide whether to include it in the set or not. We have a total of n stages, and a binary choice at each stage. Therefore the number of subsets is

$$\underbrace{2 \cdot 2 \cdots 2}_{n \text{ times}} = 2^n.$$

It should be noted that the Counting Principle remains valid even if each first-stage result leads to a different set of potential second-stage results, etc. The only requirement is that the number of possible second-stage results is constant, regardless of the first-stage result. 

In what follows, we will focus primarily on two types of counting arguments that involve the selection of k objects out of a collection of n objects. If the order of selection matters, the selection is called a **permutation**, and otherwise, it is called a **combination**. We will then discuss a more general type of counting, involving a **partition** of a collection of n objects into multiple subsets.

***k*-permutations**

We start with n distinct objects, and let k be some positive integer, with $k \leq n$. We wish to count the number of different ways that we can pick k out of these n objects and arrange them in a sequence, i.e., the number of distinct k -object sequences. We can choose any of the n objects to be the first one. Having chosen the first, there are only $n - 1$ possible choices for the second; given the choice of the first two, there only remain $n - 2$ available objects for the third stage, etc. When we are ready to select the last (the k th) object, we have already chosen $k - 1$ objects, which leaves us with $n - (k - 1)$ choices for the last one. By the Counting Principle, the number of possible sequences, called ***k*-permutations**, is

$$\begin{aligned} n(n - 1) \cdots (n - k + 1) &= \frac{n(n - 1) \cdots (n - k + 1)(n - k) \cdots 2 \cdot 1}{(n - k) \cdots 2 \cdot 1} \\ &= \frac{n!}{(n - k)!}. \end{aligned}$$

In the special case where $k = n$, the number of possible sequences, simply called **permutations**, is

$$n(n - 1)(n - 2) \cdots 2 \cdot 1 = n!.$$

(Let $k = n$ in the formula for the number of k -permutations, and recall the convention $0! = 1$.)

Example 1.28. Let us count the number of words that consist of four distinct letters. This is the problem of counting the number of **4-permutations** of the 26 letters in the alphabet. The desired number is

$$\frac{n!}{(n - k)!} = \frac{26!}{22!} = 26 \cdot 25 \cdot 24 \cdot 23 = 358,800.$$

The count for permutations can be combined with the Counting Principle to solve more complicated counting problems.

Example 1.29. You have n_1 classical music CDs, n_2 rock music CDs, and n_3 country music CDs. In how many different ways can you arrange them so that the CDs of the same type are contiguous?

We break down the problem in two stages, where we first select the order of the CD types, and then the order of the CDs of each type. There are $3!$ ordered sequences of the types of CDs (such as classical/rock/country, rock/country/classical, etc.), and there are $n_1!$ (or $n_2!$, or $n_3!$) permutations of the classical (or rock, or country, respectively) CDs. Thus for each of the $3!$ CD type sequences, there are $n_1! n_2! n_3!$ arrangements of CDs, and the desired total number is $3! n_1! n_2! n_3!$.

Combinations

There are n people and we are interested in forming a committee of k . How many different committees are possible? More abstractly, this is the same as the problem of counting the number of k -element subsets of a given n -element set. Notice that forming a combination is different than forming a k -permutation, because **in a combination there is no ordering of the selected elements**. Thus for example, whereas the 2-permutations of the letters A, B, C, and D are

$$AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC,$$

the combinations of two out of these four letters are

$$AB, AC, AD, BC, BD, CD.$$

(Since the elements of a combination are unordered, BA is not viewed as being distinct from AB.)

To count the number of combinations, we observe that selecting a k -permutation is the same as first selecting a combination of k items and then ordering them. Since there are $k!$ ways of ordering the k selected items, we see that the number $n!/(n - k)!$ of k -permutations is equal to the number of combinations times $k!$. Hence, the number of possible combinations, is equal to

$$\frac{n!}{k! (n - k)!}.$$

Let us now relate the above expression to the binomial coefficient, which was denoted by $\binom{n}{k}$ and was defined in the preceding section as the number of n -toss sequences with k heads. We note that specifying an n -toss sequence with k heads is the same as selecting k elements (those that correspond to heads) out of the n -element set of tosses, i.e., a combination of k out of n objects. Hence, the binomial coefficient is also given by the same formula and we have

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

Example 1.30. The number of combinations of two out of the four letters A, B, C, and D is found by letting $n = 4$ and $k = 2$. It is

$$\binom{4}{2} = \frac{4!}{2! 2!} = 6,$$

consistently with the listing given earlier.

It is worth observing that counting arguments sometimes lead to formulas that are rather difficult to derive algebraically. One example is the **binomial formula**

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = 1,$$

discussed in Section 1.5. In the special case where $p = 1/2$, this formula becomes

$$\sum_{k=0}^n \binom{n}{k} = 2^n,$$

and admits the following simple interpretation. Since $\binom{n}{k}$ is the number of k -element subsets of a given n -element subset, the sum over k of $\binom{n}{k}$ counts the number of subsets of all possible cardinalities. It is therefore equal to the number of all subsets of an n -element set, which is 2^n .

Partitions

Recall that a combination is a choice of k elements out of an n -element set without regard to order. Thus, a combination can be viewed as a partition of the set in two: one part contains k elements and the other contains the remaining $n - k$. We now generalize by considering partitions into more than two subsets.

We are given an n -element set and nonnegative integers n_1, n_2, \dots, n_r , whose sum is equal to n . We consider partitions of the set into r disjoint subsets, with the i th subset containing exactly n_i elements. Let us count in how many ways this can be done.

We form the subsets one at a time. We have $\binom{n}{n_1}$ ways of forming the first subset. Having formed the first subset, we are left with $n - n_1$ elements. We need to choose n_2 of them in order to form the second subset, and we have $\binom{n-n_1}{n_2}$ choices, etc. Using the Counting Principle for this r -stage process, the total number of choices is

$$\binom{n}{n_1} \binom{n-n_1}{n_2} \binom{n-n_1-n_2}{n_3} \cdots \binom{n-n_1-\cdots-n_{r-1}}{n_r},$$

which is equal to

$$\frac{n!}{n_1! (n - n_1)!} \cdot \frac{(n - n_1)!}{n_2! (n - n_1 - n_2)!} \cdots \frac{(n - n_1 - \cdots - n_{r-1})!}{(n - n_1 - \cdots - n_{r-1} - n_r)! n_r!}.$$

We note that several terms cancel and we are left with

$$\frac{n!}{n_1! n_2! \cdots n_r!}.$$

This is called the **multinomial coefficient** and is usually denoted by

$$\binom{n}{n_1, n_2, \dots, n_r}.$$

Example 1.31. Anagrams. How many different words (letter sequences) can be obtained by rearranging the letters in the word TATTOO? There are six positions to be filled by the available letters. Each rearrangement corresponds to a partition of the set of the six positions into a group of size 3 (the positions that get the letter T), a group of size 1 (the position that gets the letter A), and a group of size 2 (the positions that get the letter O). Thus, the desired number is

$$\frac{6!}{1! 2! 3!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6}{1 \cdot 1 \cdot 2 \cdot 1 \cdot 2 \cdot 3} = 60.$$

It is instructive to derive this answer using an alternative argument. (This argument can also be used to rederive the multinomial coefficient formula; see the end-of-chapter problems.) Let us write TATTOO in the form $T_1 A T_2 T_3 O_1 O_2$ pretending for a moment that we are dealing with 6 distinguishable objects. These 6 objects can be rearranged in $6!$ different ways. However, any of the $3!$ possible permutations of T_1 , T_2 , and T_3 , as well as any of the $2!$ possible permutations of O_1 and O_2 , lead to the same word. Thus, when the subscripts are removed, there are only $6!/(3! 2!)$ different words.

Example 1.32. A class consisting of 4 graduate and 12 undergraduate students is randomly divided into four groups of 4. What is the probability that each group includes a graduate student? This is the same as Example 1.11 in Section 1.3, but we will now obtain the answer using a counting argument.

We first determine the nature of the sample space. A typical outcome is a particular way of partitioning the 16 students into four groups of 4. We take the term “randomly” to mean that every possible partition is equally likely, so that the probability question can be reduced to one of counting.

According to our earlier discussion, there are

$$\binom{16}{4, 4, 4, 4} = \frac{16!}{4! 4! 4! 4!}$$

different partitions, and this is the size of the sample space.

Let us now focus on the event that each group contains a graduate student. Generating an outcome with this property can be accomplished in two stages:

- Take the four graduate students and distribute them to the four groups; there are four choices for the group of the first graduate student, three choices for the second, two for the third. Thus, there is a total of $4!$ choices for this stage.
- Take the remaining 12 undergraduate students and distribute them to the four groups (3 students in each). This can be done in

$$\binom{12}{3, 3, 3, 3} = \frac{12!}{3! 3! 3! 3!}$$

different ways.

By the Counting Principle, the event of interest can occur in

$$\frac{4! 12!}{3! 3! 3! 3!}$$

different ways. The probability of this event is

$$\frac{\frac{4! 12!}{3! 3! 3! 3!}}{\frac{16!}{4! 4! 4! 4!}}.$$

After some cancellations, we find that this is equal to

$$\frac{12 \cdot 8 \cdot 4}{15 \cdot 14 \cdot 13},$$

consistent with the answer obtained in Example 1.11.

Here is a summary of all the counting results we have developed.

Summary of Counting Results

- **Permutations** of n objects: $n!$.
 - k -**permutations** of n objects: $n!/(n - k)!$.
 - **Combinations** of k out of n objects: $\binom{n}{k} = \frac{n!}{k! (n - k)!}$.
 - **Partitions** of n objects into r groups, with the i th group having n_i objects:
- $$\binom{n}{n_1, n_2, \dots, n_r} \equiv \frac{n!}{n_1! n_2! \cdots n_r!},$$

1.7 SUMMARY AND DISCUSSION

A probability problem can usually be broken down into a few basic steps:

- (a) The description of the sample space, that is, the set of possible outcomes of a given experiment.
- (b) The (possibly indirect) specification of the probability law (the probability of each event).
- (c) The calculation of probabilities and conditional probabilities of various events of interest.

The probabilities of events must satisfy the nonnegativity, additivity, and normalization axioms. In the important special case where the set of possible outcomes is finite, one can just specify the probability of each outcome and obtain the probability of any event by adding the probabilities of the elements of the event.

Given a probability law, we are often interested in conditional probabilities, which allow us to reason based on partial information about the outcome of the experiment. We can view conditional probabilities as probability laws of a special type, under which only outcomes contained in the conditioning event can have positive conditional probability. Conditional probabilities can be derived from the (unconditional) probability law using the definition $\mathbf{P}(A | B) = \mathbf{P}(A \cap B) / \mathbf{P}(B)$. However, the reverse process is often convenient, that is, first specify some conditional probabilities that are natural for the real situation that we wish to model, and then use them to derive the (unconditional) probability law.

We have illustrated through examples three methods for calculating probabilities:

- (a) The **counting method**. This method applies to the case where the number of possible outcomes is finite, and all outcomes are equally likely. To calculate the probability of an event, we count the number of elements of the event and divide by the number of elements of the sample space.
- (b) The **sequential method**. This method applies when the experiment has a sequential character, and suitable conditional probabilities are specified or calculated along the branches of the corresponding tree (perhaps using the counting method). The probabilities of various events are then obtained by multiplying conditional probabilities along the corresponding paths of the tree, using the multiplication rule.
- (c) The **divide-and-conquer method**. Here, the probabilities $\mathbf{P}(B)$ of various events B are obtained from conditional probabilities $\mathbf{P}(B | A_i)$, where the A_i are suitable events that form a partition of the sample space and have known probabilities $\mathbf{P}(A_i)$. The probabilities $\mathbf{P}(B)$ are then obtained by using the total probability theorem.

Finally, we have focused on a few side topics that reinforce our main themes. We have discussed the use of Bayes' rule in inference, which is an important application context. We have also discussed some basic principles of counting and combinatorics, which are helpful in applying the counting method.

P R O B L E M S

SECTION 1.1. Sets

Problem 1. Consider rolling a six-sided die. Let A be the set of outcomes where the roll is an even number. Let B be the set of outcomes where the roll is greater than 3. Calculate and compare the sets on both sides of De Morgan's laws

$$(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c.$$

Problem 2. Let A and B be two sets.

(a) Show that

$$A^c = (A^c \cap B) \cup (A^c \cap B^c), \quad B^c = (A \cap B^c) \cup (A^c \cap B^c).$$

(b) Show that

$$(A \cap B)^c = (A^c \cap B) \cup (A^c \cap B^c) \cup (A \cap B^c).$$

(c) Consider rolling a six-sided die. Let A be the set of outcomes where the roll is an odd number. Let B be the set of outcomes where the roll is less than 4. Calculate the sets on both sides of the equality in part (b), and verify that the equality holds.

Problem 3.* Prove the identity

$$A \cup \left(\cap_{n=1}^{\infty} B_n \right) = \cap_{n=1}^{\infty} (A \cup B_n).$$

Solution. If x belongs to the set on the left, there are two possibilities. Either $x \in A$, in which case x belongs to all of the sets $A \cup B_n$, and therefore belongs to the set on the right. Alternatively, x belongs to all of the sets B_n in which case, it belongs to all of the sets $A \cup B_n$, and therefore again belongs to the set on the right.

Conversely, if x belongs to the set on the right, then it belongs to $A \cup B_n$ for all n . If x belongs to A , then it belongs to the set on the left. Otherwise, x must belong to every set B_n and again belongs to the set on the left.

Problem 4.* Cantor's diagonalization argument. Show that the unit interval $[0, 1]$ is uncountable, i.e., its elements cannot be arranged in a sequence.

Solution. Any number x in $[0, 1]$ can be represented in terms of its decimal expansion, e.g., $1/3 = 0.3333\cdots$. Note that most numbers have a unique decimal expansion, but there are a few exceptions. For example, $1/2$ can be represented as $0.5000\cdots$ or as $0.49999\cdots$. It can be shown that this is the only kind of exception, i.e., decimal expansions that end with an infinite string of zeroes or an infinite string of nines.

Suppose, to obtain a contradiction, that the elements of $[0, 1]$ can be arranged in a sequence x_1, x_2, x_3, \dots , so that every element of $[0, 1]$ appears in the sequence. Consider the decimal expansion of x_n :

$$x_n = 0.a_n^1 a_n^2 a_n^3 \dots,$$

where each digit a_n^i belongs to $\{0, 1, \dots, 9\}$. Consider now a number y constructed as follows. The n th digit of y can be 1 or 2, and is chosen so that it is different from the n th digit of x_n . Note that y has a unique decimal expansion since it does not end with an infinite sequence of zeroes or nines. The number y differs from each x_n , since it has a different n th digit. Therefore, the sequence x_1, x_2, \dots does not exhaust the elements of $[0, 1]$, contrary to what was assumed. The contradiction establishes that the set $[0, 1]$ is uncountable.

SECTION 1.2. Probabilistic Models

Problem 5. Out of the students in a class, 60% are geniuses, 70% love chocolate, and 40% fall into both categories. Determine the probability that a randomly selected student is neither a genius nor a chocolate lover.

Problem 6. A six-sided die is loaded in a way that each even face is twice as likely as each odd face. All even faces are equally likely, as are all odd faces. Construct a probabilistic model for a single roll of this die and find the probability that the outcome is less than 4.

Problem 7. A four-sided die is rolled repeatedly, until the first time (if ever) that an even number is obtained. What is the sample space for this experiment?

Problem 8.* Bonferroni's inequality.

(a) Prove that for any two events A and B , we have

$$\mathbf{P}(A \cap B) \geq \mathbf{P}(A) + \mathbf{P}(B) - 1.$$

(b) Generalize to the case of n events A_1, A_2, \dots, A_n , by showing that

$$\mathbf{P}(A_1 \cap A_2 \cap \dots \cap A_n) \geq \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots + \mathbf{P}(A_n) - (n - 1).$$

Solution. We have $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$ and $\mathbf{P}(A \cup B) \leq 1$, which implies part (a). For part (b), we use De Morgan's law to obtain

$$\begin{aligned} 1 - \mathbf{P}(A_1 \cap \dots \cap A_n) &= \mathbf{P}((A_1 \cap \dots \cap A_n)^c) \\ &= \mathbf{P}(A_1^c \cup \dots \cup A_n^c) \\ &\leq \mathbf{P}(A_1^c) + \dots + \mathbf{P}(A_n^c) \\ &= (1 - \mathbf{P}(A_1)) + \dots + (1 - \mathbf{P}(A_n)) \\ &= n - \mathbf{P}(A_1) - \dots - \mathbf{P}(A_n). \end{aligned}$$

Problem 9.* The inclusion-exclusion formula. Show the following generalizations of the formula

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B).$$

(a) Let A , B , and C be events. Then,

$$\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(A \cap B) - \mathbf{P}(B \cap C) - \mathbf{P}(A \cap C) + \mathbf{P}(A \cap B \cap C).$$

(b) Let A_1, A_2, \dots, A_n be events. Let $S_1 = \{i \mid 1 \leq i \leq n\}$, $S_2 = \{(i_1, i_2) \mid 1 \leq i_1 < i_2 \leq n\}$, and more generally, let S_m be the set of all m -tuples (i_1, \dots, i_m) of indices that satisfy $1 \leq i_1 < i_2 < \dots < i_m \leq n$. Then,

$$\begin{aligned} \mathbf{P}(\cup_{k=1}^n A_k) &= \sum_{i \in S_1} \mathbf{P}(A_i) - \sum_{(i_1, i_2) \in S_2} \mathbf{P}(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum_{(i_1, i_2, i_3) \in S_3} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \dots + (-1)^{n-1} \mathbf{P}(\cap_{k=1}^n A_k). \end{aligned}$$

Solution. (a) We use the formulas $\mathbf{P}(X \cup Y) = \mathbf{P}(X) + \mathbf{P}(Y) - \mathbf{P}(X \cap Y)$ and $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$. We have

$$\begin{aligned} \mathbf{P}(A \cup B \cup C) &= \mathbf{P}(A \cup B) + \mathbf{P}(C) - \mathbf{P}((A \cup B) \cap C) \\ &= \mathbf{P}(A \cup B) + \mathbf{P}(C) - \mathbf{P}((A \cap C) \cup (B \cap C)) \\ &= \mathbf{P}(A \cup B) + \mathbf{P}(C) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) + \mathbf{P}(A \cap B \cap C) \\ &= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) + \mathbf{P}(C) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) \\ &\quad + \mathbf{P}(A \cap B \cap C) \\ &= \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(A \cap B) - \mathbf{P}(B \cap C) - \mathbf{P}(A \cap C) \\ &\quad + \mathbf{P}(A \cap B \cap C). \end{aligned}$$

(b) Use induction and verify the main induction step by emulating the derivation of part (a). For a different approach, see the problems at the end of Chapter 2.

Problem 10.* Continuity property of probabilities.

- Let A_1, A_2, \dots be an infinite sequence of events, which is “monotonically increasing,” meaning that $A_n \subset A_{n+1}$ for every n . Let $A = \cup_{n=1}^{\infty} A_n$. Show that $\mathbf{P}(A) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n)$. *Hint:* Express the event A as a union of countably many disjoint sets.
- Suppose now that the events are “monotonically decreasing,” i.e., $A_{n+1} \subset A_n$ for every n . Let $A = \cap_{n=1}^{\infty} A_n$. Show that $\mathbf{P}(A) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n)$. *Hint:* Apply the result of part (a) to the complements of the events.
- Consider a probabilistic model whose sample space is the real line. Show that

$$\mathbf{P}([0, \infty)) = \lim_{n \rightarrow \infty} \mathbf{P}([0, n]) \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbf{P}([n, \infty)) = 0.$$

Solution. (a) Let $B_1 = A_1$ and, for $n \geq 2$, $B_n = A_n \cap A_{n-1}^c$. The events B_n are disjoint, and we have $\cup_{k=1}^n B_k = A_n$, and $\cup_{k=1}^{\infty} B_k = A$. We apply the additivity axiom to obtain

$$\mathbf{P}(A) = \sum_{k=1}^{\infty} \mathbf{P}(B_k) = \lim_{n \rightarrow \infty} \sum_{k=1}^n \mathbf{P}(B_k) = \lim_{n \rightarrow \infty} \mathbf{P}(\cup_{k=1}^n B_k) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n).$$

(b) Let $C_n = A_n^c$ and $C = A^c$. Since $A_{n+1} \subset A_n$, we obtain $C_n \subset C_{n+1}$, and the events C_n are increasing. Furthermore, $C = A^c = (\cap_{n=1}^{\infty} A_n)^c = \cup_{n=1}^{\infty} A_n^c = \cup_{n=1}^{\infty} C_n$. Using the result from part (a) for the sequence C_n , we obtain

$$1 - \mathbf{P}(A) = \mathbf{P}(A^c) = \mathbf{P}(C) = \lim_{n \rightarrow \infty} \mathbf{P}(C_n) = \lim_{n \rightarrow \infty} (1 - \mathbf{P}(A_n)),$$

from which we conclude that $\mathbf{P}(A) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n)$.

(c) For the first equality, use the result from part (a) with $A_n = [0, n]$ and $A = [0, \infty)$. For the second, use the result from part (b) with $A_n = [n, \infty)$ and $A = \cap_{n=1}^{\infty} A_n = \emptyset$.

SECTION 1.3. Conditional Probability

Problem 11. We roll two fair 6-sided dice. Each one of the 36 possible outcomes is assumed to be equally likely.

- (a) Find the probability that doubles are rolled.
- (b) Given that the roll results in a sum of 4 or less, find the conditional probability that doubles are rolled.
- (c) Find the probability that at least one die roll is a 6.
- (d) Given that the two dice land on different numbers, find the conditional probability that at least one die roll is a 6.

Problem 12. A coin is tossed twice. Alice claims that the event of two heads is at least as likely if we know that the first toss is a head than if we know that at least one of the tosses is a head. Is she right? Does it make a difference if the coin is fair or unfair? How can we generalize Alice's reasoning?

Problem 13. We are given three coins: one has heads in both faces, the second has tails in both faces, and the third has a head in one face and a tail in the other. We choose a coin at random, toss it, and it comes heads. What is the probability that the opposite face is tails?

Problem 14. A batch of one hundred items is inspected by testing four randomly selected items. If one of the four is defective, the batch is rejected. What is the probability that the batch is accepted if it contains five defectives?

Problem 15. Let A and B be events. Show that $\mathbf{P}(A \cap B | B) = \mathbf{P}(A | B)$, assuming that $\mathbf{P}(B) > 0$.

SECTION 1.4. Total Probability Theorem and Bayes' Rule

Problem 16. Alice searches for her term paper in her filing cabinet, which has n drawers. She knows that she left her term paper in drawer j with probability $p_j > 0$. The drawers are so messy that even if she correctly guesses that the term paper is in drawer i , the probability that she finds it is only d_i . Alice searches in a particular drawer, say drawer i , but the search is unsuccessful. Conditioned on this event, show that the probability that her paper is in drawer j , is given by

$$\frac{p_j}{1 - p_i d_i}, \quad \text{if } j \neq i, \quad \frac{p_i(1 - d_i)}{1 - p_i d_i}, \quad \text{if } j = i.$$

Problem 17. How an inferior player with a superior strategy can gain an advantage. Boris is about to play a two-game chess match with an opponent, and wants to find the strategy that maximizes his winning chances. Each game ends with either a win by one of the players, or a draw. If the score is tied at the end of the two games, the match goes into sudden-death mode, and the players continue to play until the first time one of them wins a game (and the match). Boris has two playing styles, *timid* and *bold*, and he can choose one of the two at will in each game, no matter what style he chose in previous games. With timid play, he draws with probability $p_d > 0$, and he loses with probability $1 - p_d$. With bold play, he wins with probability p_w , and he loses with probability $1 - p_w$. Boris will always play bold during sudden death, but may switch style between games 1 and 2.

- (a) Find the probability that Boris wins the match for each of the following strategies:
 - (i) Play bold in both games 1 and 2.
 - (ii) Play timid in both games 1 and 2.
 - (iii) Play timid whenever he is ahead in the score, and play bold otherwise.
- (b) Assume that $p_w < 1/2$, so Boris is the worse player, regardless of the playing style he adopts. Show that with the strategy in (iii) above, and depending on the values of p_w and p_d , Boris may have a better than a 50-50 chance to win the match. How do you explain this advantage?

Problem 18. Two players take turns removing a ball from a jar that initially contains m white and n black balls. The first player to remove a white ball wins. Develop a recursive formula that allows the convenient computation of the probability that the starting player wins.

Problem 19. Each of k jars contains m white and n black balls. A ball is randomly chosen from jar 1 and transferred to jar 2, then a ball is randomly chosen from jar 2 and transferred to jar 3, etc. Finally, a ball is randomly chosen from jar k . Show that the probability that the last ball is white is the same as the probability that the first ball is white, i.e., it is $m/(m + n)$.

Problem 20. We have two jars each containing initially n balls. We perform four successive ball exchanges. In each exchange, we pick simultaneously and at random a ball from each jar and move it to the other jar. What is the probability that at the end of the four exchanges all the balls will be in the jar where they started?

Problem 21. The prisoner's dilemma. Two out of three prisoners are to be released. One of the prisoners asks a guard to tell him the identity of a prisoner other than himself that will be released. The guard refuses with the following rationale: at your present state of knowledge, your probability of being released is $2/3$, but after you know my answer, your probability of being released will become $1/2$, since there will be two prisoners (including yourself) whose fate is unknown and exactly one of the two will be released. What is wrong with the guard's reasoning?

Problem 22. A two-envelopes puzzle. You are handed two envelopes, and you know that each contains a positive integer dollar amount and that the two amounts are different. The values of these two amounts are modeled as constants that are unknown. Without knowing what the amounts are, you select at random one of the two envelopes, and after looking at the amount inside, you may switch envelopes if you wish. A friend claims that the following strategy will increase above $1/2$ your probability of ending up with the envelope with the larger amount: toss a coin repeatedly, let X be equal to $1/2$ plus the number of tosses required to obtain heads for the first time, and switch if the amount in the envelope you selected is less than the value of X . Is your friend correct?

Problem 23. The paradox of induction. Consider a statement whose truth is unknown. If we see many examples that are compatible with it, we are tempted to view the statement as more probable. Such reasoning is often referred to as *inductive inference* (in a philosophical, rather than mathematical sense). Consider now the statement that "all cows are white." An equivalent statement is that "everything that is not white is not a cow." We then observe several black cows. Our observations are clearly compatible with the statement, but do they make the hypothesis "all cows are white" more likely?

To analyze such a situation, we consider a probabilistic model. Let us assume that there are two possible states of the world, which we model as complementary events:

$$A : \text{all cows are white,}$$

$$A^c : 50\% \text{ of all cows are white.}$$

Let p be the prior probability $\mathbf{P}(A)$ that all cows are white. We make an observation of a cow or a crow, with probability q and $1 - q$, respectively, independently of whether event A occurs or not. Assume that $0 < p < 1$, $0 < q < 1$, and that all crows are black.

- (a) Given the event $B = \{\text{a black crow was observed}\}$, what is $\mathbf{P}(A | B)$?
- (b) Given the event $C = \{\text{a white cow was observed}\}$, what is $\mathbf{P}(A | C)$?

Problem 24.* Conditional version of the total probability theorem. Show the identity

$$\mathbf{P}(A | B) = \mathbf{P}(C | B)\mathbf{P}(A | B \cap C) + \mathbf{P}(C^c | B)\mathbf{P}(A | B \cap C^c),$$

assuming all the conditioning events have positive probability.

Solution. Using the conditional probability definition and the additivity axiom on the disjoint sets $A \cap B \cap C$ and $A \cap B \cap C^c$, we obtain

$$\begin{aligned}
& \mathbf{P}(C|B)\mathbf{P}(A|B \cap C) + \mathbf{P}(C^c|B)\mathbf{P}(A|B \cap C^c) \\
&= \frac{\mathbf{P}(B \cap C)}{\mathbf{P}(B)} \cdot \frac{\mathbf{P}(A \cap B \cap C)}{\mathbf{P}(B \cap C)} + \frac{\mathbf{P}(B \cap C^c)}{\mathbf{P}(B)} \cdot \frac{\mathbf{P}(A \cap B \cap C^c)}{\mathbf{P}(B \cap C^c)} \\
&= \frac{\mathbf{P}(A \cap B \cap C) + \mathbf{P}(A \cap B \cap C^c)}{\mathbf{P}(B)} \\
&= \frac{\mathbf{P}((A \cap B \cap C) \cup (A \cap B \cap C^c))}{\mathbf{P}(B)} \\
&= \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \\
&= \mathbf{P}(A|B).
\end{aligned}$$

Problem 25.* Let A and B be events with $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$. We say that an event B *suggests* an event A if $\mathbf{P}(A|B) > \mathbf{P}(A)$, and *does not suggest* event A if $\mathbf{P}(A|B) < \mathbf{P}(A)$.

- Show that B suggests A if and only if A suggests B .
- Assume that $\mathbf{P}(B^c) > 0$. Show that B suggests A if and only if B^c does not suggest A .
- We know that a treasure is located in one of two places, with probabilities β and $1 - \beta$, respectively, where $0 < \beta < 1$. We search the first place and if the treasure is there, we find it with probability $p > 0$. Show that the event of not finding the treasure in the first place suggests that the treasure is in the second place.

Solution. (a) We have $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$, so B suggests A if and only if $\mathbf{P}(A \cap B) > \mathbf{P}(A)\mathbf{P}(B)$, which is equivalent to A suggesting B , by symmetry.

(b) Since $\mathbf{P}(B) + \mathbf{P}(B^c) = 1$, we have

$$\mathbf{P}(B)\mathbf{P}(A) + \mathbf{P}(B^c)\mathbf{P}(A) = \mathbf{P}(A) = \mathbf{P}(B)\mathbf{P}(A|B) + \mathbf{P}(B^c)\mathbf{P}(A|B^c),$$

which implies that

$$\mathbf{P}(B^c)(\mathbf{P}(A) - \mathbf{P}(A|B^c)) = \mathbf{P}(B)(\mathbf{P}(A|B) - \mathbf{P}(A)).$$

Thus, $\mathbf{P}(A|B) > \mathbf{P}(A)$ (B suggests A) if and only if $\mathbf{P}(A) > \mathbf{P}(A|B^c)$ (B^c does not suggest A).

(c) Let A and B be the events

$$A = \{\text{the treasure is in the second place}\},$$

$$B = \{\text{we don't find the treasure in the first place}\}.$$

Using the total probability theorem, we have

$$\mathbf{P}(B) = \mathbf{P}(A^c)\mathbf{P}(B|A^c) + \mathbf{P}(A)\mathbf{P}(B|A) = \beta(1 - p) + (1 - \beta)p,$$

so

$$\mathbf{P}(A | B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{1 - \beta}{\beta(1 - p) + (1 - \beta)} = \frac{1 - \beta}{1 - \beta p} > 1 - \beta = \mathbf{P}(A).$$

It follows that event B suggests event A .

SECTION 1.5. Independence

Problem 26. A hunter has two hunting dogs. One day, on the trail of some animal, the hunter comes to a place where the road diverges into two paths. He knows that each dog, independently of the other, will choose the correct path with probability p . The hunter decides to let each dog choose a path, and if they agree, take that one, and if they disagree, to randomly pick a path. Is his strategy better than just letting one of the two dogs decide on a path?

Problem 27. Communication through a noisy channel. A binary (0 or 1) symbol transmitted through a noisy communication channel is received incorrectly with probability ϵ_0 and ϵ_1 , respectively (see Fig. 1.18). Errors in different symbol transmissions are independent.

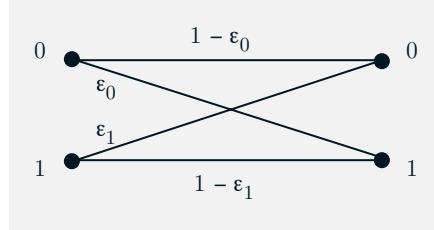


Figure 1.18: Error probabilities in a binary communication channel.

- Suppose that the channel source transmits a 0 with probability p and transmits a 1 with probability $1 - p$. What is the probability that a randomly chosen symbol is received correctly?
- Suppose that the string of symbols 1011 is transmitted. What is the probability that all the symbols in the string are received correctly?
- In an effort to improve reliability, each symbol is transmitted three times and the received symbol is decoded by majority rule. In other words, a 0 (or 1) is transmitted as 000 (or 111, respectively), and it is decoded at the receiver as a 0 (or 1) if and only if the received three-symbol string contains at least two 0s (or 1s, respectively). What is the probability that a transmitted 0 is correctly decoded?
- Suppose that the channel source transmits a 0 with probability p and transmits a 1 with probability $1 - p$, and that the scheme of part (c) is used. What is the probability that a 0 was transmitted given that the received string is 101?

Problem 28. The king's sibling. The king has only one sibling. What is the probability that the sibling is male? Assume that every birth results in a boy with probability

1/2, independent of other births. Be careful to state any additional assumptions you have to make in order to arrive at an answer.

Problem 29. Using a biased coin to make an unbiased decision. Alice and Bob want to choose between the opera and the movies by tossing a fair coin. Unfortunately, the only available coin is biased (though the bias is not known exactly). How can they use the biased coin to make a decision so that either option (opera or the movies) is equally likely to be chosen?

Problem 30. An electrical system consists of identical components that are operational with probability p , independently of other components. The components are connected in three subsystems, as shown in Fig. 1.19. The system is operational if there is a path that starts at point A , ends at point B , and consists of operational components. This is the same as requiring that all three subsystems are operational. What are the probabilities that the three subsystems, as well as the entire system, are operational?

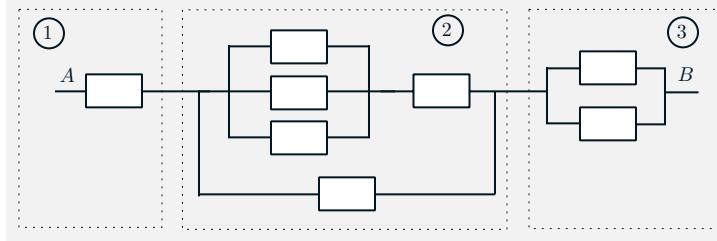


Figure 1.19: A system of identical components that consists of the three subsystems 1, 2, and 3. The system is operational if there is a path that starts at point A , ends at point B , and consists of operational components.

Problem 31. Reliability of a k -out-of- n system. A system consists of n identical components that are operational with probability p , independently of other components. The system is operational if at least k out of the n components are operational. What is the probability that the system is operational?

Problem 32. A power utility can supply electricity to a city from n different power plants. Power plant i fails with probability p_i , independently of the others.

- (a) Suppose that any one plant can produce enough electricity to supply the entire city. What is the probability that the city will experience a black-out?
- (b) Suppose that two power plants are necessary to keep the city from a black-out. Find the probability that the city will experience a black-out.

Problem 33. A cellular phone system services a population of n_1 “voice users” (those that occasionally need a voice connection) and n_2 “data users” (those that occasionally need a data connection). We estimate that at a given time, each user will need to be connected to the system with probability p_1 (for voice users) or p_2 (for data users),

independently of other users. The data rate for a voice user is r_1 bits/sec and for a data user is r_2 bits/sec. The cellular system has a total capacity of c bits/sec. What is the probability that more users want to use the system than the system can accommodate?

Problem 34. The problem of points. Telis and Wendy play a round of golf (18 holes) for a \$10 stake, and their probabilities of winning on any one hole are p and $1 - p$, respectively, independently of their results in other holes. At the end of 10 holes, with the score 4 to 6 in favor of Wendy, Telis receives an urgent call and has to report back to work. They decide to split the stake in proportion to their probabilities of winning had they completed the round, as follows. If p_T and p_W are the conditional probabilities that Telis and Wendy, respectively, are ahead in the score after 18 holes given the 4-6 score after 10 holes, then Telis should get a fraction $p_T/(p_T + p_W)$ of the stake, and Wendy should get the remaining $p_W/(p_T + p_W)$. How much money should Telis get? *Note:* This is an example of the, so-called, problem of points, which played an important historical role in the development of probability theory. The problem was posed by Chevalier de Méré in the 17th century to Pascal, who introduced the idea that the stake of an interrupted game should be divided in proportion to the players' conditional probabilities of winning given the state of the game at the time of interruption. Pascal worked out some special cases and through a correspondence with Fermat, stimulated much thinking and several probability-related investigations.

Problem 35. A particular class has had a history of low attendance. The annoyed professor decides that she will not lecture unless at least k of the n students enrolled in the class are present. Each student will independently show up with probability p_g if the weather is good, and with probability p_b if the weather is bad. Given the probability of bad weather on a given day, calculate the probability that the professor will teach her class on that day.

Problem 36. Consider a coin that comes up heads with probability p and tails with probability $1 - p$. Let q_n be the probability that after n independent tosses, there have been an even number of heads. Derive a recursion that relates q_n to q_{n-1} , and solve this recursion to establish the formula

$$q_n = (1 + (1 - 2p)^n)/2.$$

Problem 37.* Gambler's ruin. A gambler makes a sequence of independent bets. In each bet, he wins \$1 with probability p , and loses \$1 with probability $1 - p$. Initially, the gambler has \$ k , and plays until he either accumulates \$ n or has no money left. What is the probability that the gambler will end up with \$ n ?

Solution. Let us denote by A the event that he ends up with \$ n , and by F the event that he wins the first bet. Denote also by w_k the probability of event A , if he starts with \$ k . We apply the total probability theorem to obtain

$$w_k = \mathbf{P}(A|F)\mathbf{P}(F) + \mathbf{P}(A|F^c)\mathbf{P}(F^c) = p\mathbf{P}(A|F) + q\mathbf{P}(A|F^c), \quad 0 < k < n,$$

where $q = 1 - p$. By the independence of past and future bets, having won the first bet is the same as if he were just starting now but with \$($k+1$), so that $\mathbf{P}(A|F) = w_{k+1}$ and similarly $\mathbf{P}(A|F^c) = w_{k-1}$. Thus, we have $w_k = pw_{k+1} + qw_{k-1}$, which can be written as

$$w_{k+1} - w_k = r(w_k - w_{k-1}), \quad 0 < k < n,$$

where $r = q/p$. We will solve for w_k in terms of p and q using iteration, and the boundary values $w_0 = 0$ and $w_n = 1$.

We have $w_{k+1} - w_k = r^k(w_1 - w_0)$, and since $w_0 = 0$,

$$w_{k+1} = w_k + r^k w_1 = w_{k-1} + r^{k-1} w_1 + r^k w_1 = w_1 + r w_1 + \cdots + r^k w_1.$$

The sum in the right-hand side can be calculated separately for the two cases where $r = 1$ (or $p = q$) and $r \neq 1$ (or $p \neq q$). We have

$$w_k = \begin{cases} \frac{1-r^k}{1-r} w_1, & \text{if } p \neq q, \\ kw_1, & \text{if } p = q. \end{cases}$$

Since $w_n = 1$, we can solve for w_1 and therefore for w_k :

$$w_1 = \begin{cases} \frac{1-r}{1-r^n}, & \text{if } p \neq q, \\ \frac{1}{n}, & \text{if } p = q, \end{cases}$$

so that

$$w_k = \begin{cases} \frac{1-r^k}{1-r^n}, & \text{if } p \neq q, \\ \frac{k}{n}, & \text{if } p = q. \end{cases}$$

Problem 38.* Let A and B be independent events. Use the definition of independence to prove the following:

- (a) The events A and B^c are independent.
- (b) The events A^c and B^c are independent.

Solution. (a) The event A is the union of the disjoint events $A \cap B^c$ and $A \cap B$. Using the additivity axiom and the independence of A and B , we obtain

$$\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B) + \mathbf{P}(A \cap B^c).$$

It follows that

$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A)(1 - \mathbf{P}(B)) = \mathbf{P}(A)\mathbf{P}(B^c),$$

so A and B^c are independent.

- (b) Apply the result of part (a) twice: first on A and B , then on B^c and A .

Problem 39.* Let A , B , and C be independent events, with $\mathbf{P}(C) > 0$. Prove that A and B are conditionally independent given C .

Solution. We have

$$\begin{aligned} \mathbf{P}(A \cap B \mid C) &= \frac{\mathbf{P}(A \cap B \cap C)}{\mathbf{P}(C)} \\ &= \frac{\mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)}{\mathbf{P}(C)} \\ &= \mathbf{P}(A)\mathbf{P}(B) \\ &= \mathbf{P}(A \mid C)\mathbf{P}(B \mid C), \end{aligned}$$

so A and B are conditionally independent given C . In the preceding calculation, the first equality uses the definition of conditional probabilities; the second uses the assumed independence; the fourth uses the independence of A from C , and of B from C .

Problem 40.* Assume that the events A_1, A_2, A_3, A_4 are independent and that $\mathbf{P}(A_3 \cap A_4) > 0$. Show that

$$\mathbf{P}(A_1 \cup A_2 | A_3 \cap A_4) = \mathbf{P}(A_1 \cup A_2).$$

Solution. We have

$$\mathbf{P}(A_1 | A_3 \cap A_4) = \frac{\mathbf{P}(A_1 \cap A_3 \cap A_4)}{\mathbf{P}(A_3 \cap A_4)} = \frac{\mathbf{P}(A_1)\mathbf{P}(A_3)\mathbf{P}(A_4)}{\mathbf{P}(A_3)\mathbf{P}(A_4)} = \mathbf{P}(A_1).$$

We similarly obtain $\mathbf{P}(A_2 | A_3 \cap A_4) = \mathbf{P}(A_2)$ and $\mathbf{P}(A_1 \cap A_2 | A_3 \cap A_4) = \mathbf{P}(A_1 \cap A_2)$, and finally,

$$\begin{aligned} \mathbf{P}(A_1 \cup A_2 | A_3 \cap A_4) &= \mathbf{P}(A_1 | A_3 \cap A_4) + \mathbf{P}(A_2 | A_3 \cap A_4) - \mathbf{P}(A_1 \cap A_2 | A_3 \cap A_4) \\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2) - \mathbf{P}(A_1 \cap A_2) \\ &= \mathbf{P}(A_1 \cup A_2). \end{aligned}$$

Problem 41.* Laplace's rule of succession. Consider $m + 1$ boxes with the k th box containing k red balls and $m - k$ white balls, where k ranges from 0 to m . We choose a box at random (all boxes are equally likely) and then choose a ball at random from that box, n successive times (the ball drawn is replaced each time, and a new ball is selected independently). Suppose a red ball was drawn each of the n times. What is the probability that if we draw a ball one more time it will be red? Estimate this probability for large m .

Solution. We want to find the conditional probability $\mathbf{P}(E | R_n)$, where E is the event of a red ball drawn at time $n + 1$, and R_n is the event of a red ball drawn each of the n preceding times. Intuitively, the consistent draw of a red ball indicates that a box with a high percentage of red balls was chosen, so we expect that $\mathbf{P}(E | R_n)$ is closer to 1 than to 0. In fact, Laplace used this example to calculate the probability that the sun will rise tomorrow given that it has risen for the preceding 5,000 years. (It is not clear how serious Laplace was about this calculation, but the story is part of the folklore of probability theory.)

We have

$$\mathbf{P}(E | R_n) = \frac{\mathbf{P}(E \cap R_n)}{\mathbf{P}(R_n)},$$

and by using the total probability theorem, we obtain

$$\begin{aligned} \mathbf{P}(R_n) &= \sum_{k=0}^m \mathbf{P}(\text{kth box chosen}) \left(\frac{k}{m}\right)^n = \frac{1}{m+1} \sum_{k=0}^m \left(\frac{k}{m}\right)^n, \\ \mathbf{P}(E \cap R_n) &= \mathbf{P}(R_{n+1}) = \frac{1}{m+1} \sum_{k=0}^m \left(\frac{k}{m}\right)^{n+1}. \end{aligned}$$

For large m , we can view $\mathbf{P}(R_n)$ as a piecewise constant approximation to an integral:

$$\mathbf{P}(R_n) = \frac{1}{m+1} \sum_{k=0}^m \left(\frac{k}{m}\right)^n \approx \frac{1}{(m+1)m^n} \int_0^m x^n dx = \frac{1}{(m+1)m^n} \cdot \frac{m^{n+1}}{n+1} \approx \frac{1}{n+1}.$$

Similarly,

$$\mathbf{P}(E \cap R_n) = \mathbf{P}(R_{n+1}) \approx \frac{1}{n+2},$$

so that

$$\mathbf{P}(E | R_n) \approx \frac{n+1}{n+2}.$$

Thus, for large m , drawing a red ball one more time is almost certain when n is large.

Problem 42.* Binomial coefficient formula and the Pascal triangle.

- Use the definition of $\binom{n}{k}$ as the number of distinct n -toss sequences with k heads, to derive the recursion suggested by the so called Pascal triangle, given in Fig. 1.20.
- Use the recursion derived in part (a) and induction, to establish the formula

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

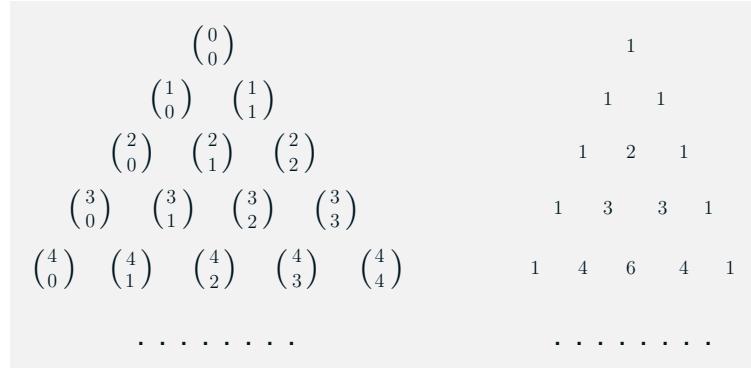


Figure 1.20: Sequential calculation method of the binomial coefficients using the Pascal triangle. Each term $\binom{n}{k}$ in the triangular array on the left is computed and placed in the triangular array on the right by adding its two neighbors in the row above it (except for the boundary terms with $k=0$ or $k=n$, which are equal to 1).

Solution. (a) Note that n -toss sequences that contain k heads (for $0 < k < n$) can be obtained in two ways:

- By starting with an $(n-1)$ -toss sequence that contains k heads and adding a tail at the end. There are $\binom{n-1}{k}$ different sequences of this type.

- (2) By starting with an $(n-1)$ -toss sequence that contains $k-1$ heads and adding a head at the end. There are $\binom{n-1}{k-1}$ different sequences of this type.

Thus,

$$\binom{n}{k} = \begin{cases} \binom{n-1}{k-1} + \binom{n-1}{k}, & \text{if } k = 1, 2, \dots, n-1, \\ 1, & \text{if } k = 0, n. \end{cases}$$

This is the formula corresponding to the Pascal triangle calculation, given in Fig. 1.20.

- (b) We now use the recursion from part (a), to demonstrate the formula

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

by induction on n . Indeed, we have from the definition $\binom{1}{0} = \binom{1}{1} = 1$, so for $n = 1$ the above formula is seen to hold as long as we use the convention $0! = 1$. If the formula holds for each index up to $n-1$, we have for $k = 1, 2, \dots, n-1$,

$$\begin{aligned} \binom{n}{k} &= \binom{n-1}{k-1} + \binom{n-1}{k} \\ &= \frac{(n-1)!}{(k-1)!(n-1-k+1)!} + \frac{(n-1)!}{k!(n-1-k)!} \\ &= \frac{k}{n} \cdot \frac{n!}{k!(n-k)!} + \frac{n-k}{n} \cdot \frac{n!}{k!(n-k)!} \\ &= \frac{n!}{k!(n-k)!}, \end{aligned}$$

and the induction is complete.

Problem 43.* The Borel-Cantelli lemma. Consider an infinite sequence of trials. The probability of success at the i th trial is some positive number p_i . Let N be the event that there is no success, and let I be the event that there is an infinite number of successes.

- (a) Assume that the trials are independent and that $\sum_{i=1}^{\infty} p_i = \infty$. Show that $\mathbf{P}(N) = 0$ and $\mathbf{P}(I) = 1$.

- (b) Assume that $\sum_{i=1}^{\infty} p_i < \infty$. Show that $\mathbf{P}(I) = 0$.

Solution. (a) The event N is a subset of the event that there were no successes in the first n trials, so that

$$\mathbf{P}(N) \leq \prod_{i=1}^n (1 - p_i).$$

Taking logarithms,

$$\log \mathbf{P}(N) \leq \sum_{i=1}^n \log(1 - p_i) \leq \sum_{i=1}^n (-p_i).$$

Taking the limit as n tends to infinity, we obtain $\log \mathbf{P}(N) = -\infty$, or $\mathbf{P}(N) = 0$.

Let now L_n be the event that there is a finite number of successes and that the last success occurs at the n th trial. We use the already established result $\mathbf{P}(N) = 0$, and apply it to the sequence of trials after trial n , to obtain $\mathbf{P}(L_n) = 0$. The event I^c (finite number of successes) is the union of the disjoint events L_n , $n \geq 1$, and N , so that

$$\mathbf{P}(I^c) = \mathbf{P}(N) + \sum_{n=1}^{\infty} \mathbf{P}(L_n) = 0,$$

and $\mathbf{P}(I) = 1$.

(b) Let S_i be the event that the i th trial is a success. Fix some number n and for every $i > n$, let F_i be the event that the first success after time n occurs at time i . Note that $F_i \subset S_i$. Finally, let A_n be the event that there is at least one success after time n . Note that $I \subset A_n$, because an infinite number of successes implies that there are successes subsequent to time n . Furthermore, the event A_n is the union of the disjoint events F_i , $i > n$. Therefore,

$$\mathbf{P}(I) \leq \mathbf{P}(A_n) = \mathbf{P}\left(\bigcup_{i=n+1}^{\infty} F_i\right) = \sum_{i=n+1}^{\infty} \mathbf{P}(F_i) \leq \sum_{i=n+1}^{\infty} \mathbf{P}(S_i) = \sum_{i=n+1}^{\infty} p_i.$$

We take the limit of both sides as $n \rightarrow \infty$. Because of the assumption $\sum_{i=1}^{\infty} p_i < \infty$, the right-hand side converges to zero. This implies that $\mathbf{P}(I) = 0$.

SECTION 1.6. Counting

Problem 44. De Méré's puzzle. A six-sided die is rolled three times independently. Which is more likely: a sum of 11 or a sum of 12? (This question was posed by the French nobleman de Méré to his friend Pascal in the 17th century.)

Problem 45. The birthday problem. Consider n people who are attending a party. We assume that every person has an equal probability of being born on any day during the year, independently of everyone else, and ignore the additional complication presented by leap years (i.e., nobody is born on February 29). What is the probability that each person has a distinct birthday?

Problem 46. An urn contains m red and n white balls.

- (a) We draw two balls randomly and simultaneously. Describe the sample space and calculate the probability that the selected balls are of different color, by using two approaches: a counting approach based on the discrete uniform law, and a sequential approach based on the multiplication rule.
- (b) We roll a fair 3-sided die whose faces are labeled 1,2,3, and if k comes up, we remove k balls from the urn at random and put them aside. Describe the sample space and calculate the probability that all of the balls drawn are red, using a divide-and-conquer approach and the total probability theorem.

Problem 47. We deal from a well-shuffled 52-card deck. Calculate the probability that the 13th card is the first king to be dealt.

Problem 48. Ninety students, including Joe and Jane, are to be split into three classes of equal size, and this is to be done at random. What is the probability that Joe and Jane end up in the same class?

Problem 49. Twenty distinct cars park in the same parking lot every day. Ten of these cars are US-made, while the other ten are foreign-made. The parking lot has exactly twenty spaces, all in a row, so the cars park side by side. However, the drivers have varying schedules, so the position any car might take on a certain day is random.

- (a) In how many different ways can the cars line up?
- (b) What is the probability that on a given day, the cars will park in such a way that they alternate (no two US-made are adjacent and no two foreign-made are adjacent)?

Problem 50. Eight rooks are placed in distinct squares of an 8×8 chessboard, with all possible placements being equally likely. Find the probability that all the rooks are safe from one another, i.e., that there is no row or column with more than one rook.

Problem 51. An academic department offers 8 lower level courses: $\{L_1, L_2, \dots, L_8\}$ and 10 higher level courses: $\{H_1, H_2, \dots, H_{10}\}$. A valid curriculum consists of 4 lower level courses, and 3 higher level courses.

- (a) How many different curricula are possible?
- (b) Suppose that $\{H_1, \dots, H_5\}$ have L_1 as a prerequisite, and $\{H_6, \dots, H_{10}\}$ have L_2 and L_3 as prerequisites, i.e., any curricula which involve, say, one of $\{H_1, \dots, H_5\}$ must also include L_1 . How many different curricula are there?

Problem 52. How many 6-word sentences can be made using each of the 26 letters of the alphabet exactly once? A word is defined as a nonempty (possibly gibberish) sequence of letters.

Problem 53. Consider a group of n persons. A club consists of a special person from the group (the club leader) and a number (possibly zero) of additional club members.

- (a) Explain why the number of possible clubs is $n2^{n-1}$.
- (b) Find an alternative way of counting the number of possible clubs and show the identity

$$\sum_{k=1}^n k \binom{n}{k} = n2^{n-1}.$$

Problem 54. We draw the top 7 cards from a well-shuffled standard 52-card deck. Find the probability that:

- (a) The 7 cards include exactly 3 aces.
- (b) The 7 cards include exactly 2 kings.
- (c) The probability that the 7 cards include exactly 3 aces or exactly 2 kings.

Problem 55. A parking lot contains 100 cars, k of which happen to be lemons. We select m of these cars at random and take them for a testdrive. Find the probability

that n of the cars tested turn out to be lemons.

Problem 56. A well-shuffled 52-card deck is dealt to 4 players. Find the probability that each of the players gets an ace.

Problem 57.* Hypergeometric probabilities. An urn contains n balls, out of which m are red. We select k of the balls at random, without replacement (i.e., selected balls are not put back into the urn before the next selection). What is the probability that i of the selected balls are red?

Solution. The sample space consists of the $\binom{n}{k}$ different ways that we can select k out of the available balls. For the event of interest to occur, we have to select i out of the m red balls, which can be done in $\binom{m}{i}$ ways, and also select $k - i$ out of the $n - m$ blue balls, which can be done in $\binom{n-m}{k-i}$ ways. Therefore, the desired probability is

$$\frac{\binom{m}{i} \binom{n-m}{k-i}}{\binom{n}{k}},$$

for $i \geq 0$ satisfying $i \leq m$, $i \leq k$, and $k - i \leq n - m$. For all other i , the probability is zero.

Problem 58.* Correcting the number of permutations for indistinguishable objects. When permuting n objects, some of which are indistinguishable, different permutations may lead to indistinguishable object sequences, so the number of distinguishable object sequences is less than $n!$. For example, there are six permutations of the letters A, B, and C:

$$\text{ABC, ACB, BAC, BCA, CAB, CBA,}$$

but only three distinguishable sequences that can be formed using the letters A, D, and D:

$$\text{ADD, DAD, DDA.}$$

- (a) Suppose that k out of the n objects are indistinguishable. Show that the number of distinguishable object sequences is $n!/k!$.
- (b) Suppose that we have r types of indistinguishable objects, and for each i , k_i objects of type i . Show that the number of distinguishable object sequences is

$$\frac{n!}{k_1! k_2! \cdots k_r!}.$$

Solution. (a) Each one of the $n!$ permutations corresponds to $k!$ duplicates which are obtained by permuting the k indistinguishable objects. Thus, the $n!$ permutations can be grouped into $n!/k!$ groups of $k!$ indistinguishable permutations that result in the same object sequence. Therefore, the number of distinguishable object sequences is $n!/k!$. For example, the three letters A, D, and D give the $3! = 6$ permutations

$$\text{ADD, ADD, DAD, DDA, DAD, DDA,}$$

obtained by replacing B and C by D in the permutations of A, B, and C given earlier. However, these 6 permutations can be divided into the $n!/k! = 3!/2! = 3$ groups

$$\{\text{ADD, ADD}\}, \{\text{DAD, DAD}\}, \{\text{DDA, DDA}\},$$

each having $k! = 2! = 2$ indistinguishable permutations.

(b) One solution is to extend the argument in (a) above: for each object type i , there are $k_i!$ indistinguishable permutations of the k_i objects. Hence, each permutation belongs to a group of $k_1!k_2!\cdots k_r!$ indistinguishable permutations, all of which yield the same object sequence.

An alternative argument goes as follows. Choosing a distinguishable object sequence is the same as starting with n slots and for each i , choosing the k_i slots to be occupied by objects of type i . This is the same as partitioning the set $\{1, \dots, n\}$ into groups of size k_1, \dots, k_r , and the number of such partitions is given by the multinomial coefficient.

Discrete Random Variables

Contents

2.1. Basic Concepts	p. 72
2.2. Probability Mass Functions	p. 74
2.3. Functions of Random Variables	p. 80
2.4. Expectation, Mean, and Variance	p. 81
2.5. Joint PMFs of Multiple Random Variables	p. 92
2.6. Conditioning	p. 98
2.7. Independence	p. 110
2.8. Summary and Discussion	p. 116
Problems	p. 119

2.1 BASIC CONCEPTS

In many probabilistic models, the outcomes are numerical, e.g., when they correspond to instrument readings or stock prices. In other experiments, the outcomes are not numerical, but they may be associated with some numerical values of interest. For example, if the experiment is the selection of students from a given population, we may wish to consider their grade point average. When dealing with such numerical values, it is often useful to assign probabilities to them. This is done through the notion of a **random variable**, the focus of the present chapter.

Given an experiment and the corresponding set of possible outcomes (the sample space), a random variable associates a particular number with each outcome; see Fig. 2.1. We refer to this number as the **numerical value** or simply the **value** of the random variable. Mathematically, **a random variable is a real-valued function of the experimental outcome.**

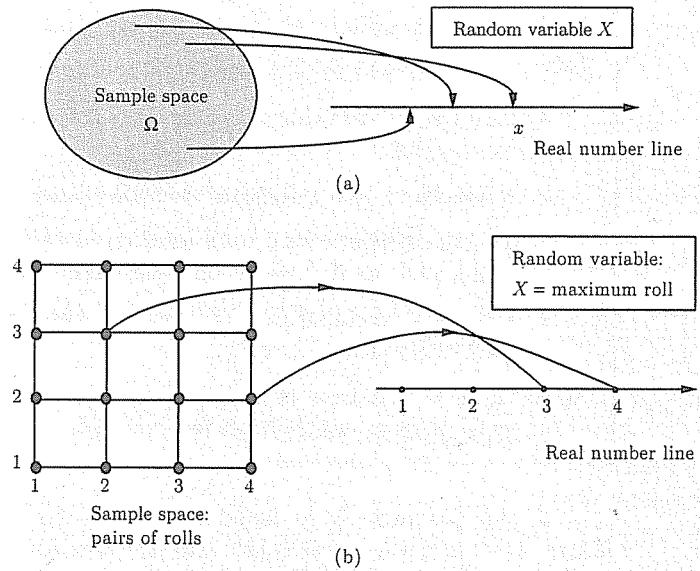


Figure 2.1: (a) Visualization of a random variable. It is a function that assigns a numerical value to each possible outcome of the experiment. (b) An example of a random variable. The experiment consists of two rolls of a 4-sided die, and the random variable is the maximum of the two rolls. If the outcome of the experiment is $(4, 2)$, the value of this random variable is 4.

Here are some examples of random variables:

- In an experiment involving a sequence of 5 tosses of a coin, the number of heads in the sequence is a random variable. However, the 5-long sequence

of heads and tails is not considered a random variable because it does not have an explicit numerical value.

- (b) In an experiment involving two rolls of a die, the following are examples of random variables:
- (i) The sum of the two rolls.
 - (ii) The number of sixes in the two rolls.
 - (iii) The second roll raised to the fifth power.
- (c) In an experiment involving the transmission of a message, the time needed to transmit the message, the number of symbols received in error, and the delay with which the message is received are all random variables.

There are several basic concepts associated with random variables, which are summarized below. These concepts will be discussed in detail in the present chapter.

Main Concepts Related to Random Variables

Starting with a probabilistic model of an experiment:

- A **random variable** is a real-valued function of the outcome of the experiment.
- A **function of a random variable** defines another random variable.
- We can associate with each random variable certain “averages” of interest, such as the **mean** and the **variance**.
- A random variable can be **conditioned** on an event or on another random variable.
- There is a notion of **independence** of a random variable from an event or from another random variable.

A random variable is called **discrete** if its range (the set of values that it can take) is finite or at most countably infinite. For example, the random variables mentioned in (a) and (b) above can take at most a finite number of numerical values, and are therefore discrete.

A random variable that can take an uncountably infinite number of values is not discrete. For an example, consider the experiment of choosing a point a from the interval $[-1, 1]$. The random variable that associates the numerical value a^2 to the outcome a is not discrete. On the other hand, the random variable that associates with a the numerical value

$$\text{sgn}(a) = \begin{cases} 1, & \text{if } a > 0, \\ 0, & \text{if } a = 0, \\ -1, & \text{if } a < 0, \end{cases}$$

is discrete.

In this chapter, we focus exclusively on discrete random variables, even though we will typically omit the qualifier “discrete.”

Concepts Related to Discrete Random Variables

Starting with a probabilistic model of an experiment:

- A discrete random variable is a real-valued function of the outcome of the experiment that can take a finite or countably infinite number of values.
- A discrete random variable has an associated probability mass function (PMF), which gives the probability of each numerical value that the random variable can take.
- A function of a discrete random variable defines another discrete random variable, whose PMF can be obtained from the PMF of the original random variable.

We will discuss each of the above concepts and the associated methodology in the following sections. In addition, we will provide examples of some important and frequently encountered random variables. In Chapter 3, we will discuss general (not necessarily discrete) random variables.

Even though this chapter may appear to be covering a lot of new ground, this is not really the case. The general line of development is to simply take the concepts from Chapter 1 (probabilities, conditioning, independence, etc.) and apply them to random variables rather than events, together with some appropriate new notation. The only genuinely new concepts relate to means and variances.

2.2 PROBABILITY MASS FUNCTIONS

The most important way to characterize a random variable is through the probabilities of the values that it can take. For a discrete random variable X , these are captured by the probability mass function (PMF for short) of X , denoted p_X . In particular, if x is any possible value of X , the probability mass of x , denoted $p_X(x)$, is the probability of the event $\{X = x\}$ consisting of all outcomes that give rise to a value of X equal to x :

$$p_X(x) = P(\{X = x\}).$$

For example, let the experiment consist of two independent tosses of a fair coin, and let X be the number of heads obtained. Then the PMF of X is

$$p_X(x) = \begin{cases} 1/4, & \text{if } x = 0 \text{ or } x = 2, \\ 1/2, & \text{if } x = 1, \\ 0, & \text{otherwise.} \end{cases}$$

In what follows, we will often omit the braces from the event/set notation, when no ambiguity can arise. In particular, we will usually write $\mathbf{P}(X = x)$ in place of the more correct notation $\mathbf{P}(\{X = x\})$. We will also adhere to the following convention throughout: we will use upper case characters to denote random variables, and lower case characters to denote real numbers such as the numerical values of a random variable.

Note that

$$\sum_x p_X(x) = 1,$$

where in the summation above, x ranges over all the possible numerical values of X . This follows from the additivity and normalization axioms, because the events $\{X = x\}$ are disjoint and form a partition of the sample space, as x ranges over all possible values of X . By a similar argument, for any set S of possible values of X , we have

$$\mathbf{P}(X \in S) = \sum_{x \in S} p_X(x).$$

For example, if X is the number of heads obtained in two independent tosses of a fair coin, as above, the probability of at least one head is

$$\mathbf{P}(X > 0) = \sum_{x=1}^2 p_X(x) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

Calculating the PMF of X is conceptually straightforward, and is illustrated in Fig. 2.2.

Calculation of the PMF of a Random Variable X

For each possible value x of X :

1. Collect all the possible outcomes that give rise to the event $\{X = x\}$.
2. Add their probabilities to obtain $p_X(x)$.

The Bernoulli Random Variable

Consider the toss of a coin, which comes up a head with probability p , and a tail with probability $1 - p$. The Bernoulli random variable takes the two values 1 and 0, depending on whether the outcome is a head or a tail:

$$X = \begin{cases} 1, & \text{if a head,} \\ 0, & \text{if a tail.} \end{cases}$$

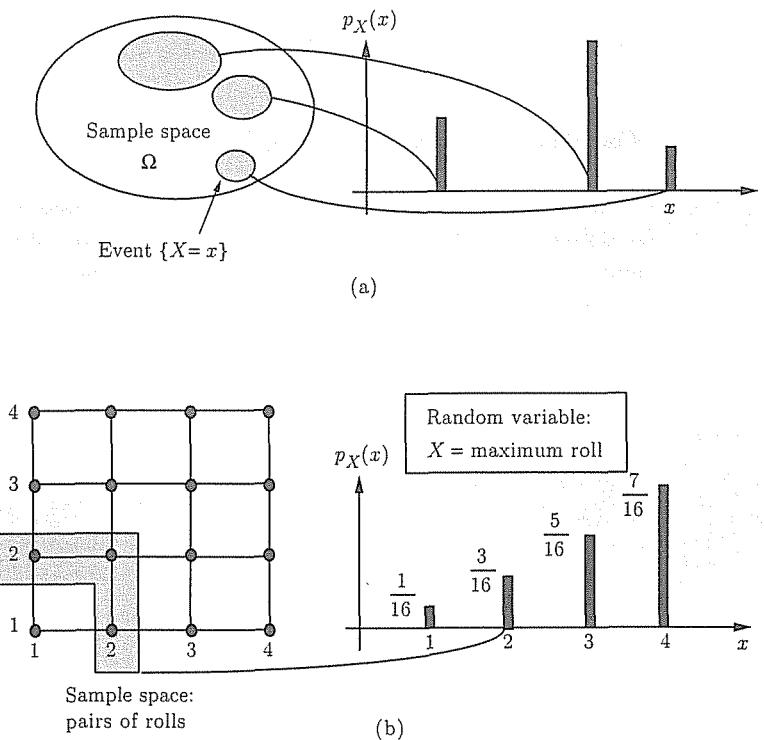


Figure 2.2: (a) Illustration of the method to calculate the PMF of a random variable X . For each possible value x , we collect all the outcomes that give rise to $X = x$ and add their probabilities to obtain $p_X(x)$. (b) Calculation of the PMF p_X of the random variable $X = \text{maximum roll}$ in two independent rolls of a fair 4-sided die. There are four possible values x , namely, 1, 2, 3, 4. To calculate $p_X(x)$ for a given x , we add the probabilities of the outcomes that give rise to x . For example, there are three outcomes that give rise to $x = 2$, namely, $(1, 2), (2, 2), (2, 1)$. Each of these outcomes has probability $1/16$, so $p_X(2) = 3/16$, as indicated in the figure.

Its PMF is

$$p_X(x) = \begin{cases} p, & \text{if } x = 1, \\ 1 - p, & \text{if } x = 0. \end{cases}$$

For all its simplicity, the Bernoulli random variable is very important. In practice, it is used to model generic probabilistic situations with just two outcomes, such as:

- (a) The state of a telephone at a given time that can be either free or busy.
- (b) A person who can be either healthy or sick with a certain disease.
- (c) The preference of a person who can be either for or against a certain political candidate.

Furthermore, by combining multiple Bernoulli random variables, one can construct more complicated random variables, such as the binomial random variable, which is discussed next.

The Binomial Random Variable

A coin is tossed n times. At each toss, the coin comes up a head with probability p , and a tail with probability $1 - p$, independently of prior tosses. Let X be the number of heads in the n -toss sequence. We refer to X as a **binomial** random variable **with parameters n and p** . The PMF of X consists of the binomial probabilities that were calculated in Section 1.5:

$$p_X(k) = \mathbf{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n.$$

(Note that here and elsewhere, we simplify notation and use k , instead of x , to denote the values of integer-valued random variables.) The normalization property, specialized to the binomial random variable, is written as

$$\sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} = 1.$$

Some special cases of the binomial PMF are sketched in Fig. 2.3.

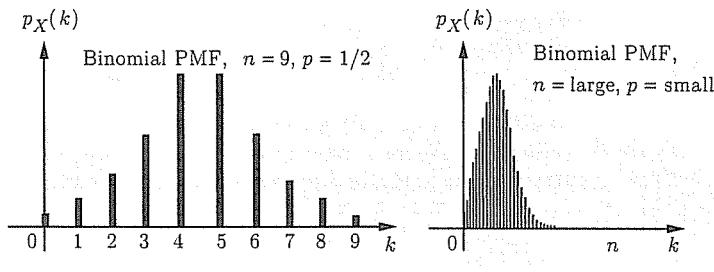


Figure 2.3: The PMF of a binomial random variable. If $p = 1/2$, the PMF is symmetric around $n/2$. Otherwise, the PMF is skewed towards 0 if $p < 1/2$, and towards n if $p > 1/2$.

The Geometric Random Variable

Suppose that we repeatedly and independently toss a coin with probability of a head equal to p , where $0 < p < 1$. The **geometric** random variable is the number X of tosses needed for a head to come up for the first time. Its PMF is given by

$$p_X(k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots,$$

since $(1-p)^{k-1}p$ is the probability of the sequence consisting of $k-1$ successive tails followed by a head; see Fig. 2.4. This is a legitimate PMF because

$$\sum_{k=1}^{\infty} p_X(k) = \sum_{k=1}^{\infty} (1-p)^{k-1}p = p \sum_{k=0}^{\infty} (1-p)^k = p \cdot \frac{1}{1-(1-p)} = 1.$$

Naturally, the use of coin tosses here is just to provide insight. More generally, we can interpret the geometric random variable in terms of repeated independent trials until the first “success.” Each trial has probability of success p and the number of trials until (and including) the first success is modeled by the geometric random variable. The meaning of “success” is context-dependent. For example, it could mean passing a test in a given try, or finding a missing item in a given search, or logging into a computer system in a given attempt, etc.

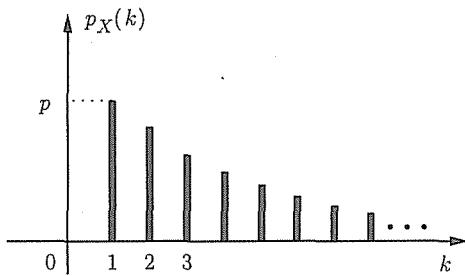


Figure 2.4: The PMF

$$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots,$$

of a geometric random variable. It decreases as a geometric progression with parameter $1-p$.

The Poisson Random Variable

A Poisson random variable has a PMF given by

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots,$$

where λ is a positive parameter characterizing the PMF, see Fig. 2.5. This is a legitimate PMF because

$$\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \right) = e^{-\lambda} e^{\lambda} = 1.$$

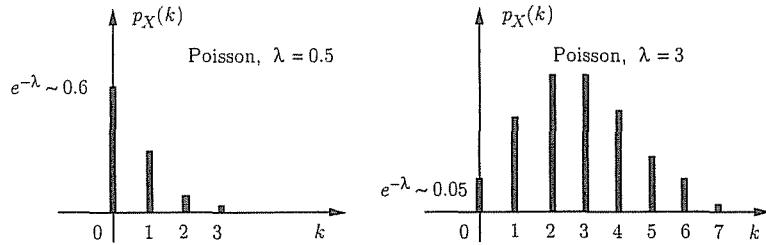


Figure 2.5: The PMF $e^{-\lambda}\lambda^k/k!$ of the Poisson random variable for different values of λ . Note that if $\lambda < 1$, then the PMF is monotonically decreasing, while if $\lambda > 1$, the PMF first increases and then decreases as the value of k increases (this is shown in the end-of-chapter problems).

To get a feel for the Poisson random variable, think of a binomial random variable with very small p and very large n . For example, consider the number of typos in a book with a total of n words, when the probability p that any one word is misspelled is very small (associate a word with a coin toss which comes a head when the word is misspelled), or the number of cars involved in accidents in a city on a given day (associate a car with a coin toss which comes a head when the car has an accident). Such random variables can be well-modeled with a Poisson PMF.

More precisely, the Poisson PMF with parameter λ is a good approximation for a binomial PMF with parameters n and p , i.e.,

$$e^{-\lambda} \frac{\lambda^k}{k!} \approx \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

provided $\lambda = np$, n is very large, and p is very small. In this case, using the Poisson PMF may result in simpler models and calculations. For example, let $n = 100$ and $p = 0.01$. Then the probability of $k = 5$ successes in $n = 100$ trials is calculated using the binomial PMF as

$$\frac{100!}{95!5!} \cdot 0.01^5 (1-0.01)^{95} = 0.00290.$$

Using the Poisson PMF with $\lambda = np = 100 \cdot 0.01 = 1$, this probability is approximated by

$$e^{-1} \frac{1}{5!} = 0.00306.$$

We provide a formal justification of the Poisson approximation property in the end-of-chapter problems and also in Chapter 5, where we will further interpret it, extend it, and use it in the context of the Poisson process.

2.3 FUNCTIONS OF RANDOM VARIABLES

Given a random variable X , one may generate other random variables by applying various transformations on X . As an example, let the random variable X be today's temperature in degrees Celsius, and consider the transformation $Y = 1.8X + 32$, which gives the temperature in degrees Fahrenheit. In this example, Y is a **linear** function of X , of the form

$$Y = g(X) = aX + b,$$

where a and b are scalars. We may also consider nonlinear functions of the general form

$$Y = g(X).$$

For example, if we wish to display temperatures on a logarithmic scale, we would want to use the function $g(X) = \log X$.

If $Y = g(X)$ is a function of a random variable X , then Y is also a random variable, since it provides a numerical value for each possible outcome. This is because every outcome in the sample space defines a numerical value x for X and hence also the numerical value $y = g(x)$ for Y . If X is discrete with PMF p_X , then Y is also discrete, and its PMF p_Y can be calculated using the PMF of X . In particular, to obtain $p_Y(y)$ for any y , we add the probabilities of all values of x such that $g(x) = y$:

$$p_Y(y) = \sum_{\{x \mid g(x)=y\}} p_X(x).$$

Example 2.1. Let $Y = |X|$ and let us apply the preceding formula for the PMF p_Y to the case where

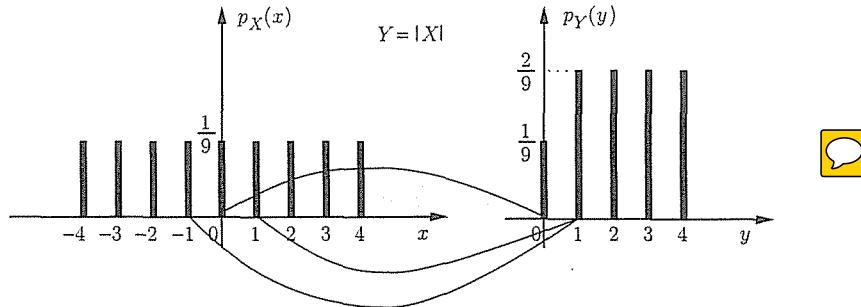
$$p_X(x) = \begin{cases} 1/9, & \text{if } x \text{ is an integer in the range } [-4, 4], \\ 0, & \text{otherwise;} \end{cases}$$

see Fig. 2.6 for an illustration. The possible values of Y are $y = 0, 1, 2, 3, 4$. To compute $p_Y(y)$ for some given value y from this range, we must add $p_X(x)$ over all values x such that $|x| = y$. In particular, there is only one value of X that corresponds to $y = 0$, namely $x = 0$. Thus,

$$p_Y(0) = p_X(0) = \frac{1}{9}.$$

Also, there are two values of X that correspond to each $y = 1, 2, 3, 4$, so for example,

$$p_Y(1) = p_X(-1) + p_X(1) = \frac{2}{9}.$$

Figure 2.6: The PMFs of X and $Y = |X|$ in Example 2.1.

Thus, the PMF of Y is

$$p_Y(y) = \begin{cases} 2/9, & \text{if } y = 1, 2, 3, 4, \\ 1/9, & \text{if } y = 0, \\ 0, & \text{otherwise.} \end{cases}$$

For another related example, let $Z = X^2$. To obtain the PMF of Z , we can view it either as the square of the random variable X or as the square of the random variable $Y = |X|$. By applying the formula $p_Z(z) = \sum_{\{x \mid x^2=z\}} p_X(x)$ or the formula $p_Z(z) = \sum_{\{y \mid y^2=z\}} p_Y(y)$, we obtain

$$p_Z(z) = \begin{cases} 2/9, & \text{if } z = 1, 4, 9, 16, \\ 1/9, & \text{if } z = 0, \\ 0, & \text{otherwise.} \end{cases}$$

2.4 EXPECTATION, MEAN, AND VARIANCE

The PMF of a random variable X provides us with several numbers, the probabilities of all the possible values of X . It would be desirable to summarize this information in a single representative number. This is accomplished by the expectation of X , which is a weighted (in proportion to probabilities) average of the possible values of X .

As motivation, suppose you spin a wheel of fortune many times. At each spin, one of the numbers m_1, m_2, \dots, m_n comes up with corresponding probability p_1, p_2, \dots, p_n , and this is your monetary reward from that spin. What is the amount of money that you “expect” to get “per spin”? The terms “expect” and “per spin” are a little ambiguous, but here is a reasonable interpretation.

Suppose that you spin the wheel k times, and that k_i is the number of times that the outcome is m_i . Then, the total amount received is $m_1 k_1 + m_2 k_2 + \dots + m_n k_n$. The amount received per spin is

$$M = \frac{m_1 k_1 + m_2 k_2 + \dots + m_n k_n}{k}.$$

If the number of spins k is very large, and if we are willing to interpret probabilities as relative frequencies, it is reasonable to anticipate that m_i comes up a fraction of times that is roughly equal to p_i :

$$\frac{k_i}{k} \approx p_i, \quad i = 1, \dots, n.$$

Thus, the amount of money per spin that you “expect” to receive is

$$M = \frac{m_1 k_1 + m_2 k_2 + \dots + m_n k_n}{k} \approx m_1 p_1 + m_2 p_2 + \dots + m_n p_n.$$

Motivated by this example, we introduce the following definition.[†]

Expectation

We define the **expected value** (also called the **expectation** or the **mean**) of a random variable X , with PMF p_X , by

$$\mathbf{E}[X] = \sum_x x p_X(x).$$

Example 2.2. Consider two independent coin tosses, each with a $3/4$ probability of a head, and let X be the number of heads obtained. This is a binomial random variable with parameters $n = 2$ and $p = 3/4$. Its PMF is

$$p_X(k) = \begin{cases} (1/4)^2, & \text{if } k = 0, \\ 2 \cdot (1/4) \cdot (3/4), & \text{if } k = 1, \\ (3/4)^2, & \text{if } k = 2, \end{cases}$$

so the mean is

$$\mathbf{E}[X] = 0 \cdot \left(\frac{1}{4}\right)^2 + 1 \cdot \left(2 \cdot \frac{1}{4} \cdot \frac{3}{4}\right) + 2 \cdot \left(\frac{3}{4}\right)^2 = \frac{24}{16} = \frac{3}{2}.$$

[†] When dealing with random variables that take a countably infinite number of values, one has to deal with the possibility that the infinite sum $\sum_x x p_X(x)$ is not well-defined. More concretely, we will say that the expectation is well-defined if $\sum_x |x| p_X(x) < \infty$. In that case, it is known that the infinite sum $\sum_x x p_X(x)$ converges to a finite value that is independent of the order in which the various terms are summed.

For an example where the expectation is not well-defined, consider a random variable X that takes the value 2^k with probability 2^{-k} , for $k = 1, 2, \dots$. For a more subtle example, consider the random variable X that takes the values 2^k and -2^k with probability 2^{-k} , for $k = 2, 3, \dots$. The expectation is again undefined, even though the PMF is symmetric around zero and one might be tempted to say that $\mathbf{E}[X]$ is zero.

Throughout this book, in lack of an indication to the contrary, we implicitly assume that the expected value of the random variables of interest is well-defined.

It is useful to view the mean of X as a “representative” value of X , which lies somewhere in the middle of its range. We can make this statement more precise, by viewing the mean as the **center of gravity** of the PMF, in the sense explained in Fig. 2.7.

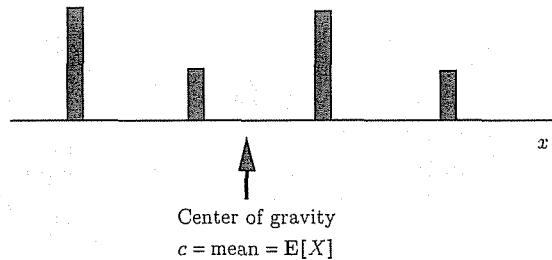


Figure 2.7: Interpretation of the mean as a center of gravity. Given a bar with a weight $p_X(x)$ placed at each point x with $p_X(x) > 0$, the center of gravity c is the point at which the sum of the torques from the weights to its left is equal to the sum of the torques from the weights to its right:

$$\sum_x (x - c)p_X(x) = 0.$$

Thus, $c = \sum_x x p_X(x)$, i.e., the center of gravity is equal to the mean $E[X]$.

Variance, Moments, and the Expected Value Rule

Besides the mean, there are several other quantities that we can associate with a random variable and its PMF. For example, we define the **2nd moment** of the random variable X as the expected value of the random variable X^2 . More generally, we define the **n th moment** as $E[X^n]$, the expected value of the random variable X^n . With this terminology, the 1st moment of X is just the mean.

The most important quantity associated with a random variable X , other than the mean, is its **variance**, which is denoted by $\text{var}(X)$ and is defined as the expected value of the random variable $(X - E[X])^2$, i.e.,

$$\text{var}(X) = E[(X - E[X])^2].$$

Since $(X - E[X])^2$ can only take nonnegative values, the variance is always nonnegative.

The variance provides a measure of dispersion of X around its mean. Another measure of dispersion is the **standard deviation** of X , which is defined as the square root of the variance and is denoted by σ_X :

$$\sigma_X = \sqrt{\text{var}(X)}.$$

The standard deviation is often easier to interpret, because it has the same units as X . For example, if X measures length in meters, the units of variance are square meters, while the units of the standard deviation are meters.

One way to calculate $\text{var}(X)$, is to use the definition of expected value, after calculating the PMF of the random variable $(X - \mathbf{E}[X])^2$. This latter random variable is a function of X , and its PMF can be obtained in the manner discussed in the preceding section.

Example 2.3. Consider the random variable X of Example 2.1, which has the PMF

$$p_X(x) = \begin{cases} 1/9, & \text{if } x \text{ is an integer in the range } [-4, 4], \\ 0, & \text{otherwise.} \end{cases}$$

The mean $\mathbf{E}[X]$ is equal to 0. This can be seen from the symmetry of the PMF of X around 0, and can also be verified from the definition:

$$\mathbf{E}[X] = \sum_x x p_X(x) = \frac{1}{9} \sum_{x=-4}^4 x = 0.$$

Let $Z = (X - \mathbf{E}[X])^2 = X^2$. As in Example 2.1, we obtain

$$p_Z(z) = \begin{cases} 2/9, & \text{if } z = 1, 4, 9, 16, \\ 1/9, & \text{if } z = 0, \\ 0, & \text{otherwise.} \end{cases}$$

The variance of X is then obtained by

$$\text{var}(X) = \mathbf{E}[Z] = \sum_z z p_Z(z) = 0 \cdot \frac{1}{9} + 1 \cdot \frac{2}{9} + 4 \cdot \frac{2}{9} + 9 \cdot \frac{2}{9} + 16 \cdot \frac{2}{9} = \frac{60}{9}.$$

It turns out that there is an easier method to calculate $\text{var}(X)$, which uses the PMF of X but does not require the PMF of $(X - \mathbf{E}[X])^2$. This method is based on the following rule.

Expected Value Rule for Functions of Random Variables

Let X be a random variable with PMF p_X , and let $g(X)$ be a function of X . Then, the expected value of the random variable $g(X)$ is given by

$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x).$$

To verify this rule, we let $Y = g(X)$ and use the formula

$$p_Y(y) = \sum_{\{x \mid g(x)=y\}} p_X(x)$$

derived in the preceding section. We have

$$\begin{aligned} \mathbf{E}[g(X)] &= \mathbf{E}[Y] \\ &= \sum_y y p_Y(y) \\ &= \sum_y y \sum_{\{x \mid g(x)=y\}} p_X(x) \\ &= \sum_y \sum_{\{x \mid g(x)=y\}} y p_X(x) \\ &= \sum_y \sum_{\{x \mid g(x)=y\}} g(x) p_X(x) \\ &= \sum_x g(x) p_X(x). \end{aligned}$$

Using the expected value rule, we can write the variance of X as

$$\text{var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2] = \sum_x (x - \mathbf{E}[X])^2 p_X(x).$$

Similarly, the n th moment is given by

$$\mathbf{E}[X^n] = \sum_x x^n p_X(x),$$

and there is no need to calculate the PMF of X^n .

Example 2.3 (continued). For the random variable X with PMF

$$p_X(x) = \begin{cases} 1/9, & \text{if } x \text{ is an integer in the range } [-4, 4], \\ 0, & \text{otherwise,} \end{cases}$$

we have

$$\begin{aligned} \text{var}(X) &= \mathbf{E}[(X - \mathbf{E}[X])^2] \\ &= \sum_x (x - \mathbf{E}[X])^2 p_X(x) \\ &= \frac{1}{9} \sum_{x=-4}^4 x^2 \quad (\text{since } \mathbf{E}[X] = 0) \\ &= \frac{1}{9} (16 + 9 + 4 + 1 + 0 + 1 + 4 + 9 + 16) \\ &= \frac{60}{9}, \end{aligned}$$

which is consistent with the result obtained earlier.

As we have noted earlier, the variance is always nonnegative, but could it be zero? Since every term in the formula $\sum_x (x - \mathbf{E}[X])^2 p_X(x)$ for the variance is nonnegative, the sum is zero if and only if $(x - \mathbf{E}[X])^2 p_X(x) = 0$ for every x . This condition implies that for any x with $p_X(x) > 0$, we must have $x = \mathbf{E}[X]$ and the random variable X is not really “random”: its value is equal to the mean $\mathbf{E}[X]$, with probability 1.

Variance

The variance $\text{var}(X)$ of a random variable X is defined by

$$\text{var}(X) = \mathbf{E} \left[(X - \mathbf{E}[X])^2 \right],$$

and can be calculated as

$$\text{var}(X) = \sum_x (x - \mathbf{E}[X])^2 p_X(x).$$

It is always nonnegative. Its square root is denoted by σ_X and is called the standard deviation.

Properties of Mean and Variance

We will now use the expected value rule in order to derive some important properties of the mean and the variance. We start with a random variable X and define a new random variable Y , of the form

$$Y = aX + b,$$

where a and b are given scalars. Let us derive the mean and the variance of the linear function Y . We have

$$\mathbf{E}[Y] = \sum_x (ax + b)p_X(x) = a \sum_x xp_X(x) + b \sum_x p_X(x) = a\mathbf{E}[X] + b.$$

Furthermore,

$$\begin{aligned} \text{var}(Y) &= \sum_x (ax + b - \mathbf{E}[aX + b])^2 p_X(x) \\ &= \sum_x (ax + b - a\mathbf{E}[X] - b)^2 p_X(x) \\ &= a^2 \sum_x (x - \mathbf{E}[X])^2 p_X(x) \\ &= a^2 \text{var}(X). \end{aligned}$$

Mean and Variance of a Linear Function of a Random Variable

Let X be a random variable and let

$$Y = aX + b,$$

where a and b are given scalars. Then,

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \quad \text{var}(Y) = a^2 \text{var}(X).$$

Let us also give a convenient formula for the variance of a random variable X with given PMF.

Variance in Terms of Moments Expression

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

This expression is verified as follows:

$$\begin{aligned} \text{var}(X) &= \sum_x (x - \mathbf{E}[X])^2 p_X(x) \\ &= \sum_x (x^2 - 2x\mathbf{E}[X] + (\mathbf{E}[X])^2) p_X(x) \\ &= \sum_x x^2 p_X(x) - 2\mathbf{E}[X] \sum_x x p_X(x) + (\mathbf{E}[X])^2 \sum_x p_X(x) \\ &= \mathbf{E}[X^2] - 2(\mathbf{E}[X])^2 + (\mathbf{E}[X])^2 \\ &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2. \end{aligned}$$

We finally illustrate by example a common pitfall: unless $g(X)$ is a linear function, it is not generally true that $\mathbf{E}[g(X)]$ is equal to $g(\mathbf{E}[X])$.

Example 2.4. Average Speed Versus Average Time. If the weather is good (which happens with probability 0.6), Alice walks the 2 miles to class at a speed of $V = 5$ miles per hour, and otherwise rides her motorcycle at a speed of $V = 30$ miles per hour. What is the mean of the time T to get to class?

A correct way to solve the problem is to first derive the PMF of T ,

$$p_T(t) = \begin{cases} 0.6, & \text{if } t = 2/5 \text{ hours,} \\ 0.4, & \text{if } t = 2/30 \text{ hours,} \end{cases}$$

and then calculate its mean by

$$\mathbf{E}[T] = 0.6 \cdot \frac{2}{5} + 0.4 \cdot \frac{2}{30} = \frac{4}{15} \text{ hours.}$$

However, it is wrong to calculate the mean of the speed V ,

$$\mathbf{E}[V] = 0.6 \cdot 5 + 0.4 \cdot 30 = 15 \text{ miles per hour,}$$

and then claim that the mean of the time T is

$$\frac{2}{\mathbf{E}[V]} = \frac{2}{15} \text{ hours.}$$

To summarize, in this example we have

$$T = \frac{2}{V}, \quad \text{and} \quad \mathbf{E}[T] = \mathbf{E}\left[\frac{2}{V}\right] \neq \frac{2}{\mathbf{E}[V]}.$$

Mean and Variance of Some Common Random Variables

We will now derive formulas for the mean and the variance of a few important random variables. These formulas will be used repeatedly in a variety of contexts throughout the text.

Example 2.5. Mean and Variance of the Bernoulli. Consider the experiment of tossing a coin, which comes up a head with probability p and a tail with probability $1 - p$, and the Bernoulli random variable X with PMF

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1 - p, & \text{if } k = 0. \end{cases}$$

Its mean, second moment, and variance are given by the following calculations:

$$\begin{aligned} \mathbf{E}[X] &= 1 \cdot p + 0 \cdot (1 - p) = p, \\ \mathbf{E}[X^2] &= 1^2 \cdot p + 0 \cdot (1 - p) = p, \\ \text{var}(X) &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = p - p^2 = p(1 - p). \end{aligned}$$

Example 2.6. Discrete Uniform Random Variable. What is the mean and variance of the roll of a fair six-sided die? If we view the result of the roll as a random variable X , its PMF is

$$p_X(k) = \begin{cases} 1/6, & \text{if } k = 1, 2, 3, 4, 5, 6, \\ 0, & \text{otherwise.} \end{cases}$$

Since the PMF is symmetric around 3.5, we conclude that $\mathbf{E}[X] = 3.5$. Regarding the variance, we have

$$\begin{aligned}\text{var}(X) &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\ &= \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) - (3.5)^2,\end{aligned}$$

which yields $\text{var}(X) = 35/12$.

The above random variable is a special case of a **discrete uniformly distributed** random variable (or **discrete uniform** for short), which by definition, takes one out of a range of contiguous integer values, with equal probability. More precisely, this random variable has a PMF of the form

$$p_X(k) = \begin{cases} \frac{1}{b-a+1}, & \text{if } k = a, a+1, \dots, b, \\ 0, & \text{otherwise,} \end{cases}$$

where a and b are two integers with $a < b$; see Fig. 2.8.

The mean is

$$\mathbf{E}[X] = \frac{a+b}{2},$$

as can be seen by inspection, since the PMF is symmetric around $(a+b)/2$. To calculate the variance of X , we first consider the simpler case where $a = 1$ and $b = n$. It can be verified by induction on n that

$$\mathbf{E}[X^2] = \frac{1}{n} \sum_{k=1}^n k^2 = \frac{1}{6}(n+1)(2n+1).$$

We leave the verification of this as an exercise for the reader. The variance can now be obtained in terms of the first and second moments

$$\begin{aligned}\text{var}(X) &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\ &= \frac{1}{6}(n+1)(2n+1) - \frac{1}{4}(n+1)^2 \\ &= \frac{1}{12}(n+1)(4n+2-3n-3) \\ &= \frac{n^2-1}{12}.\end{aligned}$$

For the case of general integers a and b , we note that a random variable which is uniformly distributed over the interval $[a, b]$ has the same variance as one which is uniformly distributed over $[1, b-a+1]$, since the PMF of the second is just a shifted version of the PMF of the first. Therefore, the desired variance is given by the above formula with $n = b-a+1$, which yields

$$\text{var}(X) = \frac{(b-a+1)^2 - 1}{12} = \frac{(b-a)(b-a+2)}{12}.$$

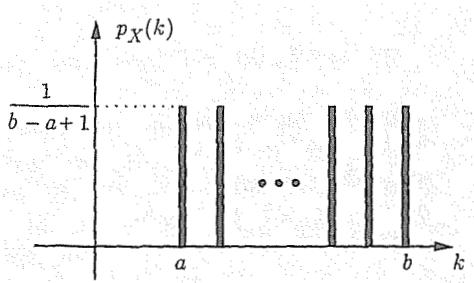


Figure 2.8: PMF of the discrete random variable that is uniformly distributed between two integers a and b . Its mean and variance are

$$\mathbb{E}[X] = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

Example 2.7. The Mean of the Poisson. The mean of the Poisson PMF

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots,$$

can be calculated as follows:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \quad (\text{the } k=0 \text{ term is zero}) \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{m=0}^{\infty} e^{-\lambda} \frac{\lambda^m}{m!} \quad (\text{let } m=k-1) \\ &= \lambda. \end{aligned}$$

The last equality is obtained by noting that

$$\sum_{m=0}^{\infty} e^{-\lambda} \frac{\lambda^m}{m!} = \sum_{m=0}^{\infty} p_X(m) = 1$$

is the normalization property for the Poisson PMF.

A similar calculation shows that the variance of a Poisson random variable is also λ (see the end-of-chapter problems). We will derive this fact in a number of different ways in later chapters.

Decision Making Using Expected Values

Expected values often provide a convenient vehicle for formulating optimization problems where we have to choose between several candidate decisions that result in random rewards. If we view the expected reward of a decision as its “average payoff over a large number of trials,” it is reasonable to choose a decision with maximum expected reward. The following is an example.

Example 2.8. The Quiz Problem. This example, when generalized appropriately, is a prototypical model for optimal scheduling of a collection of tasks that have uncertain outcomes.

Consider a quiz game where a person is given two questions and must decide which question to answer first. Question 1 will be answered correctly with probability 0.8, and the person will then receive as prize \$100, while question 2 will be answered correctly with probability 0.5, and the person will then receive as prize \$200. If the first question attempted is answered incorrectly, the quiz terminates, i.e., the person is not allowed to attempt the second question. If the first question is answered correctly, the person is allowed to attempt the second question. Which question should be answered first to maximize the expected value of the total prize money received?

The answer is not obvious because there is a tradeoff: attempting first the more valuable but also more difficult question 2 carries the risk of never getting a chance to attempt the easier question 1. Let us view the total prize money received as a random variable X , and calculate the expected value $E[X]$ under the two possible question orders (cf. Fig. 2.9):

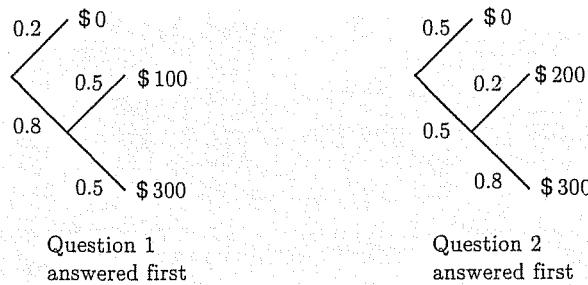


Figure 2.9: Sequential description of the sample space of the quiz problem for the two cases where we answer question 1 or question 2 first.

(a) *Answer question 1 first:* Then the PMF of X is (cf. the left side of Fig. 2.9)

$$p_X(0) = 0.2, \quad p_X(100) = 0.8 \cdot 0.5, \quad p_X(300) = 0.8 \cdot 0.5,$$

and we have

$$E[X] = 0.8 \cdot 0.5 \cdot 100 + 0.8 \cdot 0.5 \cdot 300 = \$160.$$

(b) *Answer question 2 first:* Then the PMF of X is (cf. the right side of Fig. 2.9)

$$p_X(0) = 0.5, \quad p_X(200) = 0.5 \cdot 0.2, \quad p_X(300) = 0.5 \cdot 0.8,$$

and we have

$$E[X] = 0.5 \cdot 0.2 \cdot 200 + 0.5 \cdot 0.8 \cdot 300 = \$140.$$

Thus, it is preferable to attempt the easier question 1 first.

Let us now generalize the analysis. Denote by p_1 and p_2 the probabilities of correctly answering questions 1 and 2, respectively, and by v_1 and v_2 the corresponding prizes. If question 1 is answered first, we have

$$E[X] = p_1(1 - p_2)v_1 + p_1p_2(v_1 + v_2) = p_1v_1 + p_1p_2v_2,$$

while if question 2 is answered first, we have

$$E[X] = p_2(1 - p_1)v_2 + p_2p_1(v_2 + v_1) = p_2v_2 + p_2p_1v_1.$$

It is thus optimal to answer question 1 first if and only if

$$p_1v_1 + p_1p_2v_2 \geq p_2v_2 + p_2p_1v_1,$$

or equivalently, if

$$\frac{p_1v_1}{1 - p_1} \geq \frac{p_2v_2}{1 - p_2}.$$

Therefore, it is optimal to order the questions in decreasing value of the expression $pv/(1 - p)$, which provides a convenient index of quality for a question with probability of correct answer p and value v . Interestingly, this rule generalizes to the case of more than two questions (see the end-of-chapter problems).

2.5 JOINT PMFS OF MULTIPLE RANDOM VARIABLES

Probabilistic models often involve several random variables. For example, in a medical diagnosis context, the results of several tests may be significant, or in a networking context, the workloads of several gateways may be of interest. All of these random variables are associated with the same experiment, sample space, and probability law, and their values may relate in interesting ways. This motivates us to consider probabilities involving simultaneously the numerical values of several random variables and to investigate their mutual couplings. In this section, we will extend the concepts of PMF and expectation developed so far to multiple random variables. Later on, we will also develop notions of conditioning and independence that closely parallel the ideas discussed in Chapter 1.

Consider two discrete random variables X and Y associated with the same experiment. The probabilities of the values that X and Y can take are captured

by the **joint PMF** of X and Y , denoted $p_{X,Y}$. In particular, if (x, y) is a pair of possible values of X and Y , the probability mass of (x, y) is the probability of the event $\{X = x, Y = y\}$:

$$p_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y).$$

Here and elsewhere, we will use the abbreviated notation $\mathbf{P}(X = x, Y = y)$ instead of the more precise notations $\mathbf{P}(\{X = x\} \cap \{Y = y\})$ or $\mathbf{P}(X = x \text{ and } Y = y)$.

The joint PMF determines the probability of any event that can be specified in terms of the random variables X and Y . For example if A is the set of all pairs (x, y) that have a certain property, then

$$\mathbf{P}((X, Y) \in A) = \sum_{(x, y) \in A} p_{X,Y}(x, y).$$

In fact, we can calculate the PMFs of X and Y by using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad p_Y(y) = \sum_x p_{X,Y}(x, y).$$

The formula for $p_X(x)$ can be verified using the calculation

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ &= \sum_y \mathbf{P}(X = x, Y = y) \\ &= \sum_y p_{X,Y}(x, y), \end{aligned}$$

where the second equality follows by noting that the event $\{X = x\}$ is the union of the disjoint events $\{X = x, Y = y\}$ as y ranges over all the different values of Y . The formula for $p_Y(y)$ is verified similarly. We sometimes refer to p_X and p_Y as the **marginal PMFs**, to distinguish them from the joint PMF.

We can calculate the marginal PMFs from the joint PMF by using the **tabular method**. Here, the joint PMF of X and Y is arranged in a two-dimensional table, and the **marginal PMF of X or Y at a given value** is obtained by adding the table entries along a corresponding column or row, respectively. This is illustrated in the following example and in Fig. 2.10.

Example 2.9. Consider two random variables, X and Y , described by the joint PMF shown in Fig. 2.10. The marginal PMFs are calculated by adding the table entries along the columns (for the marginal PMF of X) and along the rows (for the marginal PMF of Y), as indicated.

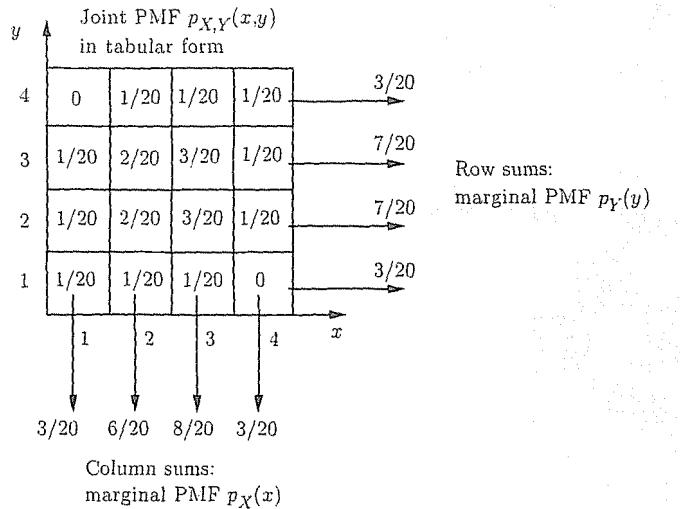


Figure 2.10: Illustration of the tabular method for calculating the marginal PMFs from the joint PMF in Example 2.9. The joint PMF is represented by the table, where the number in each square (x, y) gives the value of $p_{X,Y}(x, y)$. To calculate the marginal PMF $p_X(x)$ for a given value of x , we add the numbers in the column corresponding to x . For example $p_X(2) = 6/20$. Similarly, to calculate the marginal PMF $p_Y(y)$ for a given value of y , we add the numbers in the row corresponding to y . For example $p_Y(2) = 7/20$.

Functions of Multiple Random Variables

When there are multiple random variables of interest, it is possible to generate new random variables by considering functions involving several of these random variables. In particular, a function $Z = g(X, Y)$ of the random variables X and Y defines another random variable. Its PMF can be calculated from the joint PMF $p_{X,Y}$ according to

$$p_Z(z) = \sum_{\{(x,y) \mid g(x,y)=z\}} p_{X,Y}(x,y).$$

Furthermore, the expected value rule for functions naturally extends and takes the form

$$\mathbf{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y).$$

The verification of this is very similar to the earlier case of a function of a single random variable. In the special case where g is linear and of the form $aX + bY + c$, where a , b , and c are given scalars, we have

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$

Example 2.9 (continued). Consider again the random variables X and Y whose joint PMF is given in Fig. 2.10, and a new random variable Z defined by

$$Z = X + 2Y.$$

The PMF of Z can be calculated using the formula

$$p_Z(z) = \sum_{\{(x,y) \mid x+2y=z\}} p_{X,Y}(x,y),$$

and we have, using the PMF given in Fig. 2.10,

$$p_Z(3) = \frac{1}{20}, \quad p_Z(4) = \frac{1}{20}, \quad p_Z(5) = \frac{2}{20}, \quad p_Z(6) = \frac{2}{20}, \quad p_Z(7) = \frac{4}{20},$$

$$p_Z(8) = \frac{3}{20}, \quad p_Z(9) = \frac{3}{20}, \quad p_Z(10) = \frac{2}{20}, \quad p_Z(11) = \frac{1}{20}, \quad p_Z(12) = \frac{1}{20}.$$

The expected value of Z can be obtained from its PMF:

$$\begin{aligned} \mathbf{E}[Z] &= \sum z p_Z(z) \\ &= 3 \cdot \frac{1}{20} + 4 \cdot \frac{1}{20} + 5 \cdot \frac{2}{20} + 6 \cdot \frac{2}{20} + 7 \cdot \frac{4}{20} \\ &\quad + 8 \cdot \frac{3}{20} + 9 \cdot \frac{3}{20} + 10 \cdot \frac{2}{20} + 11 \cdot \frac{1}{20} + 12 \cdot \frac{1}{20} \\ &= 7.55. \end{aligned}$$

Alternatively, we can obtain $\mathbf{E}[Z]$ using the formula

$$\mathbf{E}[Z] = \mathbf{E}[X] + 2\mathbf{E}[Y].$$

From the marginal PMFs, given in Fig. 2.10, we have

$$\mathbf{E}[X] = 1 \cdot \frac{3}{20} + 2 \cdot \frac{6}{20} + 3 \cdot \frac{8}{20} + 4 \cdot \frac{3}{20} = \frac{51}{20},$$

$$\mathbf{E}[Y] = 1 \cdot \frac{3}{20} + 2 \cdot \frac{7}{20} + 3 \cdot \frac{7}{20} + 4 \cdot \frac{3}{20} = \frac{50}{20},$$

so

$$\mathbf{E}[Z] = \frac{51}{20} + 2 \cdot \frac{50}{20} = 7.55.$$

More than Two Random Variables

The joint PMF of three random variables X , Y , and Z is defined in analogy with the above as

$$p_{X,Y,Z}(x, y, z) = \mathbf{P}(X = x, Y = y, Z = z),$$

for all possible triplets of numerical values (x, y, z) . Corresponding marginal PMFs are analogously obtained by equations such as

$$p_{X,Y}(x, y) = \sum_z p_{X,Y,Z}(x, y, z),$$

and

$$p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x, y, z).$$

The expected value rule for functions is given by

$$\mathbf{E}[g(X, Y, Z)] = \sum_x \sum_y \sum_z g(x, y, z) p_{X,Y,Z}(x, y, z),$$

and if g is linear and has the form $aX + bY + cZ + d$, then

$$\mathbf{E}[aX + bY + cZ + d] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c\mathbf{E}[Z] + d.$$

Furthermore, there are obvious generalizations of the above to more than three random variables. For example, for any random variables X_1, X_2, \dots, X_n and any scalars a_1, a_2, \dots, a_n , we have

$$\mathbf{E}[a_1 X_1 + a_2 X_2 + \dots + a_n X_n] = a_1 \mathbf{E}[X_1] + a_2 \mathbf{E}[X_2] + \dots + a_n \mathbf{E}[X_n].$$

Example 2.10. Mean of the Binomial. Your probability class has 300 students and each student has probability $1/3$ of getting an A, independently of any other student. What is the mean of X , the number of students that get an A? Let

$$X_i = \begin{cases} 1, & \text{if the } i\text{th student gets an A,} \\ 0, & \text{otherwise.} \end{cases}$$

Thus X_1, X_2, \dots, X_n are Bernoulli random variables with common mean $p = 1/3$. Their sum

$$X = X_1 + X_2 + \dots + X_n$$

is the number of students that get an A. Since X is the number of “successes” in n independent trials, it is a binomial random variable with parameters n and p .

Using the linearity of X as a function of the X_i , we have

$$\mathbf{E}[X] = \sum_{i=1}^{300} \mathbf{E}[X_i] = \sum_{i=1}^{300} \frac{1}{3} = 300 \cdot \frac{1}{3} = 100.$$

If we repeat this calculation for a general number of students n and probability of A equal to p , we obtain

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{E}[X_i] = \sum_{i=1}^n p = np.$$

Example 2.11. The Hat Problem. Suppose that n people throw their hats in a box and then each picks one hat at random. (Each hat can be picked by only one person, and each assignment of hats to persons is equally likely.) What is the expected value of X , the number of people that get back their own hat?

For the i th person, we introduce a random variable X_i that takes the value 1 if the person selects his/her own hat, and takes the value 0 otherwise. Since $\mathbf{P}(X_i = 1) = 1/n$ and $\mathbf{P}(X_i = 0) = 1 - 1/n$, the mean of X_i is

$$\mathbf{E}[X_i] = 1 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n}.$$

We now have

$$X = X_1 + X_2 + \cdots + X_n,$$

so that

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] = n \cdot \frac{1}{n} = 1.$$

Summary of Facts About Joint PMFs

Let X and Y be random variables associated with the same experiment.

- The **joint PMF** $p_{X,Y}$ of X and Y is defined by

$$p_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y).$$

- The **marginal PMFs** of X and Y can be obtained from the joint PMF, using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad p_Y(y) = \sum_x p_{X,Y}(x, y).$$

- A function $g(X, Y)$ of X and Y defines another random variable, and

$$\mathbf{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y).$$

If g is linear, of the form $aX + bY + c$, we have

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c.$$

- The above have natural extensions to the case where more than two random variables are involved.

2.6 CONDITIONING

Similar to our discussion in Chapter 1, conditional probabilities can be used to capture the information conveyed by various events about the different possible values of a random variable. We are thus motivated to introduce conditional PMFs, given the occurrence of a certain event or given the value of another random variable. We develop this idea in this section and we discuss the properties of conditional PMFs. In reality though, there is not much that is new, only an elaboration of concepts that are familiar from Chapter 1, together with some new notation.

Conditioning a Random Variable on an Event

The **conditional PMF** of a random variable X , conditioned on a particular event A with $\mathbf{P}(A) > 0$, is defined by

$$p_{X|A}(x) = \mathbf{P}(X = x | A) = \frac{\mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)}.$$

Note that the events $\{X = x\} \cap A$ are disjoint for different values of x , their union is A , and, therefore,

$$\mathbf{P}(A) = \sum_x \mathbf{P}(\{X = x\} \cap A).$$

Combining the above two formulas, we see that

$$\sum_x p_{X|A}(x) = 1,$$

so $p_{X|A}$ is a legitimate PMF.

The conditional PMF is calculated similar to its unconditional counterpart: to obtain $p_{X|A}(x)$, we add the probabilities of the outcomes that give rise to

$X = x$ and belong to the conditioning event A , and then normalize by dividing with $\mathbf{P}(A)$.

Example 2.12. Let X be the roll of a fair six-sided die and let A be the event that the roll is an even number. Then, by applying the preceding formula, we obtain

$$\begin{aligned} p_{X|A}(k) &= \mathbf{P}(X = k \mid \text{roll is even}) \\ &= \frac{\mathbf{P}(X = k \text{ and } X \text{ is even})}{\mathbf{P}(\text{roll is even})} \\ &= \begin{cases} 1/3, & \text{if } k = 2, 4, 6, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Example 2.13. A student will take a certain test repeatedly, up to a maximum of n times, each time with a probability p of passing, independently of the number of previous attempts. What is the PMF of the number of attempts, given that the student passes the test?

Let A be the event that the student passes the test (with at most n attempts). We introduce the random variable X , which is the number of attempts that would be needed if an unlimited number of attempts were allowed. Then, X is a geometric random variable with parameter p , and $A = \{X \leq n\}$. We have

$$\mathbf{P}(A) = \sum_{m=1}^n (1-p)^{m-1} p,$$

and

$$p_{X|A}(k) = \begin{cases} \frac{(1-p)^{k-1} p}{\sum_{m=1}^n (1-p)^{m-1} p}, & \text{if } k = 1, \dots, n, \\ 0, & \text{otherwise,} \end{cases}$$

as illustrated in Fig. 2.11.

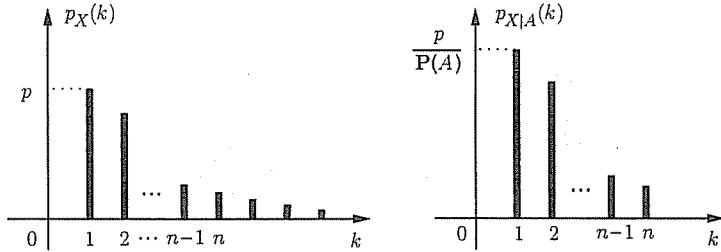


Figure 2.11: Visualization and calculation of the conditional PMF $p_{X|A}(k)$ in Example 2.13. We start with the PMF of X , we set to zero the PMF values for all k that do not belong to the conditioning event A , and we normalize the remaining values by dividing with $\mathbf{P}(A)$.

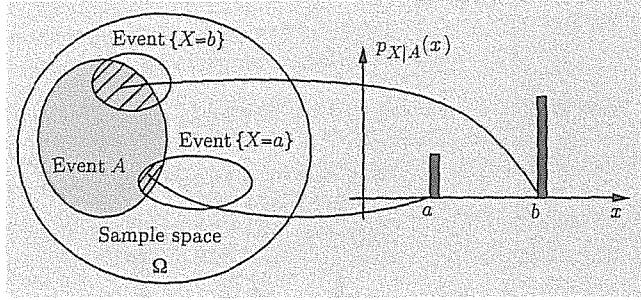


Figure 2.12: Visualization and calculation of the conditional PMF $p_{X|A}(x)$. For each x , we add the probabilities of the outcomes in the intersection $\{X = x\} \cap A$, and normalize by dividing with $\mathbf{P}(A)$.

Figure 2.12 provides a more abstract visualization of the construction of the conditional PMF.

Conditioning one Random Variable on Another

Let X and Y be two random variables associated with the same experiment. If we know that the value of Y is some particular y [with $p_Y(y) > 0$], this provides partial knowledge about the value of X . This knowledge is captured by the **conditional PMF** $p_{X|Y}$ of X given Y , which is defined by specializing the definition of $p_{X|A}$ to events A of the form $\{Y = y\}$:

$$p_{X|Y}(x | y) = \mathbf{P}(X = x | Y = y).$$

Using the definition of conditional probabilities, we have

$$p_{X|Y}(x | y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Let us fix some y with $p_Y(y) > 0$, and consider $p_{X|Y}(x | y)$ as a function of x . This function is a valid PMF for X : it assigns nonnegative values to each possible x , and these values add to 1. Furthermore, this function of x , has the same shape as $p_{X,Y}(x, y)$ except that it is divided by $p_Y(y)$, which enforces the normalization property

$$\sum_x p_{X|Y}(x | y) = 1.$$

Figure 2.13 provides a visualization of the conditional PMF.

The conditional PMF is often convenient for the calculation of the joint PMF, using a sequential approach and the formula

$$p_{X,Y}(x, y) = p_Y(y)p_{X|Y}(x | y),$$

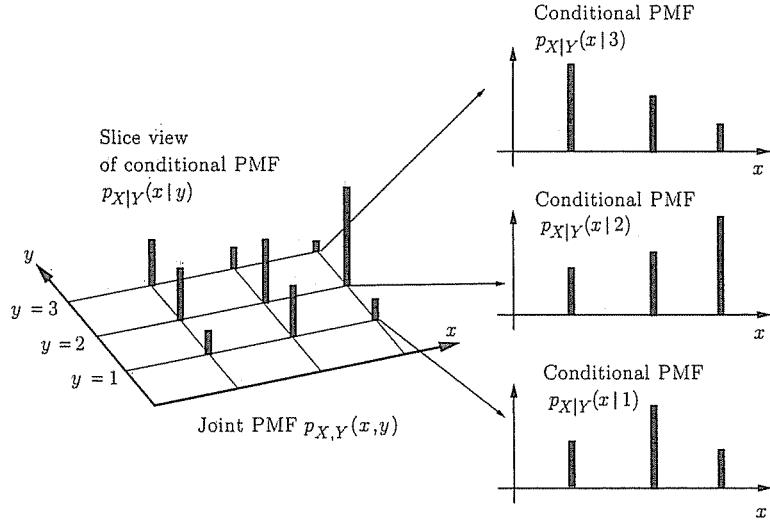


Figure 2.13: Visualization of the conditional PMF $p_{X|Y}(x|y)$. For each y , we view the joint PMF along the slice $Y = y$ and renormalize so that

$$\sum_x p_{X|Y}(x|y) = 1.$$

or its counterpart

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x).$$

This method is entirely similar to the use of the multiplication rule from Chapter 1. The following example provides an illustration.

Example 2.14. Professor May B. Right often has her facts wrong, and answers each of her students' questions incorrectly with probability $1/4$, independently of other questions. In each lecture May is asked 0, 1, or 2 questions with equal probability $1/3$. Let X and Y be the number of questions May is asked and the number of questions she answers wrong in a given lecture, respectively. To construct the joint PMF $p_{X,Y}(x,y)$, we need to calculate all the probabilities $\mathbf{P}(X = x, Y = y)$ for all combinations of values of x and y . This can be done by using a sequential description of the experiment and the multiplication rule, as shown in Fig. 2.14. For example, for the case where one question is asked and is answered wrong, we have

$$p_{X,Y}(1,1) = p_X(1)p_{Y|X}(1|1) = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}.$$

The joint PMF can be represented by a two-dimensional table, as shown in Fig. 2.14. It can be used to calculate the probability of any event of interest. For

instance, we have

$$\begin{aligned} \mathbf{P}(\text{at least one wrong answer}) &= p_{X,Y}(1,1) + p_{X,Y}(2,1) + p_{X,Y}(2,2) \\ &= \frac{4}{48} + \frac{6}{48} + \frac{1}{48}. \end{aligned}$$

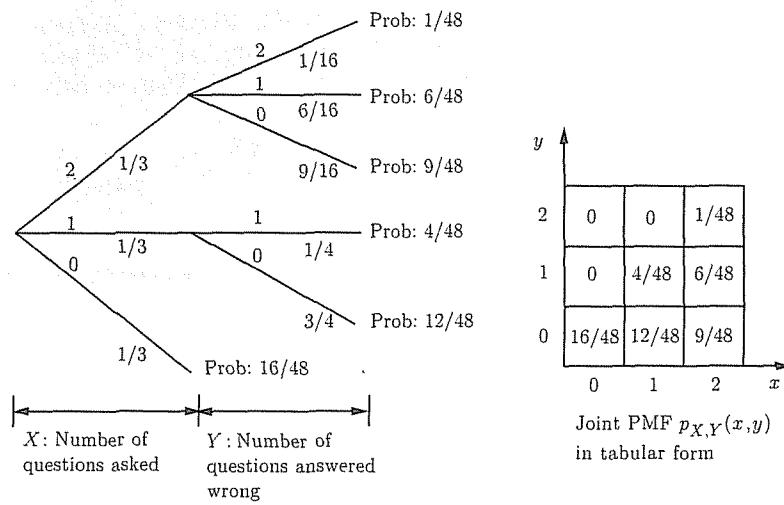


Figure 2.14: Calculation of the joint PMF $p_{X,Y}(x,y)$ in Example 2.14.

The conditional PMF can also be used to calculate the marginal PMFs. In particular, we have by using the definitions,

$$p_X(x) = \sum_y p_{X,Y}(x,y) = \sum_y p_Y(y)p_{X|Y}(x|y).$$

This formula provides a divide-and-conquer method for calculating marginal PMFs. It is in essence identical to the total probability theorem given in Chapter 1, but cast in different notation. The following example provides an illustration.

Example 2.15. Consider a transmitter that is sending messages over a computer network. Let us define the following two random variables:

X : the travel time of a given message, Y : the length of the given message.

We know the PMF of the travel time of a message that has a given length, and we know the PMF of the message length. We want to find the (unconditional) PMF of the travel time of a message.

We assume that the length of a message can take two possible values: $y = 10^2$ bytes with probability $5/6$, and $y = 10^4$ bytes with probability $1/6$, so that

$$p_Y(y) = \begin{cases} 5/6, & \text{if } y = 10^2, \\ 1/6, & \text{if } y = 10^4. \end{cases}$$

We assume that the travel time X of the message depends on its length Y and the congestion in the network at the time of transmission. In particular, the travel time is $10^{-4}Y$ seconds with probability $1/2$, $10^{-3}Y$ seconds with probability $1/3$, and $10^{-2}Y$ seconds with probability $1/6$. Thus, we have

$$p_{X|Y}(x|10^2) = \begin{cases} 1/2, & \text{if } x = 10^{-2}, \\ 1/3, & \text{if } x = 10^{-1}, \\ 1/6, & \text{if } x = 1, \end{cases} \quad p_{X|Y}(x|10^4) = \begin{cases} 1/2, & \text{if } x = 1, \\ 1/3, & \text{if } x = 10, \\ 1/6, & \text{if } x = 100. \end{cases}$$

To find the PMF of X , we use the total probability formula

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x|y).$$

We obtain

$$\begin{aligned} p_X(10^{-2}) &= \frac{5}{6} \cdot \frac{1}{2}, & p_X(10^{-1}) &= \frac{5}{6} \cdot \frac{1}{3}, & p_X(1) &= \frac{5}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{2}, \\ p_X(10) &= \frac{1}{6} \cdot \frac{1}{3}, & p_X(100) &= \frac{1}{6} \cdot \frac{1}{6}. \end{aligned}$$

Note finally that one can define conditional PMFs involving more than two random variables, such as $p_{X,Y|Z}(x,y|z)$ or $p_{X|Y,Z}(x|y,z)$. The concepts and methods described above generalize easily.

Summary of Facts About Conditional PMFs

Let X and Y be random variables associated with the same experiment.

- Conditional PMFs are similar to ordinary PMFs, but refer to a universe where the conditioning event is known to have occurred.
- The conditional PMF of X given an event A with $\mathbf{P}(A) > 0$, is defined by

$$p_{X|A}(x) = \mathbf{P}(X = x | A)$$

and satisfies

$$\sum_x p_{X|A}(x) = 1.$$

- The conditional PMF of X given $Y = y$ is related to the joint PMF by

$$p_{X,Y}(x, y) = p_Y(y)p_{X|Y}(x|y).$$

This is analogous to the multiplication rule for calculating probabilities and can be used to calculate the joint PMF from the conditional PMF.

- The conditional PMF of X given Y can be used to calculate the marginal PMF of X with the formula

$$p_X(x) = \sum_y p_Y(y)p_{X|Y}(x|y).$$

This is analogous to the divide-and-conquer approach for calculating probabilities using the total probability theorem.

- There are natural extensions to the above involving more than two random variables.

Conditional Expectation

A conditional PMF can be thought of as an ordinary PMF over a new universe determined by the conditioning event. In the same spirit, a conditional expectation is the same as an ordinary expectation, except that it refers to the new universe, and all probabilities and PMFs are replaced by their conditional counterparts. (Conditional variances can also be treated similarly.) We list the main definitions and relevant facts below.

Summary of Facts About Conditional Expectations

Let X and Y be random variables associated with the same experiment.

- The conditional expectation of X given an event A with $\mathbf{P}(A) > 0$, is defined by

$$\mathbf{E}[X | A] = \sum_x x p_{X|A}(x).$$

For a function $g(X)$, we have

$$\mathbf{E}[g(X) | A] = \sum_x g(x)p_{X|A}(x).$$

- The conditional expectation of X given a value y of Y is defined by

$$\mathbf{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y).$$

- We have

$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X | Y = y].$$

This is the **total expectation theorem**.

- Let A_1, \dots, A_n be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$ for all i . Then,

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X | A_i].$$

- Let A_1, \dots, A_n be disjoint events that form a partition of an event B , and assume that $\mathbf{P}(A_i \cap B) > 0$ for all i . Then,

$$\mathbf{E}[X | B] = \sum_{i=1}^n \mathbf{P}(A_i | B) \mathbf{E}[X | A_i \cap B].$$

Let us verify the total expectation theorem, which basically says that “the unconditional average can be obtained by averaging the conditional averages.” The theorem is derived using the total probability formula

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x | y),$$

and the calculation

$$\begin{aligned} \mathbf{E}[X] &= \sum_x x p_X(x) \\ &= \sum_x x \sum_y p_Y(y) p_{X|Y}(x | y) \\ &= \sum_y p_Y(y) \sum_x x p_{X|Y}(x | y) \\ &= \sum_y p_Y(y) \mathbf{E}[X | Y = y]. \end{aligned}$$

The relation $\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X | A_i]$ can be verified by viewing it as a special case of the total expectation theorem. Let us introduce the random

variable Y that takes the value i if and only if the event A_i occurs. Its PMF is given by

$$p_Y(i) = \begin{cases} \mathbf{P}(A_i), & \text{if } i = 1, 2, \dots, n, \\ 0, & \text{otherwise.} \end{cases}$$

The total expectation theorem yields

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X | Y = i],$$

and since the event $\{Y = i\}$ is just A_i , we obtain the desired expression

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X | A_i].$$

Finally, the last relation in the table is essentially the same as the preceding one, but applied to a universe where event B is known to have occurred.

The total expectation theorem is analogous to the total probability theorem. It can be used to calculate the unconditional expectation $\mathbf{E}[X]$ from the conditional PMF or expectation, using a divide-and-conquer approach.

Example 2.16. Messages transmitted by a computer in Boston through a data network are destined for New York with probability 0.5, for Chicago with probability 0.3, and for San Francisco with probability 0.2. The transit time X of a message is random. Its mean is 0.05 seconds if it is destined for New York, 0.1 seconds if it is destined for Chicago, and 0.3 seconds if it is destined for San Francisco. Then, $\mathbf{E}[X]$ is easily calculated using the total expectation theorem as

$$\mathbf{E}[X] = 0.5 \cdot 0.05 + 0.3 \cdot 0.1 + 0.2 \cdot 0.3 = 0.115 \text{ seconds.}$$

Example 2.17. Mean and Variance of the Geometric. You write a software program over and over, and each time there is probability p that it works correctly, independently from previous attempts. What is the mean and variance of X , the number of tries until the program works correctly?

We recognize X as a geometric random variable with PMF

$$p_X(k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots$$

The mean and variance of X are given by

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k(1 - p)^{k-1} p, \quad \text{var}(X) = \sum_{k=1}^{\infty} (k - \mathbf{E}[X])^2 (1 - p)^{k-1} p,$$

but evaluating these infinite sums is somewhat tedious. As an alternative, we will apply the total expectation theorem, with $A_1 = \{X = 1\} = \{\text{first try is a success}\}$,

$A_2 = \{X > 1\} = \{\text{first try is a failure}\}$, and end up with a much simpler calculation.

If the first try is successful, we have $X = 1$, and

$$\mathbf{E}[X | X = 1] = 1.$$

If the first try fails ($X > 1$), we have wasted one try, and we are back where we started. So, the expected number of remaining tries is $\mathbf{E}[X]$, and

$$\mathbf{E}[X | X > 1] = 1 + \mathbf{E}[X].$$

Thus,

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{P}(X = 1)\mathbf{E}[X | X = 1] + \mathbf{P}(X > 1)\mathbf{E}[X | X > 1] \\ &= p + (1 - p)(1 + \mathbf{E}[X]), \end{aligned}$$

from which we obtain

$$\mathbf{E}[X] = \frac{1}{p}.$$

With similar reasoning, we also have

$$\mathbf{E}[X^2 | X = 1] = 1, \quad \mathbf{E}[X^2 | X > 1] = \mathbf{E}[(1 + X)^2] = 1 + 2\mathbf{E}[X] + \mathbf{E}[X^2],$$

so that

$$\mathbf{E}[X^2] = p \cdot 1 + (1 - p)(1 + 2\mathbf{E}[X] + \mathbf{E}[X^2]),$$

from which we obtain

$$\mathbf{E}[X^2] = \frac{1 + 2(1 - p)\mathbf{E}[X]}{p},$$

and, using the formula $\mathbf{E}[X] = 1/p$ derived above,

$$\mathbf{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}.$$

We conclude that

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1 - p}{p^2}.$$

Example 2.18. The Two-Envelopes Paradox. This is a much discussed puzzle that involves a subtle mathematical point regarding conditional expectations.

You are handed two envelopes, and you are told that one of them contains m times as much money as the other, where m is an integer with $m > 1$. You open one of the envelopes and look at the amount inside. You may now keep this amount, or you may switch envelopes and keep the amount in the other envelope. What is the best strategy?

Here is a line of reasoning that argues in favor of switching. Let A be the envelope you open and B be the envelope that you may switch to. Let also x and

y be the amounts in A and B, respectively. Then, as the argument goes, either $y = x/m$ or $y = mx$, with equal probability 1/2, so given x , the expected value of y is

$$\frac{1}{2} \cdot \frac{x}{m} + \frac{1}{2} \cdot mx = \frac{1}{2} \left(\frac{1}{m} + m \right) x = \frac{1+m^2}{2m} x > x,$$

since $1+m^2 > 2m$ for $m > 1$. Therefore, you should always switch to envelope B! But then, since you should switch regardless of the amount found in A, you might as well open B to begin with; but once you do, you should switch again, etc.

There are two assumptions, both flawed to some extent, that underlie this paradoxical line of reasoning.

- (a) You have no a priori knowledge about the amounts in the envelopes, so given x , the only thing you know about y , is that it is either $1/m$ or m times x , and there is no reason to assume that one is more likely than the other.
- (b) Given two random variables X and Y , representing monetary amounts, if

$$\mathbf{E}[Y | X = x] > x,$$

for all possible values x of X , then the strategy that always switches to Y yields a higher expected monetary gain.

Let us scrutinize these assumptions.

Assumption (a) is flawed because it relies on an incompletely specified probabilistic model. Indeed, in any correct model, all events, including the possible values of X and Y , must have well-defined probabilities. With such probabilistic knowledge about X and Y , the value of X may reveal a great deal of information about Y . For example, assume the following probabilistic model: someone chooses an integer dollar amount Z from a known range $[\underline{z}, \bar{z}]$ according to some distribution, places this amount in a randomly chosen envelope, and places m times this amount in the other envelope. You then choose to open one of the two envelopes (with equal probability), and look at the enclosed amount X . If X turns out to be larger than the upper range limit \bar{z} , you know that X is the larger of the two amounts, and hence you should not switch. On the other hand, for some other values of X , such as the lower range limit \underline{z} , you should switch envelopes. Thus, in this model, the choice to switch or not should depend on the value of X . Roughly speaking, if you have an idea about the range and likelihood of the values of X , you can judge whether the amount X found in A is relatively small or relatively large, and accordingly switch or not switch envelopes.

Mathematically, in a correct probabilistic model, we must have a joint PMF for the random variables X and Y , the amounts in envelopes A and B, respectively. This joint PMF is specified by introducing a PMF p_Z for the random variable Z , the minimum of the amounts in the two envelopes. Then, for all z ,

$$p_{X,Y}(mz, z) = p_{X,Y}(z, mz) = \frac{1}{2} p_Z(z),$$

and

$$p_{X,Y}(x, y) = 0,$$

for every (x, y) that is not of the form (mz, z) or (z, mz) . With this specification of $p_{X,Y}(x, y)$, and given that $X = x$, one can use the rule

switch if and only if $\mathbf{E}[Y | X = x] > x$.

According to this decision rule, one may or may not switch envelopes, depending on the value of X , as indicated earlier.

Is it true that, with the above described probabilistic model and decision rule, you should be switching for some values x but not for others? Ordinarily yes, as illustrated from the earlier example where Z takes values in a bounded range. However, here is a devilish example where because of a subtle mathematical quirk, you will always switch!

A fair coin is tossed until it comes up heads. Let N be the number of tosses. Then, m^N dollars are placed in one envelope and m^{N-1} dollars are placed in the other. Let X be the amount in the envelope you open (envelope A), and let Y be the amount in the other envelope (envelope B).

Now, if A contains \$1, clearly B contains $\$m$, so you should switch envelopes. If, on the other hand, A contains m^n dollars, where $n > 0$, then B contains either m^{n-1} or m^{n+1} dollars. Since N has a geometric PMF, we have

$$\frac{\mathbf{P}(Y = m^{n+1} | X = m^n)}{\mathbf{P}(Y = m^{n-1} | X = m^n)} = \frac{\mathbf{P}(Y = m^{n+1}, X = m^n)}{\mathbf{P}(Y = m^{n-1}, X = m^n)} = \frac{\mathbf{P}(N = n+1)}{\mathbf{P}(N = n)} = \frac{1}{2}.$$

Thus

$$\mathbf{P}(Y = m^{n-1} | X = m^n) = \frac{2}{3}, \quad \mathbf{P}(Y = m^{n+1} | X = m^n) = \frac{1}{3},$$

and

$$\mathbf{E}[\text{amount in B} | X = m^n] = \frac{2}{3} \cdot m^{n-1} + \frac{1}{3} \cdot m^{n+1} = \frac{2 + m^2}{3m} \cdot m^n.$$

We have $(2 + m^2)/3m > 1$ if and only if $m^2 - 3m + 2 > 0$ or $(m-1)(m-2) > 0$. Thus if $m > 2$, then

$$\mathbf{E}[\text{amount in B} | X = m^n] > m^n,$$

and to maximize the expected monetary gain you should always switch to B!

What is happening in this example is that you switch for all values of x because

$$\mathbf{E}[Y | X = x] > x, \quad \text{for all } x.$$

A naive application of the total expectation theorem might seem to indicate that $\mathbf{E}[Y] > \mathbf{E}[X]$. However, this cannot be true, since X and Y have identical PMFs. Instead, we have

$$\mathbf{E}[Y] = \mathbf{E}[X] = \infty,$$

which is not necessarily inconsistent with the relation $\mathbf{E}[Y | X = x] > x$ for all x .

The conclusion is that the decision rule that switches if and only if $\mathbf{E}[Y | X = x] > x$ does not improve the expected monetary gain in the case where $\mathbf{E}[Y] = \mathbf{E}[X] = \infty$, and the apparent paradox is resolved.

2.7 INDEPENDENCE

We now discuss concepts of independence related to random variables. These concepts are analogous to the concepts of independence between events (cf. Chapter 1). They are developed by simply introducing suitable events involving the possible values of various random variables, and by considering the independence of these events.

Independence of a Random Variable from an Event

The independence of a random variable from an event is similar to the independence of two events. The idea is that knowing the occurrence of the conditioning event provides no new information on the value of the random variable. More formally, we say that the random variable X is **independent of the event A** if

$$\mathbf{P}(X = x \text{ and } A) = \mathbf{P}(X = x)\mathbf{P}(A) = p_X(x)\mathbf{P}(A), \quad \text{for all } x,$$

which is the same as requiring that the two events $\{X = x\}$ and A be independent, for any choice x . From the definition of the conditional PMF, we have

$$\mathbf{P}(X = x \text{ and } A) = p_{X|A}(x)\mathbf{P}(A),$$

so that as long as $\mathbf{P}(A) > 0$, independence is the same as the condition

$$p_{X|A}(x) = p_X(x), \quad \text{for all } x.$$

Example 2.19. Consider two independent tosses of a fair coin. Let X be the number of heads and let A be the event that the number of heads is even. The (unconditional) PMF of X is

$$p_X(x) = \begin{cases} 1/4, & \text{if } x = 0, \\ 1/2, & \text{if } x = 1, \\ 1/4, & \text{if } x = 2, \end{cases}$$

and $\mathbf{P}(A) = 1/2$. The conditional PMF is obtained from the definition $p_{X|A}(x) = \mathbf{P}(X = x \text{ and } A)/\mathbf{P}(A)$:

$$p_{X|A}(x) = \begin{cases} 1/2, & \text{if } x = 0, \\ 0, & \text{if } x = 1, \\ 1/2, & \text{if } x = 2. \end{cases}$$

Clearly, X and A are not independent, since the PMFs p_X and $p_{X|A}$ are different. For an example of a random variable that is independent of A , consider the random variable that takes the value 0 if the first toss is a head, and the value 1 if the first toss is a tail. This is intuitively clear and can also be verified by using the definition of independence.

Independence of Random Variables

The notion of independence of two random variables is similar to the independence of a random variable from an event. We say that two **random variables** X and Y are **independent** if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \text{for all } x,y.$$

This is the same as requiring that the two events $\{X = x\}$ and $\{Y = y\}$ be independent for every x and y . Finally, the formula $p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y)$ shows that independence is equivalent to the condition

$$p_{X|Y}(x|y) = p_X(x), \quad \text{for all } y \text{ with } p_Y(y) > 0 \text{ and all } x.$$

Intuitively, independence means that the value of Y provides no information on the value of X .

There is a similar notion of conditional independence of two random variables, given an event A with $\mathbf{P}(A) > 0$. The conditioning event A defines a new universe and all probabilities (or PMFs) have to be replaced by their conditional counterparts. For example, X and Y are said to be **conditionally independent**, given a positive probability event A , if

$$\mathbf{P}(X = x, Y = y | A) = \mathbf{P}(X = x | A)\mathbf{P}(Y = y | A), \quad \text{for all } x \text{ and } y,$$

or, in this chapter's notation,

$$p_{X,Y|A}(x,y) = p_{X|A}(x)p_{Y|A}(y), \quad \text{for all } x \text{ and } y.$$

Once more, this is equivalent to

$$p_{X|Y,A}(x|y) = p_{X|A}(x) \quad \text{for all } x \text{ and } y \text{ such that } p_{Y|A}(y) > 0.$$

As in the case of events (Section 1.5), conditional independence may not imply unconditional independence and vice versa. This is illustrated by the example in Fig. 2.15.

If X and Y are independent random variables, then

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y],$$

as shown by the following calculation:

$$\begin{aligned} \mathbf{E}[XY] &= \sum_x \sum_y xy p_{X,Y}(x,y) \\ &= \sum_x \sum_y xy p_X(x)p_Y(y) \quad (\text{by independence}) \\ &= \sum_x x p_X(x) \sum_y y p_Y(y) \\ &= \mathbf{E}[X]\mathbf{E}[Y]. \end{aligned}$$

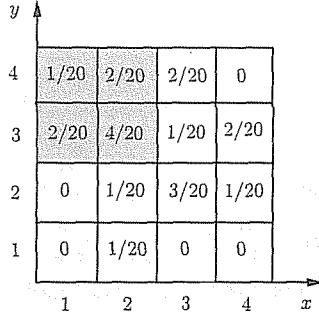


Figure 2.15: Example illustrating that conditional independence may not imply unconditional independence. For the PMF shown, the random variables X and Y are not independent. For example, we have

$$p_{X|Y}(1|1) = P(X = 1|Y = 1) = 0 \neq P(X = 1) = p_X(1).$$

On the other hand, conditional on the event $A = \{X \leq 2, Y \geq 3\}$ (the shaded set in the figure), the random variables X and Y can be seen to be independent. In particular, we have

$$p_{X|Y,A}(x|y) = \begin{cases} 1/3, & \text{if } x = 1, \\ 2/3, & \text{if } x = 2, \end{cases}$$

for both values $y = 3$ and $y = 4$.

A very similar calculation also shows that if X and Y are independent, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)],$$

for any functions g and h . In fact, this follows immediately once we realize that if X and Y are independent, then the same is true for $g(X)$ and $h(Y)$. This is intuitively clear and its formal verification is left as an end-of-chapter problem.

Consider now the sum $X + Y$ of two independent random variables X and Y , and let us calculate its variance. Since the variance of a random variable is unchanged when the random variable is shifted by a constant, it is convenient to work with the zero-mean random variables $\tilde{X} = X - E[X]$ and $\tilde{Y} = Y - E[Y]$. We have

$$\begin{aligned} \text{var}(X + Y) &= \text{var}(\tilde{X} + \tilde{Y}) \\ &= E[(\tilde{X} + \tilde{Y})^2] \\ &= E[\tilde{X}^2 + 2\tilde{X}\tilde{Y} + \tilde{Y}^2] \\ &= E[\tilde{X}^2] + 2E[\tilde{X}\tilde{Y}] + E[\tilde{Y}^2] \\ &= E[\tilde{X}^2] + E[\tilde{Y}^2] \\ &= \text{var}(\tilde{X}) + \text{var}(\tilde{Y}) \\ &= \text{var}(X) + \text{var}(Y). \end{aligned}$$

We have used above the property $\mathbf{E}[\tilde{X} \tilde{Y}] = 0$, which is justified as follows. The random variables $\tilde{X} = X - \mathbf{E}[X]$ and $\tilde{Y} = Y - \mathbf{E}[Y]$ are independent (because they are functions of the independent random variables X and Y), and since they also have zero-mean, we obtain

$$\mathbf{E}[\tilde{X} \tilde{Y}] = \mathbf{E}[\tilde{X}] \mathbf{E}[\tilde{Y}] = 0.$$

In conclusion, the variance of the sum of two **independent** random variables is equal to the sum of their variances. As an interesting contrast, note that the mean of the sum of two random variables is **always** equal to the sum of their means, even if they are not independent.

Summary of Facts About Independent Random Variables

Let A be an event, with $\mathbf{P}(A) > 0$, and let X and Y be random variables associated with the same experiment.

- X is independent of the event A if

$$p_{X|A}(x) = p_X(x), \quad \text{for all } x,$$

that is, if for all x , the events $\{X = x\}$ and A are independent.

- X and Y are independent if for all possible pairs (x, y) , the events $\{X = x\}$ and $\{Y = y\}$ are independent, or equivalently

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \quad \text{for all } x, y.$$

- If X and Y are independent random variables, then

$$\mathbf{E}[XY] = \mathbf{E}[X] \mathbf{E}[Y].$$

Furthermore, for any functions g and h , the random variables $g(X)$ and $h(Y)$ are independent, and we have

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

- If X and Y are independent, then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

Independence of Several Random Variables

The preceding discussion extends naturally to the case of more than two random variables. For example, three random variables X , Y , and Z are said to be

independent if

$$p_{X,Y,Z}(x,y,z) = p_X(x)p_Y(y)p_Z(z), \quad \text{for all } x,y,z.$$

If X , Y , and Z are independent random variables, then any three random variables of the form $f(X)$, $g(Y)$, and $h(Z)$, are also independent. Similarly, any two random variables of the form $g(X, Y)$ and $h(Z)$ are independent. On the other hand, two random variables of the form $g(X, Y)$ and $h(Y, Z)$ are usually not independent, because they are both affected by Y . Properties such as the above are intuitively clear if we interpret independence in terms of noninteracting (sub)experiments. They can be formally verified but this is sometimes tedious. Fortunately, there is general agreement between intuition and what is mathematically correct. This is basically a testament that our definitions of independence adequately reflect the intended interpretation.

Variance of the Sum of Independent Random Variables

Sums of independent random variables are especially important in a variety of contexts. For example, they arise in statistical applications where we “average” a number of independent measurements, with the aim of minimizing the effects of measurement errors. They also arise when dealing with the cumulative effect of several independent sources of randomness. We provide some illustrations in the examples that follow and we will also return to this theme in later chapters.

In the examples below, we will make use of the following key property. If X_1, X_2, \dots, X_n are independent random variables, then

$$\text{var}(X_1 + X_2 + \dots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n).$$

This can be verified by repeated use of the formula $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$ for two independent random variables X and Y .

Example 2.20. Variance of the Binomial. We consider n independent coin tosses, with each toss having probability p of coming up a head. For each i , we let X_i be the Bernoulli random variable which is equal to 1 if the i th toss comes up a head, and is 0 otherwise. Then, $X = X_1 + X_2 + \dots + X_n$ is a binomial random variable. By the independence of the coin tosses, the random variables X_1, \dots, X_n are independent, and

$$\text{var}(X) = \sum_{i=1}^n \text{var}(X_i) = np(1-p).$$

The formulas for the mean and variance of a weighted sum of random variables form the basis for many statistical procedures that estimate the mean of a random variable by averaging many independent samples. A typical case is illustrated in the following example.

Example 2.21. Mean and Variance of the Sample Mean. We wish to estimate the approval rating of a president, to be called B. To this end, we ask n persons drawn at random from the voter population, and we let X_i be a random variable that encodes the response of the i th person:

$$X_i = \begin{cases} 1, & \text{if the } i\text{th person approves B's performance,} \\ 0, & \text{if the } i\text{th person disapproves B's performance.} \end{cases}$$

We model X_1, X_2, \dots, X_n as independent Bernoulli random variables with common mean p and variance $p(1-p)$. Naturally, we view p as the true approval rating of B. We “average” the responses and compute the **sample mean** S_n , defined as

$$S_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Thus, the random variable S_n is the approval rating of B within our n -person sample.

We have, using the linearity of S_n as a function of the X_i ,

$$\mathbf{E}[S_n] = \sum_{i=1}^n \frac{1}{n} \mathbf{E}[X_i] = \frac{1}{n} \sum_{i=1}^n p = p,$$

and making use of the independence of X_1, \dots, X_n ,

$$\text{var}(S_n) = \sum_{i=1}^n \frac{1}{n^2} \text{var}(X_i) = \frac{p(1-p)}{n}.$$

The sample mean S_n can be viewed as a “good” estimate of the approval rating. This is because it has the correct expected value, which is the approval rating p , and its accuracy, as reflected by its variance, improves as the sample size n increases.

Note that even if the random variables X_i are not Bernoulli, the same calculation yields

$$\text{var}(S_n) = \frac{\text{var}(X)}{n},$$

as long as the X_i are independent, with common mean $\mathbf{E}[X]$ and variance $\text{var}(X)$. Thus, again, the sample mean becomes a very good estimate (in terms of variance) of the true mean $\mathbf{E}[X]$, as the sample size n increases. We will revisit the properties of the sample mean and discuss them in much greater detail in Chapter 7, in connection with the laws of large numbers.

Example 2.22. Estimating Probabilities by Simulation. In many practical situations, the analytical calculation of the probability of some event of interest is very difficult. However, if we have a physical or computer model that can generate outcomes of a given experiment in accordance with their true probabilities, we can use simulation to calculate with high accuracy the probability of any given event A . In particular, we independently generate with our model n outcomes, we record the number m of outcomes that belong to the event A of interest, and we approximate $\mathbf{P}(A)$ by m/n . For example, to calculate the probability $p =$

$\mathbf{P}(\text{Heads})$ of a coin, we toss the coin n times, and we approximate p with the ratio (number of heads recorded)/ n .

To see how accurate this process is, consider n independent Bernoulli random variables X_1, \dots, X_n , each with PMF

$$p_{X_i}(k) = \begin{cases} \mathbf{P}(A), & \text{if } k = 1, \\ 1 - \mathbf{P}(A), & \text{if } k = 0. \end{cases}$$

In a simulation context, X_i corresponds to the i th outcome, and takes the value 1 if the i th outcome belongs to the event A . The value of the random variable

$$X = \frac{X_1 + X_2 + \dots + X_n}{n}$$

is the estimate of $\mathbf{P}(A)$ provided by the simulation. According to Example 2.21, X has mean $\mathbf{P}(A)$ and variance $\mathbf{P}(A)(1 - \mathbf{P}(A))/n$, so that for large n , it provides an accurate estimate of $\mathbf{P}(A)$.

2.8 SUMMARY AND DISCUSSION

Random variables provide the natural tools for dealing with probabilistic models in which the outcome determines certain numerical values of interest. In this chapter, we focused on discrete random variables, and developed a conceptual framework and some relevant tools.

In particular, we introduced concepts such as the PMF, the mean, and the variance, which describe in various degrees of detail the probabilistic character of a discrete random variable. We showed how to use the PMF of a random variable X to calculate the mean and the variance of a related random variable $Y = g(X)$ without calculating the PMF of Y . In the special case where g is a linear function, $Y = aX + b$, the means and the variances of X and Y are related by

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \quad \text{var}(Y) = a^2\text{var}(X).$$

We also discussed several special random variables, and derived their PMF, mean, and variance, as summarized in the table that follows.

Summary of Results for Special Random Variables

Discrete Uniform over $[a, b]$:

$$p_X(k) = \begin{cases} \frac{1}{b - a + 1}, & \text{if } k = a, a + 1, \dots, b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a + b}{2}, \quad \text{var}(X) = \frac{(b - a)(b - a + 2)}{12}.$$

Bernoulli with Parameter p : (Describes the success or failure in a single trial.)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1 - p, & \text{if } k = 0, \end{cases}$$

$$\mathbf{E}[X] = p, \quad \text{var}(X) = p(1 - p).$$

Binomial with Parameters p and n : (Describes the number of successes in n independent Bernoulli trials.)

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n,$$

$$\mathbf{E}[X] = np, \quad \text{var}(X) = np(1 - p).$$

Geometric with Parameter p : (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \quad \text{var}(X) = \frac{1 - p}{p^2}.$$

Poisson with Parameter λ : (Approximates the binomial PMF when n is large, p is small, and $\lambda = np$.)

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \dots,$$

$$\mathbf{E}[X] = \lambda, \quad \text{var}(X) = \lambda.$$

We also considered multiple random variables, and introduced their joint and conditional PMFs, and associated expected values. Conditional PMFs are often the starting point in probabilistic models and can be used to calculate other quantities of interest, such as marginal or joint PMFs and expectations, through a sequential or a divide-and-conquer approach. In particular, given the conditional PMF $p_{X|Y}(x|y)$:

- (a) The joint PMF can be calculated by

$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x|y).$$

This can be extended to the case of three or more random variables, as in

$$p_{X,Y,Z}(x, y, z) = p_Z(z) p_{Y|Z}(y|z) p_{X|Y,Z}(x|y, z),$$

and is analogous to the sequential tree-based calculation method using the multiplication rule, discussed in Chapter 1.

- (b) The marginal PMF can be calculated by

$$p_X(x) = \sum_y p_Y(y)p_{X|Y}(x|y),$$

which generalizes the divide-and-conquer calculation method we discussed in Chapter 1.

- (c) The divide-and-conquer calculation method in (b) above can be extended to compute expected values using the total expectation theorem:

$$\mathbf{E}[X] = \sum_y p_Y(y)\mathbf{E}[X|Y=y].$$

We introduced the notion of independence of random variables, in analogy with the notion of independence of events. Among other topics, we focused on random variables X obtained by adding several independent random variables X_1, \dots, X_n :

$$X = X_1 + \dots + X_n.$$

We argued that the mean and the variance of the sum are equal to the sum of the means and the sum of the variances, respectively:

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \dots + \mathbf{E}[X_n], \quad \text{var}(X) = \text{var}(X_1) + \dots + \text{var}(X_n).$$

The formula for the mean does not require independence of the X_i , but the formula for the variance does.

The concepts and methods of this chapter extend appropriately to general random variables (see the next chapter), and are fundamental for our subject.

P R O B L E M S

SECTION 2.2. Probability Mass Functions

Problem 1. The MIT soccer team has 2 games scheduled for one weekend. It has a 0.4 probability of not losing the first game, and a 0.7 probability of not losing the second game, independently of the first. If it does not lose a particular game, the team is equally likely to win or tie, independently of what happens in the other game. The MIT team will receive 2 points for a win, 1 for a tie, and 0 for a loss. Find the PMF of the number of points that the team earns over the weekend.

Problem 2. You go to a party with 500 guests. What is the probability that exactly one other guest has the same birthday as you? Calculate this exactly and also approximately by using the Poisson PMF. (For simplicity, exclude birthdays on February 29.)

Problem 3. Fischer and Spassky play a chess match in which the first player to win a game wins the match. After 10 successive draws, the match is declared drawn. Each game is won by Fischer with probability 0.4, is won by Spassky with probability 0.3, and is a draw with probability 0.3, independently of previous games.

- (a) What is the probability that Fischer wins the match?
- (b) What is the PMF of the duration of the match?

Problem 4. An internet service provider uses 50 modems to serve the needs of 1000 customers. It is estimated that at a given time, each customer will need a connection with probability 0.01, independently of the other customers.

- (a) What is the PMF of the number of modems in use at the given time?
- (b) Repeat part (a) by approximating the PMF of the number of customers that need a connection with a Poisson PMF.
- (c) What is the probability that there are more customers needing a connection than there are modems? Provide an exact, as well as an approximate formula based on the Poisson approximation of part (b).

Problem 5. A packet communication system consists of a buffer that stores packets from some source, and a communication line that retrieves packets from the buffer and transmits them to a receiver. The system operates in time-slot pairs. In the first slot, the system stores a number of packets that are generated by the source according to a Poisson PMF with parameter λ ; however, the maximum number of packets that can be stored is a given integer b , and packets arriving to a full buffer are discarded. In the second slot, the system transmits either all the stored packets or c packets (whichever is less). Here, c is a given integer with $0 < c < b$.

- (a) Assuming that at the beginning of the first slot the buffer is empty, find the PMF of the number of packets stored at the end of the first slot and at the end of the second slot.
- (b) What is the probability that some packets get discarded during the first slot?

Problem 6. The Celtics and the Lakers are set to play a playoff series of n basketball games, where n is odd. The Celtics have a probability p of winning any one game, independently of other games.

- (a) Find the values of p for which $n = 5$ is better for the Celtics than $n = 3$.
- (b) Generalize part (a), i.e., for any $k > 0$, find the values for p for which $n = 2k + 1$ is better for the Celtics than $n = 2k - 1$.

Problem 7. You just rented a large house and the realtor gave you 5 keys, one for each of the 5 doors of the house. Unfortunately, all keys look identical, so to open the front door, you try them at random.

- (a) Find the PMF of the number of trials you will need to open the door, under the following alternative assumptions: (1) after an unsuccessful trial, you mark the corresponding key, so that you never try it again, and (2) at each trial you are equally likely to choose any key.
- (b) Repeat part (a) for the case where the realtor gave you an extra duplicate key for each of the 5 doors.

Problem 8. Recursive computation of the binomial PMF. Let X be a binomial random variable with parameters n and p . Show that its PMF can be computed by starting with $p_X(0) = (1 - p)^n$, and by using the recursive formula

$$p_X(k+1) = \frac{p}{1-p} \cdot \frac{n-k}{k+1} \cdot p_X(k), \quad k = 0, 1, \dots, n-1.$$

Problem 9. Form of the binomial PMF. Consider a binomial random variable X with parameters n and p . Let k^* be the largest integer that is less than or equal to $(n+1)p$. Show that the PMF $p_X(k)$ is monotonically nondecreasing with k in the range from 0 to k^* , and is monotonically decreasing with k for $k \geq k^*$.

Problem 10. Form of the Poisson PMF. Let X be a Poisson random variable with parameter λ . Show that the PMF $p_X(k)$ increases monotonically with k up to the point where k reaches the largest integer not exceeding λ , and after that point decreases monotonically with k .

Problem 11.* The matchbox problem – inspired by Banach's smoking habits. A smoker mathematician carries one matchbox in his right pocket and one in his left pocket. Each time he wants to light a cigarette, he selects a matchbox from either pocket with probability $p = 1/2$, independently of earlier selections. The two matchboxes have initially n matches each. What is the PMF of the number of remaining matches at the moment when the mathematician reaches for a match and discovers that the corresponding matchbox is empty? How can we generalize to the case where the probabilities of a left and a right pocket selection are p and $1 - p$, respectively?

Solution. Let X be the number of matches that remain when a matchbox is found empty. For $k = 0, 1, \dots, n$, let L_k (or R_k) be the event that an empty box is first discovered in the left (respectively, right) pocket while the number of matches in the right (respectively, left) pocket is k at that time. The PMF of X is

$$p_X(k) = \mathbf{P}(L_k) + \mathbf{P}(R_k), \quad k = 0, 1, \dots, n.$$

Viewing a left and a right pocket selection as a “success” and a “failure,” respectively, $\mathbf{P}(L_k)$ is the probability that there are n successes in the first $2n - k$ trials, and trial $2n - k + 1$ is a success, or

$$\mathbf{P}(L_k) = \frac{1}{2} \binom{2n - k}{n} \left(\frac{1}{2}\right)^{2n-k}, \quad k = 0, 1, \dots, n.$$

By symmetry, $\mathbf{P}(L_k) = \mathbf{P}(R_k)$, so

$$p_X(k) = \mathbf{P}(L_k) + \mathbf{P}(R_k) = \binom{2n - k}{n} \left(\frac{1}{2}\right)^{2n-k}, \quad k = 0, 1, \dots, n.$$

In the more general case, where the probabilities of a left and a right pocket selection are p and $1 - p$, using a similar reasoning, we obtain

$$\mathbf{P}(L_k) = p \binom{2n - k}{n} p^n (1 - p)^{n-k}, \quad k = 0, 1, \dots, n,$$

and

$$\mathbf{P}(R_k) = (1 - p) \binom{2n - k}{n} p^{n-k} (1 - p)^n, \quad k = 0, 1, \dots, n,$$

which yields

$$\begin{aligned} p_X(k) &= \mathbf{P}(L_k) + \mathbf{P}(R_k) \\ &= \binom{2n - k}{n} (p^{n+1} (1 - p)^{n-k} + p^{n-k} (1 - p)^{n+1}), \quad k = 0, 1, \dots, n. \end{aligned}$$

Problem 12.* Justification of the Poisson approximation property. Consider the PMF of a binomial random variable with parameters n and p . Show that asymptotically, as

$$n \rightarrow \infty, \quad p \rightarrow 0,$$

while np is fixed at a given value λ , this PMF approaches the PMF of a Poisson random variable with parameter λ .

Solution. Using the equation $\lambda = np$, write the binomial PMF as

$$\begin{aligned} p_X(k) &= \frac{n!}{(n - k)! k!} p^k (1 - p)^{n-k} \\ &= \frac{n(n - 1) \cdots (n - k + 1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}. \end{aligned}$$

Fix k and let $n \rightarrow \infty$. We have, for $j = 1, \dots, k$,

$$\frac{n-k+j}{n} \rightarrow 1, \quad \left(1 - \frac{\lambda}{n}\right)^{-k} \rightarrow 1, \quad \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}.$$

Thus, for each fixed k , as $n \rightarrow \infty$ we obtain

$$p_X(k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

SECTION 2.3. Functions of Random Variables

Problem 13. A family has 5 natural children and has adopted 2 girls. Each natural child has equal probability of being a girl or a boy, independently of the other children. Find the PMF of the number of girls out of the 7 children.

Problem 14. Let X be a random variable that takes values from 0 to 9 with equal probability $1/10$.

- (a) Find the PMF of the random variable $Y = X \bmod(3)$.
- (b) Find the PMF of the random variable $Y = 5 \bmod(X + 1)$.

Problem 15. Let K be a random variable that takes, with equal probability $1/(2n+1)$, the integer values in the interval $[-n, n]$. Find the PMF of the random variable $Y = \ln X$, where $X = a^{|K|}$, and a is a positive number.

SECTION 2.4. Expectation, Mean, and Variance

Problem 16. Let X be a random variable with PMF

$$p_X(x) = \begin{cases} x^2/a, & \text{if } x = -3, -2, -1, 0, 1, 2, 3, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Find a and $E[X]$.
- (b) What is the PMF of the random variable $Z = (X - E[X])^2$?
- (c) Using part (b), compute the variance of X .
- (d) Compute the variance of X using the formula $\text{var}(X) = \sum_x (x - E[X])^2 p_X(x)$.

Problem 17. A city's temperature is modeled as a random variable with mean and standard deviation both equal to 10 degrees Celsius. A day is described as "normal" if the temperature during that day ranges within one standard deviation from the mean. What would be the temperature range for a normal day if temperature were expressed in degrees Fahrenheit?

Problem 18. Let a and b be positive integers with $a \leq b$, and let X be a random variable that takes as values, with equal probability, the powers of 2 in the interval $[2^a, 2^b]$. Find the expected value and the variance of X .

Problem 19. A prize is randomly placed in one of ten boxes, numbered from 1 to 10. You search for the prize by asking yes-no questions. Find the expected number of questions until you are sure about the location of the prize, under each of the following strategies.

- (a) An enumeration strategy: you ask questions of the form “is it in box k ?”.
- (b) A bisection strategy: you eliminate as close to half of the remaining boxes as possible by asking questions of the form “is it in a box numbered less than or equal to k ?”.

Problem 20. As an advertising campaign, a chocolate factory places golden tickets in some of its candy bars, with the promise that a golden ticket is worth a trip through the chocolate factory, and all the chocolate you can eat for life. If the probability of finding a golden ticket is p , find the mean and the variance of the number of candy bars you need to eat to find a ticket.

Problem 21. St. Petersburg paradox. You toss independently a fair coin and you count the number of tosses until the first tail appears. If this number is n , you receive 2^n dollars. What is the expected amount that you will receive? How much would you be willing to pay to play this game?

Problem 22. Two coins are simultaneously tossed until one of them comes up a head and the other a tail. The first coin comes up a head with probability p and the second with probability q . All tosses are assumed independent.

- (a) Find the PMF, the expected value, and the variance of the number of tosses.
- (b) What is the probability that the last toss of the first coin is a head?

Problem 23.

- (a) A fair coin is tossed successively until two consecutive heads or two consecutive tails appear. Find the PMF, the expected value, and the variance of the number of tosses.
- (b) Assume now that the coin is tossed successively until we obtain a tail that is immediately preceded by a head. Find the PMF and the expected value of the number of tosses.

Problem 24.* Variance of the Poisson. Consider the Poisson random variable with parameter λ . Compute its second moment and variance.

Solution. As shown in the text, the mean is given by

$$\mathbf{E}[X] = \lambda.$$

Also

$$\begin{aligned}\mathbf{E}[X^2] &= \sum_{k=1}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \lambda \sum_{k=1}^{\infty} k \frac{e^{-\lambda} \lambda^{k-1}}{(k-1)!}\end{aligned}$$

$$\begin{aligned}
&= \lambda \sum_{m=0}^{\infty} (m+1) \frac{e^{-\lambda} \lambda^m}{m!} \\
&= \lambda (\mathbf{E}[X] + 1) \\
&= \lambda(\lambda + 1),
\end{aligned}$$

from which

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda.$$

SECTION 2.5. Joint PMFs of Multiple Random Variables

Problem 25. A stock market trader buys 100 shares of stock A and 200 shares of stock B. Let X and Y be the price changes of A and B, respectively, over a certain time period, and assume that the joint PMF of X and Y is uniform over the set of integers x and y satisfying

$$-2 \leq x \leq 4, \quad -1 \leq y - x \leq 1.$$

- (a) Find the marginal PMFs and the means of X and Y .
- (b) Find the mean of the trader's profit.

Problem 26. A class of n students takes a test consisting of m questions. Suppose that student i submitted answers to the first m_i questions.

- (a) The grader randomly picks one answer, call it (I, J) , where I is the student ID number (taking values $1, \dots, n$) and J is the question number (taking values $1, \dots, m$). Assume that all answers are equally likely to be picked. Calculate the joint and the marginal PMFs of I and J .
- (b) Assume that an answer to question j , if submitted by student i , is correct with probability p_{ij} . Each answer gets a points if it is correct and gets b points otherwise. Calculate the expected value of the score of student i .

Problem 27. PMF of the minimum of several random variables. On a given day, your golf score takes values from the range 101 to 110, with probability 0.1, independently from other days. Determined to improve your score, you decide to play on three different days and declare as your score the minimum X of the scores X_1 , X_2 , and X_3 on the different days.

- (a) Calculate the PMF of X .
- (b) By how much has your expected score improved as a result of playing on three days?

Problem 28.* The quiz problem. Consider a quiz contest where a person is given a list of n questions and can answer these questions in any order he or she chooses. Question i will be answered correctly with probability p_i , and the person will then receive a reward v_i . At the first incorrect answer, the quiz terminates and the person is allowed to keep his or her previous rewards. The problem is to choose the ordering of questions so as to maximize the expected value of the total reward obtained. Show that it is optimal to answer questions in a nonincreasing order of $p_i v_i / (1 - p_i)$.

Solution. We will use a so-called interchange argument, which is often useful in scheduling-like problems. Let i and j be the k th and $(k+1)$ st questions in an optimally ordered list

$$L = (i_1, \dots, i_{k-1}, i, j, i_{k+2}, \dots, i_n).$$

Consider the list

$$L' = (i_1, \dots, i_{k-1}, j, i, i_{k+2}, \dots, i_n)$$

obtained from L by interchanging the order of questions i and j . We compute the expected values of the rewards of L and L' , and argue that since L is optimally ordered, we have

$$\mathbf{E}[\text{reward of } L] \geq \mathbf{E}[\text{reward of } L'].$$

Define the *weight* of question i to be

$$w(i) = \frac{p_i v_i}{(1 - p_i)}.$$

We will show that any permutation of the questions in a nonincreasing order of weights maximizes the expected reward.

If $L = (i_1, \dots, i_n)$ is a permutation of the questions, define $L^{(k)}$ to be the permutation obtained from L by interchanging questions i_k and i_{k+1} . Let us first compute the difference between the expected reward of L and that of $L^{(k)}$. We have

$$\mathbf{E}[\text{reward of } L] = p_{i_1} v_{i_1} + p_{i_1} p_{i_2} v_{i_2} + \dots + p_{i_1} \dots p_{i_n} v_{i_n},$$

and

$$\begin{aligned} \mathbf{E}[\text{reward of } L^{(k)}] &= p_{i_1} v_{i_1} + p_{i_1} p_{i_2} v_{i_2} + \dots + p_{i_1} \dots p_{i_{k-1}} v_{i_{k-1}} \\ &\quad + p_{i_1} \dots p_{i_{k-1}} p_{i_{k+1}} v_{i_{k+1}} + p_{i_1} \dots p_{i_{k-1}} p_{i_{k+1}} p_{i_k} v_{i_k} \\ &\quad + p_{i_1} \dots p_{i_{k+1}} v_{i_{k+1}} + \dots + p_{i_1} \dots p_{i_n} v_{i_n}. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbf{E}[\text{reward of } L^{(k)}] - \mathbf{E}[\text{reward of } L] &= p_{i_1} \dots p_{i_{k-1}} (p_{i_{k+1}} v_{i_{k+1}} + p_{i_{k+1}} p_{i_k} v_{i_k} \\ &\quad - p_{i_k} v_{i_k} - p_{i_k} p_{i_{k+1}} v_{i_{k+1}}) \\ &= p_{i_1} \dots p_{i_{k-1}} (1 - p_{i_k}) (1 - p_{i_{k+1}}) (w(i_{k+1}) - w(i_k)). \end{aligned}$$

Now, let us go back to our problem. Consider any permutation L of the questions. If $w(i_k) < w(i_{k+1})$ for some k , it follows from the above equation that the permutation $L^{(k)}$ has an expected reward larger than that of L . So, an optimal permutation of the questions must be in a nonincreasing order of weights.

Let us finally show that any two such permutations have equal expected rewards. Assume that L is such a permutation and say that $w(i_k) = w(i_{k+1})$ for some k . We know that interchanging i_k and i_{k+1} preserves the expected reward. So, the expected reward of any permutation L' in a non-increasing order of weights is equal to that of L , because L' can be obtained from L by repeatedly interchanging adjacent questions having equal weights.

Problem 29.* The inclusion-exclusion formula. Let A_1, A_2, \dots, A_n be events. Let $S_1 = \{i \mid 1 \leq i \leq n\}$, $S_2 = \{(i_1, i_2) \mid 1 \leq i_1 < i_2 \leq n\}$, and more generally, let S_m be the set of all m -tuples (i_1, \dots, i_m) of indices that satisfy $1 \leq i_1 < i_2 < \dots < i_m \leq n$. Show that

$$\begin{aligned}\mathbf{P}(\cup_{k=1}^n A_k) &= \sum_{i \in S_1} \mathbf{P}(A_i) - \sum_{(i_1, i_2) \in S_2} \mathbf{P}(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum_{(i_1, i_2, i_3) \in S_3} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \dots + (-1)^{n-1} \mathbf{P}(\cap_{k=1}^n A_k).\end{aligned}$$

Hint: Let X_i be a binary random variable which is equal to 1 when A_i occurs, and equal to 0 otherwise. Relate the event of interest to the random variable $(1 - X_1)(1 - X_2) \dots (1 - X_n)$.

Solution. Let us express the event $B = \cup_{k=1}^n A_k$ in terms of the random variables X_1, \dots, X_n . The event B^c occurs when all of the random variables X_1, \dots, X_n are zero, which happens when the random variable $Y = (1 - X_1)(1 - X_2) \dots (1 - X_n)$ is equal to 1. Note that Y can only take values in the set $\{0, 1\}$. Thus,

$$\mathbf{P}(B^c) = \mathbf{P}(Y = 1) = \mathbf{E}[Y].$$

Therefore,

$$\begin{aligned}\mathbf{P}(B) &= 1 - \mathbf{E}[(1 - X_1)(1 - X_2) \dots (1 - X_n)] \\ &= \mathbf{E}[X_1 + \dots + X_n] - \mathbf{E}\left[\sum_{(i_1, i_2) \in S_2} X_{i_1} X_{i_2}\right] + \dots + (-1)^{n-1} \mathbf{E}[X_1 \dots X_n].\end{aligned}$$

We note that

$$\begin{aligned}\mathbf{E}[X_i] &= \mathbf{P}(A_i), & \mathbf{E}[X_{i_1} X_{i_2}] &= \mathbf{P}(A_{i_1} \cap A_{i_2}), \\ \mathbf{E}[X_{i_1} X_{i_2} X_{i_3}] &= \mathbf{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}), & \mathbf{E}[X_1 X_2 \dots X_n] &= \mathbf{P}(\cap_{k=1}^n A_k),\end{aligned}$$

etc., from which the desired formula follows.

Problem 30.* Alvin's database of friends contains n entries, but due to a software glitch, the addresses correspond to the names in a totally random fashion. Alvin writes a holiday card to each of his friends and sends it to the (software-corrupted) address. What is the probability that at least one of his friends will get the correct card? *Hint:* Use the inclusion-exclusion formula.

Solution. Let A_k be the event that the k th card is sent to the correct address. We have for any k, j, i ,

$$\begin{aligned}\mathbf{P}(A_k) &= \frac{1}{n} = \frac{(n-1)!}{n!}, \\ \mathbf{P}(A_k \cap A_j) &= \mathbf{P}(A_k) \mathbf{P}(A_j \mid A_k) = \frac{1}{n} \cdot \frac{1}{n-1} = \frac{(n-2)!}{n!}, \\ \mathbf{P}(A_k \cap A_j \cap A_i) &= \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} = \frac{(n-3)!}{n!},\end{aligned}$$

etc., and

$$\mathbf{P}(\cap_{k=1}^n A_k) = \frac{1}{n!}.$$

Applying the inclusion-exclusion formula,

$$\begin{aligned} \mathbf{P}(\cup_{k=1}^n A_k) &= \sum_{i \in S_1} \mathbf{P}(A_i) - \sum_{(i_1, i_2) \in S_2} \mathbf{P}(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum_{(i_1, i_2, i_3) \in S_3} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n-1} \mathbf{P}(\cap_{k=1}^n A_k), \end{aligned}$$

we obtain the desired probability

$$\begin{aligned} \mathbf{P}(\cup_{k=1}^n A_k) &= \binom{n}{1} \frac{(n-1)!}{n!} - \binom{n}{2} \frac{(n-2)!}{n!} + \binom{n}{3} \frac{(n-3)!}{n!} - \cdots + (-1)^{n-1} \frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1} \frac{1}{n!}. \end{aligned}$$

When n is large, this probability can be approximated by $1 - e^{-1}$.

SECTION 2.6. Conditioning

Problem 31. Consider four independent rolls of a 6-sided die. Let X be the number of 1's and let Y be the number of 2's obtained. What is the joint PMF of X and Y ?

Problem 32. D. Bernoulli's problem of joint lives. Consider $2m$ persons forming m couples who live together at a given time. Suppose that at some later time, the probability of each person being alive is p , independently of other persons. At that later time, let A be the number of persons that are alive and let S be the number of couples in which both partners are alive. For any survivor number a , find $\mathbf{E}[S | A = a]$.

Problem 33.* A coin that has probability of heads equal to p , is tossed successively and independently until a head comes twice in a row or a tail comes twice in a row. Find the expected value of the number of tosses.

Solution. One possibility here is to calculate the PMF of X , the number of tosses until the game is over, and use it to compute $\mathbf{E}[X]$. However, with an unfair coin, this turns out to be cumbersome, so we argue by using the total expectation theorem and a suitable partition of the sample space. Let H_k (or T_k) be the event that a head (or a tail, respectively) comes at the k th toss, and let p (respectively, q) be the probability of H_k (respectively, T_k). Since H_1 and T_1 form a partition of the sample space, and $\mathbf{P}(H_1) = p$ and $\mathbf{P}(T_1) = q$, we have

$$\mathbf{E}[X] = p\mathbf{E}[X | H_1] + q\mathbf{E}[X | T_1].$$

Using again the total expectation theorem, we have

$$\mathbf{E}[X | H_1] = p\mathbf{E}[X | H_1 \cap H_2] + q\mathbf{E}[X | H_1 \cap T_2] = 2p + q(1 + \mathbf{E}[X | T_1]),$$

where we have used the fact

$$\mathbf{E}[X | H_1 \cap H_2] = 2$$

(since the game ends after two successive heads), and

$$\mathbf{E}[X | H_1 \cap T_2] = 1 + \mathbf{E}[X | T_1]$$

(since if the game is not over, only the last toss matters in determining the number of additional tosses up to termination). Similarly, we obtain

$$\mathbf{E}[X | T_1] = 2q + p(1 + \mathbf{E}[X | H_1]).$$

Combining the above two relations, collecting terms, and using the fact $p + q = 1$, we obtain after some calculation

$$\mathbf{E}[X | T_1] = \frac{2 + p^2}{1 - pq},$$

and similarly

$$\mathbf{E}[X | H_1] = \frac{2 + q^2}{1 - pq}.$$

Thus,

$$\mathbf{E}[X] = p \cdot \frac{2 + q^2}{1 - pq} + q \cdot \frac{2 + p^2}{1 - pq},$$

and finally, using the fact $p + q = 1$,

$$\mathbf{E}[X] = \frac{2 + pq}{1 - pq}.$$

In the case of a fair coin ($p = q = 1/2$), we obtain $\mathbf{E}[X] = 3$. It can also be verified that $2 \leq \mathbf{E}[X] \leq 3$ for all values of p .

Problem 34.* A spider and a fly move along a straight line. At each second, the fly moves a unit step to the right or to the left with equal probability p , and stays where it is with probability $1 - 2p$. The spider always takes a unit step in the direction of the fly. The spider and the fly start D units apart, where D is a random variable taking positive integer values with a given PMF. If the spider lands on top of the fly, it's the end. What is the expected value of the time it takes for this to happen?

Solution. Let T be the time at which the spider lands on top of the fly. We define

A_d : the event that initially the spider and the fly are d units apart,

B_d : the event that after one second the spider and the fly are d units apart.

Our approach will be to first apply the (conditional version of the) total expectation theorem to compute $\mathbf{E}[T | A_1]$, then use the result to compute $\mathbf{E}[T | A_2]$, and similarly compute sequentially $\mathbf{E}[T | A_d]$ for all relevant values of d . We will then apply the (unconditional version of the) total expectation theorem to compute $\mathbf{E}[T]$.

We have

$$A_d = (A_d \cap B_d) \cup (A_d \cap B_{d-1}) \cup (A_d \cap B_{d-2}), \quad \text{if } d > 1.$$

This is because if the spider and the fly are at a distance $d > 1$ apart, then one second later their distance will be d (if the fly moves away from the spider) or $d - 1$ (if the fly

does not move) or $d - 2$ (if the fly moves towards the spider). We also have, for the case where the spider and the fly start one unit apart,

$$A_1 = (A_1 \cap B_1) \cup (A_1 \cap B_0).$$

Using the total expectation theorem, we obtain

$$\begin{aligned} \mathbf{E}[T | A_d] &= \mathbf{P}(B_d | A_d) \mathbf{E}[T | A_d \cap B_d] \\ &\quad + \mathbf{P}(B_{d-1} | A_d) \mathbf{E}[T | A_d \cap B_{d-1}] \\ &\quad + \mathbf{P}(B_{d-2} | A_d) \mathbf{E}[T | A_d \cap B_{d-2}], \quad \text{if } d > 1, \end{aligned}$$

and

$$\mathbf{E}[T | A_1] = \mathbf{P}(B_1 | A_1) \mathbf{E}[T | A_1 \cap B_1] + \mathbf{P}(B_0 | A_1) \mathbf{E}[T | A_1 \cap B_0], \quad \text{if } d = 1.$$

It can be seen based on the problem data that

$$\mathbf{P}(B_1 | A_1) = 2p, \quad \mathbf{P}(B_0 | A_1) = 1 - 2p,$$

$$\mathbf{E}[T | A_1 \cap B_1] = 1 + \mathbf{E}[T | A_1], \quad \mathbf{E}[T | A_1 \cap B_0] = 1,$$

so by applying the formula for the case $d = 1$, we obtain

$$\mathbf{E}[T | A_1] = 2p(1 + \mathbf{E}[T | A_1]) + (1 - 2p),$$

or

$$\mathbf{E}[T | A_1] = \frac{1}{1 - 2p}.$$

By applying the formula with $d = 2$, we obtain

$$\mathbf{E}[T | A_2] = p \mathbf{E}[T | A_2 \cap B_2] + (1 - 2p) \mathbf{E}[T | A_2 \cap B_1] + p \mathbf{E}[T | A_2 \cap B_0].$$

We have

$$\begin{aligned} \mathbf{E}[T | A_2 \cap B_0] &= 1, \\ \mathbf{E}[T | A_2 \cap B_1] &= 1 + \mathbf{E}[T | A_1], \\ \mathbf{E}[T | A_2 \cap B_2] &= 1 + \mathbf{E}[T | A_2], \end{aligned}$$

so by substituting these relations in the expression for $\mathbf{E}[T | A_2]$, we obtain

$$\begin{aligned} \mathbf{E}[T | A_2] &= p(1 + \mathbf{E}[T | A_2]) + (1 - 2p)(1 + \mathbf{E}[T | A_1]) + p \\ &= p(1 + \mathbf{E}[T | A_2]) + (1 - 2p) \left(1 + \frac{1}{1 - 2p} \right) + p. \end{aligned}$$

This equation yields after some calculation

$$\mathbf{E}[T | A_2] = \frac{2}{1 - p}.$$

Generalizing, we obtain for $d > 2$,

$$\mathbf{E}[T | A_d] = p(1 + \mathbf{E}[T | A_d]) + (1 - 2p)(1 + \mathbf{E}[T | A_{d-1}]) + p(1 + \mathbf{E}[T | A_{d-2}]).$$

Thus, $\mathbf{E}[T | A_d]$ can be generated recursively for any initial distance d , using as initial conditions the values of $\mathbf{E}[T | A_1]$ and $\mathbf{E}[T | A_2]$ obtained earlier.

Finally, the expected value of T can be obtained using the given PMF for the initial distance D and the total expectation theorem:

$$\mathbf{E}[T] = \sum_d p_D(d) \mathbf{E}[T | A_d].$$

Problem 35.* Verify the expected value rule

$$\mathbf{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y),$$

using the expected value rule for a function of a single random variable. Then, use the rule for the special case of a linear function, to verify the formula

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y],$$

where a and b are given scalars.

Solution. We use the total expectation theorem to reduce the problem to the case of a single random variable. In particular, we have

$$\begin{aligned} \mathbf{E}[g(X, Y)] &= \sum_y p_Y(y) \mathbf{E}[g(X, Y) | Y = y] \\ &= \sum_y p_Y(y) \mathbf{E}[g(X, y) | Y = y] \\ &= \sum_y p_Y(y) \sum_x g(x, y) p_{X|Y}(x | y) \\ &= \sum_x \sum_y g(x, y) p_{X,Y}(x, y), \end{aligned}$$

as desired. Note that the third equality above used the expected value rule for the function $g(X, y)$ of a single random variable X .

For the linear special case, the expected value rule gives

$$\begin{aligned} \mathbf{E}[aX + bY] &= \sum_x \sum_y (ax + by) p_{X,Y}(x, y) \\ &= a \sum_x x \sum_y p_{X,Y}(x, y) + b \sum_y y \sum_x p_{X,Y}(x, y) \\ &= a \sum_x x p_X(x) + b \sum_y y p_Y(y) \\ &= a\mathbf{E}[X] + b\mathbf{E}[Y]. \end{aligned}$$

Problem 36.* The multiplication rule for conditional PMFs. Let X , Y , and Z be random variables.

(a) Show that

$$p_{X,Y,Z}(x, y, z) = p_X(x)p_{Y|X}(y|x)p_{Z|X,Y}(z|x, y).$$

(b) How can we interpret this formula as a special case of the multiplication rule given in Section 1.3?

(c) Generalize to the case of more than three random variables.

Solution. (a) We have

$$\begin{aligned} p_{X,Y,Z}(x, y, z) &= \mathbf{P}(X = x, Y = y, Z = z) \\ &= \mathbf{P}(X = x)\mathbf{P}(Y = y, Z = z | X = x) \\ &= \mathbf{P}(X = x)\mathbf{P}(Y = y | X = x)\mathbf{P}(Z = z | X = x, Y = y) \\ &= p_X(x)p_{Y|X}(y|x)p_{Z|X,Y}(z|x, y). \end{aligned}$$

(b) The formula can be written as

$$\mathbf{P}(X = x, Y = y, Z = z) = \mathbf{P}(X = x)\mathbf{P}(Y = y | X = x)\mathbf{P}(Z = z | X = x, Y = y),$$

which is a special case of the multiplication rule.

(c) The generalization is

$$\begin{aligned} p_{X_1, \dots, X_n}(x_1, \dots, x_n) \\ = p_{X_1}(x_1)p_{X_2|X_1}(x_2|x_1) \cdots p_{X_n|X_1, \dots, X_{n-1}}(x_n|x_1, \dots, x_{n-1}). \end{aligned}$$

Problem 37.* Splitting a Poisson random variable. A transmitter sends out either a 1 with probability p , or a 0 with probability $1 - p$, independently of earlier transmissions. If the number of transmissions within a given time interval has a Poisson PMF with parameter λ , show that the number of 1's transmitted in that same time interval has a Poisson PMF with parameter $p\lambda$.

Solution. Let X and Y be the numbers of 1's and 0's transmitted, respectively. Let $Z = X + Y$ be the total number of symbols transmitted. We have

$$\begin{aligned} \mathbf{P}(X = n, Y = m) &= \mathbf{P}(X = n, Y = m | Z = n + m)\mathbf{P}(Z = n + m) \\ &= \binom{n+m}{n} p^n (1-p)^m \cdot \frac{e^{-\lambda} \lambda^{n+m}}{(n+m)!} \\ &= \frac{e^{-\lambda p} (\lambda p)^n}{n!} \cdot \frac{e^{-\lambda(1-p)} (\lambda(1-p))^m}{m!}. \end{aligned}$$

Thus,

$$\begin{aligned}
 \mathbf{P}(X = n) &= \sum_{m=0}^{\infty} \mathbf{P}(X = n, Y = m) \\
 &= \frac{e^{-\lambda p} (\lambda p)^n}{n!} e^{-\lambda(1-p)} \sum_{m=0}^{\infty} \frac{(\lambda(1-p))^m}{m!} \\
 &= \frac{e^{-\lambda p} (\lambda p)^n}{n!} e^{-\lambda(1-p)} e^{\lambda(1-p)} \\
 &= \frac{e^{-\lambda p} (\lambda p)^n}{n!},
 \end{aligned}$$

so that X is Poisson with parameter λp .

SECTION 2.7. Independence

Problem 38. Alice passes through four traffic lights on her way to work, and each light is equally likely to be green or red, independently of the others.

- (a) What is the PMF, the mean, and the variance of the number of red lights that Alice encounters?
- (b) Suppose that each red light delays Alice by exactly two minutes. What is the variance of Alice's commuting time?

Problem 39. Each morning, Hungry Harry eats some eggs. On any given morning, the number of eggs he eats is equally likely to be 1, 2, 3, 4, 5, or 6, independently of what he has done in the past. Let X be the number of eggs that Harry eats in 10 days. Find the mean and variance of X .

Problem 40. A particular professor is known for his arbitrary grading policies. Each paper receives a grade from the set $\{A, A-, B+, B, B-, C+\}$, with equal probability, independently of other papers. How many papers do you expect to hand in before you receive each possible grade at least once?

Problem 41. You drive to work 5 days a week for a full year (50 weeks), and with probability $p = 0.02$ you get a traffic ticket on any given day, independently of other days. Let X be the total number of tickets you get in the year.

- (a) What is the probability that the number of tickets you get is exactly equal to the expected value of X ?
- (b) Calculate approximately the probability in (a) using a Poisson approximation.
- (c) Any one of the tickets is \$10 or \$20 or \$50 with respective probabilities 0.5, 0.3, and 0.2, and independently of other tickets. Find the mean and the variance of the amount of money you pay in traffic tickets during the year.
- (d) Suppose you don't know the probability p of getting a ticket, but you got 5 tickets during the year, and you estimate p by the sample mean

$$\hat{p} = \frac{5}{250} = 0.02.$$

What is the range of possible values of p assuming that the difference between p and the sample mean \hat{p} is within 5 times the standard deviation of the sample mean?

Problem 42. Computational problem. Here is a probabilistic method for computing the area of a given subset S of the unit square. The method uses a sequence of independent random selections of points in the unit square $[0, 1] \times [0, 1]$, according to a uniform probability law. If the i th point belongs to the subset S the value of a random variable X_i is set to 1, and otherwise it is set to 0. Let X_1, X_2, \dots be the sequence of random variables thus defined, and for any n , let

$$S_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

- (a) Show that $E[S_n]$ is equal to the area of the subset S , and that $\text{var}(S_n)$ diminishes to 0 as n increases.
- (b) Show that to calculate S_n , it is sufficient to know S_{n-1} and X_n , so the past values of X_k , $k = 1, \dots, n-1$, do not need to be remembered. Give a formula.
- (c) Write a computer program to generate S_n for $n = 1, 2, \dots, 10000$, using the computer's random number generator, for the case where the subset S is the circle inscribed within the unit square. How can you use your program to measure experimentally the value of π ?
- (d) Use a similar computer program to calculate approximately the area of the set of all (x, y) that lie within the unit square and satisfy $0 \leq \cos \pi x + \sin \pi y \leq 1$.

Problem 43.* Suppose that X and Y are independent, identically distributed, geometric random variables with parameter p . Show that

$$P(X = i | X + Y = n) = \frac{1}{n-1}, \quad i = 1, \dots, n-1.$$

Solution. We can interpret $P(X = i | X + Y = n)$ as the probability that a coin will come up a head for the first time on the i th toss given that it came up a head for the second time on the n th toss. We can then argue, intuitively, that given that the second head occurred on the n th toss, the first head is equally likely to have come up at any toss between 1 and $n-1$. To establish this precisely, note that we have

$$P(X = i | X + Y = n) = \frac{P(X = i, X + Y = n)}{P(X + Y = n)} = \frac{P(X = i)P(Y = n - i)}{P(X + Y = n)}.$$

Also

$$P(X = i) = p(1-p)^{i-1}, \quad \text{for } i \geq 1,$$

and

$$P(Y = n - i) = p(1-p)^{n-i-1}, \quad \text{for } n - i \geq 1.$$

It follows that

$$P(X = i)P(Y = n - i) = \begin{cases} p^2(1-p)^{n-2}, & \text{if } i = 1, \dots, n-1, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, for any i and j in the range $[1, n - 1]$, we have

$$\mathbf{P}(X = i | X + Y = n) = \mathbf{P}(X = j | X + Y = n).$$

Hence

$$\mathbf{P}(X = i | X + Y = n) = \frac{1}{n-1}, \quad i = 1, \dots, n-1.$$

Problem 44.* Let X and Y be two random variables with given joint PMF, and let g and h be two functions of X and Y , respectively. Show that if X and Y are independent, then the same is true for the random variables $g(X)$ and $h(Y)$.

Solution. Let $U = g(X)$ and $V = h(Y)$. Then, we have

$$\begin{aligned} p_{U,V}(u, v) &= \sum_{\{(x,y) \mid g(x)=u, h(y)=v\}} p_{X,Y}(x, y) \\ &= \sum_{\{(x,y) \mid g(x)=u, h(y)=v\}} p_X(x)p_Y(y) \\ &= \sum_{\{x \mid g(x)=u\}} p_X(x) \sum_{\{y \mid h(y)=v\}} p_Y(y) \\ &= p_U(u)p_V(v), \end{aligned}$$

so U and V are independent.

Problem 45.* Variability extremes. Let X_1, \dots, X_n be independent random variables and let $X = X_1 + \dots + X_n$ be their sum.

- (a) Suppose that each X_i is Bernoulli with parameter p_i , and that p_1, \dots, p_n are chosen so that the mean of X is a given $\mu > 0$. Show that the variance of X is maximized if the p_i are chosen to be all equal to μ/n .
- (b) Suppose that each X_i is geometric with parameter p_i , and that p_1, \dots, p_n are chosen so that the mean of X is a given $\mu > 0$. Show that the variance of X is minimized if the p_i are chosen to be all equal to n/μ . [Note the strikingly different character of the results of parts (a) and (b).]

Solution. (a) We have

$$\text{var}(X) = \sum_{i=1}^n \text{var}(X_i) = \sum_{i=1}^n p_i(1-p_i) = \mu - \sum_{i=1}^n p_i^2.$$

Thus maximizing the variance is equivalent to minimizing $\sum_{i=1}^n p_i^2$. It can be seen (using the constraint $\sum_{i=1}^n p_i = \mu$) that

$$\sum_{i=1}^n p_i^2 = \sum_{i=1}^n (\mu/n)^2 + \sum_{i=1}^n (p_i - \mu/n)^2,$$

so $\sum_{i=1}^n p_i^2$ is minimized when $p_i = \mu/n$ for all i .

(b) We have

$$\mu = \sum_{i=1}^n \mathbf{E}[X_i] = \sum_{i=1}^n \frac{1}{p_i},$$

and

$$\text{var}(X) = \sum_{i=1}^n \text{var}(X_i) = \sum_{i=1}^n \frac{1-p_i}{p_i^2}.$$

Introducing the change of variables $y_i = 1/p_i = \mathbf{E}[X_i]$, we see that the constraint becomes

$$\sum_{i=1}^n y_i = \mu,$$

and that we must minimize

$$\sum_{i=1}^n y_i(y_i - 1) = \sum_{i=1}^n y_i^2 - \mu,$$

subject to that constraint. This is the same problem as the one of part (a), so the method of proof given there applies.

Problem 46.* Entropy and uncertainty. Consider a random variable X that can take n values, x_1, \dots, x_n , with corresponding probabilities p_1, \dots, p_n . The **entropy** of X is defined to be

$$H(X) = - \sum_{i=1}^n p_i \log p_i.$$

(All logarithms in this problem are with respect to base two.) The entropy $H(X)$ provides a measure of the uncertainty about the value of X . To get a sense of this, note that $H(X) \geq 0$ and that $H(X)$ is very close to 0 when X is “nearly deterministic,” i.e., takes one of its possible values with probability very close to 1 (since we have $p \log p \approx 0$ if either $p \approx 0$ or $p \approx 1$).

The notion of entropy is fundamental in information theory, which originated with C. Shannon’s famous work and is described in many specialized textbooks. For example, it can be shown that $H(X)$ is a lower bound to the average number of yes-no questions (such as “is $X = x_1$?” or “is $X < x_5$?”) that must be asked in order to determine the value of X . Furthermore, if k is the average number of questions required to determine the value of a string of independent identically distributed random variables X_1, X_2, \dots, X_n , then, with a suitable strategy, k/n can be made as close to $H(X)$ as desired, when n is large.

(a) Show that if q_1, \dots, q_n are nonnegative numbers such that $\sum_{i=1}^n q_i = 1$, then

$$H(X) \leq - \sum_{i=1}^n p_i \log q_i,$$

with equality if and only if $p_i = q_i$ for all i . As a special case, show that $H(X) \leq \log n$, with equality if and only if $p_i = 1/n$ for all i . *Hint:* Use the inequality $\ln \alpha \leq \alpha - 1$, for $\alpha > 0$, which holds with equality if and only if $\alpha = 1$.

- (b) Let X and Y be random variables taking a finite number of values, and having joint PMF $p_{X,Y}(x,y)$. Define

$$I(X, Y) = \sum_x \sum_y p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \right).$$

Show that $I(X, Y) \geq 0$, and that $I(X, Y) = 0$ if and only if X and Y are independent.

- (c) Show that

$$I(X, Y) = H(X) + H(Y) - H(X, Y),$$

where

$$H(X, Y) = - \sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y),$$

$$H(X) = - \sum_x p_X(x) \log p_X(x), \quad H(Y) = - \sum_y p_Y(y) \log p_Y(y).$$

- (d) Show that

$$I(X, Y) = H(X) - H(X|Y),$$

where

$$H(X|Y) = - \sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log p_{X|Y}(x|y).$$

[Note that $H(X|Y)$ may be viewed as the conditional entropy of X given Y , that is, the entropy of the conditional distribution of X , given that $Y = y$, averaged over all possible values y . Thus, the quantity $I(X, Y) = H(X) - H(X|Y)$ is the reduction in the entropy (uncertainty) on X , when Y becomes known. It can be therefore interpreted as the information about X that is conveyed by Y , and is called the **mutual information** of X and Y .]

Solution. (a) We will use the inequality $\ln \alpha \leq \alpha - 1$. (To see why this inequality is true, write $\ln \alpha = \int_1^\alpha \beta^{-1} d\beta < \int_1^\alpha d\beta = \alpha - 1$ for $\alpha > 1$, and write $\ln \alpha = - \int_\alpha^1 \beta^{-1} d\beta < - \int_\alpha^1 d\beta = \alpha - 1$ for $0 < \alpha < 1$.)

We have

$$- \sum_{i=1}^n p_i \ln p_i + \sum_{i=1}^n p_i \ln q_i = \sum_{i=1}^n p_i \ln \left(\frac{q_i}{p_i} \right) \leq \sum_{i=1}^n p_i \left(\frac{q_i}{p_i} - 1 \right) = 0,$$

with equality if and only if $p_i = q_i$ for all i . Since $\ln p = \log p \ln 2$, we obtain the desired relation $H(X) \leq - \sum_{i=1}^n p_i \log q_i$. The inequality $H(X) \leq \log n$ is obtained by setting $q_i = 1/n$ for all i .

- (b) The numbers $p_X(x)p_Y(y)$ satisfy $\sum_x \sum_y p_X(x)p_Y(y) = 1$, so by part (a), we have

$$\sum_x \sum_y p_{X,Y}(x,y) \log(p_{X,Y}(x,y)) \geq \sum_x \sum_y p_{X,Y}(x,y) \log(p_X(x)p_Y(y)),$$

with equality if and only if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \text{for all } x \text{ and } y,$$

which is equivalent to X and Y being independent.

(c) We have

$$I(X, Y) = \sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y) - \sum_x \sum_y p_{X,Y}(x,y) \log(p_X(x)p_Y(y)),$$

and

$$\sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y) = -H(X, Y),$$

$$\begin{aligned} -\sum_x \sum_y p_{X,Y}(x,y) \log(p_X(x)p_Y(y)) &= -\sum_x \sum_y p_{X,Y}(x,y) \log p_X(x) \\ &\quad - \sum_x \sum_y p_{X,Y}(x,y) \log p_Y(y) \\ &= -\sum_x p_X(x) \log p_X(x) - \sum_y p_Y(y) \log p_Y(y) \\ &= H(X) + H(Y). \end{aligned}$$

Combining the above three relations, we obtain $I(X, Y) = H(X) + H(Y) - H(X, Y)$.

(d) From the calculation in part (c), we have

$$\begin{aligned} I(X, Y) &= \sum_x \sum_y p_{X,Y}(x,y) \log p_{X,Y}(x,y) - \sum_x p_X(x) \log p_X(x) \\ &\quad - \sum_x \sum_y p_{X,Y}(x,y) \log p_Y(y) \\ &= H(X) + \sum_x \sum_y p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_Y(y)} \right) \\ &= H(X) + \sum_x \sum_y p_Y(y) p_{X|Y}(x|y) \log p_{X|Y}(x|y) \\ &= H(X) - H(X|Y). \end{aligned}$$

General Random Variables

Contents

3.1. Continuous Random Variables and PDFs	p. 140
3.2. Cumulative Distribution Functions	p. 148
3.3. Normal Random Variables	p. 152
3.4. Conditioning on an Event	p. 158
3.5. Multiple Continuous Random Variables	p. 164
3.6. Derived Distributions	p. 179
3.7. Summary and Discussion	p. 190
Problems	p. 192

Random variables with a continuous range of possible values are quite common; the velocity of a vehicle traveling along the highway could be one example. If the velocity is measured by a digital speedometer, we may view the speedometer's reading as a discrete random variable. But if we wish to model the exact velocity, a continuous random variable is called for. Models involving continuous random variables can be useful for several reasons. Besides being finer-grained and possibly more accurate, they allow the use of powerful tools from calculus and often admit an insightful analysis that would not be possible under a discrete model.

All of the concepts and methods introduced in Chapter 2, such as expectation, PMFs, and conditioning, have continuous counterparts. Developing and interpreting these counterparts is the subject of this chapter.

3.1 CONTINUOUS RANDOM VARIABLES AND PDFS

A random variable X is called **continuous** if there is a nonnegative function f_X , called the **probability density function of X** , or PDF for short, such that

$$\mathbb{P}(X \in B) = \int_B f_X(x) dx,$$

for every subset B of the real line.[†] In particular, the probability that the value of X falls within an interval is

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx,$$

and can be interpreted as the area under the graph of the PDF (see Fig. 3.1). For any single value a , we have $\mathbb{P}(X = a) = \int_a^a f_X(x) dx = 0$. For this reason, including or excluding the endpoints of an interval has no effect on its probability:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b).$$

Note that to qualify as a PDF, a function f_X must be nonnegative, i.e., $f_X(x) \geq 0$ for every x , and must also have the normalization property

$$\int_{-\infty}^{\infty} f_X(x) dx = \mathbb{P}(-\infty < X < \infty) = 1.$$

[†] The integral $\int_B f_X(x) dx$ is to be interpreted in the usual calculus/Riemann sense and we implicitly assume that it is well-defined. For highly unusual functions and sets, this integral can be harder – or even impossible – to define, but such issues belong to a more advanced treatment of the subject. In any case, it is comforting to know that mathematical subtleties of this type do not arise if f_X is a piecewise continuous function with a finite or countable number of points of discontinuity, and B is the union of a finite or countable number of intervals.

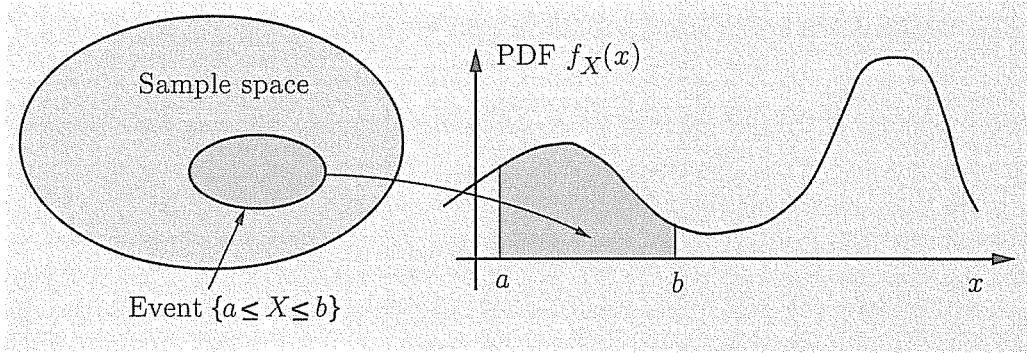


Figure 3.1: Illustration of a PDF. The probability that X takes a value in an interval $[a, b]$ is $\int_a^b f_X(x) dx$, which is the shaded area in the figure.

Graphically, this means that the entire area under the graph of the PDF must be equal to 1.

To interpret the PDF, note that for an interval $[x, x + \delta]$ with very small length δ , we have

$$\mathbf{P}([x, x + \delta]) = \int_x^{x+\delta} f_X(t) dt \approx f_X(x) \cdot \delta,$$

so we can view $f_X(x)$ as the “probability mass per unit length” near x (cf. Fig. 3.2). It is important to realize that even though a PDF is used to calculate event probabilities, $f_X(x)$ is not the probability of any particular event. In particular, it is not restricted to be less than or equal to one.

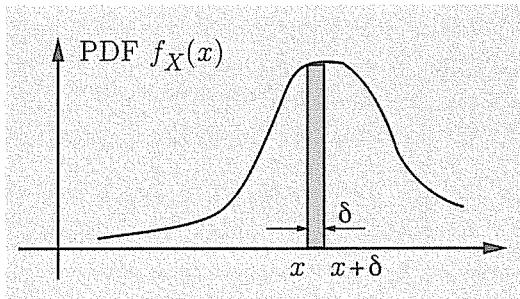


Figure 3.2: Interpretation of the PDF $f_X(x)$ as “probability mass per unit length” around x . If δ is very small, the probability that X takes a value in the interval $[x, x + \delta]$ is the shaded area in the figure, which is approximately equal to $f_X(x) \cdot \delta$.

Example 3.1. Continuous Uniform Random Variable. A gambler spins a wheel of fortune, continuously calibrated between 0 and 1, and observes the resulting number. Assuming that any two subintervals of $[0,1]$ of the same length have the same probability, this experiment can be modeled in terms of a random variable X with PDF

$$f_X(x) = \begin{cases} c, & \text{if } 0 \leq x \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

for some constant c . This constant can be determined by using the normalization property

$$1 = \int_{-\infty}^{\infty} f_X(x) dx = \int_0^1 c dx = c \int_0^1 dx = c,$$

so that $c = 1$.

More generally, we can consider a random variable X that takes values in an interval $[a, b]$, and again assume that any two subintervals of the same length have the same probability. We refer to this type of random variable as **uniform** or **uniformly distributed**. Its PDF has the form

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise,} \end{cases}$$

(cf. Fig. 3.3). The constant value of the PDF within $[a, b]$ is determined from the normalization property. Indeed, we have

$$1 = \int_{-\infty}^{\infty} f_X(x) dx = \int_a^b \frac{1}{b-a} dx.$$

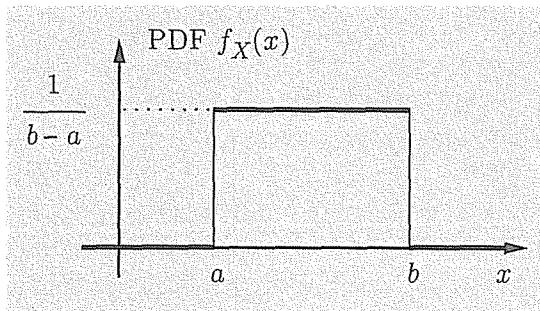


Figure 3.3: The PDF of a uniform random variable.

Example 3.2. Piecewise Constant PDF. Alvin's driving time to work is between 15 and 20 minutes if the day is sunny, and between 20 and 25 minutes if the day is rainy, with all times being equally likely in each case. Assume that a day is sunny with probability $2/3$ and rainy with probability $1/3$. What is the PDF of the driving time, viewed as a random variable X ?

We interpret the statement that "all times are equally likely" in the sunny and the rainy cases, to mean that the PDF of X is constant in each of the intervals $[15, 20]$ and $[20, 25]$. Furthermore, since these two intervals contain all possible driving times, the PDF should be zero everywhere else:

$$f_X(x) = \begin{cases} c_1, & \text{if } 15 \leq x < 20, \\ c_2, & \text{if } 20 \leq x \leq 25, \\ 0, & \text{otherwise,} \end{cases}$$

where c_1 and c_2 are some constants. We can determine these constants by using the given probabilities of a sunny and of a rainy day:

$$\frac{2}{3} = \mathbb{P}(\text{sunny day}) = \int_{15}^{20} f_X(x) dx = \int_{15}^{20} c_1 dx = 5c_1,$$

$$\frac{1}{3} = \mathbb{P}(\text{rainy day}) = \int_{20}^{25} f_X(x) dx = \int_{20}^{25} c_2 dx = 5c_2,$$

so that

$$c_1 = \frac{2}{15}, \quad c_2 = \frac{1}{15}.$$

Generalizing this example, consider a random variable X whose PDF has the piecewise constant form

$$f_X(x) = \begin{cases} c_i, & \text{if } a_i \leq x < a_{i+1}, \quad i = 1, 2, \dots, n-1, \\ 0, & \text{otherwise,} \end{cases}$$

where a_1, a_2, \dots, a_n are some scalars with $a_i < a_{i+1}$ for all i , and c_1, c_2, \dots, c_n are some nonnegative constants (cf. Fig. 3.4). The constants c_i may be determined by additional problem data, as in the preceding driving context. Generally, the c_i must be such that the normalization property holds:

$$1 = \int_{a_1}^{a_n} f_X(x) dx = \sum_{i=1}^{n-1} \int_{a_i}^{a_{i+1}} c_i dx = \sum_{i=1}^{n-1} c_i (a_{i+1} - a_i).$$

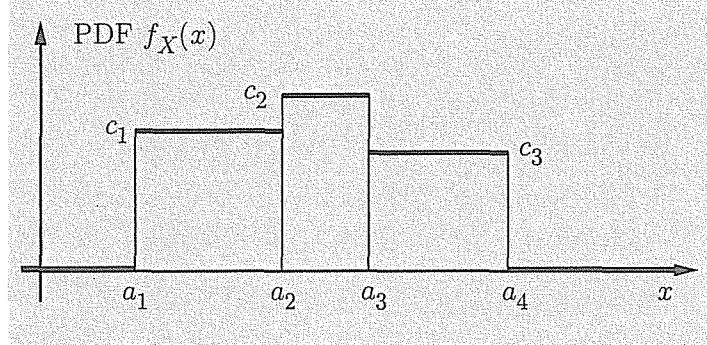


Figure 3.4: A piecewise constant PDF involving three intervals.

Example 3.3. A PDF Can Take Arbitrarily Large Values. Consider a random variable X with PDF

$$f_X(x) = \begin{cases} \frac{1}{2\sqrt{x}}, & \text{if } 0 < x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Even though $f_X(x)$ becomes infinitely large as x approaches zero, this is still a valid PDF, because

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_0^1 \frac{1}{2\sqrt{x}} dx = \sqrt{x} \Big|_0^1 = 1.$$

Summary of PDF Properties

Let X be a continuous random variable with PDF f_X .

- $f_X(x) \geq 0$ for all x .
- $\int_{-\infty}^{\infty} f_X(x) dx = 1$.
- If δ is very small, then $\mathbb{P}([x, x + \delta]) \approx f_X(x) \cdot \delta$.
- For any subset B of the real line,

$$\mathbb{P}(X \in B) = \int_B f_X(x) dx.$$

Expectation

The **expected value** or **mean** of a continuous random variable X is defined by[†]

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

This is similar to the discrete case except that the PMF is replaced by the PDF, and summation is replaced by integration. As in Chapter 2, $\mathbb{E}[X]$ can be interpreted as the “center of gravity” of the PDF and, also, as the anticipated average value of X in a large number of independent repetitions of the experiment. Its mathematical properties are similar to the discrete case – after all, an integral is just a limiting form of a sum.

If X is a continuous random variable with given PDF, any real-valued function $Y = g(X)$ of X is also a random variable. Note that Y can be a continuous random variable: for example, consider the trivial case where $Y = g(X) = X$. But Y can also turn out to be discrete. For example, suppose that

[†] One has to deal with the possibility that the integral $\int_{-\infty}^{\infty} x f_X(x) dx$ is infinite or undefined. More concretely, we will say that the expectation is well-defined if $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$. In that case, it is known that the integral $\int_{-\infty}^{\infty} x f_X(x) dx$ takes a finite and unambiguous value.

For an example where the expectation is not well-defined, consider a random variable X with PDF $f_X(x) = c/(1 + x^2)$, where c is a constant chosen to enforce the normalization condition. The expression $|x| f_X(x)$ can be approximated by $c/|x|$ when $|x|$ is large. Using the fact $\int_1^{\infty} (1/x) dx = \infty$, one can show that $\int_{-\infty}^{\infty} |x| f_X(x) dx = \infty$. Thus, $\mathbb{E}[X]$ is left undefined, despite the symmetry of the PDF around zero.

Throughout this book, in the absence of an indication to the contrary, we implicitly assume that the expected value of any random variable of interest is well-defined.

$g(x) = 1$ for $x > 0$, and $g(x) = 0$, otherwise. Then $Y = g(X)$ is a discrete random variable taking values in the finite set $\{0, 1\}$. In either case, the mean of $g(X)$ satisfies the expected value rule

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

in complete analogy with the discrete case; see the end-of-chapter problems.

The n th moment of a continuous random variable X is defined as $\mathbb{E}[X^n]$, the expected value of the random variable X^n . The variance, denoted by $\text{var}(X)$, is defined as the expected value of the random variable $(X - \mathbb{E}[X])^2$.

We now summarize this discussion and list a number of additional facts that are practically identical to their discrete counterparts.

Expectation of a Continuous Random Variable and its Properties

Let X be a continuous random variable with PDF f_X .

- The expectation of X is defined by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

- The expected value rule for a function $g(X)$ has the form

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

- The variance of X is defined by

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x) dx.$$

- We have

$$0 \leq \text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

- If $Y = aX + b$, where a and b are given scalars, then

$$\mathbb{E}[Y] = a\mathbb{E}[X] + b, \quad \text{var}(Y) = a^2\text{var}(X).$$

Example 3.4. Mean and Variance of the Uniform Random Variable.

Consider the case of a uniform PDF over an interval $[a, b]$, as in Example 3.1. We

have

$$\begin{aligned}
 \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\
 &= \int_a^b x \cdot \frac{1}{b-a} dx \\
 &= \frac{1}{b-a} \cdot \frac{1}{2} x^2 \Big|_a^b \\
 &= \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} \\
 &= \frac{a+b}{2},
 \end{aligned}$$

as one expects based on the symmetry of the PDF around $(a+b)/2$.

To obtain the variance, we first calculate the second moment. We have

$$\begin{aligned}
 \mathbb{E}[X^2] &= \int_a^b \frac{x^2}{b-a} dx \\
 &= \frac{1}{b-a} \int_a^b x^2 dx \\
 &= \frac{1}{b-a} \cdot \frac{1}{3} x^3 \Big|_a^b \\
 &= \frac{b^3 - a^3}{3(b-a)} \\
 &= \frac{a^2 + ab + b^2}{3}.
 \end{aligned}$$

Thus, the variance is obtained as

$$\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12},$$

after some calculation.

Exponential Random Variable

An exponential random variable has a PDF of the form

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where λ is a positive parameter characterizing the PDF (see Fig. 3.5). This is a legitimate PDF because

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_0^{\infty} = 1.$$

Note that the probability that X exceeds a certain value decreases exponentially. Indeed, for any $a \geq 0$, we have

$$\mathbb{P}(X \geq a) = \int_a^{\infty} \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_a^{\infty} = e^{-\lambda a}.$$

An exponential random variable can, for example, be a good model for the amount of time until a piece of equipment breaks down, until a light bulb burns out, or until an accident occurs. It will play a major role in our study of random processes in Chapter 5, but for the time being we will simply view it as a special random variable that is fairly tractable analytically.

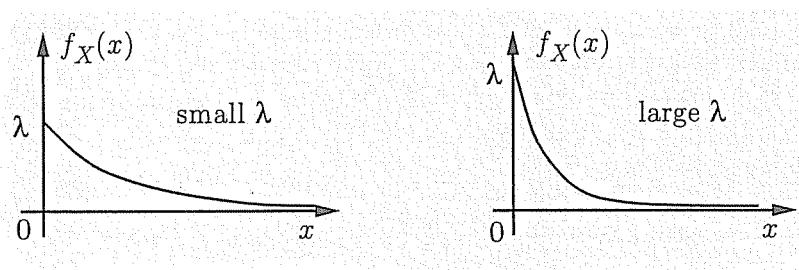


Figure 3.5: The PDF $\lambda e^{-\lambda x}$ of an exponential random variable.

The mean and the variance can be calculated to be

$$\mathbb{E}[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2}.$$

These formulas can be verified by straightforward calculation, as we now show. We have, using integration by parts,

$$\begin{aligned} \mathbb{E}[X] &= \int_0^{\infty} x \lambda e^{-\lambda x} dx \\ &= (-xe^{-\lambda x}) \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx \\ &= 0 - \frac{e^{-\lambda x}}{\lambda} \Big|_0^{\infty} \\ &= \frac{1}{\lambda}. \end{aligned}$$

Using again integration by parts, the second moment is

$$\begin{aligned}\mathbb{E}[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= (-x^2 e^{-\lambda x}) \Big|_0^\infty + \int_0^\infty 2x e^{-\lambda x} dx \\ &= 0 + \frac{2}{\lambda} \mathbb{E}[X] \\ &= \frac{2}{\lambda^2}.\end{aligned}$$

Finally, using the formula $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, we obtain

$$\text{var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Example 3.5. The time until a small meteorite first lands anywhere in the Sahara desert is modeled as an exponential random variable with a mean of 10 days. The time is currently midnight. What is the probability that a meteorite first lands some time between 6 a.m. and 6 p.m. of the first day?

Let X be the time elapsed until the event of interest, measured in days. Then, X is exponential, with mean $1/\lambda = 10$, which yields $\lambda = 1/10$. The desired probability is

$$\mathbb{P}(1/4 \leq X \leq 3/4) = \mathbb{P}(X \geq 1/4) - \mathbb{P}(X > 3/4) = e^{-1/40} - e^{-3/40} = 0.0476,$$

where we have used the formula $\mathbb{P}(X \geq a) = \mathbb{P}(X > a) = e^{-\lambda a}$.

.2 CUMULATIVE DISTRIBUTION FUNCTIONS

We have been dealing with discrete and continuous random variables in a somewhat different manner, using PMFs and PDFs, respectively. It would be desirable to describe all kinds of random variables with a single mathematical concept. This is accomplished with the **cumulative distribution function**, or CDF for short. The CDF of a random variable X is denoted by F_X and provides the probability $\mathbb{P}(X \leq x)$. In particular, for every x we have

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} \sum_{k \leq x} p_X(k), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^x f_X(t) dt, & \text{if } X \text{ is continuous.} \end{cases}$$

Loosely speaking, the CDF $F_X(x)$ “accumulates” probability “up to” the value x .

Any random variable associated with a given probability model has a CDF, regardless of whether it is discrete, continuous, or other. This is because $\{X \leq x\}$ is always an event and therefore has a well-defined probability. In what follows, any unambiguous specification of the probabilities of all events of the form $\{X \leq x\}$, be it through a PMF, PDF, or CDF, will be referred to as the **probability law** of the random variable X .

Figures 3.6 and 3.7 illustrate the CDFs of various discrete and continuous random variables. From these figures, as well as from the definition, some general properties of the CDF can be observed.

Properties of a CDF

The CDF F_X of a random variable X is defined by

$$F_X(x) = \mathbb{P}(X \leq x), \quad \text{for all } x,$$

and has the following properties.

- F_X is monotonically nondecreasing:

$$\text{if } x \leq y, \text{ then } F_X(x) \leq F_X(y).$$

- $F_X(x)$ tends to 0 as $x \rightarrow -\infty$, and to 1 as $x \rightarrow \infty$.
- If X is discrete, then $F_X(x)$ is a piecewise constant function of x .
- If X is continuous, then $F_X(x)$ is a continuous function of x .
- If X is discrete and takes integer values, the PMF and the CDF can be obtained from each other by summing or differencing:

$$F_X(k) = \sum_{i=-\infty}^k p_X(i),$$

$$p_X(k) = \mathbb{P}(X \leq k) - \mathbb{P}(X \leq k-1) = F_X(k) - F_X(k-1),$$

for all integers k .

- If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt, \quad f_X(x) = \frac{dF_X}{dx}(x).$$

(The second equality is valid for those x for which the CDF has a derivative.)

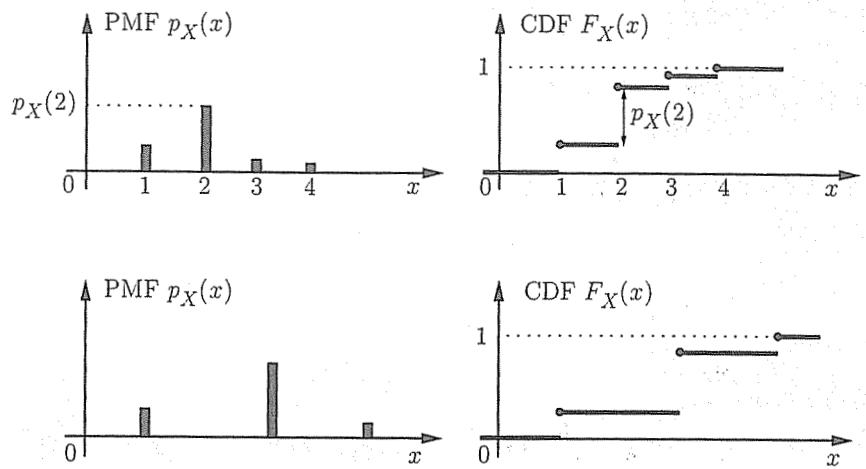


Figure 3.6: CDFs of some discrete random variables. The CDF is related to the PMF through the formula

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{k \leq x} p_X(k)$$

and has a staircase form, with jumps occurring at the values of positive probability mass. Note that at the points where a jump occurs, the value of F_X is the larger of the two corresponding values (i.e., F_X is continuous from the right).

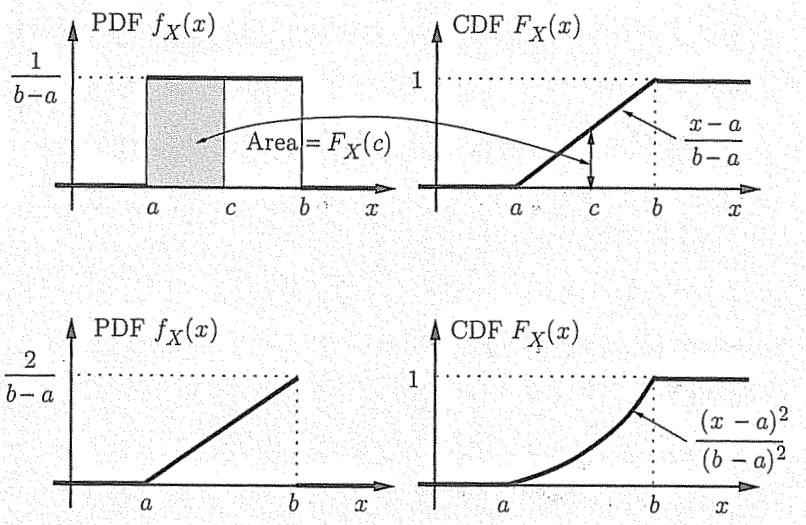


Figure 3.7: CDFs of some continuous random variables. The CDF is related to the PDF through the formula

$$F_X(x) = \mathbf{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt.$$

Thus, the PDF f_X can be obtained from the CDF by differentiation:

$$f_X(x) = \frac{dF_X}{dx}(x).$$

For a continuous random variable, the CDF has no jumps, i.e., it is continuous.

Sometimes, in order to calculate the PMF or PDF of a discrete or continuous random variable, respectively, it is more convenient to first calculate the CDF. The systematic use of this approach for the case of a continuous random variable will be discussed in Section 3.6. The following is a discrete example.

Example 3.6. The Maximum of Several Random Variables. You are allowed to take a certain test three times, and your final score will be the maximum of the test scores. Thus,

$$X = \max\{X_1, X_2, X_3\},$$

where X_1, X_2, X_3 are the three test scores and X is the final score. Assume that your score in each test takes one of the values from 1 to 10 with equal probability $1/10$, independently of the scores in other tests. What is the PMF p_X of the final score?

We calculate the PMF indirectly. We first compute the CDF F_X and then obtain the PMF as

$$p_X(k) = F_X(k) - F_X(k-1), \quad k = 1, \dots, 10.$$

We have

$$\begin{aligned} F_X(k) &= \mathbf{P}(X \leq k) \\ &= \mathbf{P}(X_1 \leq k, X_2 \leq k, X_3 \leq k) \\ &= \mathbf{P}(X_1 \leq k)\mathbf{P}(X_2 \leq k)\mathbf{P}(X_3 \leq k) \\ &= \left(\frac{k}{10}\right)^3, \end{aligned}$$

where the third equality follows from the independence of the events $\{X_1 \leq k\}$, $\{X_2 \leq k\}$, $\{X_3 \leq k\}$. Thus, the PMF is given by

$$p_X(k) = \left(\frac{k}{10}\right)^3 - \left(\frac{k-1}{10}\right)^3, \quad k = 1, \dots, 10.$$

The Geometric and Exponential CDFs

Because the CDF is defined for any type of random variable, it provides a convenient means for exploring the relations between continuous and discrete random variables. A particularly interesting case in point is the relation between geometric and exponential random variables.

Let X be a geometric random variable with parameter p ; that is, X is the number of trials until the first success in a sequence of independent Bernoulli trials, where the probability of success is p . Thus, for $k = 1, 2, \dots$, we have $\mathbf{P}(X = k) = p(1-p)^{k-1}$ and the CDF is given by

$$F_{\text{geo}}(n) = \sum_{k=1}^n p(1-p)^{k-1} = p \frac{1 - (1-p)^n}{1 - (1-p)} = 1 - (1-p)^n, \quad \text{for } n = 1, 2, \dots$$

Suppose now that X is an exponential random variable with parameter $\lambda > 0$. Its CDF is given by

$$F_{\text{exp}}(x) = \mathbb{P}(X \leq x) = 0, \quad \text{for } x \leq 0,$$

and

$$F_{\text{exp}}(x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t} \Big|_0^x = 1 - e^{-\lambda x}, \quad \text{for } x > 0.$$

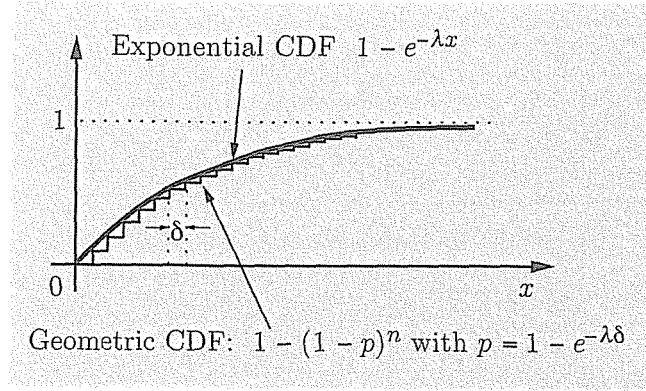


Figure 3.8: Relation of the geometric and the exponential CDFs. We have

$$F_{\text{exp}}(n\delta) = F_{\text{geo}}(n), \quad n = 1, 2, \dots,$$

where the interval δ is such that $e^{-\lambda\delta} = 1 - p$. As δ approaches 0, the exponential random variable can be interpreted as a “limit” of the geometric.

To compare the two CDFs above, let us define $\delta = -\ln(1 - p)/\lambda$, so that

$$e^{-\lambda\delta} = 1 - p.$$

Then we see that the values of the exponential and the geometric CDFs are equal for all $x = n\delta$, where $n = 1, 2, \dots$, i.e.,

$$F_{\text{exp}}(n\delta) = F_{\text{geo}}(n), \quad n = 1, 2, \dots$$

For other values of x , the two CDFs are close to each other, as illustrated in Fig. 3.8. This relation between the geometric and the exponential random variables will play an important role when we study the Bernoulli and Poisson processes in Chapter 5.

3.3 NORMAL RANDOM VARIABLES

A continuous random variable X is said to be **normal** or **Gaussian** if it has a PDF of the form (see Fig. 3.9)

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2},$$

where μ and σ are two scalar parameters characterizing the PDF, with σ assumed positive. It can be verified that the normalization property

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = 1$$

holds (see the end-of-chapter problems).

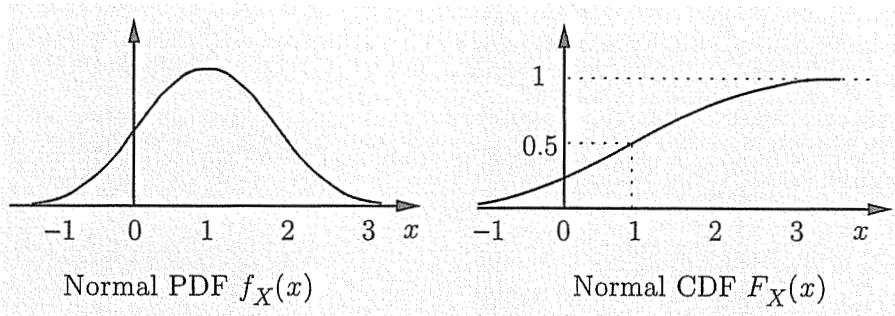


Figure 3.9: A normal PDF and CDF, with $\mu = 1$ and $\sigma^2 = 1$. We observe that the PDF is symmetric around its mean μ , and has a characteristic bell shape. As x gets further from μ , the term $e^{-(x-\mu)^2/2\sigma^2}$ decreases very rapidly. In this figure, the PDF is very close to zero outside the interval $[-1, 3]$.

The mean and the variance can be calculated to be

$$\mathbb{E}[X] = \mu, \quad \text{var}(X) = \sigma^2.$$

To see this, note that the PDF is symmetric around μ , so the mean can only be μ . Furthermore, the variance is given by

$$\text{var}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx.$$

Using the change of variables $y = (x-\mu)/\sigma$ and integration by parts, we have

$$\begin{aligned} \text{var}(X) &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \left(-ye^{-y^2/2} \right) \Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= \sigma^2. \end{aligned}$$

The last equality above is obtained by using the fact

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = 1,$$

which is just the normalization property of the normal PDF for the case where $\mu = 0$ and $\sigma = 1$.

The normal random variable has several special properties. The following one is particularly important and will be justified in Section 3.6.

Normality is Preserved by Linear Transformations

If X is a normal random variable with mean μ and variance σ^2 , and if $a \neq 0$, b are scalars, then the random variable

$$Y = aX + b$$

is also normal, with mean and variance

$$\mathbb{E}[Y] = a\mu + b, \quad \text{var}(Y) = a^2\sigma^2.$$

The Standard Normal Random Variable

A normal random variable Y with zero mean and unit variance is said to be a **standard normal**. Its CDF is denoted by Φ ,

$$\Phi(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Y < y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-t^2/2} dt.$$

It is recorded in a table (given in the next page), and is a very useful tool for calculating various probabilities involving normal random variables; see also Fig. 3.10.

Note that the table only provides the values of $\Phi(y)$ for $y \geq 0$, because the omitted values can be found using the symmetry of the PDF. For example, if Y is a standard normal random variable, we have

$$\begin{aligned} \Phi(-0.5) &= \mathbb{P}(Y \leq -0.5) = \mathbb{P}(Y \geq 0.5) = 1 - \mathbb{P}(Y < 0.5) \\ &= 1 - \Phi(0.5) = 1 - .6915 = 0.3085. \end{aligned}$$

More generally, we have

$$\Phi(-y) = 1 - \Phi(y), \quad \text{for all } y.$$

Let X be a normal random variable with mean μ and variance σ^2 . We “standardize” X by defining a new random variable Y given by

$$Y = \frac{X - \mu}{\sigma}.$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

The standard normal table. The entries in this table provide the numerical values of $\Phi(y) = P(Y \leq y)$, where Y is a standard normal random variable, for y between 0 and 3.49. For example, to find $\Phi(1.71)$, we look at the row corresponding to 1.7 and the column corresponding to 0.01, so that $\Phi(1.71) = .9564$. When y is negative, the value of $\Phi(y)$ can be found using the formula $\Phi(y) = 1 - \Phi(-y)$.

Since Y is a linear function of X , it is normal. Furthermore,

$$\mathbb{E}[Y] = \frac{\mathbb{E}[X] - \mu}{\sigma} = 0, \quad \text{var}(Y) = \frac{\text{var}(X)}{\sigma^2} = 1.$$

Thus, Y is a standard normal random variable. This fact allows us to calculate the probability of any event defined in terms of X : we redefine the event in terms of Y , and then use the standard normal table.

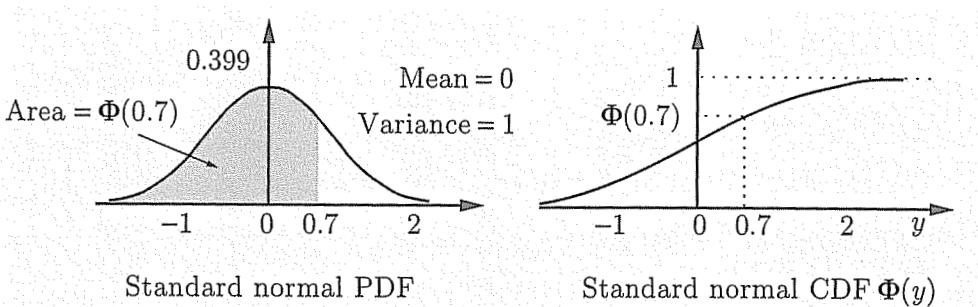


Figure 3.10: The PDF

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

of the standard normal random variable. The corresponding CDF, which is denoted by Φ , is recorded in a table.

Example 3.7. Using the Normal Table. The annual snowfall at a particular geographic location is modeled as a normal random variable with a mean of $\mu = 60$ inches and a standard deviation of $\sigma = 20$. What is the probability that this year's snowfall will be at least 80 inches?

Let X be the snow accumulation, viewed as a normal random variable, and let

$$Y = \frac{X - \mu}{\sigma} = \frac{X - 60}{20},$$

be the corresponding standard normal random variable. We want to find

$$\mathbb{P}(X \geq 80) = \mathbb{P}\left(\frac{X - 60}{20} \geq \frac{80 - 60}{20}\right) = \mathbb{P}\left(Y \geq \frac{80 - 60}{20}\right) = \mathbb{P}(Y \geq 1) = 1 - \Phi(1),$$

where Φ is the CDF of the standard normal. We read the value $\Phi(1)$ from the table:

$$\Phi(1) = 0.8413,$$

so that

$$\mathbb{P}(X \geq 80) = 1 - \Phi(1) = 0.1587.$$

Generalizing the approach in the preceding example, we have the following procedure.

CDF Calculation of the Normal Random Variable

The CDF of a normal random variable X with mean μ and variance σ^2 is obtained using the standard normal table as

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \mathbb{P}\left(Y \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

where Y is a standard normal random variable.

The normal random variable is often used in signal processing and communications engineering to model noise and unpredictable distortions of signals. The following is a typical example.

Example 3.8. Signal Detection. A binary message is transmitted as a signal S , which is either -1 or $+1$. The communication channel corrupts the transmission with additive normal noise with mean $\mu = 0$ and variance σ^2 . The receiver concludes that the signal -1 (or $+1$) was transmitted if the value received is < 0 (or ≥ 0 , respectively); see Fig. 3.11. What is the probability of error?

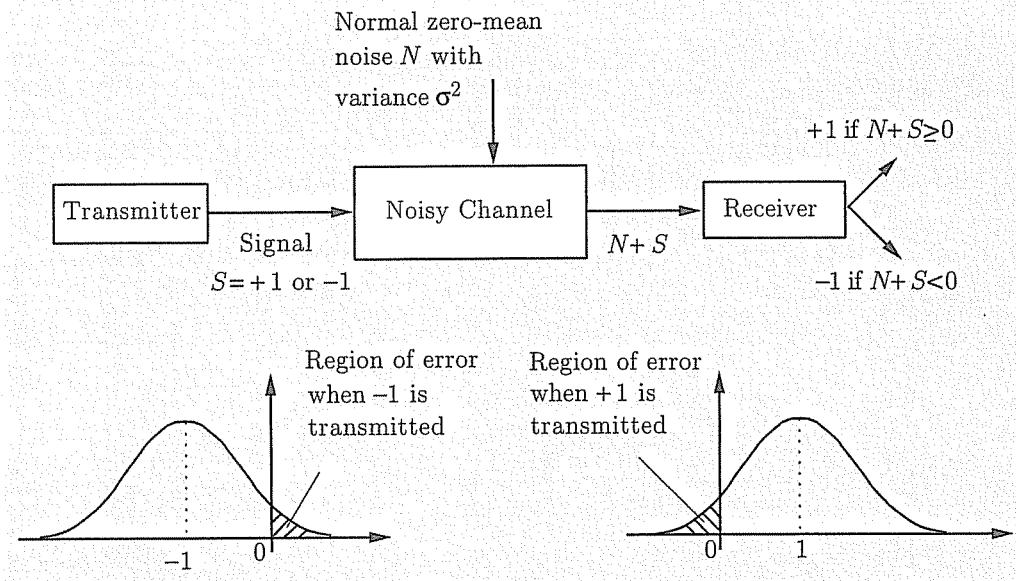


Figure 3.11: The signal detection scheme of Example 3.8. The area of the shaded region gives the probability of error in the two cases where -1 and $+1$ is transmitted.

An error occurs whenever -1 is transmitted and the noise N is at least 1 so that $N + S = N - 1 \geq 0$, or whenever $+1$ is transmitted and the noise N is smaller

than -1 so that $N + S = N + 1 < 0$. In the former case, the probability of error is

$$\begin{aligned}\mathbf{P}(N \geq 1) &= 1 - \mathbf{P}(N < 1) = 1 - \mathbf{P}\left(\frac{N - \mu}{\sigma} < \frac{1 - \mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{1 - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{1}{\sigma}\right).\end{aligned}$$

In the latter case, the probability of error is the same, by symmetry. The value of $\Phi(1/\sigma)$ can be obtained from the normal table. For $\sigma = 1$, we have $\Phi(1/\sigma) = \Phi(1) = 0.8413$, and the probability of error is 0.1587.

The normal random variable plays an important role in a broad range of probabilistic models. The main reason is that, generally speaking, it models well the additive effect of many independent factors in a variety of engineering, physical, and statistical contexts. Mathematically, the key fact is that *the sum of a large number of independent and identically distributed (not necessarily normal) random variables has an approximately normal CDF, regardless of the CDF of the individual random variables*. This property is captured in the celebrated *central limit theorem*, which will be discussed in Chapter 7.

3.4 CONDITIONING ON AN EVENT

The **conditional PDF** of a continuous random variable X , given an event $\{X \in A\}$ with $\mathbf{P}(X \in A) > 0$, is defined by:

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathbf{P}(X \in A)}, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

It can be used to calculate the conditional probability of events of the form $\{X \in B\}$ because

$$\mathbf{P}(X \in B | X \in A) = \frac{\mathbf{P}(X \in B, X \in A)}{\mathbf{P}(X \in A)} = \frac{\int_{A \cap B} f_X(x) dx}{\mathbf{P}(X \in A)} = \int_B f_{X|A}(x) dx.$$

As in the discrete case, the conditional PDF is zero outside the conditioning set. Within the conditioning set, the conditional PDF has exactly the same shape as the unconditional one, except that it is scaled by the constant factor $1/\mathbf{P}(X \in A)$. This normalization ensures that $f_{X|A}$ integrates to 1, which makes it a legitimate PDF; see Fig. 3.12. Thus, the conditional PDF is similar to an ordinary PDF, except that it refers to a new universe in which the event $\{X \in A\}$ is known to have occurred.

In the more general case where we wish to condition on a positive probability event A that cannot be described in terms of the random variable X , the

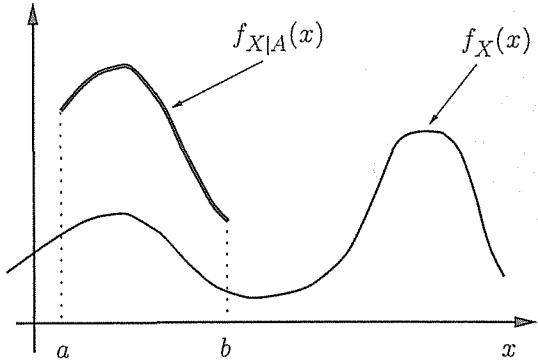


Figure 3.12: The unconditional PDF f_X and the conditional PDF $f_{X|A}$, where A is the interval $[a, b]$. Note that within the conditioning event A , $f_{X|A}$ retains the same shape as f_X , except that it is scaled along the vertical axis.

conditional PDF of X given A is defined as a nonnegative function $f_{X|A}$ that satisfies

$$\mathbf{P}(X \in B | A) = \int_B f_{X|A}(x) dx,$$

for any subset B of the real line although, in general, there is no simple formula for $f_{X|A}(x)$ in terms of $f_X(x)$.[†]

Example 3.9. The Exponential Random Variable is Memoryless. The time T until a new light bulb burns out is an exponential random variable with parameter λ . Ariadne turns the light on, leaves the room, and when she returns, t time units later, finds that the light bulb is still on, which corresponds to the event $A = \{T > t\}$. Let X be the additional time until the light bulb burns out. What is the conditional CDF of X , given the event A ?

We have, for $x \geq 0$,

$$\begin{aligned} \mathbf{P}(X > x | A) &= \mathbf{P}(T > t + x | T > t) \\ &= \frac{\mathbf{P}(T > t + x \text{ and } T > t)}{\mathbf{P}(T > t)} \\ &= \frac{\mathbf{P}(T > t + x)}{\mathbf{P}(T > t)} \\ &= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} \\ &= e^{-\lambda x}, \end{aligned}$$

[†] According to the notation introduced here, when we condition on an event of the form $\{X \in A\}$, we should be using the notation $f_{X| \{X \in A\}}(x)$. However, in such cases, we will be reverting to the simpler notation $f_{X|A}(x)$, introduced in the beginning of this section.

where we have used the expression for the CDF of an exponential random variable derived in Section 3.2.

Thus, the conditional CDF of X is exponential with parameter λ , regardless of the time t that elapsed between the lighting of the bulb and Ariadne's arrival. This is known as the *memorylessness property* of the exponential. Generally, if we model the time to complete a certain operation by an exponential random variable X , this property implies that as long as the operation has not been completed, the remaining time up to completion has the same exponential CDF, no matter when the operation started.

For a continuous random variable, we define the **conditional expectation** $\mathbb{E}[X | A]$ and the **conditional variance** $\text{var}(X | A)$, given an event A , similar to the unconditional case, except that we now need to use the conditional PDF. We summarize the discussion so far, together with some additional properties in the table that follows.

Conditional PDF and Expectation Given an Event

- The conditional PDF $f_{X|A}$ of a continuous random variable X given an event A with $\mathbb{P}(A) > 0$, satisfies

$$\mathbb{P}(X \in B | A) = \int_B f_{X|A}(x) dx.$$

- If A is a subset of the real line with $\mathbb{P}(X \in A) > 0$, then

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathbb{P}(X \in A)}, & \text{if } x \in A, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\mathbb{P}(X \in B | X \in A) = \int_B f_{X|A}(x) dx,$$

for any set B .

- The corresponding conditional expectation is defined by

$$\mathbb{E}[X | A] = \int_{-\infty}^{\infty} x f_{X|A}(x) dx.$$

- The expected value rule remains valid:

$$\mathbb{E}[g(X) | A] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x) dx.$$

- If A_1, A_2, \dots, A_n are disjoint events with $\mathbf{P}(A_i) > 0$ for each i , that form a partition of the sample space, then

$$f_X(x) = \sum_{i=1}^n \mathbf{P}(A_i) f_{X|A_i}(x)$$

(a version of the total probability theorem), and

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X | A_i]$$

(the total expectation theorem). Similarly,

$$\mathbf{E}[g(X)] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[g(X) | A_i].$$

To justify the above version of the total probability theorem, we use the total probability theorem from Chapter 1, to obtain

$$\mathbf{P}(X \leq x) = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{P}(X \leq x | A_i).$$

This formula can be rewritten as

$$\int_{-\infty}^x f_X(t) dt = \sum_{i=1}^n \mathbf{P}(A_i) \int_{-\infty}^x f_{X|A_i}(t) dt.$$

We take the derivative of both sides, with respect to x , and obtain the desired relation

$$f_X(x) = \sum_{i=1}^n \mathbf{P}(A_i) f_{X|A_i}(x).$$

If we now multiply both sides by x and then integrate from $-\infty$ to ∞ , we obtain the total expectation theorem for continuous random variables.

The total expectation theorem can often facilitate the calculation of the mean, variance, and other moments of a random variable, using a divide-and-conquer approach.

Example 3.10. Mean and Variance of a Piecewise Constant PDF. Suppose that the random variable X has the piecewise constant PDF

$$f_X(x) = \begin{cases} 1/3, & \text{if } 0 \leq x \leq 1, \\ 2/3, & \text{if } 1 < x \leq 2, \\ 0, & \text{otherwise,} \end{cases}$$

(see Fig. 3.13). Consider the events

$$A_1 = \{X \text{ lies in the first interval } [0, 1]\},$$

$$A_2 = \{X \text{ lies in the second interval } (1, 2]\}.$$

We have from the given PDF,

$$\mathbf{P}(A_1) = \int_0^1 f_X(x) dx = \frac{1}{3}, \quad \mathbf{P}(A_2) = \int_1^2 f_X(x) dx = \frac{2}{3}.$$

Furthermore, the conditional mean and second moment of X , conditioned on A_1 and A_2 , are easily calculated since the corresponding conditional PDFs $f_{X|A_1}$ and $f_{X|A_2}$ are uniform. We recall from Example 3.4 that the mean of a uniform random variable over an interval $[a, b]$ is $(a+b)/2$ and its second moment is $(a^2 + ab + b^2)/3$. Thus,

$$\mathbf{E}[X | A_1] = \frac{1}{2}, \quad \mathbf{E}[X | A_2] = \frac{3}{2},$$

$$\mathbf{E}[X^2 | A_1] = \frac{1}{3}, \quad \mathbf{E}[X^2 | A_2] = \frac{7}{3}.$$

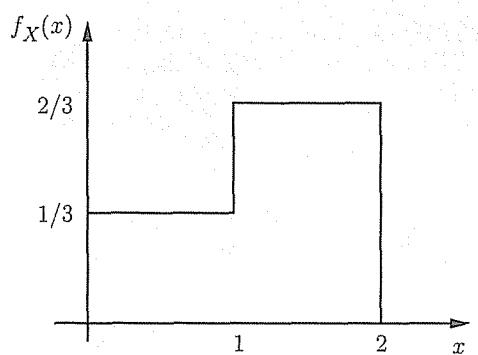


Figure 3.13: Piecewise constant PDF for Example 3.10.

We now use the total expectation theorem to obtain

$$\mathbf{E}[X] = \mathbf{P}(A_1)\mathbf{E}[X | A_1] + \mathbf{P}(A_2)\mathbf{E}[X | A_2] = \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{3}{2} = \frac{7}{6},$$

$$\mathbf{E}[X^2] = \mathbf{P}(A_1)\mathbf{E}[X^2 | A_1] + \mathbf{P}(A_2)\mathbf{E}[X^2 | A_2] = \frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot \frac{7}{3} = \frac{15}{9}.$$

The variance is given by

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \frac{15}{9} - \frac{49}{36} = \frac{11}{36}.$$

Note that this approach to the mean and variance calculation is easily generalized to piecewise constant PDFs with more than two pieces.

The next example illustrates a divide-and-conquer approach that uses the total probability theorem to calculate a PDF.

Example 3.11. The metro train arrives at the station near your home every quarter hour starting at 6:00 a.m. You walk into the station every morning between 7:10 and 7:30 a.m., with the time in this interval being a uniform random variable. What is the PDF of the time you have to wait for the first train to arrive?

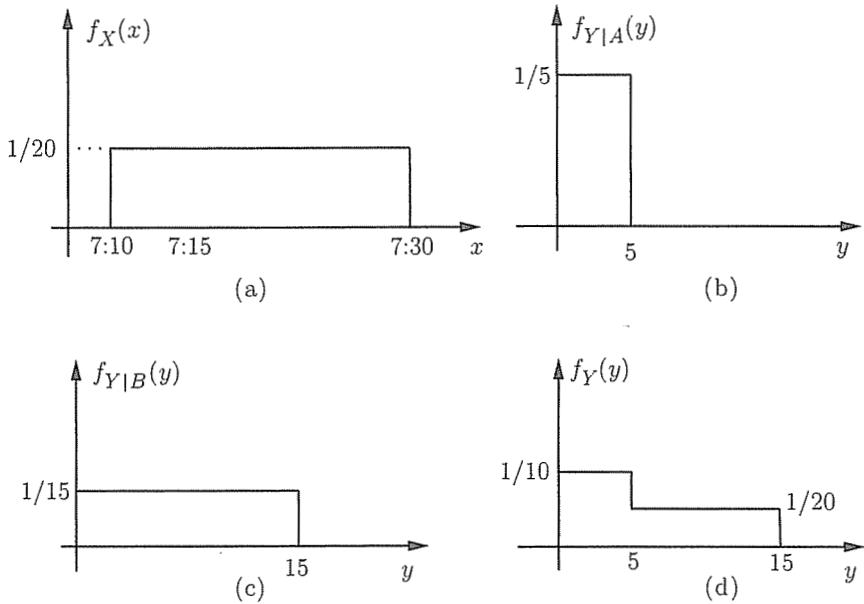


Figure 3.14: The PDFs f_X , $f_{Y|A}$, $f_{Y|B}$, and f_Y in Example 3.11.

The time of your arrival, denoted by X , is a uniform random variable over the interval from 7:10 to 7:30; see Fig. 3.14(a). Let Y be the waiting time. We calculate the PDF f_Y using a divide-and-conquer strategy. Let A and B be the events

$$A = \{7:10 \leq X \leq 7:15\} = \{\text{you board the 7:15 train}\},$$

$$B = \{7:15 < X \leq 7:30\} = \{\text{you board the 7:30 train}\}.$$

Conditioned on the event A , your arrival time is uniform over the interval from 7:10 to 7:15. In that case, the waiting time Y is also uniform and takes values between 0 and 5 minutes; see Fig. 3.14(b). Similarly, conditioned on B , Y is uniform and takes values between 0 and 15 minutes; see Fig. 3.14(c). The PDF of Y is obtained using the total probability theorem,

$$f_Y(y) = \mathbf{P}(A)f_{Y|A}(y) + \mathbf{P}(B)f_{Y|B}(y),$$

and is shown in Fig. 3.14(d). In particular,

$$f_Y(y) = \frac{1}{4} \cdot \frac{1}{5} + \frac{3}{4} \cdot \frac{1}{15} = \frac{1}{10}, \quad \text{for } 0 \leq y \leq 5,$$

and

$$f_Y(y) = \frac{1}{4} \cdot 0 + \frac{3}{4} \cdot \frac{1}{15} = \frac{1}{20}, \quad \text{for } 5 < y \leq 15.$$

3.5 MULTIPLE CONTINUOUS RANDOM VARIABLES

We will now extend the notion of a PDF to the case of multiple random variables. In complete analogy with discrete random variables, we introduce joint, marginal, and conditional PDFs. Their intuitive interpretation as well as their main properties parallel the discrete case.

We say that two continuous random variables associated with the same experiment are **jointly continuous** and can be described in terms of a **joint PDF** $f_{X,Y}$, if $f_{X,Y}$ is a nonnegative function that satisfies

$$\mathbf{P}((X, Y) \in B) = \iint_{(x,y) \in B} f_{X,Y}(x, y) dx dy,$$

for every subset B of the two-dimensional plane. The notation above means that the integration is carried over the set B . In the particular case where B is a rectangle of the form $B = \{(x, y) \mid a \leq x \leq b, c \leq y \leq d\}$, we have

$$\mathbf{P}(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x, y) dx dy.$$

Furthermore, by letting B be the entire two-dimensional plane, we obtain the normalization property

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1.$$

To interpret the joint PDF, we let δ be a small positive number and consider the probability of a small rectangle. We have

$$\mathbf{P}(a \leq X \leq a + \delta, c \leq Y \leq c + \delta) = \int_c^{c+\delta} \int_a^{a+\delta} f_{X,Y}(x, y) dx dy \approx f_{X,Y}(a, c) \cdot \delta^2,$$

so we can view $f_{X,Y}(a, c)$ as the “probability per unit area” in the vicinity of (a, c) .

The joint PDF contains all conceivable probabilistic information on the random variables X and Y , as well as their dependencies. It allows us to calculate the probability of any event that can be defined in terms of these two random variables. As a special case, it can be used to calculate the probability of an

event involving only one of them. For example, let A be a subset of the real line and consider the event $\{X \in A\}$. We have

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A \text{ and } Y \in (-\infty, \infty)) = \int_A \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy dx.$$

Comparing with the formula

$$\mathbb{P}(X \in A) = \int_A f_X(x) dx,$$

we see that the **marginal PDF** f_X of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx.$$

Example 3.12. Two-Dimensional Uniform PDF. Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour (recall the example given in Section 1.2). Let X and Y denote the delays of Romeo and Juliet, respectively. Assuming that no pairs (x, y) in the unit square are more likely than others, a natural model involves a joint PDF of the form

$$f_{X,Y}(x, y) = \begin{cases} c, & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where c is a constant. For this PDF to satisfy the normalization property

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = \int_0^1 \int_0^1 c dx dy = 1,$$

we must have

$$c = 1.$$

This is an example of a uniform joint PDF. More generally, let us fix some subset S of the two-dimensional plane. The corresponding uniform joint PDF on S is defined to be

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\text{area of } S}, & \text{if } (x, y) \in S, \\ 0, & \text{otherwise.} \end{cases}$$

For any set $A \subset S$, the probability that (X, Y) lies in A is

$$\mathbb{P}((X, Y) \in A) = \iint_{(x,y) \in A} f_{X,Y}(x, y) dx dy = \frac{1}{\text{area of } S} \iint_{(x,y) \in A \cap S} dx dy = \frac{\text{area of } A \cap S}{\text{area of } S}.$$

Example 3.13. We are told that the joint PDF of the random variables X and Y is a constant c on the set S shown in Fig. 3.15 and is zero outside. We wish to determine the value of c and the marginal PDFs of X and Y .

The area of the set S is equal to 4 and, therefore, $f_{X,Y}(x, y) = c = 1/4$, for $(x, y) \in S$. To find the marginal PDF $f_X(x)$ for some particular x , we integrate (with respect to y) the joint PDF over the vertical line corresponding to that x . The resulting PDF is shown in the figure. We can compute f_Y similarly.

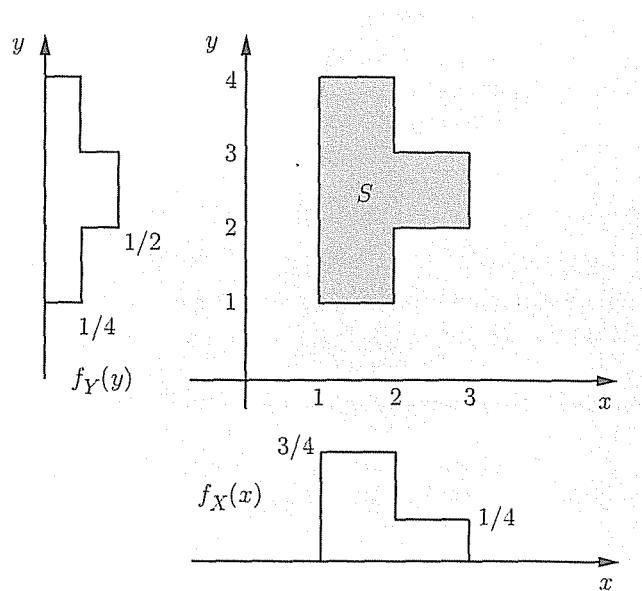


Figure 3.15: The joint PDF in Example 3.13 and the resulting marginal PDFs.

Example 3.14. Buffon's Needle. [†] This is a famous example, which marks the origin of the subject of geometrical probability, that is, the analysis of the

[†] This problem was posed and solved in 1777 by the French naturalist Buffon. A number of variants of the problem have been investigated, including the case where the surface is ruled with two sets of perpendicular lines (Laplace, 1812); see the end-of-chapter problems. The problem has long fascinated scientists, and has been used as a basis for experimental evaluations of π (among others, it has been reported that a captain named Fox measured π experimentally using needles, while recovering from wounds suffered in the American Civil War). The internet contains several graphical simulation programs for computing π using Buffon's ideas.

geometrical configuration of randomly placed objects.

A surface is ruled with parallel lines, which are at distance d from each other (see Fig. 3.16). Suppose that we throw a needle of length l on the surface at random. What is the probability that the needle will intersect one of the lines?

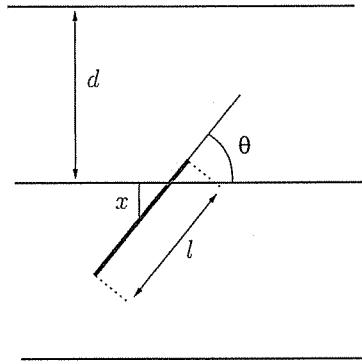


Figure 3.16: Buffon's needle. The length of the line segment between the midpoint of the needle and the point of intersection of the axis of the needle with the closest parallel line is $x/\sin \theta$. The needle will intersect the closest parallel line if and only if this length is less than $l/2$.

We assume here that $l < d$ so that the needle cannot intersect two lines simultaneously. Let X be the vertical distance from the midpoint of the needle to the nearest of the parallel lines, and let Θ be the acute angle formed by the axis of the needle and the parallel lines (see Fig. 3.16). We model the pair of random variables (X, Θ) with a uniform joint PDF over the rectangular set $\{(x, \theta) \mid 0 \leq x \leq d/2, 0 \leq \theta \leq \pi/2\}$, so that

$$f_{X,\Theta}(x, \theta) = \begin{cases} 4/(\pi d), & \text{if } x \in [0, d/2] \text{ and } \theta \in [0, \pi/2], \\ 0, & \text{otherwise.} \end{cases}$$

As can be seen from Fig. 3.16, the needle will intersect one of the lines if and only if

$$X \leq \frac{l}{2} \sin \Theta,$$

so the probability of intersection is

$$\begin{aligned} \mathbf{P}(X \leq (l/2) \sin \Theta) &= \iint_{\substack{x \leq (l/2) \sin \theta \\ x \in [0, d/2] \\ \theta \in [0, \pi/2]}} f_{X,\Theta}(x, \theta) dx d\theta \\ &= \frac{4}{\pi d} \int_0^{\pi/2} \int_0^{(l/2) \sin \theta} dx d\theta \\ &= \frac{4}{\pi d} \int_0^{\pi/2} \frac{l}{2} \sin \theta d\theta \\ &= \frac{2l}{\pi d} (-\cos \theta) \Big|_0^{\pi/2} \\ &= \frac{2l}{\pi d}. \end{aligned}$$

The probability of intersection can be empirically estimated, by repeating the experiment a large number of times. Since it is equal to $2l/\pi d$, this provides us with a method for the experimental evaluation of π .

Expectation

If X and Y are jointly continuous random variables and g is some function, then $Z = g(X, Y)$ is also a random variable. We will see in Section 3.6 methods for computing the PDF of Z , if it has one. For now, let us note that the expected value rule is still applicable and

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy.$$

As an important special case, for any scalars a, b , we have

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Conditioning One Random Variable on Another

Let X and Y be continuous random variables with joint PDF $f_{X,Y}$. For any fixed y with $f_Y(y) > 0$, the conditional PDF of X given that $Y = y$, is defined by

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

This definition is analogous to the formula $p_{X|Y}(x | y) = p_{X,Y}(x, y)/p_Y(y)$ for the discrete case.

When thinking about the conditional PDF, it is best to view y as a fixed number and consider $f_{X|Y}(x | y)$ as a function of the single variable x . As a function of x , the conditional PDF $f_{X|Y}(x | y)$ has the same shape as the joint PDF $f_{X,Y}(x, y)$, because the normalizing factor $f_Y(y)$ does not depend on x ; see Fig. 3.17. Note that the normalization ensures that

$$\int_{-\infty}^{\infty} f_{X|Y}(x | y) dx = 1,$$

so for any fixed y , $f_{X|Y}(x | y)$ is a legitimate PDF.

Example 3.15. Circular Uniform PDF. Ben throws a dart at a circular target of radius r (see Fig. 3.18). We assume that he always hits the target, and that all points of impact (x, y) are equally likely, so that the joint PDF of the random

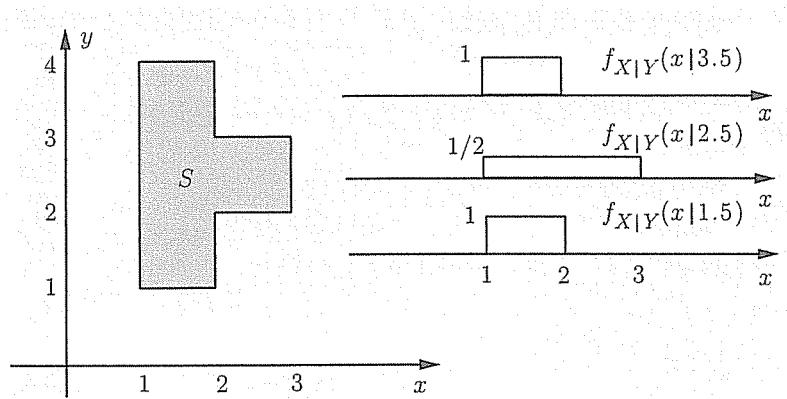


Figure 3.17: Visualization of the conditional PDF $f_{X|Y}(x|y)$. Let X and Y have a joint PDF which is uniform on the set S . For each fixed y , we consider the joint PDF along the slice $Y = y$ and normalize it so that it integrates to 1.

variables X and Y is uniform. Following Example 3.12, and since the area of the circle is πr^2 , we have

$$\begin{aligned} f_{X,Y}(x,y) &= \begin{cases} \frac{1}{\text{area of the circle}}, & \text{if } (x,y) \text{ is in the circle,} \\ 0, & \text{otherwise,} \end{cases} \\ &= \begin{cases} \frac{1}{\pi r^2}, & \text{if } x^2 + y^2 \leq r^2, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

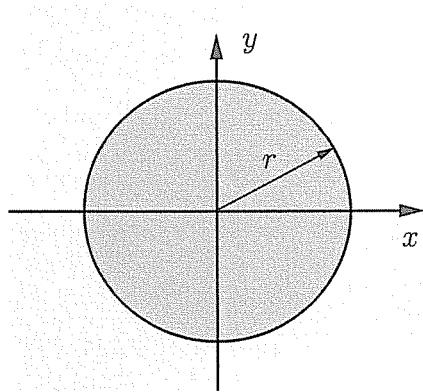


Figure 3.18: Circular target for Example 3.15.

To calculate the conditional PDF $f_{X|Y}(x|y)$, let us first calculate the marginal PDF $f_Y(y)$. For $|y| > r$, it is zero. For $|y| \leq r$, it can be calculated as follows:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx \\ &= \frac{1}{\pi r^2} \int_{x^2+y^2 \leq r^2} dx \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\pi r^2} \int_{-\sqrt{r^2 - y^2}}^{\sqrt{r^2 - y^2}} dx \\
&= \frac{2}{\pi r^2} \sqrt{r^2 - y^2}, \quad \text{if } |y| \leq r.
\end{aligned}$$

Note that the marginal PDF f_Y is not uniform.

The conditional PDF is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{\pi r^2}}{\frac{2}{\pi r^2} \sqrt{r^2 - y^2}} = \frac{1}{2\sqrt{r^2 - y^2}}, \quad \text{if } x^2 + y^2 \leq r^2.$$

Thus, for a fixed value of y , the conditional PDF $f_{X|Y}$ is uniform.

To interpret the conditional PDF, let us fix some small positive numbers δ_1 and δ_2 , and condition on the event $B = \{y \leq Y \leq y + \delta_2\}$. We have

$$\begin{aligned}
\mathbf{P}(x \leq X \leq x + \delta_1 | y \leq Y \leq y + \delta_2) &= \frac{\mathbf{P}(x \leq X \leq x + \delta_1 \text{ and } y \leq Y \leq y + \delta_2)}{\mathbf{P}(y \leq Y \leq y + \delta_2)} \\
&\approx \frac{f_{X,Y}(x,y)\delta_1\delta_2}{f_Y(y)\delta_2} \\
&= f_{X|Y}(x|y)\delta_1.
\end{aligned}$$

In words, $f_{X|Y}(x|y)\delta_1$ provides us with the probability that X belongs to a small interval $[x, x + \delta_1]$, given that Y belongs to a small interval $[y, y + \delta_2]$. Since $f_{X|Y}(x|y)\delta_1$ does not depend on δ_2 , we can think of the limiting case where δ_2 decreases to zero and write

$$\mathbf{P}(x \leq X \leq x + \delta_1 | Y = y) \approx f_{X|Y}(x|y)\delta_1, \quad (\delta_1 \text{ small}),$$

and, more generally,

$$\mathbf{P}(X \in A | Y = y) = \int_A f_{X|Y}(x|y) dx.$$

Conditional probabilities, given the zero probability event $\{Y = y\}$, were left undefined in Chapter 1. But the above formula provides a natural way of defining such conditional probabilities in the present context. In addition, it allows us to view the conditional PDF $f_{X|Y}(x|y)$ (as a function of x) as a description of the probability law of X , given that the event $\{Y = y\}$ has occurred.

As in the discrete case, the conditional PDF $f_{X|Y}$, together with the marginal PDF f_Y are sometimes used to calculate the joint PDF. Furthermore, this approach can be also used for modeling: instead of directly specifying $f_{X,Y}$, it is often natural to provide a probability law for Y , in terms of a PDF f_Y ,

and then provide a conditional PDF $f_{X|Y}(x|y)$ for X , given any possible value y of Y .

Example 3.16. The speed of a typical vehicle that drives past a police radar is modeled as an exponentially distributed random variable X with mean 50 miles per hour. The police radar's measurement Y of the vehicle's speed has an error which is modeled as a normal random variable with zero mean and standard deviation equal to one tenth of the vehicle's speed. What is the joint PDF of X and Y ?

We have $f_X(x) = (1/50)e^{-x/50}$, for $x \geq 0$. Also, conditioned on $X = x$, the measurement Y has a normal PDF with mean x and variance $x^2/100$. Therefore,

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}(x/10)} e^{-(y-x)^2/(2x^2/100)}.$$

Thus,

$$f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x) = \frac{1}{50}e^{-x/50} \frac{10}{\sqrt{2\pi}x} e^{-50(y-x)^2/x^2},$$

for all $x \geq 0$ and all y .

Having defined a conditional probability law, we can also define a corresponding conditional expectation by letting

$$\mathbb{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx.$$

The properties of unconditional expectations carry through, with the obvious modifications, to conditional expectations. For example, we have the following conditional version of the expected value rule:

$$\mathbb{E}[g(X)|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx.$$

Summary of Facts About Multiple Continuous Random Variables

Let X and Y be jointly continuous random variables with joint PDF $f_{X,Y}$.

- The joint, marginal, and conditional PDFs are related to each other by the formulas

$$f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x|y),$$

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x|y) dy.$$

The conditional PDF $f_{X|Y}(x|y)$ is defined only for those y for which $f_Y(y) > 0$.

- These PDFs can be used to calculate probabilities:

$$\begin{aligned}\mathbb{P}((X, Y) \in B) &= \iint_{(x,y) \in B} f_{X,Y}(x, y) dx dy, \\ \mathbb{P}(X \in A) &= \int_A f_X(x) dx, \\ \mathbb{P}(X \in A | Y = y) &= \int_A f_{X|Y}(x | y) dx.\end{aligned}$$

- They can also be used to calculate expectations:

$$\begin{aligned}\mathbb{E}[g(X)] &= \int g(x) f_X(x) dx, \\ \mathbb{E}[g(X, Y)] &= \iint g(x, y) f_{X,Y}(x, y) dx dy, \\ \mathbb{E}[g(X) | Y = y] &= \int g(x) f_{X|Y}(x | y) dx, \\ \mathbb{E}[g(X, Y) | Y = y] &= \int g(x, y) f_{X|Y}(x | y) dx.\end{aligned}$$

- For any event A , we have the following version of the total probability theorem:

$$\mathbb{P}(A) = \int \mathbb{P}(A | Y = y) f_Y(y) dy.$$

- We have the following versions of the total expectation theorem:

$$\begin{aligned}\mathbb{E}[X] &= \int \mathbb{E}[X | Y = y] f_Y(y) dy, \\ \mathbb{E}[g(X)] &= \int \mathbb{E}[g(X) | Y = y] f_Y(y) dy, \\ \mathbb{E}[g(X, Y)] &= \int \mathbb{E}[g(X, Y) | Y = y] f_Y(y) dy.\end{aligned}$$

To justify the first version of the total expectation theorem, we observe that

$$\begin{aligned}\int \mathbb{E}[X | Y = y] f_Y(y) dy &= \int \left[\int x f_{X|Y}(x | y) dx \right] f_Y(y) dy \\ &= \int \int x f_{X|Y}(x | y) f_Y(y) dx dy\end{aligned}$$

$$\begin{aligned}
&= \int \int x f_{X,Y}(x, y) dx dy \\
&= \int x \left[\int f_{X,Y}(x, y) dy \right] dx \\
&= \int x f_X(x) dx \\
&= \mathbb{E}[X].
\end{aligned}$$

The other two versions are justified similarly. The total probability equation $\mathbb{P}(A) = \int \mathbb{P}(A | Y = y) f_Y(y) dy$ is a special case of the total expectation theorem: we let X be the random variable that takes the value 1 if $X \in A$ and the value 0 otherwise, in which case, $\mathbb{E}[X] = \mathbb{P}(A)$ and $\mathbb{E}[X | Y = y] = \mathbb{P}(A | Y = y)$.

Independence

In full analogy with the discrete case, we say that two continuous random variables X and Y are **independent** if their joint PDF is the product of the marginal PDFs:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y), \quad \text{for all } x, y.$$

Comparing with the formula $f_{X,Y}(x, y) = f_{X|Y}(x | y) f_Y(y)$, we see that independence is the same as the condition

$$f_{X|Y}(x | y) = f_X(x), \quad \text{for all } x \text{ and all } y \text{ with } f_Y(y) > 0,$$

or, symmetrically,

$$f_{Y|X}(y | x) = f_Y(y), \quad \text{for all } y \text{ and all } x \text{ with } f_X(x) > 0.$$

If X and Y are independent, then any two events of the form $\{X \in A\}$ and $\{Y \in B\}$ are independent. Indeed,

$$\begin{aligned}
\mathbb{P}(X \in A \text{ and } Y \in B) &= \int_{x \in A} \int_{y \in B} f_{X,Y}(x, y) dy dx \\
&= \int_{x \in A} \int_{y \in B} f_X(x) f_Y(y) dy dx \\
&= \int_{x \in A} f_X(x) dx \int_{y \in B} f_Y(y) dy \\
&= \mathbb{P}(X \in A) \mathbb{P}(Y \in B).
\end{aligned}$$

A converse statement is also true; see the end-of-chapter problems.

An argument similar to the discrete case shows that if X and Y are independent, then

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)],$$

for any two functions g and h . Finally, the variance of the sum of *independent* random variables is again equal to the sum of the variances.

Independence of Continuous Random Variables

Suppose that X and Y are independent, that is,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \text{for all } x,y.$$

We then have the following properties.

- The random variables $g(X)$ and $h(Y)$ are independent, for any functions g and h .
- We have

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y],$$

and, more generally,

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

- We have

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y).$$

Joint CDFs

If X and Y are two random variables associated with the same experiment, we define their joint CDF by

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y).$$

As in the case of one random variable, the advantage of working with the CDF is that it applies equally well to discrete and continuous random variables. In particular, if X and Y are described by a joint PDF $f_{X,Y}$, then

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

Conversely, the PDF can be recovered from the CDF by differentiating:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y).$$

Example 3.17. Let X and Y be described by a uniform PDF on the unit square. The joint CDF is given by

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y) = xy, \quad \text{for } 0 \leq x, y \leq 1.$$

We then verify that

$$\frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x, y) = \frac{\partial^2 (xy)}{\partial x \partial y}(x, y) = 1 = f_{X,Y}(x, y),$$

for all (x, y) in the unit square.

Inference and the Continuous Bayes' Rule

In many situations, we have a model of an underlying but unobserved phenomenon, represented by a random variable X with PDF f_X , and we make a noisy measurement Y , which is modeled in terms of a conditional PDF $f_{Y|X}$. Once the value of Y is measured, what information does it provide on the unknown value of X ?

This setting is similar to the one of Section 1.4, where we introduced Bayes' rule and used it to solve inference problems; see Fig. 3.19. The only difference is that we are now dealing with continuous random variables.

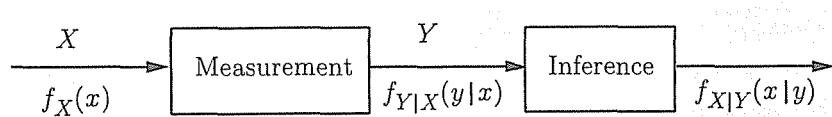


Figure 3.19: Schematic description of the inference problem. We have an unobserved random variable X with known PDF and we obtain a measurement Y , according to a conditional PDF $f_{Y|X}$. Given an observed value y of Y , the inference problem is to evaluate the conditional PDF $f_{X|Y}(x|y)$.

Note that whatever information is provided by the event $\{Y = y\}$ is captured by the conditional PDF $f_{X|Y}(x|y)$. It thus suffices to evaluate this PDF. A calculation analogous to the original derivation of Bayes' rule, based on the formulas $f_X f_{Y|X} = f_{X,Y} = f_Y f_{X|Y}$, yields

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)} = \frac{f_X(x) f_{Y|X}(y|x)}{\int f_X(t) f_{Y|X}(y|t) dt},$$

which is the desired formula.

Example 3.18. A lightbulb produced by the General Illumination Company is known to have an exponentially distributed lifetime Y . However, the company has been experiencing quality control problems. On any given day, the parameter λ of the PDF of Y is actually a random variable, uniformly distributed in the interval $[1, 3/2]$. We test a lightbulb and record its lifetime. What can we say about the underlying parameter λ ?

We model the parameter λ in terms of a uniform random variable Λ with PDF

$$f_{\Lambda}(\lambda) = 2, \quad \text{for } 1 \leq \lambda \leq \frac{3}{2}.$$

The available information about Λ is captured by the conditional PDF $f_{\Lambda|Y}(\lambda | y)$, which using the continuous Bayes' rule, is given by

$$f_{\Lambda|Y}(\lambda | y) = \frac{f_{\Lambda}(\lambda) f_{Y|\Lambda}(y | \lambda)}{\int f_{\Lambda}(t) f_{Y|\Lambda}(y | t) dt} = \frac{2\lambda e^{-\lambda y}}{\int_1^{3/2} 2te^{-ty} dt}, \quad \text{for } 1 \leq \lambda \leq \frac{3}{2}.$$

In some cases, the unobserved phenomenon is inherently discrete. For some examples, consider a binary signal which is observed in the presence of normally distributed noise, or a medical diagnosis that is made on the basis of continuous measurements such as temperature and blood counts. In such cases, a somewhat different version of Bayes' rule applies.

We first consider the case where the unobserved phenomenon is described in terms of an event A whose occurrence is unknown. Let $\mathbf{P}(A)$ be the probability of event A . Let Y be a continuous random variable, and assume that the conditional PDFs $f_{Y|A}(y)$ and $f_{Y|A^c}(y)$ are known. We are interested in the conditional probability $\mathbf{P}(A | Y = y)$ of the event A , given the value y of Y .

Instead of working with the conditioning event $\{Y = y\}$, which has zero probability, let us instead condition on the event $\{y \leq Y \leq y + \delta\}$, where δ is a small positive number, and then take the limit as δ tends to zero. We have, using Bayes' rule,

$$\begin{aligned} \mathbf{P}(A | Y = y) &\approx \mathbf{P}(A | y \leq Y \leq y + \delta) \\ &= \frac{\mathbf{P}(A) \mathbf{P}(y \leq Y \leq y + \delta | A)}{\mathbf{P}(y \leq Y \leq y + \delta)} \\ &\approx \frac{\mathbf{P}(A) f_{Y|A}(y) \delta}{f_Y(y) \delta} \\ &= \frac{\mathbf{P}(A) f_{Y|A}(y)}{f_Y(y)}. \end{aligned}$$

The denominator can be evaluated using the following version of the total probability theorem:

$$f_Y(y) = \mathbf{P}(A) f_{Y|A}(y) + \mathbf{P}(A^c) f_{Y|A^c}(y),$$

so that

$$\mathbf{P}(A | Y = y) = \frac{\mathbf{P}(A) f_{Y|A}(y)}{\mathbf{P}(A) f_{Y|A}(y) + \mathbf{P}(A^c) f_{Y|A^c}(y)}.$$

In a variant of this formula, we consider an event A of the form $\{N = n\}$, where N is a discrete random variable that represents the different discrete

possibilities for the unobserved phenomenon of interest. Let p_N be the PMF of N . Let also Y be a continuous random variable which, for any given value n of N , is described by a conditional PDF $f_{Y|N}(y|n)$. The above formula becomes

$$\mathbf{P}(N = n | Y = y) = \frac{p_N(n)f_{Y|N}(y|n)}{f_Y(y)}.$$

The denominator can be evaluated using the following version of the total probability theorem:

$$f_Y(y) = \sum_i p_N(i)f_{Y|N}(y|i),$$

so that

$$\mathbf{P}(N = n | Y = y) = \frac{p_N(n)f_{Y|N}(y|n)}{\sum_i p_N(i)f_{Y|N}(y|i)}.$$

Example 3.19. Signal Detection. A binary signal S is transmitted, and we are given that $\mathbf{P}(S = 1) = p$ and $\mathbf{P}(S = -1) = 1 - p$. The received signal is $Y = N + S$, where N is normal noise, with zero mean and unit variance, independent of S . What is the probability that $S = 1$, as a function of the observed value y of Y ?

Conditioned on $S = s$, the random variable Y has a normal distribution with mean s and unit variance. Applying the formulas given above, we obtain

$$\mathbf{P}(S = 1 | Y = y) = \frac{p_S(1)f_{Y|S}(y|1)}{f_Y(y)} = \frac{\frac{p}{\sqrt{2\pi}}e^{-(y-1)^2/2}}{\frac{p}{\sqrt{2\pi}}e^{-(y-1)^2/2} + \frac{1-p}{\sqrt{2\pi}}e^{-(y+1)^2/2}},$$

which simplifies to

$$\mathbf{P}(S = 1 | Y = y) = \frac{pe^y}{pe^y + (1-p)e^{-y}}.$$

Note that the probability $\mathbf{P}(S = 1 | Y = y)$ goes to zero as y decreases to $-\infty$, goes to 1 as y increases to ∞ , and is monotonically increasing in between, which is consistent with intuition.

Bayes' Rule for Continuous Random Variables

Let Y be a continuous random variable.

- If X is a continuous random variable, we have

$$f_{X|Y}(x|y)f_Y(y) = f_X(x)f_{Y|X}(y|x),$$

and

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{\int f_X(t)f_{Y|X}(y|t)dt}.$$

- If A is an event, we have

$$\mathbb{P}(A | Y = y) f_Y(y) = \mathbb{P}(A) f_{Y|A}(y),$$

and

$$\mathbb{P}(A | Y = y) = \frac{\mathbb{P}(A) f_{Y|A}(y)}{\mathbb{P}(A) f_{Y|A}(y) + \mathbb{P}(A^c) f_{Y|A^c}(y)}.$$

- If N is a discrete random variable, we have

$$\mathbb{P}(N = n | Y = y) f_Y(y) = p_N(n) f_{Y|N}(y | n),$$

and

$$\mathbb{P}(N = n | Y = y) = \frac{p_N(n) f_{Y|N}(y | n)}{\sum_i p_N(i) f_{Y|N}(y | i)}.$$

More than Two Random Variables

The joint PDF of three random variables X , Y , and Z is defined in analogy with the case of two random variables. For example, we have

$$\mathbb{P}((X, Y, Z) \in B) = \int \int \int_{(x,y,z) \in B} f_{X,Y,Z}(x, y, z) dx dy dz,$$

for any set B . We also have relations such as

$$f_{X,Y}(x, y) = \int f_{X,Y,Z}(x, y, z) dz,$$

and

$$f_X(x) = \int \int f_{X,Y,Z}(x, y, z) dy dz.$$

One can also define conditional PDFs by formulas such as

$$f_{X,Y|Z}(x, y | z) = \frac{f_{X,Y,Z}(x, y, z)}{f_Z(z)}, \quad \text{for } f_Z(z) > 0,$$

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

There is an analog of the multiplication rule:

$$f_{X,Y,Z}(x, y, z) = f_{X|Y,Z}(x | y, z) f_{Y|Z}(y | z) f_Z(z).$$

Finally, we say that the three random variables X , Y , and Z are independent if

$$f_{X,Y,Z}(x, y, z) = f_X(x)f_Y(y)f_Z(z), \quad \text{for all } x, y, z.$$

The expected value rule for functions takes the form

$$\mathbb{E}[g(X, Y, Z)] = \int \int \int g(x, y, z) f_{X,Y,Z}(x, y, z) dx dy dz,$$

and if g is linear and of the form $aX + bY + cZ$, then

$$\mathbb{E}[aX + bY + cZ] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c\mathbb{E}[Z].$$

Furthermore, there are obvious generalizations of the above to the case of more than three random variables. For example, for any random variables X_1, X_2, \dots, X_n and any scalars a_1, a_2, \dots, a_n , we have

$$\mathbb{E}[a_1X_1 + a_2X_2 + \dots + a_nX_n] = a_1\mathbb{E}[X_1] + a_2\mathbb{E}[X_2] + \dots + a_n\mathbb{E}[X_n].$$

3.6 DERIVED DISTRIBUTIONS

In this section, we consider functions $Y = g(X)$ of a continuous random variable X . We discuss techniques whereby, given the PDF of X , we calculate the PDF of Y (also called a *derived distribution*). The principal method for doing so is the following two-step approach.

Calculation of the PDF of a Function $Y = g(X)$ of a Continuous Random Variable X

1. Calculate the CDF F_Y of Y using the formula

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \int_{\{x \mid g(x) \leq y\}} f_X(x) dx.$$

2. Differentiate to obtain the PDF of Y :

$$f_Y(y) = \frac{dF_Y}{dy}(y).$$

Example 3.20. Let X be uniform on $[0, 1]$, and let $Y = \sqrt{X}$. We note that for every $y \in [0, 1]$, we have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\sqrt{X} \leq y) = \mathbb{P}(X \leq y^2) = y^2.$$

We then differentiate and obtain

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \frac{d(y^2)}{dy} = 2y, \quad 0 \leq y \leq 1.$$

Outside the range $[0, 1]$, the CDF $F_Y(y)$ is constant, with $F_Y(y) = 0$ for $y \leq 0$, and $F_Y(y) = 1$ for $y \geq 1$. By differentiating, we see that $f_Y(y) = 0$ for y outside $[0, 1]$.

Example 3.21. John Slow is driving from Boston to the New York area, a distance of 180 miles at a constant speed, whose value is uniformly distributed between 30 and 60 miles per hour. What is the PDF of the duration of the trip?

Let X be the speed and let $Y = g(X)$ be the trip duration:

$$g(X) = \frac{180}{X}.$$

To find the CDF of Y , we must calculate

$$\mathbf{P}(Y \leq y) = \mathbf{P}\left(\frac{180}{X} \leq y\right) = \mathbf{P}\left(\frac{180}{y} \leq X\right).$$

We use the given uniform PDF of X , which is

$$f_X(x) = \begin{cases} 1/30, & \text{if } 30 \leq x \leq 60, \\ 0, & \text{otherwise,} \end{cases}$$

and the corresponding CDF, which is

$$F_X(x) = \begin{cases} 0, & \text{if } x \leq 30, \\ (x - 30)/30, & \text{if } 30 \leq x \leq 60, \\ 1, & \text{if } 60 \leq x. \end{cases}$$

Thus,

$$\begin{aligned} F_Y(y) &= \mathbf{P}\left(\frac{180}{y} \leq X\right) \\ &= 1 - F_X\left(\frac{180}{y}\right) \\ &= \begin{cases} 0, & \text{if } y \leq 180/60, \\ 1 - \frac{\frac{180}{y} - 30}{30}, & \text{if } 180/60 \leq y \leq 180/30, \\ 1, & \text{if } 180/30 \leq y, \end{cases} \\ &= \begin{cases} 0, & \text{if } y \leq 3, \\ 2 - (6/y), & \text{if } 3 \leq y \leq 6, \\ 1, & \text{if } 6 \leq y, \end{cases} \end{aligned}$$

(see Fig. 3.20). Differentiating this expression, we obtain the PDF of Y :

$$f_Y(y) = \begin{cases} 0, & \text{if } y < 3, \\ 6/y^2, & \text{if } 3 < y < 6, \\ 0, & \text{if } 6 < y. \end{cases}$$

Example 3.22. Let $Y = g(X) = X^2$, where X is a random variable with known PDF. For any $y \geq 0$, we have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}), \end{aligned}$$

and therefore, by differentiating and using the chain rule,

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}), \quad y \geq 0.$$

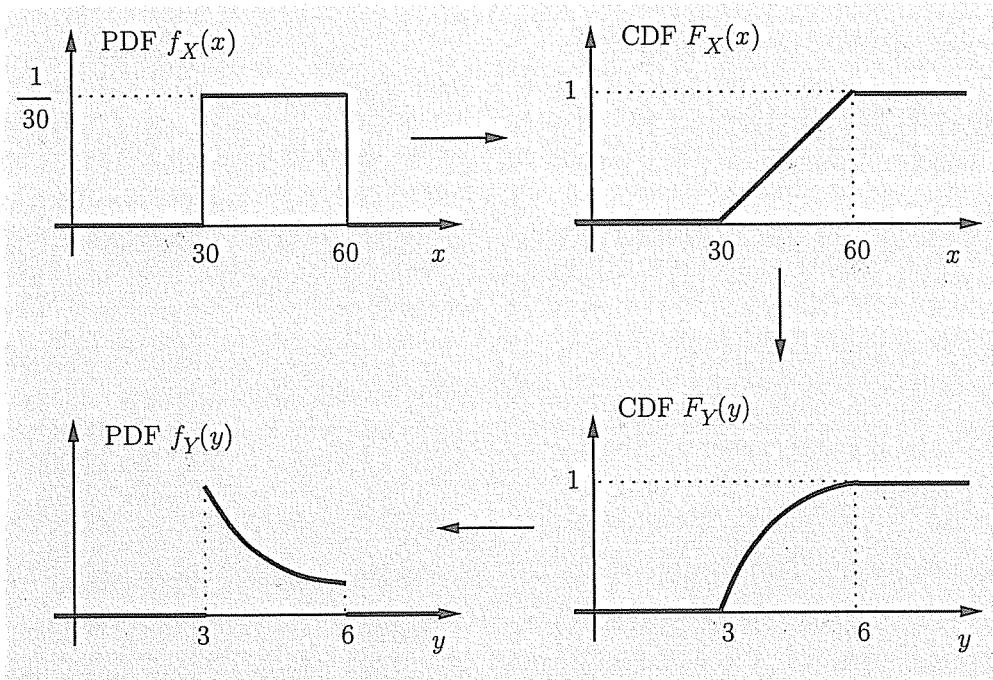


Figure 3.20: The calculation of the PDF of $Y = 180/X$ in Example 3.21. The arrows indicate the flow of the calculation.

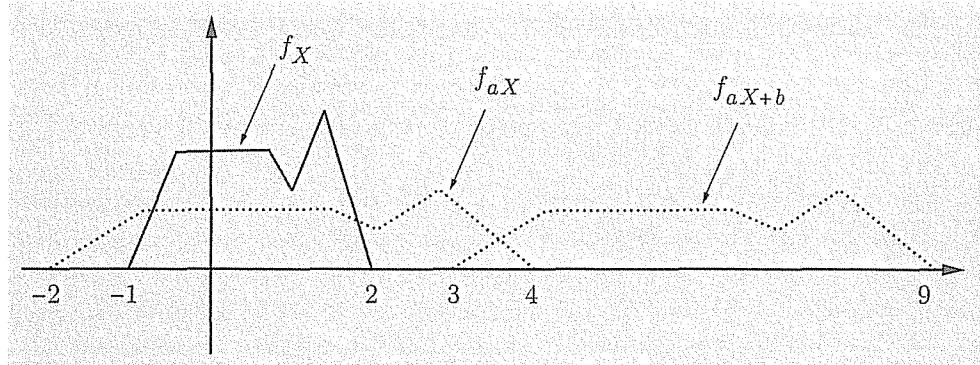


Figure 3.21: The PDF of $aX + b$ in terms of the PDF of X . In this figure, $a = 2$ and $b = 5$. As a first step, we obtain the PDF of aX . The range of Y is wider than the range of X , by a factor of a . Thus, the PDF f_X must be stretched (scaled horizontally) by this factor. But in order to keep the total area under the PDF equal to 1, we need to scale down the PDF (vertically) by the same factor a . The random variable $aX + b$ is the same as aX except that its values are shifted by b . Accordingly, we take the PDF of aX and shift it (horizontally) by b . The end result of these operations is the PDF of $Y = aX + b$ and is given mathematically by

$$f_Y(y) = \frac{1}{|a|} f_X \left(\frac{y - b}{a} \right).$$

If a were negative, the procedure would be the same except that the PDF of X would first need to be reflected around the vertical axis (“flipped”) yielding f_{-X} . Then a horizontal and vertical scaling (by a factor of $|a|$ and $1/|a|$, respectively) yields the PDF of $-|a|X = aX$. Finally, a horizontal shift of b would again yield the PDF of $aX + b$.

The Linear Case

We now focus on the important special case where Y is a linear function of X ; see Fig. 3.21 for a graphical interpretation.

The PDF of a Linear Function of a Random Variable

Let X be a continuous random variable with PDF f_X , and let

$$Y = aX + b,$$

where a and b are scalars with $a \neq 0$. Then,

$$f_Y(y) = \frac{1}{|a|} f_X \left(\frac{y - b}{a} \right).$$

To verify this formula, we first calculate the CDF of Y and then differentiate. We only show the steps for the case where $a > 0$; the case $a < 0$ is similar. We have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(aX + b \leq y) \\ &= \mathbb{P}\left(X \leq \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right). \end{aligned}$$

We now differentiate this equality and use the chain rule, to obtain

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

Example 3.23. A Linear Function of an Exponential Random Variable. Suppose that X is an exponential random variable with PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where λ is a positive parameter. Let $Y = aX + b$. Then,

$$f_Y(y) = \begin{cases} \frac{\lambda}{|a|} e^{-\lambda(y-b)/a}, & \text{if } (y-b)/a \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note that if $b = 0$ and $a > 0$, then Y is an exponential random variable with parameter λ/a . In general, however, Y need not be exponential. For example, if $a < 0$ and $b = 0$, then the range of Y is the negative real axis.

Example 3.24. A Linear Function of a Normal Random Variable is Normal. Suppose that X is a normal random variable with mean μ and variance σ^2 , and let $Y = aX + b$, where a and b are scalars with $a \neq 0$. We have

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

Therefore,

$$\begin{aligned} f_Y(y) &= \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{1}{|a|} \frac{1}{\sqrt{2\pi}\sigma} e^{-((y-b)/a-\mu)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}|a|\sigma} e^{-(y-b-a\mu)^2/2a^2\sigma^2}. \end{aligned}$$

We recognize this as a normal PDF with mean $a\mu + b$ and variance $a^2\sigma^2$. In particular, Y is a normal random variable.

The Monotonic Case

The calculation and the formula for the linear case can be generalized to the case where g is a monotonic function. Let X be a continuous random variable and suppose that its range is contained in a certain interval I , in the sense that $f_X(x) = 0$ for $x \notin I$. We consider the random variable $Y = g(X)$, and assume that g is **strictly monotonic** over the interval I , so that either

- (a) $g(x) < g(x')$ for all $x, x' \in I$ satisfying $x < x'$ (monotonically increasing case), or
- (b) $g(x) > g(x')$ for all $x, x' \in I$ satisfying $x < x'$ (monotonically decreasing case).

Furthermore, we assume that the function g is differentiable. Its derivative will necessarily be nonnegative in the increasing case and nonpositive in the decreasing case.

An important fact is that a strictly monotonic function can be “inverted” in the sense that there is some function h , called the inverse of g , such that for all $x \in I$, we have $y = g(x)$ if and only if $x = h(y)$. For example, the inverse of the function $g(x) = 180/x$ considered in Example 3.21 is $h(y) = 180/y$, because we have $y = 180/x$ if and only if $x = 180/y$. Other such examples of pairs of inverse functions include

$$g(x) = ax + b, \quad h(y) = \frac{y - b}{a},$$

where a and b are scalars with $a \neq 0$, and

$$g(x) = e^{ax}, \quad h(y) = \frac{\ln y}{a},$$

where a is a nonzero scalar.

For strictly monotonic functions g , the following is a convenient analytical formula for the PDF of the function $Y = g(X)$.

PDF Formula for a Strictly Monotonic Function of a Continuous Random Variable

Suppose that g is strictly monotonic and that for some function h and all x in the range of X we have

$$y = g(x) \quad \text{if and only if} \quad x = h(y).$$

Assume that h is differentiable. Then, the PDF of Y in the region where $f_Y(y) > 0$ is given by

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|.$$

For a verification of the above formula, assume first that g is monotonically increasing. Then, we have

$$F_Y(y) = \mathbf{P}(g(X) \leq y) = \mathbf{P}(X \leq h(y)) = F_X(h(y)),$$

where the second equality can be justified using the monotonically increasing property of g (see Fig. 3.22). By differentiating this relation, using also the chain rule, we obtain

$$f_Y(y) = \frac{dF_Y}{dy}(y) = f_X(h(y)) \frac{dh}{dy}(y).$$

Because g is monotonically increasing, h is also monotonically increasing, so its derivative is nonnegative:

$$\frac{dh}{dy}(y) = \left| \frac{dh}{dy}(y) \right|.$$

This justifies the PDF formula for a monotonically increasing function g . The justification for the case of monotonically decreasing function is similar: we differentiate instead the relation

$$F_Y(y) = \mathbf{P}(g(X) \leq y) = \mathbf{P}(X \geq h(y)) = 1 - F_X(h(y)),$$

and use the chain rule.

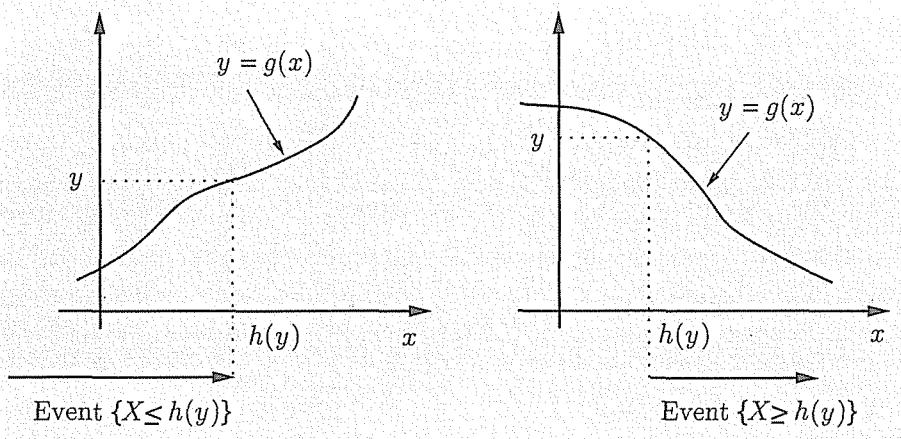


Figure 3.22: Calculating the probability $\mathbf{P}(g(X) \leq y)$. When g is monotonically increasing (left figure), the event $\{g(X) \leq y\}$ is the same as the event $\{X \leq h(y)\}$. When g is monotonically decreasing (right figure), the event $\{g(X) \leq y\}$ is the same as the event $\{X \geq h(y)\}$.

Example 3.21 (continued). To check the PDF formula, let us apply it to the problem of Example 3.21. In the region of interest, $x \in [30, 60]$, we have $h(y) = 180/y$, and

$$f_X(h(y)) = \frac{1}{30}, \quad \left| \frac{dh}{dy}(y) \right| = \frac{180}{y^2}.$$

Thus, in the region of interest $y \in [3, 6]$, the PDF formula yields

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right| = \frac{1}{30} \cdot \frac{180}{y^2} = \frac{6}{y^2},$$

consistent with the expression obtained earlier.

Example 3.25. Let $Y = g(X) = X^2$, where X is a continuous uniform random variable on the interval $(0, 1]$. Within this interval, g is strictly monotonic, and its inverse is $h(y) = \sqrt{y}$. Thus, for any $y \in (0, 1]$, we have

$$f_X(\sqrt{y}) = 1, \quad \left| \frac{dh}{dy}(y) \right| = \frac{1}{2\sqrt{y}},$$

and

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}, & \text{if } y \in (0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

We finally note that if we interpret PDFs in terms of probabilities of small intervals, the content of our formulas becomes pretty intuitive; see Fig. 3.23.

Functions of Two Random Variables

The two-step procedure that first calculates the CDF and then differentiates to obtain the PDF also applies to functions of more than one random variable.

Example 3.26. Two archers shoot at a target. The distance of each shot from the center of the target is uniformly distributed from 0 to 1, independently of the other shot. What is the PDF of the distance of the losing shot from the center?

Let X and Y be the distances from the center of the first and second shots, respectively. Let also Z be the distance of the losing shot:

$$Z = \max\{X, Y\}.$$

We know that X and Y are uniformly distributed over $[0, 1]$, so that for all $z \in [0, 1]$, we have

$$\mathbf{P}(X \leq z) = \mathbf{P}(Y \leq z) = z.$$

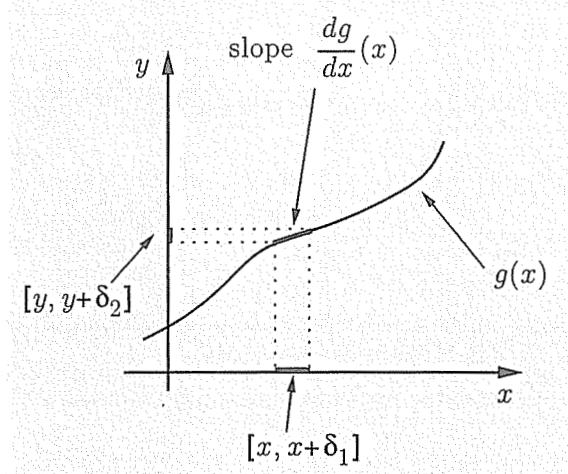


Figure 3.23: Illustration of the PDF formula for a monotonically increasing function g . Consider an interval $[x, x + \delta_1]$, where δ_1 is a small number. Under the mapping g , the image of this interval is another interval $[y, y + \delta_2]$. Since $(dg/dx)(x)$ is the slope of g , we have

$$\frac{\delta_2}{\delta_1} \approx \frac{dg}{dx}(x),$$

or in terms of the inverse function,

$$\frac{\delta_1}{\delta_2} \approx \frac{dh}{dy}(y),$$

We now note that the event $\{x \leq X \leq x + \delta_1\}$ is the same as the event $\{y \leq Y \leq y + \delta_2\}$. Thus,

$$\begin{aligned} f_Y(y)\delta_2 &\approx \mathbf{P}(y \leq Y \leq y + \delta_2) \\ &= \mathbf{P}(x \leq X \leq x + \delta_1) \\ &\approx f_X(x)\delta_1. \end{aligned}$$

We move δ_1 to the left-hand side and use our earlier formula for the ratio δ_2/δ_1 , to obtain

$$f_Y(y) \frac{dg}{dx}(x) = f_X(x).$$

Alternatively, if we move δ_2 to the right-hand side and use the formula for δ_1/δ_2 , we obtain

$$f_Y(y) = f_X(h(y)) \frac{dh}{dy}(y).$$

Thus, using the independence of X and Y , we have for all $z \in [0, 1]$,

$$\begin{aligned} F_Z(z) &= \mathbf{P}(\max\{X, Y\} \leq z) \\ &= \mathbf{P}(X \leq z, Y \leq z) \\ &= \mathbf{P}(X \leq z)\mathbf{P}(Y \leq z) \\ &= z^2. \end{aligned}$$

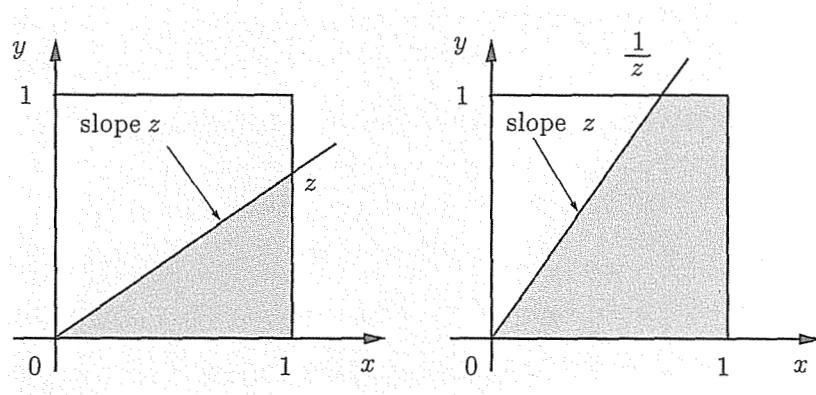


Figure 3.24: The calculation of the CDF of $Z = Y/X$ in Example 3.27. The value $P(Y/X \leq z)$ is equal to the shaded subarea of the unit square. The figure on the left deals with the case where $0 \leq z \leq 1$ and the figure on the right refers to the case where $z > 1$.

Differentiating, we obtain

$$f_Z(z) = \begin{cases} 2z, & \text{if } 0 \leq z \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Example 3.27. Let X and Y be independent random variables that are uniformly distributed on the interval $[0, 1]$. What is the PDF of the random variable $Z = Y/X$?

We will find the PDF of Z by first finding its CDF and then differentiating. We consider separately the cases $0 \leq z \leq 1$ and $z > 1$. As shown in Fig. 3.24, we have

$$F_Z(z) = P\left(\frac{Y}{X} \leq z\right) = \begin{cases} z/2, & \text{if } 0 \leq z \leq 1, \\ 1 - 1/(2z), & \text{if } z > 1, \\ 0, & \text{otherwise.} \end{cases}$$

By differentiating, we obtain

$$f_Z(z) = \begin{cases} 1/2, & \text{if } 0 \leq z \leq 1, \\ 1/(2z^2), & \text{if } z > 1, \\ 0, & \text{otherwise.} \end{cases}$$

Example 3.28. Romeo and Juliet have a date at a given time, and each, independently, will be late by an amount of time that is exponentially distributed with parameter λ . What is the PDF of the difference between their times of arrival?

Let us denote by X and Y the amounts by which Romeo and Juliet are late, respectively. We want to find the PDF of $Z = X - Y$, assuming that X and Y are independent and exponentially distributed with parameter λ . We will first calculate the CDF $F_Z(z)$ by considering separately the cases $z \geq 0$ and $z < 0$ (see Fig. 3.25).

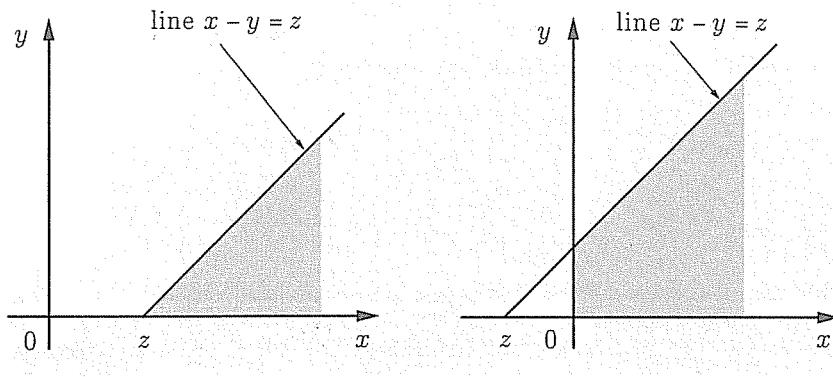


Figure 3.25: The calculation of the CDF of $Z = X - Y$ in Example 3.28. To obtain the value $\mathbf{P}(X - Y > z)$ we must integrate the joint PDF $f_{X,Y}(x,y)$ over the shaded area in the above figures, which correspond to $z \geq 0$ (left side) and $z < 0$ (right side).

For $z \geq 0$, we have (see the left side of Fig. 3.25)

$$\begin{aligned}
 F_Z(z) &= \mathbf{P}(X - Y \leq z) \\
 &= 1 - \mathbf{P}(X - Y > z) \\
 &= 1 - \int_0^\infty \left(\int_{z+y}^\infty f_{X,Y}(x,y) dx \right) dy \\
 &= 1 - \int_0^\infty \lambda e^{-\lambda y} \left(\int_{z+y}^\infty \lambda e^{-\lambda x} dx \right) dy \\
 &= 1 - \int_0^\infty \lambda e^{-\lambda y} e^{-\lambda(z+y)} dy \\
 &= 1 - e^{-\lambda z} \int_0^\infty \lambda e^{-2\lambda y} dy \\
 &= 1 - \frac{1}{2} e^{-\lambda z}.
 \end{aligned}$$

For the case $z < 0$, we can use a similar calculation, but we can also argue using symmetry. Indeed, the symmetry of the situation implies that the random variables $Z = X - Y$ and $-Z = Y - X$ have the same distribution. We have

$$F_Z(z) = \mathbf{P}(Z \leq z) = \mathbf{P}(-Z \geq -z) = \mathbf{P}(Z \geq -z) = 1 - F_Z(-z).$$

With $z < 0$, we have $-z > 0$ and using the formula derived earlier,

$$F_Z(z) = 1 - F_Z(-z) = 1 - \left(1 - \frac{1}{2} e^{-\lambda(-z)} \right) = \frac{1}{2} e^{\lambda z}.$$

Combining the two cases $z \geq 0$ and $z < 0$, we obtain

$$F_Z(z) = \begin{cases} 1 - \frac{1}{2} e^{-\lambda z}, & \text{if } z \geq 0, \\ \frac{1}{2} e^{\lambda z}, & \text{if } z < 0, \end{cases}$$

We now calculate the PDF of Z by differentiating its CDF. We have

$$f_Z(z) = \begin{cases} \frac{\lambda}{2} e^{-\lambda z}, & \text{if } z \geq 0, \\ \frac{\lambda}{2} e^{\lambda z}, & \text{if } z < 0, \end{cases}$$

or

$$f_Z(z) = \frac{\lambda}{2} e^{-\lambda|z|}.$$

This is known as a **two-sided exponential PDF**, also called the **Laplace PDF**.

3.7 SUMMARY AND DISCUSSION

Continuous random variables are characterized by PDFs, which are used to calculate event probabilities. This is similar to the use of PMFs for the discrete case, except that now we need to integrate instead of summing. Joint PDFs are similar to joint PMFs and are used to determine the probability of events that are defined in terms of multiple random variables. Furthermore, conditional PDFs are similar to conditional PMFs and are used to calculate conditional probabilities, given the value of the conditioning random variable. An important application is in problems of inference, using various forms of Bayes' rule that were developed in this chapter.

There are several special continuous random variables which frequently arise in probabilistic models. We introduced some of them, and derived their mean and variance. A summary is provided in the table that follows.

Summary of Results for Special Random Variables

Continuous Uniform Over $[a, b]$:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbb{E}[X] = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)^2}{12}.$$

Exponential with Parameter λ :

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases} \quad F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2}.$$

Normal with Parameters μ and $\sigma^2 > 0$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2},$$

$$\mathbb{E}[X] = \mu, \quad \text{var}(X) = \sigma^2.$$

We have also introduced CDFs, which can be used to characterize general random variables that are neither discrete nor continuous. CDFs are related to PMFs and PDFs, but are more general. In the case of a discrete random variable, we can obtain the PMF by differencing the CDF, and in the case of a continuous random variable, we can obtain the PDF by differentiating the CDF.

Calculating the PDF of a function $g(X)$ of a continuous random variable X can be challenging. The concept of a CDF is useful in this calculation. In particular, the PDF of $g(X)$ is typically obtained by calculating and differentiating the corresponding CDF. In some cases, such as when the function g is strictly monotonic, the calculation is facilitated through the use of some special formulas.

 PROBLEMS

SECTION 3.1. Continuous Random Variables and PDFs

Problem 1. Let X be uniformly distributed in the unit interval $[0, 1]$. Consider the random variable $Y = g(X)$, where

$$g(x) = \begin{cases} 1, & \text{if } x \leq 1/3, \\ 2, & \text{if } x > 1/3. \end{cases}$$

Find the expected value of Y by first deriving its PMF. Verify the result using the expected value rule.

Problem 2. Laplace random variable. Let X have the PDF

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda|x|},$$

where λ is a positive scalar. Verify that f_X satisfies the normalization condition, and evaluate the mean and variance of X .

Problem 3.* Show that the expected value of a continuous random variable X satisfies

$$\mathbf{E}[X] = \int_0^\infty \mathbf{P}(X > x) dx - \int_0^\infty \mathbf{P}(X < -x) dx.$$

Solution. We have

$$\begin{aligned} \int_0^\infty \mathbf{P}(X > x) dx &= \int_0^\infty \left(\int_x^\infty f_X(y) dy \right) dx \\ &= \int_0^\infty \left(\int_0^y f_X(y) dx \right) dy \\ &= \int_0^\infty f_X(y) \left(\int_0^y dx \right) dy \\ &= \int_0^\infty y f_X(y) dy, \end{aligned}$$

where for the second equality we have reversed the order of integration by writing the set $\{(x, y) \mid 0 \leq x < \infty, x \leq y < \infty\}$ as $\{(x, y) \mid 0 \leq x \leq y, 0 \leq y < \infty\}$. Similarly, we can show that

$$\int_0^\infty \mathbf{P}(X < -x) dx = - \int_{-\infty}^0 y f_X(y) dy.$$

Combining the two relations above, we obtain the desired result.

Problem 4.* Establish the validity of the expected value rule

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

where X is a continuous random variable with PDF f_X .

Solution. Let us express the function g as the difference of two nonnegative functions,

$$g(x) = g^+(x) - g^-(x),$$

where $g^+(x) = \max\{g(x), 0\}$, and $g^-(x) = \max\{-g(x), 0\}$. In particular, for any $t \geq 0$, we have $g(x) > t$ if and only if $g^+(x) > t$.

We will use the result

$$\mathbf{E}[g(X)] = \int_0^{\infty} \mathbf{P}(g(X) > t) dt - \int_0^{\infty} \mathbf{P}(g(X) < -t) dt$$

from the preceding exercise. The first term in the right-hand side is equal to

$$\int_0^{\infty} \int_{\{x \mid g(x) > t\}} f_X(x) dx dt = \int_{-\infty}^{\infty} \int_{\{t \mid 0 \leq t < g(x)\}} f_X(x) dt dx = \int_{-\infty}^{\infty} g^+(x) f_X(x) dx.$$

By a symmetrical argument, the term in the left-hand side is given by

$$\int_0^{\infty} \mathbf{P}(g(X) < -t) dt = \int_{-\infty}^{\infty} g^-(x) f_X(x) dx.$$

Combining the above equalities, we obtain

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g^+(x) f_X(x) dx - \int_{-\infty}^{\infty} g^-(x) f_X(x) dx = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

SECTION 3.2. Cumulative Distribution Functions

Problem 5. Consider a triangle and a point chosen within the triangle according to the uniform probability law. Let X be the distance from the point to the base of the triangle. Given the height of the triangle, find the CDF and the PDF of X .

Problem 6. Calamity Jane goes to the bank to make a withdrawal, and is equally likely to find 0 or 1 customers ahead of her. The service time of the customer ahead, if present, is exponentially distributed with parameter λ . What is the CDF of Jane's waiting time?

Problem 7. Consider two continuous random variables Y and Z , and a random variable X that is equal to Y with probability p and to Z with probability $1 - p$.

(a) Show that the PDF of X is given by

$$f_X(x) = p f_Y(x) + (1 - p) f_Z(x).$$

- (b) Calculate the CDF of the two-sided exponential random variable that has PDF given by

$$f_X(x) = \begin{cases} p\lambda e^{\lambda x}, & \text{if } x < 0, \\ (1-p)\lambda e^{-\lambda x}, & \text{if } x \geq 0, \end{cases}$$

where $\lambda > 0$ and $0 < p < 1$.

Problem 8.* Simulating a continuous random variable. A computer has a subroutine that can generate values of a random variable U that is uniformly distributed in the interval $[0, 1]$. Such a subroutine can be used to generate values of a continuous random variable with given CDF $F(x)$ as follows. If U takes a value u , we let the value of X be a number x that satisfies $F(x) = u$. For simplicity, we assume that the given CDF is strictly increasing over the range S of values of interest, where $S = \{x \mid 0 < F(x) < 1\}$. This condition guarantees that for any $u \in (0, 1)$, there is a unique x that satisfies $F(x) = u$.

- (a) Show that the CDF of the random variable X thus generated is indeed equal to the given CDF.
- (b) Describe how this procedure can be used to simulate an exponential random variable with parameter λ .
- (c) How can this procedure be generalized to simulate a discrete integer-valued random variable?

Solution. (a) By definition, the random variables X and U satisfy the relation $F(X) = U$. Since F is monotonic, we have for every x ,

$$X \leq x \quad \text{if and only if} \quad F(X) \leq F(x).$$

Therefore,

$$\mathbf{P}(X \leq x) = \mathbf{P}(F(X) \leq F(x)) = \mathbf{P}(U \leq F(x)) = F(x),$$

where the last equality follows because U is uniform. Thus, X has the desired CDF.

- (b) The exponential CDF has the form $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. Thus, to generate values of X , we should generate values $u \in (0, 1)$ of a uniformly distributed random variable U , and set X to the value for which $1 - e^{-\lambda x} = u$, or $x = -\ln(1 - u)/\lambda$.
- (c) Let again F be the desired CDF. To any $u \in (0, 1)$, there corresponds a unique integer x_u such that $F(x_u - 1) < u \leq F(x_u)$. This correspondence defines a random variable X as a function of the random variable U . We then have, for every integer k ,

$$\mathbf{P}(X = k) = \mathbf{P}(F(k - 1) < U \leq F(k)) = F(k) - F(k - 1).$$

Therefore, the CDF of X is equal to F , as desired.

SECTION 3.3. Normal Random Variables

Problem 9. Let X and Y be normal random variables with means 0 and 1, respectively, and variances 1 and 4, respectively.

- (a) Find $\mathbf{P}(X \leq 1.5)$ and $\mathbf{P}(X \leq -1)$.

- (b) Find the PDF of $(Y - 1)/2$.
 (c) Find $\mathbb{P}(-1 \leq Y \leq 1)$.

Problem 10. Let X be a normal random variable with zero mean and standard deviation σ . Use the normal tables to compute the probabilities of the events $\{X \geq k\sigma\}$ and $\{|X| \leq k\sigma\}$ for $k = 1, 2, 3$.

Problem 11. A city's temperature is modeled as a normal random variable with mean and standard deviation both equal to 10 degrees Celsius. What is the probability that the temperature at a randomly chosen time will be less than or equal to 59 degrees Fahrenheit?

Problem 12.* Show that the normal PDF satisfies the normalization property. *Hint:* The integral $\int_{-\infty}^{\infty} e^{-x^2/2} dx$ is equal to the square root of

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-y^2/2} dx dy,$$

and the latter integral can be evaluated by transforming to polar coordinates.

Solution. We note that

$$\begin{aligned} \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \right)^2 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} dx \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy \\ &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta \\ &= \int_0^{\infty} e^{-r^2/2} r dr \\ &= \int_0^{\infty} e^{-u} du \\ &= -e^{-u} \Big|_0^{\infty} \\ &= 1, \end{aligned}$$

where for the third equality, we use a transformation into polar coordinates, and for the fifth equality, we use the change of variables $u = r^2/2$. Thus, we have

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1,$$

because the integral is positive. Using the change of variables $u = (x - \mu)/\sigma$, it follows that

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = 1.$$

SECTION 3.4. Conditioning on an Event

Problem 13. Let X be a random variable with PDF

$$f_X(x) = \begin{cases} x/4, & \text{if } 1 < x \leq 3, \\ 0, & \text{otherwise,} \end{cases}$$

and let A be the event $\{X \geq 2\}$.

- (a) Find $\mathbb{E}[X]$, $\mathbb{P}(A)$, $f_{X|A}(x)$, and $\mathbb{E}[X | A]$.
- (b) Let $Y = X^2$. Find $\mathbb{E}[Y]$ and $\text{var}(Y)$.

Problem 14. The random variable X has the PDF

$$f_X(x) = \begin{cases} cx^{-2}, & \text{if } 1 \leq x \leq 2, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Determine the value of c .
- (b) Let A be the event $\{X > 1.5\}$. Calculate $\mathbb{P}(A)$ and the conditional PDF of X given that A has occurred.
- (c) Let $Y = X^2$. Calculate the conditional expectation and the conditional variance of Y given A .

Problem 15. An absent-minded professor schedules two student appointments for the same time. The appointment durations are independent and exponentially distributed with mean thirty minutes. The first student arrives on time, but the second student arrives five minutes late. What is the expected value of the time between the arrival of the first student and the departure of the second student?

Problem 16. Alvin throws darts at a circular target of radius r and is equally likely to hit any point in the target. Let X be the distance of Alvin's hit from the center.

- (a) Find the PDF, the mean, and the variance of X .
- (b) The target has an inner circle of radius t . If $X \leq t$, Alvin gets a score of $S = 1/X$. Otherwise his score is $S = 0$. Find the CDF of S . Is S a continuous random variable?

Problem 17.* Consider the following two-sided exponential PDF

$$f_X(x) = \begin{cases} p\lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ (1-p)\lambda e^{\lambda x}, & \text{if } x < 0, \end{cases}$$

where λ and p are scalars with $\lambda > 0$ and $p \in [0, 1]$. Find the mean and the variance of X in two ways:

- (a) By straightforward calculation of the associated expected values.
- (b) By using a divide-and-conquer strategy, and the mean and variance of the (one-sided) exponential random variable.

Solution. (a)

$$\begin{aligned}
 \mathbf{E}[X] &= \int_{-\infty}^{\infty} xf_X(x) dx \\
 &= \int_{-\infty}^0 x(1-p)\lambda e^{\lambda x} dx + \int_0^{\infty} xp\lambda e^{-\lambda x} dx \\
 &= -\frac{1-p}{\lambda} + \frac{p}{\lambda} \\
 &= \frac{2p-1}{\lambda}, \\
 \mathbf{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\
 &= \int_{-\infty}^0 x^2(1-p)\lambda e^{\lambda x} dx + \int_0^{\infty} x^2 p\lambda e^{-\lambda x} dx \\
 &= \frac{2(1-p)}{\lambda^2} + \frac{2p}{\lambda^2} \\
 &= \frac{2}{\lambda^2},
 \end{aligned}$$

and

$$\text{var}(X) = \frac{2}{\lambda^2} - \left(\frac{2p-1}{\lambda}\right)^2.$$

(b) Let A be the event $\{X \geq 0\}$, and note that $\mathbf{P}(A) = p$. Conditioned on A , the random variable X has a (one-sided) exponential distribution with parameter λ . Also, conditioned on A^c , the random variable $-X$ has the same one-sided exponential distribution. Thus,

$$\mathbf{E}[X | A] = \frac{1}{\lambda}, \quad \mathbf{E}[X | A^c] = -\frac{1}{\lambda},$$

and

$$\mathbf{E}[X^2 | A] = \mathbf{E}[X^2 | A^c] = \frac{2}{\lambda^2}.$$

It follows that

$$\begin{aligned}
 \mathbf{E}[X] &= \mathbf{P}(A)\mathbf{E}[X | A] + \mathbf{P}(A^c)\mathbf{E}[X | A^c] \\
 &= \frac{p}{\lambda} - \frac{1-p}{\lambda} \\
 &= \frac{2p-1}{\lambda},
 \end{aligned}$$

$$\begin{aligned}
 \mathbf{E}[X^2] &= \mathbf{P}(A)\mathbf{E}[X^2 | A] + \mathbf{P}(A^c)\mathbf{E}[X^2 | A^c] \\
 &= \frac{2p}{\lambda^2} + \frac{2(1-p)}{\lambda^2} \\
 &= \frac{2}{\lambda^2},
 \end{aligned}$$

and

$$\text{var}(X) = \frac{2}{\lambda^2} - \left(\frac{2p-1}{\lambda}\right)^2.$$

Problem 18.* Mixed random variables. Probabilistic models sometimes involve random variables that can be viewed as a mixture of a discrete random variable Y and a continuous random variable Z . By this we mean that the value of X is obtained according to the probability law of Y with a given probability p , and according to the probability law of Z with the complementary probability $1 - p$. Then, X is called a *mixed random variable* and its CDF is given, using the total probability theorem, by

$$\begin{aligned} F_X(x) &= \mathbf{P}(X \leq x) \\ &= p\mathbf{P}(Y \leq x) + (1 - p)\mathbf{P}(Z \leq x) \\ &= pF_Y(x) + (1 - p)F_Z(x). \end{aligned}$$

Its expected value is defined in a way that conforms to the total expectation theorem:

$$\mathbf{E}[X] = p\mathbf{E}[Y] + (1 - p)\mathbf{E}[Z].$$

The taxi stand and the bus stop near Al's home are in the same location. Al goes there at a given time and if a taxi is waiting (this happens with probability $2/3$) he boards it. Otherwise he waits for a taxi or a bus to come, whichever comes first. The next taxi will arrive in a time that is uniformly distributed between 0 and 10 minutes, while the next bus will arrive in exactly 5 minutes. Find the CDF and the expected value of Al's waiting time.

Solution. Let A be the event that Al will find a taxi waiting or will be picked up by the bus after 5 minutes. Note that the probability of boarding the next bus, given that Al has to wait, is

$$\mathbf{P}(\text{a taxi will take more than 5 minutes to arrive}) = \frac{1}{2}.$$

Al's waiting time, call it X , is a mixed random variable. With probability

$$\mathbf{P}(A) = \frac{2}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{6},$$

it is equal to its discrete component Y (corresponding to either finding a taxi waiting, or boarding the bus), which has PMF

$$\begin{aligned} p_Y(y) &= \begin{cases} \frac{2}{3\mathbf{P}(A)}, & \text{if } y = 0, \\ \frac{1}{6\mathbf{P}(A)}, & \text{if } y = 5, \end{cases} \\ &= \begin{cases} \frac{12}{15}, & \text{if } y = 0, \\ \frac{3}{15}, & \text{if } y = 5. \end{cases} \end{aligned}$$

[This equation follows from the calculation

$$p_Y(0) = \mathbf{P}(Y = 0 \mid A) = \frac{\mathbf{P}(Y = 0, A)}{\mathbf{P}(A)} = \frac{2}{3\mathbf{P}(A)}.$$

The calculation for $p_Y(5)$ is similar.] With the complementary probability $1 - \mathbf{P}(A)$, the waiting time is equal to its continuous component Z (corresponding to boarding a taxi after having to wait for some time less than 5 minutes), which has PDF

$$f_Z(z) = \begin{cases} 1/5, & \text{if } 0 \leq z \leq 5, \\ 0, & \text{otherwise.} \end{cases}$$

The CDF is given by $F_X(x) = \mathbf{P}(A)F_Y(x) + (1 - \mathbf{P}(A))F_Z(x)$, from which

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0, \\ \frac{5}{6} \cdot \frac{12}{15} + \frac{1}{6} \cdot \frac{x}{5}, & \text{if } 0 \leq x < 5, \\ 1, & \text{if } 5 \leq x. \end{cases}$$

The expected value of the waiting time is

$$\mathbf{E}[X] = \mathbf{P}(A)\mathbf{E}[Y] + (1 - \mathbf{P}(A))\mathbf{E}[Z] = \frac{5}{6} \cdot \frac{3}{15} \cdot 5 + \frac{1}{6} \cdot \frac{5}{2} = \frac{15}{12}.$$

SECTION 3.5. Multiple Continuous Random Variables

Problem 19. We start with a stick of length ℓ . We break it at a point which is chosen according to a uniform distribution and keep the piece, of length X , that contains the left end of the stick. We then repeat the same process on the piece that we were left with, and let Y be the length of the remaining piece after breaking for the second time.

- (a) Find the joint PDF of X and Y .
- (b) Find the marginal PDF of Y .
- (c) Use the PDF of Y to evaluate $\mathbf{E}[Y]$.
- (d) Evaluate $\mathbf{E}[Y]$, by exploiting the relation $Y = X \cdot (Y/X)$.

Problem 20. We have a stick of unit length, and we consider breaking it in three pieces using one of the following three methods.

- (i) We choose randomly and independently two points on the stick using a uniform PDF, and we break the stick at these two points.
- (ii) We break the stick at a random point chosen by using a uniform PDF, and then we break the piece that contains the right end of the stick, at a random point chosen by using a uniform PDF.
- (iii) We break the stick at a random point chosen by using a uniform PDF, and then we break the larger of the two pieces at a random point chosen by using a uniform PDF.

For each of the methods (i), (ii), and (iii), what is the probability that the three pieces we are left with can form a triangle?

Problem 21. A point is chosen at random (according to a uniform PDF) within the semicircle of the form $\{(x, y) \mid x^2 + y^2 \leq r, y \geq 0\}$, for some given $r > 0$.

- (a) Find the joint PDF of the coordinates X and Y of the chosen point.
- (b) Find the marginal PDF of Y and use it to find $\mathbb{E}[Y]$.
- (c) Check your answer in (b) by computing $\mathbb{E}[Y]$ directly without using the marginal PDF of Y .

Problem 22. Consider the following variant of the Buffon needle problem (Example 3.14), which was investigated by Laplace. A needle of length l is dropped on a plane surface that is partitioned in rectangles by horizontal lines that are a apart and vertical lines that are b apart. Suppose that the needle's length l satisfies $l < a$ and $l < b$. What is the expected number of rectangle sides crossed by the needle? What is the probability that the needle will cross at least one side of some rectangle?

Problem 23. A defective coin minting machine produces coins whose probability of heads is a random variable P with PDF

$$f_P(p) = \begin{cases} pe^p, & p \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

A coin produced by this machine is selected and tossed repeatedly, with successive tosses assumed independent.

- (a) Find the probability that a coin toss results in heads.
- (b) Given that a coin toss resulted in heads, find the conditional PDF of P .
- (c) Given that a first coin toss resulted in heads, find the conditional probability of heads on the next toss.

Problem 24.* Let X , Y , and Z be three random variables with joint PDF $f_{X,Y,Z}$. Show the multiplication rule:

$$f_{X,Y,Z}(x, y, z) = f_{X|Y,Z}(x|y, z)f_{Y|Z}(y|z)f_Z(z).$$

Solution. We have, using the definition of conditional density,

$$f_{X|Y,Z}(x|y, z) = \frac{f_{X,Y,Z}(x, y, z)}{f_{Y,Z}(y, z)},$$

and

$$f_{Y,Z}(y, z) = f_{Y|Z}(y|z)f_Z(z).$$

Combining these two relations, we obtain the multiplication rule.

Problem 25.* Estimating an expected value by simulation. Let $f_X(x)$ be a PDF such that for some nonnegative scalars a , b , and c , we have $f_X(x) = 0$ for all x outside the interval $[a, b]$, and $xf_X(x) \leq c$ for all x . Let Y_i , $i = 1, \dots, n$, be independent random variables with values generated as follows: a point (V_i, W_i) is chosen at random (according to a uniform PDF) within the rectangle whose corners are $(a, 0)$, $(b, 0)$, (a, c) , and (b, c) , and if $W_i \leq V_i f_X(V_i)$, the value of Y_i is set to 1, and otherwise it is set to 0. Consider the random variable

$$Z = \frac{Y_1 + \dots + Y_n}{n}.$$

Show that

$$\mathbb{E}[Z] = \frac{\mathbb{E}[X]}{c(b-a)}$$

and

$$\text{var}(Z) \leq \frac{1}{4n}.$$

In particular, we have $\text{var}(Z) \rightarrow 0$ as $n \rightarrow \infty$.

Solution. We have

$$\begin{aligned} \mathbb{P}(Y_i = 1) &= \mathbb{P}(W_i \leq V_i f_X(V_i)) \\ &= \int_a^b \int_0^{v f_X(v)} \frac{1}{c(b-a)} dw dv \\ &= \frac{\int_a^b v f_X(v) dv}{c(b-a)} \\ &= \frac{\mathbb{E}[X]}{c(b-a)}. \end{aligned}$$

The random variable Z has mean $\mathbb{P}(Y_i = 1)$ and variance

$$\text{var}(Z) = \frac{\mathbb{P}(Y_i = 1)(1 - \mathbb{P}(Y_i = 1))}{n}.$$

Since $0 \leq (1 - 2p)^2 = 1 - 4p(1 - p)$, we have $p(1 - p) \leq 1/4$ for any p in $[0, 1]$, so it follows that $\text{var}(Z) \leq 1/(4n)$.

Problem 26.* Estimating an expected value by simulation using samples of another random variable. Let Y_1, \dots, Y_n be independent random variables drawn from a common and known PDF f_Y . Let S be the set of all possible values of Y_i , $S = \{y \mid f_Y(y) > 0\}$. Let X be a random variable with known PDF f_X , such that $f_X(y) = 0$, for all $y \notin S$. Consider the random variable

$$Z = \frac{1}{n} \sum_{i=1}^n Y_i \frac{f_X(Y_i)}{f_Y(Y_i)}.$$

Show that

$$\mathbb{E}[Z] = \mathbb{E}[X].$$

Solution. We have

$$\mathbb{E} \left[Y_i \frac{f_X(Y_i)}{f_Y(Y_i)} \right] = \int_S y \frac{f_X(y)}{f_Y(y)} f_Y(y) dy = \int_S y f_X(y) dy = \mathbb{E}[X].$$

Thus,

$$\mathbb{E}[Z] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[Y_i \frac{f_X(Y_i)}{f_Y(Y_i)} \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X] = \mathbb{E}[X].$$

Problem 27.* Let X and Y be independent continuous random variables with PDFs f_X and f_Y , respectively, and let $Z = X + Y$.

- (a) Show that $f_{Z|X}(z|x) = f_Y(z-x)$. *Hint:* Write an expression for the conditional CDF of Z given X , and differentiate.
- (b) Assume that X and Y are exponentially distributed with parameter λ . Find the conditional PDF of X , given that $Z = z$.

Solution. (a) We have

$$\begin{aligned}\mathbf{P}(Z \leq z | X = x) &= \mathbf{P}(X + Y \leq z | X = x) \\ &= \mathbf{P}(x + Y \leq z | X = x) \\ &= \mathbf{P}(x + Y \leq z) \\ &= \mathbf{P}(Y \leq z - x),\end{aligned}$$

where the third equality follows from the independence of X and Y . By differentiating both sides with respect to z , the result follows.

(b) We have, for $0 \leq x \leq z$,

$$f_{X|Z}(x|z) = \frac{f_{Z|X}(z|x)f_X(x)}{f_Z(z)} = \frac{f_Y(z-x)f_X(x)}{f_Z(z)} = \frac{\lambda e^{-\lambda(z-x)}\lambda e^{-\lambda x}}{f_Z(z)} = \frac{\lambda^2 e^{-\lambda z}}{f_Z(z)}.$$

Since this is the same for all x , it follows that the conditional distribution of X is uniform on the interval $[0, z]$, with PDF $f_{X|Z}(x|z) = 1/z$.

Problem 28.* Consider two continuous random variables with joint PDF $f_{X,Y}$. Suppose that for any subsets A and B of the real line, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent. Show that the random variables X and Y are independent.

Solution. For any two real numbers x and y , using the independence of the events $\{X \leq x\}$ and $\{Y \leq y\}$, we have

$$F_{X,Y}(x, y) = \mathbf{P}(X \leq x, Y \leq y) = \mathbf{P}(X \leq x)\mathbf{P}(Y \leq y) = F_X(x)F_Y(y).$$

Taking derivatives of both sides, we obtain

$$f_{X,Y}(x, y) = \frac{\partial F_{X,Y}}{\partial x \partial y}(x, y) = \frac{\partial F_X}{\partial x}(x) \frac{\partial F_Y}{\partial y}(y) = f_X(x)f_Y(y),$$

which establishes that X and Y are independent.

Problem 29.* The sum of a random number of random variables. You visit a random number N of stores and in the i th store, you spend a random amount of money X_i . Let

$$T = X_1 + X_2 + \cdots + X_N$$

be the total amount of money that you spend. We assume that N is a discrete random variable with a given PMF, and that the X_i are random variables with the same

mean $\mathbb{E}[X]$ and variance $\text{var}(X)$. Furthermore, we assume that N and all the X_i are independent. Show that

$$\mathbb{E}[T] = \mathbb{E}[X] \cdot \mathbb{E}[N] \quad \text{and} \quad \text{var}(T) = \text{var}(X)\mathbb{E}[N] + (\mathbb{E}[X])^2 \text{var}(N).$$

Solution. We have for all i ,

$$\mathbb{E}[T | N = i] = i\mathbb{E}[X],$$

since conditional on $N = i$, you will visit exactly i stores, and you will spend an expected amount of money $\mathbb{E}[X]$ in each.

We now apply the total expectation theorem. We have

$$\begin{aligned} \mathbb{E}[T] &= \sum_{i=1}^{\infty} \mathbb{P}(N = i) \mathbb{E}[T | N = i] \\ &= \sum_{i=1}^{\infty} \mathbb{P}(N = i) i\mathbb{E}[X] \\ &= \mathbb{E}[X] \sum_{i=1}^{\infty} i\mathbb{P}(N = i) \\ &= \mathbb{E}[X] \cdot \mathbb{E}[N]. \end{aligned}$$

Similarly, using also the independence of the X_i , which implies that $\mathbb{E}[X_i X_j] = (\mathbb{E}[X])^2$ if $i \neq j$, the second moment of T is calculated as

$$\begin{aligned} \mathbb{E}[T^2] &= \sum_{i=1}^{\infty} \mathbb{P}(N = i) \mathbb{E}[T^2 | N = i] \\ &= \sum_{i=1}^{\infty} \mathbb{P}(N = i) \mathbb{E}[(X_1 + \dots + X_N)^2 | N = i] \\ &= \sum_{i=1}^{\infty} \mathbb{P}(N = i) (i\mathbb{E}[X^2] + i(i-1)(\mathbb{E}[X])^2) \\ &= \mathbb{E}[X^2] \sum_{i=1}^{\infty} i\mathbb{P}(N = i) + (\mathbb{E}[X])^2 \sum_{i=1}^{\infty} i(i-1)\mathbb{P}(N = i) \\ &= \mathbb{E}[X^2]\mathbb{E}[N] + (\mathbb{E}[X])^2 (\mathbb{E}[N^2] - \mathbb{E}[N]) \\ &= \text{var}(X)\mathbb{E}[N] + (\mathbb{E}[X])^2 \mathbb{E}[N^2]. \end{aligned}$$

The variance is then obtained by

$$\begin{aligned} \text{var}(T) &= \mathbb{E}[T^2] - (\mathbb{E}[T])^2 \\ &= \text{var}(X)\mathbb{E}[N] + (\mathbb{E}[X])^2 \mathbb{E}[N^2] - (\mathbb{E}[X])^2 (\mathbb{E}[N])^2 \\ &= \text{var}(X)\mathbb{E}[N] + (\mathbb{E}[X])^2 (\mathbb{E}[N^2] - (\mathbb{E}[N])^2), \end{aligned}$$

so finally

$$\text{var}(T) = \text{var}(X)\mathbb{E}[N] + (\mathbb{E}[X])^2 \text{var}(N).$$

Note: The formulas for $\mathbb{E}[T]$ and $\text{var}(T)$ will also be obtained in Chapter 4, using a more abstract approach.

SECTION 3.6. Derived Distributions

Problem 30. If X is a random variable that is uniformly distributed between -1 and 1 , find the PDF of $\sqrt{|X|}$ and the PDF of $-\ln|X|$.

Problem 31. Find the PDF of e^X in terms of the PDF of X . Specialize the answer to the case where X is uniformly distributed between 0 and 1 .

Problem 32. Find the PDFs of $|X|^{1/3}$ and $|X|^{1/4}$ in terms of the PDF of X .

Problem 33. The metro train arrives at the station near your home every quarter hour starting at 6:00 a.m. You walk into the station every morning between 7:10 and 7:30 a.m., with the time in this interval being a random variable with given PDF (cf. Example 3.11). Let X be the elapsed time, in minutes, between 7:10 and the time of your arrival. Let Y be the time that you have to wait until you board a train. Calculate the CDF of Y in terms of the CDF of X and differentiate to obtain a formula for the PDF of Y .

Problem 34. Let X and Y be independent random variables, uniformly distributed in the interval $[0, 1]$. Find the CDF and the PDF of $|X - Y|$.

Problem 35. Let X and Y be the Cartesian coordinates of a randomly chosen point (according to a uniform PDF) in the triangle with vertices at $(0, 1)$, $(0, -1)$, and $(1, 0)$. Find the CDF and the PDF of $|X - Y|$.

Problem 36. Two points are chosen randomly and independently from the interval $[0, 1]$ according to a uniform distribution. Show that the expected distance between the two points is $1/3$.

Problem 37. Consider the same problem as in Example 3.28, but assume that the random variables X and Y are independent and exponentially distributed with different parameters λ and μ , respectively. Find the PDF of $X - Y$.

Problem 38.* Cauchy random variable.

- (a) Let X be a random variable that is uniformly distributed between $-1/2$ and $1/2$. Show that the PDF of $Y = \tan(\pi X)$ is

$$f_Y(y) = \frac{1}{\pi(1+y^2)}, \quad -\infty < y < \infty.$$

(Y is called a *Cauchy random variable*.)

- (b) Let Y be a Cauchy random variable. Find the PDF of the random variable X , which is equal to the angle between $-\pi/2$ and $\pi/2$ whose tangent is Y .

Solution. (a) We first note that Y is a continuous, strictly monotonically increasing function of X , which takes values between $-\infty$ and ∞ , as X ranges over the interval $[-1/2, 1/2]$. Therefore, we have for all scalars y ,

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(\tan(\pi X) \leq y) = \mathbf{P}(\pi X \leq \tan^{-1} y) = \frac{1}{\pi} \tan^{-1} y + \frac{1}{2},$$

where the last equality follows using the CDF of X , which is uniformly distributed in the interval $[-1/2, 1/2]$. Therefore, by differentiation, using the formula $d/dy(\tan^{-1} y) = 1/(1+y^2)$, we have for all y ,

$$f_Y(y) = \frac{1}{\pi(1+y^2)}.$$

(b) We first compute the CDF of X and then differentiate to obtain its PDF. We have for $-\pi/2 \leq x \leq \pi/2$,

$$\begin{aligned} \mathbf{P}(X \leq x) &= \mathbf{P}(\tan^{-1} Y \leq x) \\ &= \mathbf{P}(Y \leq \tan x) \\ &= \frac{1}{\pi} \int_{-\infty}^{\tan x} \frac{1}{1+y^2} dy \\ &= \frac{1}{\pi} \tan^{-1} y \Big|_{-\infty}^{\tan x} \\ &= \frac{1}{\pi} \left(x + \frac{\pi}{2} \right). \end{aligned}$$

For $x < -\pi/2$, we have $\mathbf{P}(X \leq x) = 0$, and for $\pi/2 < x$, we have $\mathbf{P}(X \leq x) = 1$. Taking the derivative of the CDF $\mathbf{P}(X \leq x)$, we find that X is uniformly distributed on the interval $[-\pi/2, \pi/2]$.

Note: An interesting property of the Cauchy random variable is that it satisfies

$$\int_0^\infty \frac{y}{\pi(1+y^2)} dy = - \int_{-\infty}^0 \frac{y}{\pi(1+y^2)} dy = \infty,$$

as can be easily verified. As a result, the Cauchy random variable does not have a well-defined expected value, despite the symmetry of its PDF around 0.

Problem 39.* Competing exponentials. The lifetimes of two lightbulbs are modeled as independent and exponential random variables X and Y , with parameters λ and μ , respectively. The time at which a lightbulb first burns out is

$$Z = \min\{X, Y\}.$$

Show that Z is an exponential random variable with parameter $\lambda + \mu$.

Solution. For all $z \geq 0$, we have, using the independence of X and Y , and the form of the exponential CDF,

$$\begin{aligned} F_Z(z) &= \mathbf{P}(\min\{X, Y\} \leq z) \\ &= 1 - \mathbf{P}(\min\{X, Y\} > z) \\ &= 1 - \mathbf{P}(X > z, Y > z) \\ &= 1 - \mathbf{P}(X > z)\mathbf{P}(Y > z) \\ &= 1 - e^{-\lambda z} e^{-\mu z} \\ &= 1 - e^{-(\lambda+\mu)z}. \end{aligned}$$

This is recognized as the exponential CDF with parameter $\lambda + \mu$. Thus, the minimum of two independent exponentials with parameters λ and μ is an exponential with parameter $\lambda + \mu$.

Problem 40.* Let X and Y be independent standard normal random variables. The pair (X, Y) can be described in polar coordinates in terms of random variables $R \geq 0$ and $\Theta \in [0, 2\pi]$, so that

$$X = R \cos \Theta, \quad Y = R \sin \Theta.$$

(a) Show that Θ is uniformly distributed in $[0, 2\pi]$, that R has the PDF

$$f_R(r) = r e^{-r^2/2}, \quad r \geq 0,$$

and that R and Θ are independent. (The random variable R is said to have a **Rayleigh distribution**.)

(b) Show that R^2 has an exponential distribution, with parameter $1/2$.

Note: Using the results in this problem, we see that samples of a normal random variable can be generated using samples of independent uniform and exponential random variables.

Solution. (a) The joint PDF of X and Y is

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}.$$

Consider a subset A of the two-dimensional plane, and let B be a representation of the same subset in polar coordinates, i.e.,

$$(x, y) \in A \quad \text{if and only if } (r, \theta) \in B,$$

where (x, y) and (r, θ) are related by

$$x = r \cos \theta, \quad y = r \sin \theta.$$

In particular, the events $(X, Y) \in A$ and $(R, \Theta) \in B$ coincide. Therefore,

$$\mathbf{P}((R, \Theta) \in B) = \mathbf{P}((X, Y) \in A) = \frac{1}{2\pi} \iint_{(x,y) \in A} e^{-(x^2+y^2)/2} dx dy = \frac{1}{2\pi} \iint_{(r,\theta) \in B} e^{-r^2/2} r dr d\theta,$$

where the third equality is obtained by transforming to polar coordinates. By comparing with the formula

$$\mathbf{P}((R, \Theta) \in B) = \iint_{(r,\theta) \in B} f_{R,\Theta}(r, \theta) dr d\theta,$$

we see that

$$f_{R,\Theta}(r, \theta) = \frac{r}{2\pi} e^{-r^2/2}, \quad r \geq 0, \theta \in [0, 2\pi].$$

We then obtain

$$f_R(r) = \int_0^{2\pi} \frac{r}{2\pi} e^{-r^2/2} d\theta = r e^{-r^2/2}, \quad r \geq 0,$$

and

$$f_\Theta(\theta) = \int_0^\infty \frac{r}{2\pi} e^{-r^2/2} dr = \frac{1}{2\pi}, \quad \theta \in [0, 2\pi].$$

Since $f_{R,\Theta}(r, \theta) = f_R(r)f_\Theta(\theta)$, we see that R and Θ are independent.

(b) Let $t \geq 0$. We have

$$\mathbf{P}(R^2 \geq t) = \mathbf{P}(R \geq \sqrt{t}) = \int_{\sqrt{t}}^\infty r e^{-r^2/2} dr = \int_{t/2}^\infty e^{-u} du = e^{-t/2},$$

where we have used the change of variables $u = r^2/2$. By differentiating, we obtain

$$f_{R^2}(t) = \frac{1}{2} e^{-t/2}, \quad t \geq 0.$$

Further Topics on Random Variables

Contents

4.1. Transforms	p. 210
4.2. Sums of Independent Random Variables - Convolution	p. 221
4.3. More on Conditional Expectation and Variance	p. 225
4.4. Sum of a Random Number of Independent Random Variables .	p. 232
4.5. Covariance and Correlation	p. 236
4.6. Least Squares Estimation	p. 240
4.7. The Bivariate Normal Distribution	p. 247
4.8. Summary and Discussion	p. 255
Problems	p. 257

In this chapter, we develop a number of more advanced topics. We introduce methods that are useful in:

- (a) dealing with the sum of independent random variables, including the case where the number of random variables is itself random;
- (b) addressing problems of estimation or prediction of an unknown random variable on the basis of observed values of other random variables.

With these goals in mind, we introduce a number of tools, including transforms and convolutions, and we refine our understanding of the concept of conditional expectation. We end the chapter with a discussion of the bivariate normal distribution.

The material in this chapter may be viewed as optional in a first reading of the book. It is not used anywhere in the main text of other chapters, although it is used in the solutions of a few problems. On the other hand, the concepts and methods discussed here constitute essential background for a more advanced treatment of probability and stochastic processes, and provide powerful tools in several disciplines that rely on probabilistic models.

4.1 TRANSFORMS

In this section, we introduce the transform associated with a random variable. The transform provides us with an alternative representation of a probability law (PMF or PDF). It is not particularly intuitive, but it is often convenient for certain types of mathematical manipulations.

The **transform** associated with a random variable X (also referred to as the associated **moment generating function**) is a function $M_X(s)$ of a scalar parameter s , defined by

$$M_X(s) = \mathbf{E}[e^{sX}].$$

The simpler notation $M(s)$ can also be used whenever the underlying random variable X is clear from the context. In more detail, when X is a discrete random variable, the corresponding transform is given by

$$M(s) = \sum_x e^{sx} p_X(x),$$

while in the continuous case, we have[†]

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx.$$

Let us now provide some examples of transforms.

Example 4.1. Let

$$p_X(x) = \begin{cases} 1/2, & \text{if } x = 2, \\ 1/6, & \text{if } x = 3, \\ 1/3, & \text{if } x = 5. \end{cases}$$

Then, the corresponding transform is

$$M(s) = \frac{1}{2}e^{2s} + \frac{1}{6}e^{3s} + \frac{1}{3}e^{5s}.$$

Example 4.2. The Transform Associated with a Poisson Random Variable. Let X be a Poisson random variable with parameter λ :

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots$$

The corresponding transform is given by

$$M(s) = \sum_{x=0}^{\infty} e^{sx} \frac{\lambda^x e^{-\lambda}}{x!}.$$

We let $a = e^s \lambda$ and obtain

$$M(s) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{a^x}{x!} = e^{-\lambda} e^a = e^{a-\lambda} = e^{\lambda(e^s - 1)}.$$

Example 4.3. The Transform Associated with an Exponential Random Variable. Let X be an exponential random variable with parameter λ :

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Then,

$$\begin{aligned} M(s) &= \lambda \int_0^{\infty} e^{sx} e^{-\lambda x} dx \\ &= \lambda \int_0^{\infty} e^{(s-\lambda)x} dx \\ &= \lambda \frac{e^{(s-\lambda)x}}{s-\lambda} \Big|_0^{\infty} \quad (\text{if } s < \lambda) \\ &= \frac{\lambda}{\lambda - s}. \end{aligned}$$

† The reader who is familiar with Laplace transforms may recognize that the transform associated with a continuous random variable is essentially the same as the Laplace transform of its PDF, the only difference being that Laplace transforms usually involve e^{-sx} rather than e^{sx} . For the discrete case, a variable z is sometimes used in place of e^s and the resulting transform $M(z) = \sum_x z^x p_X(x)$ is known as the **z-transform**. However, we will not be using z-transforms in this book.

The above calculation and the formula for $M(s)$ is correct only if the integrand $e^{(s-\lambda)x}$ decays as x increases, which is the case if and only if $s < \lambda$; otherwise, the integral is infinite.

It is important to realize that the transform is not a number but rather a *function* of a parameter s . Thus, we are dealing with a transformation that starts with a function, e.g., a PDF, and results in a new function. Strictly speaking, $M(s)$ is only defined for those values of s for which $\mathbf{E}[e^{sX}]$ is finite, as noted in the preceding example.

Example 4.4. The Transform Associated with a Linear Function of a Random Variable. Let $M_X(s)$ be the transform associated with a random variable X . Consider a new random variable $Y = aX + b$. We then have

$$M_Y(s) = \mathbf{E}[e^{s(aX+b)}] = e^{sb} \mathbf{E}[e^{saX}] = e^{sb} M_X(sa).$$

For example, if X is exponential with parameter $\lambda = 1$, so that $M_X(s) = 1/(1-s)$, and if $Y = 2X + 3$, then

$$M_Y(s) = e^{3s} \frac{1}{1-2s}.$$

Example 4.5. The Transform Associated with a Normal Random Variable. Let X be a normal random variable with mean μ and variance σ^2 . To calculate the corresponding transform, we first consider the special case of the standard normal random variable Y , where $\mu = 0$ and $\sigma^2 = 1$, and then use the formula derived in the preceding example. The PDF of the standard normal is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2},$$

and the associated transform is

$$\begin{aligned} M_Y(s) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} e^{sy} dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)+sy} dy \\ &= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)+sy-(s^2/2)} dy \\ &= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y-s)^2/2} dy \\ &= e^{s^2/2}, \end{aligned}$$

where the last equality follows by using the normalization property of a normal PDF with mean s and unit variance.

A general normal random variable with mean μ and variance σ^2 is obtained from the standard normal via the linear transformation

$$X = \sigma Y + \mu.$$

The transform associated with the standard normal is $M_Y(s) = e^{s^2/2}$, as verified above. By applying the formula of Example 4.4, we obtain

$$M_X(s) = e^{s\mu} M_Y(s\sigma) = e^{(\sigma^2 s^2/2) + \mu s}.$$

From Transforms to Moments

The reason behind the alternative name “moment generating function” is that the moments of a random variable are easily computed once a formula for the associated transform is available. To see this, let us consider a continuous random variable X , and let us take the derivative of both sides of the definition

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx,$$

with respect to s . We obtain

$$\begin{aligned} \frac{d}{ds} M(s) &= \frac{d}{ds} \int_{-\infty}^{\infty} e^{sx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} \frac{d}{ds} e^{sx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} x e^{sx} f_X(x) dx. \end{aligned}$$

This equality holds for all values of s .[†] By considering the special case where $s = 0$, we obtain

$$\left. \frac{d}{ds} M(s) \right|_{s=0} = \int_{-\infty}^{\infty} x f_X(x) dx = \mathbf{E}[X].$$

[†] This derivation involves an interchange of differentiation and integration. The interchange turns out to be justified for all of the applications to be considered in this book. Furthermore, the derivation remains valid for general random variables, including discrete ones. In fact, it could be carried out more abstractly, in the form

$$\frac{d}{ds} M(s) = \frac{d}{ds} \mathbf{E}[e^{sX}] = \mathbf{E} \left[\frac{d}{ds} e^{sX} \right] = \mathbf{E}[X e^{sX}],$$

leading to the same conclusion.

More generally, if we differentiate n times the function $M(s)$ with respect to s , a similar calculation yields

$$\frac{d^n}{ds^n} M(s) \Big|_{s=0} = \int_{-\infty}^{\infty} x^n f_X(x) dx = \mathbf{E}[X^n].$$

Example 4.6. We saw earlier (Example 4.1) that the PMF

$$p_X(x) = \begin{cases} 1/2, & \text{if } x = 2, \\ 1/6, & \text{if } x = 3, \\ 1/3, & \text{if } x = 5, \end{cases}$$

is associated with the transform

$$M(s) = \frac{1}{2}e^{2s} + \frac{1}{6}e^{3s} + \frac{1}{3}e^{5s}.$$

Thus,

$$\begin{aligned} \mathbf{E}[X] &= \frac{d}{ds} M(s) \Big|_{s=0} \\ &= \frac{1}{2} \cdot 2e^{2s} + \frac{1}{6} \cdot 3e^{3s} + \frac{1}{3} \cdot 5e^{5s} \Big|_{s=0} \\ &= \frac{1}{2} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{3} \cdot 5 \\ &= \frac{19}{6}. \end{aligned}$$

Also,

$$\begin{aligned} \mathbf{E}[X^2] &= \frac{d^2}{ds^2} M(s) \Big|_{s=0} \\ &= \frac{1}{2} \cdot 4e^{2s} + \frac{1}{6} \cdot 9e^{3s} + \frac{1}{3} \cdot 25e^{5s} \Big|_{s=0} \\ &= \frac{1}{2} \cdot 4 + \frac{1}{6} \cdot 9 + \frac{1}{3} \cdot 25 \\ &= \frac{71}{6}. \end{aligned}$$

For an exponential random variable with PDF

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

we found earlier that

$$M(s) = \frac{\lambda}{\lambda - s}.$$

Thus,

$$\frac{d}{ds} M(s) = \frac{\lambda}{(\lambda - s)^2}, \quad \frac{d^2}{ds^2} M(s) = \frac{2\lambda}{(\lambda - s)^3}.$$

By setting $s = 0$, we obtain

$$\mathbf{E}[X] = \frac{1}{\lambda}, \quad \mathbf{E}[X^2] = \frac{2}{\lambda^2},$$

which agrees with the formulas derived in Chapter 3.

We close by noting two more useful and generic properties of transforms. For any random variable X , we have

$$M_X(0) = \mathbf{E}[e^{0X}] = \mathbf{E}[1] = 1,$$

and if X takes only nonnegative integer values, then

$$\lim_{s \rightarrow -\infty} M_X(s) = \mathbf{P}(X = 0)$$

(see the end-of-chapter problems).

Inversion of Transforms

A very important property of the transform $M_X(s)$ is that it can be inverted, i.e., it can be used to determine the probability law of the random variable X . Some appropriate mathematical conditions are required, which are satisfied in all of our examples that make use of the inversion property. The following is a more precise statement. Its proof is beyond our scope.

Inversion Property

The transform $M_X(s)$ associated with a random variable X uniquely determines the CDF of X , assuming that $M_X(s)$ is finite for all s in an interval $[-a, a]$, where a is a positive number.

There exist explicit formulas that allow us to recover the PMF or PDF of a random variable starting from the associated transform, but they are quite difficult to use. In practice, transforms are usually inverted by “pattern matching,” based on tables of known distribution-transform pairs. We will see a number of such examples shortly.

Example 4.7. We are told that the transform associated with a random variable X is

$$M(s) = \frac{1}{4}e^{-s} + \frac{1}{2} + \frac{1}{8}e^{4s} + \frac{1}{8}e^{5s}.$$

Since $M(s)$ is a sum of terms of the form e^{sx} , we can compare with the general formula

$$M(s) = \sum_x e^{sx} p_X(x),$$

and infer that X is a discrete random variable. The different values that X can take can be read from the corresponding exponents, and are $-1, 0, 4$, and 5 . The probability of each value x is given by the coefficient multiplying the corresponding e^{sx} term. In our case,

$$\mathbf{P}(X = -1) = \frac{1}{4}, \quad \mathbf{P}(X = 0) = \frac{1}{2}, \quad \mathbf{P}(X = 4) = \frac{1}{8}, \quad \mathbf{P}(X = 5) = \frac{1}{8}.$$

Generalizing from the last example, the distribution of a finite-valued discrete random variable can be always found by inspection of the corresponding transform. The same procedure also works for discrete random variables with an infinite range, as in the example that follows.

Example 4.8. The Transform Associated with a Geometric Random Variable. We are told that the transform associated with a random variable X is of the form

$$M(s) = \frac{pe^s}{1 - (1 - p)e^s},$$

where p is a constant in the range $0 < p \leq 1$. We wish to find the distribution of X . We recall the formula for the geometric series:

$$\frac{1}{1 - \alpha} = 1 + \alpha + \alpha^2 + \dots,$$

which is valid whenever $|\alpha| < 1$. We use this formula with $\alpha = (1 - p)e^s$, and for s sufficiently close to zero so that $(1 - p)e^s < 1$. We obtain

$$M(s) = pe^s \left(1 + (1 - p)e^s + (1 - p)^2 e^{2s} + (1 - p)^3 e^{3s} + \dots \right).$$

As in the previous example, we infer that this is a discrete random variable that takes positive integer values. The probability $\mathbf{P}(X = k)$ is found by reading the coefficient of the term e^{ks} . In particular, $\mathbf{P}(X = 1) = p$, $\mathbf{P}(X = 2) = p(1 - p)$, and

$$\mathbf{P}(X = k) = p(1 - p)^{k-1}, \quad k = 1, 2, \dots$$

We recognize this as the geometric distribution with parameter p .

Note that

$$\frac{d}{ds} M(s) = \frac{pe^s}{1 - (1 - p)e^s} + \frac{(1 - p)pe^s}{(1 - (1 - p)e^s)^2}.$$

For $s = 0$, the right-hand side is equal to $1/p$, which agrees with the formula for $\mathbf{E}[X]$ derived in Chapter 2.

Example 4.9. The Transform Associated with a Mixture of Two Distributions. The neighborhood bank has three tellers, two of them fast, one slow. The time to assist a customer is exponentially distributed with parameter $\lambda = 6$ at the fast tellers, and $\lambda = 4$ at the slow teller. Jane enters the bank and chooses a

teller at random, each one with probability 1/3. Find the PDF of the time it takes to assist Jane and the associated transform.

We have

$$f_X(x) = \frac{2}{3} \cdot 6e^{-6x} + \frac{1}{3} \cdot 4e^{-4x}, \quad x \geq 0.$$

Then,

$$\begin{aligned} M(s) &= \int_0^\infty e^{sx} \left(\frac{2}{3}6e^{-6x} + \frac{1}{3}4e^{-4x} \right) dx \\ &= \frac{2}{3} \int_0^\infty e^{sx} 6e^{-6x} dx + \frac{1}{3} \int_0^\infty e^{sx} 4e^{-4x} dx \\ &= \frac{2}{3} \cdot \frac{6}{6-s} + \frac{1}{3} \cdot \frac{4}{4-s} \quad (\text{for } s < 4). \end{aligned}$$

More generally, let X_1, \dots, X_n be continuous random variables with PDFs f_{X_1}, \dots, f_{X_n} . The value y of a random variable Y is generated as follows: an index i is chosen with a corresponding probability p_i , and y is taken to be equal to the value of X_i . Then,

$$f_Y(y) = p_1 f_{X_1}(y) + \dots + p_n f_{X_n}(y),$$

and

$$M_Y(s) = p_1 M_{X_1}(s) + \dots + p_n M_{X_n}(s).$$

The steps in this problem can be reversed. For example, we may be given that the transform associated with a random variable Y is of the form

$$\frac{1}{2} \cdot \frac{1}{2-s} + \frac{3}{4} \cdot \frac{1}{1-s}.$$

We can then rewrite it as

$$\frac{1}{4} \cdot \frac{2}{2-s} + \frac{3}{4} \cdot \frac{1}{1-s},$$

and recognize that Y is the mixture of two exponential random variables with parameters 2 and 1, which are selected with probabilities 1/4 and 3/4, respectively.

Sums of Independent Random Variables

Transform methods are particularly convenient when dealing with a sum of random variables. The reason is that *addition of independent random variables corresponds to multiplication of transforms*, as we now show.

Let X and Y be independent random variables, and let $W = X + Y$. The transform associated with W is, by definition,

$$M_W(s) = \mathbf{E}[e^{sW}] = \mathbf{E}[e^{s(X+Y)}] = \mathbf{E}[e^{sX}e^{sY}].$$

Since X and Y are independent, e^{sX} and e^{sY} are independent random variables, for any fixed value of s . Hence, the expectation of their product is the product of the expectations, and

$$M_W(s) = \mathbf{E}[e^{sX}]\mathbf{E}[e^{sY}] = M_X(s)M_Y(s).$$

By the same argument, if X_1, \dots, X_n is a collection of independent random variables, and

$$W = X_1 + \dots + X_n,$$

then

$$M_W(s) = M_{X_1}(s) \cdots M_{X_n}(s).$$

Example 4.10. The Transform Associated with the Binomial. Let X_1, \dots, X_n be independent Bernoulli random variables with a common parameter p . Then,

$$M_{X_i}(s) = (1-p)e^{0s} + pe^{1s} = 1 - p + pe^s, \quad \text{for all } i.$$

The random variable $Y = X_1 + \dots + X_n$ is binomial with parameters n and p . The corresponding transform is given by

$$M_Y(s) = (1 - p + pe^s)^n.$$

Example 4.11. The Sum of Independent Poisson Random Variables is Poisson. Let X and Y be independent Poisson random variables with means λ and μ , respectively, and let $W = X + Y$. Then,

$$M_X(s) = e^{\lambda(e^s - 1)}, \quad M_Y(s) = e^{\mu(e^s - 1)},$$

and

$$M_W(s) = M_X(s)M_Y(s) = e^{\lambda(e^s - 1)}e^{\mu(e^s - 1)} = e^{(\lambda + \mu)(e^s - 1)}.$$

Thus, the transform associated with W is the same as the transform associated with a Poisson random variable with mean $\lambda + \mu$. By the uniqueness property of transforms, W is Poisson with mean $\lambda + \mu$.

Example 4.12. The Sum of Independent Normal Random Variables is Normal. Let X and Y be independent normal random variables with means μ_x , μ_y , and variances σ_x^2 , σ_y^2 , respectively, and let $W = X + Y$. Then,

$$M_X(s) = e^{\frac{\sigma_x^2 s^2}{2} + \mu_x s}, \quad M_Y(s) = e^{\frac{\sigma_y^2 s^2}{2} + \mu_y s},$$

and

$$M_W(s) = e^{\frac{(\sigma_x^2 + \sigma_y^2)s^2}{2} + (\mu_x + \mu_y)s}.$$

Thus, the transform associated with W is the same as the transform associated with a normal random variable with mean $\mu_x + \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$. By the uniqueness property of transforms, W is normal with these parameters.

Summary of Transforms and their Properties

- The transform associated with a random variable X is given by

$$M_X(s) = \mathbf{E}[e^{sX}] = \begin{cases} \sum_x e^{sx} p_X(x), & X \text{ discrete,} \\ \int_{-\infty}^{\infty} e^{sx} f_X(x) dx, & X \text{ continuous.} \end{cases}$$

- The distribution of a random variable is completely determined by the corresponding transform.
- Moment generating properties:

$$M_X(0) = 1, \quad \left. \frac{d}{ds} M_X(s) \right|_{s=0} = \mathbf{E}[X], \quad \left. \frac{d^n}{ds^n} M_X(s) \right|_{s=0} = \mathbf{E}[X^n].$$

- If $Y = aX + b$, then $M_Y(s) = e^{sb} M_X(as)$.
- If X and Y are independent, then $M_{X+Y}(s) = M_X(s)M_Y(s)$.

We have obtained formulas for the transforms associated with a few common random variables. We can derive such formulas with a moderate amount of algebra for many other distributions (see the end-of-chapter problems for the case of the uniform distribution). We summarize the most useful ones in the tables that follow.

Transforms Associated with Joint Distributions

If two random variables X and Y are described by some joint distribution (e.g., a joint PDF), then each one is associated with a transform $M_X(s)$ or $M_Y(s)$. These are the transforms of the marginal distributions and do not convey information on the dependence between the two random variables. Such information is contained in a multivariate transform, which we now define.

Consider n random variables X_1, \dots, X_n related to the same experiment. Let s_1, \dots, s_n be scalar free parameters. The associated **multivariate transform** is a function of these n parameters and is defined by

$$M_{X_1, \dots, X_n}(s_1, \dots, s_n) = \mathbf{E}[e^{s_1 X_1 + \dots + s_n X_n}].$$

Transforms for Common Discrete Random Variables
Bernoulli(p) ($k = 0, 1$)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1 - p, & \text{if } k = 0, \end{cases} \quad M_X(s) = 1 - p + pe^s.$$

Binomial(n, p) ($k = 0, 1, \dots, n$)

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad M_X(s) = (1 - p + pe^s)^n.$$

Geometric(p) ($k = 1, 2, \dots$)

$$p_X(k) = p(1 - p)^{k-1}, \quad M_X(s) = \frac{pe^s}{1 - (1 - p)e^s}.$$

Poisson(λ) ($k = 0, 1, \dots$)

$$p_X(k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad M_X(s) = e^{\lambda(e^s - 1)}.$$

Uniform(a, b) ($k = a, a + 1, \dots, b$)

$$p_X(k) = \frac{1}{b - a + 1}, \quad M_X(s) = \frac{e^{as}}{b - a + 1} \cdot \frac{e^{(b-a+1)s} - 1}{e^s - 1}.$$

Transforms for Common Continuous Random Variables
Uniform(a, b) ($a \leq x \leq b$)

$$f_X(x) = \frac{1}{b - a}, \quad M_X(s) = \frac{1}{b - a} \cdot \frac{e^{sb} - e^{sa}}{s}.$$

Exponential(λ) ($x \geq 0$)

$$f_X(x) = \lambda e^{-\lambda x}, \quad M_X(s) = \frac{\lambda}{\lambda - s}, \quad (s < \lambda).$$

Normal(μ, σ^2) ($-\infty < x < \infty$)

$$f_X(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad M_X(s) = e^{(\sigma^2 s^2/2) + \mu s}.$$

The inversion property of transforms discussed earlier extends to the multivariate case. That is, if Y_1, \dots, Y_n is another set of random variables and

$M_{X_1, \dots, X_n}(s_1, \dots, s_n)$, $M_{Y_1, \dots, Y_n}(s_1, \dots, s_n)$ are the same functions of s_1, \dots, s_n , then the joint distribution of X_1, \dots, X_n is the same as the joint distribution of Y_1, \dots, Y_n .

4.2 SUMS OF INDEPENDENT RANDOM VARIABLES — CONVOLUTION

If X and Y are independent random variables, the distribution of their sum $W = X + Y$ can be obtained by computing and then inverting the transform $M_W(s) = M_X(s)M_Y(s)$. But it can also be obtained directly, using the method developed in this section.

The Discrete Case

Let $W = X + Y$, where X and Y are independent integer-valued random variables with PMFs p_X and p_Y , respectively. Then, for any integer w ,

$$\begin{aligned} p_W(w) &= \mathbf{P}(X + Y = w) \\ &= \sum_{\{(x,y) \mid x+y=w\}} \mathbf{P}(X = x, Y = y) \\ &= \sum_x \mathbf{P}(X = x, Y = w - x) \\ &= \sum_x p_X(x)p_Y(w - x). \end{aligned}$$

The resulting PMF p_W is called the **convolution** of the PMFs of X and Y . See Fig. 4.1 for an illustration.

Example 4.13. Let X and Y be independent random variables with PMFs

$$p_X(x) = \begin{cases} 1/3, & \text{if } x = 1, 2, 3, \\ 0, & \text{otherwise,} \end{cases} \quad p_Y(y) = \begin{cases} 1/2, & \text{if } y = 0, \\ 1/3, & \text{if } y = 1, \\ 1/6, & \text{if } y = 2, \\ 0, & \text{otherwise.} \end{cases}$$

To calculate the PMF of $W = X + Y$ by convolution, we first note that the range of possible values of W are the integers from the range $[1, 5]$. Thus we have

$$p_W(w) = 0, \quad \text{if } w \neq 1, 2, 3, 4, 5.$$

We calculate $p_W(w)$ for each of the values $w = 1, 2, 3, 4, 5$, using the convolution formula. We have

$$p_W(1) = \sum_x p_X(x)p_Y(1 - x) = p_X(1)p_Y(0) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6},$$

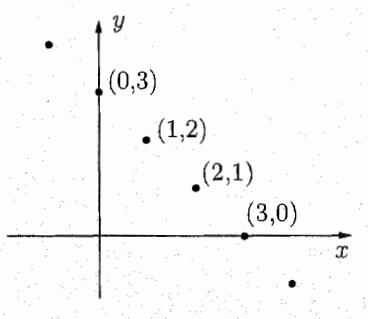


Figure 4.1: The probability $p_W(3)$ that $X+Y = 3$ is the sum of the probabilities of all pairs (x, y) such that $x+y = 3$, which are the points indicated in the figure. The probability of a generic such point is of the form

$$p_{X,Y}(x, 3-x) = p_X(x)p_Y(3-x).$$

where the second equality above is based on the fact that for $x \neq 1$ either $p_X(x)$ or $p_Y(1-x)$ (or both) is zero. Similarly, we obtain

$$p_W(2) = p_X(1)p_Y(1) + p_X(2)p_Y(0) = \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{18},$$

$$p_W(3) = p_X(1)p_Y(2) + p_X(2)p_Y(1) + p_X(3)p_Y(0) = \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{3},$$

$$p_W(4) = p_X(2)p_Y(2) + p_X(3)p_Y(1) = \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{6},$$

$$p_W(5) = p_X(3)p_Y(2) = \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{18}.$$

The Continuous Case

Let X and Y be independent continuous random variables with PDFs f_X and f_Y , respectively. We wish to find the PDF of $W = X + Y$. Towards this goal, we will first find the joint PDF of X and W , and then integrate to find the PDF of W .

We first note that

$$\begin{aligned} \mathbf{P}(W \leq w \mid X = x) &= \mathbf{P}(X + Y \leq w \mid X = x) \\ &= \mathbf{P}(x + Y \leq w) \\ &= \mathbf{P}(Y \leq w - x), \end{aligned}$$

where the second equality follows from the independence of X and Y . By differentiating both sides with respect to w , we see that $f_{W|X}(w|x) = f_Y(w-x)$. Using the multiplication rule, we have

$$f_{X,W}(x,w) = f_X(x)f_{W|X}(w|x) = f_X(x)f_Y(w-x),$$

from which we finally obtain

$$f_W(w) = \int_{-\infty}^{\infty} f_{X,W}(x,w) dx = \int_{-\infty}^{\infty} f_X(x)f_Y(w-x) dx.$$

This formula is entirely analogous to the formula for the discrete case, except that the summation is replaced by an integral and the PMFs are replaced by PDFs. For an intuitive understanding of this formula, see Fig. 4.2.

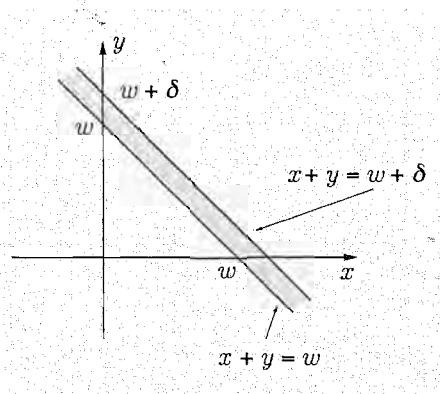


Figure 4.2: Illustration of the convolution formula for the case of continuous random variables (compare with Fig. 4.1). For small δ , the probability of the strip indicated in the figure is $P(w \leq X + Y \leq w + \delta) \approx f_W(w)\delta$. Thus,

$$\begin{aligned} f_W(w)\delta &= P(w \leq X + Y \leq w + \delta) \\ &= \int_{-\infty}^{\infty} \int_{w-x}^{w-x+\delta} f_X(x)f_Y(y) dy dx \\ &\approx \int_{-\infty}^{\infty} f_X(x)f_Y(w-x)\delta dx. \end{aligned}$$

The desired formula follows by canceling δ from both sides.

Example 4.14. The random variables X and Y are independent and uniformly distributed in the interval $[0, 1]$. The PDF of $W = X + Y$ is

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x)f_Y(w-x) dx.$$

The integrand $f_X(x)f_Y(w-x)$ is nonzero (and equal to 1) for $0 \leq x \leq 1$ and $0 \leq w-x \leq 1$. Combining these two inequalities, the integrand is nonzero for $\max\{0, w-1\} \leq x \leq \min\{1, w\}$. Thus,

$$f_W(w) = \begin{cases} \min\{1, w\} - \max\{0, w-1\}, & 0 \leq w \leq 2, \\ 0, & \text{otherwise,} \end{cases}$$

which has the triangular shape shown in Fig. 4.3.

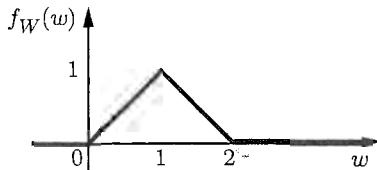


Figure 4.3: The PDF of the sum of two independent uniform random variables in $[0, 1]$.

The calculation in the last example was based on a literal application of the convolution formula. The most delicate step was to determine the correct limits for the integration. This is often tedious and error prone, but can be bypassed using a graphical method described next.

Graphical Calculation of Convolutions

We use a dummy variable t as the argument of the different functions involved in this discussion; see also Fig. 4.4. Consider two PDFs $f_X(t)$ and $f_Y(t)$. For a fixed value of w , the graphical evaluation of the convolution

$$f_W(w) = \int_{-\infty}^{\infty} f_X(t)f_Y(w-t) dt$$

consists of the following steps:

- We plot $f_Y(w-t)$ as a function of t . This plot has the same shape as the plot of $f_Y(t)$ except that it is first “flipped” and then shifted by an amount w . If $w > 0$, this is a shift to the right, if $w < 0$, this is a shift to the left.
- We place the plots of $f_X(t)$ and $f_Y(w-t)$ on top of each other, and form their product.
- We calculate the value of $f_W(w)$ by calculating the integral of the product of these two plots.

By varying the amount w by which we are shifting, we obtain $f_W(w)$ for any w .

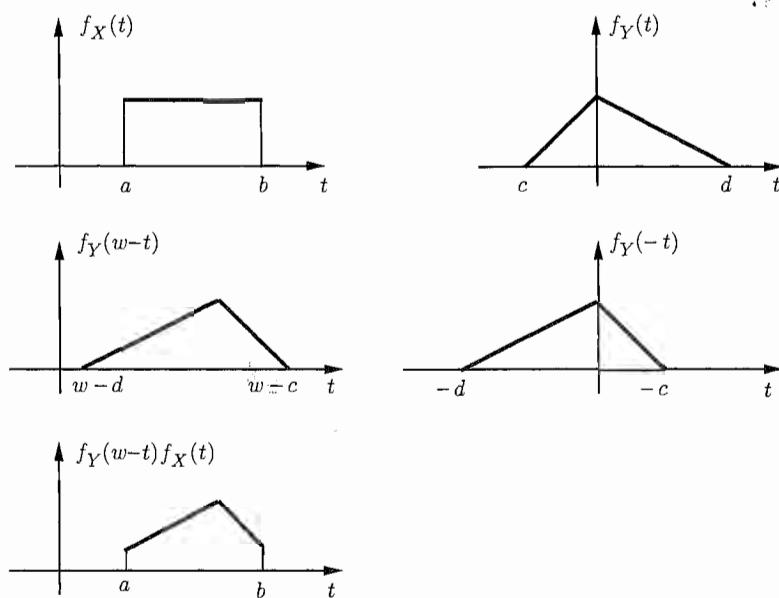


Figure 4.4: Illustration of the convolution calculation. For the value of w under consideration, $f_W(w)$ is equal to the integral of the function shown in the last plot.

4.3 MORE ON CONDITIONAL EXPECTATION AND VARIANCE

The value of the conditional expectation $\mathbf{E}[X | Y = y]$ of a random variable X , given another random variable Y , depends on the value y of Y . This makes $\mathbf{E}[X | Y]$ a function of Y , and therefore a random variable. In this section, we study the expectation and variance of $\mathbf{E}[X | Y]$. In the process, we obtain some useful formulas (the **law of iterated expectations** and the **law of total variance**) that are often convenient for the calculation of expected values and variances.

Recall that the conditional expectation $\mathbf{E}[X | Y = y]$ is defined by

$$\mathbf{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y), \quad X \text{ discrete},$$

and

$$\mathbf{E}[X | Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx, \quad X \text{ continuous}.$$

Once a value of y is given, the above summation or integration yields a numerical value for $\mathbf{E}[X | Y = y]$.

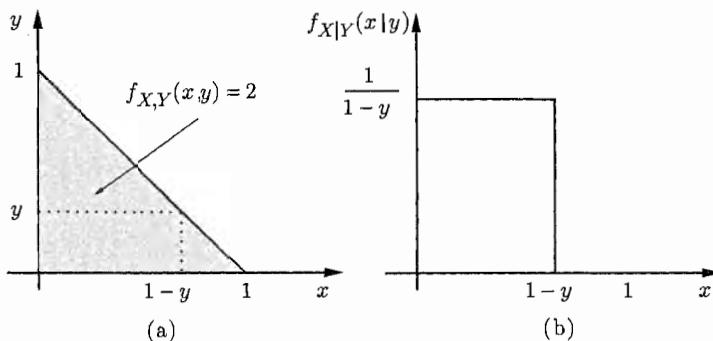


Figure 4.5: (a) The joint PDF in Example 4.15. (b) The conditional density of X .

Example 4.15. Let the random variables X and Y have a joint PDF which is equal to 2 for (x, y) belonging to the triangle indicated in Fig. 4.5(a), and zero everywhere else. In order to compute $\mathbf{E}[X | Y = y]$, we first obtain the conditional density of X given $Y = y$.

We have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx = \int_0^{1-y} 2 dx = 2(1-y), \quad 0 \leq y \leq 1,$$

and

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{1-y}, \quad 0 \leq x \leq 1-y.$$

The conditional density is shown in Fig. 4.5(b).

Intuitively, since the joint PDF is constant, the conditional PDF (which is a “slice” of the joint, at some fixed y) is also constant. Therefore, the conditional PDF must be a uniform distribution. Given that $Y = y$, X ranges from 0 to $1 - y$. Therefore, for the PDF to integrate to 1, its height must be equal to $1/(1 - y)$, in agreement with Fig. 4.5(b).

For $y > 1$ or $y < 0$, the conditional PDF is undefined, since these values of y are impossible. For $0 \leq y < 1$, the conditional mean $\mathbf{E}[X | Y = y]$ is obtained using the uniform PDF in Fig. 4.5(b), and we have

$$\mathbf{E}[X | Y = y] = \frac{1-y}{2}, \quad 0 \leq y < 1.$$

For $y = 1$, X must be equal to 0, with certainty, so $\mathbf{E}[X | Y = 1] = 0$. Thus, the above formula is also valid when $y = 1$. The conditional expectation is undefined when y is outside $[0, 1]$.

For any number y , $\mathbf{E}[X | Y = y]$ is also a number. As y varies, so does $\mathbf{E}[X | Y = y]$, and we can therefore view $\mathbf{E}[X | Y = y]$ as a function of y . Since y

is the value of the random variable Y , we are dealing with a function of a random variable, hence a new random variable. More precisely, we **define** $\mathbf{E}[X|Y]$ to be the random variable whose value is $\mathbf{E}[X|Y=y]$ when the value of Y is y .

Example 4.15 (continued). We saw that $\mathbf{E}[X|Y=y] = (1-y)/2$. Hence, $\mathbf{E}[X|Y]$ is the random variable $(1-Y)/2$:

$$\mathbf{E}[X|Y] = \frac{1-Y}{2}.$$

Since $\mathbf{E}[X|Y]$ is a random variable, it has an expectation $\mathbf{E}[\mathbf{E}[X|Y]]$ of its own, which can be calculated using the expected value rule:

$$\mathbf{E}[\mathbf{E}[X|Y]] = \begin{cases} \sum_y \mathbf{E}[X|Y=y] p_Y(y), & Y \text{ discrete,} \\ \int_{-\infty}^{\infty} \mathbf{E}[X|Y=y] f_Y(y) dy, & Y \text{ continuous.} \end{cases}$$

Both expressions in the right-hand side are familiar from Chapters 2 and 3, respectively. By the corresponding versions of the total expectation theorem, they are equal to $\mathbf{E}[X]$. This brings us to the following conclusion, which is actually valid for every type of random variable Y (discrete, continuous, or mixed), as long as X has a well-defined and finite expectation $\mathbf{E}[X]$.

Law of Iterated Expectations: $\mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}[X]$.

Example 4.15 (continued). In Example 4.15, we found $\mathbf{E}[X|Y] = (1-Y)/2$. Taking expectations of both sides, and using the law of iterated expectations to evaluate the left-hand side, we obtain $\mathbf{E}[X] = (1 - \mathbf{E}[Y])/2$. Because of symmetry, we must have $\mathbf{E}[X] = \mathbf{E}[Y]$. Therefore, $\mathbf{E}[X] = (1 - \mathbf{E}[X])/2$, which yields $\mathbf{E}[X] = 1/3$.

Example 4.16. We start with a stick of length ℓ . We break it at a point which is chosen randomly and uniformly over its length, and keep the piece that contains the left end of the stick. We then repeat the same process on the piece that we were left with. What is the expected length of the piece that we are left with after breaking twice?

Let Y be the length of the piece after we break for the first time. Let X be the length after we break for the second time. We have $\mathbf{E}[X|Y] = Y/2$, since the breakpoint is chosen uniformly over a piece of length Y . For a similar reason, we also have $\mathbf{E}[Y] = \ell/2$. Thus,

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}\left[\frac{Y}{2}\right] = \frac{\mathbf{E}[Y]}{2} = \frac{\ell}{4}.$$

Example 4.17. Averaging Quiz Scores by Section. A class has n students and the quiz score of student i is x_i . The average quiz score is

$$m = \frac{1}{n} \sum_{i=1}^n x_i.$$

The students are divided into k disjoint subsets A_1, \dots, A_k , and are accordingly assigned to different sections. We use n_s to denote the number of students in section s . The average score in section s is

$$m_s = \frac{1}{n_s} \sum_{i \in A_s} x_i.$$

The average score over the whole class can be computed by taking the average score m_s of each section, and then forming a *weighted average*; the weight given to section s is proportional to the number of students in that section, and is n_s/n . We verify that this gives the correct result:

$$\begin{aligned} \sum_{s=1}^k \frac{n_s}{n} m_s &= \sum_{s=1}^k \frac{n_s}{n} \cdot \frac{1}{n_s} \sum_{i \in A_s} x_i \\ &= \frac{1}{n} \sum_{s=1}^k \sum_{i \in A_s} x_i \\ &= \frac{1}{n} \sum_{i=1}^n x_i. \\ &= m. \end{aligned}$$

How is this related to conditional expectations? Consider an experiment in which a student is selected at random, each student having probability $1/n$ of being selected. Consider the following two random variables:

X = quiz score of a student,

Y = section of a student, ($Y \in \{1, \dots, k\}$).

We then have

$$\mathbf{E}[X] = m.$$

Conditioning on $Y = s$ is the same as assuming that the selected student is in section s . Conditional on that event, every student in that section has the same probability $1/n_s$ of being chosen. Therefore,

$$\mathbf{E}[X | Y = s] = \frac{1}{n_s} \sum_{i \in A_s} x_i = m_s.$$

A randomly selected student belongs to section s with probability n_s/n , i.e., $\mathbf{P}(Y = s) = n_s/n$. Hence,

$$m = \mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X | Y]] = \sum_{s=1}^k \mathbf{E}[X | Y = s] \mathbf{P}(Y = s) = \sum_{s=1}^k \frac{n_s}{n} m_s.$$

Thus, averaging by section can be viewed as a special case of the law of iterated expectations.

Example 4.18. Forecast Revisions. Let Y be the sales of a company in the first semester of the coming year, and let X be the sales over the entire year. The company has constructed a statistical model of sales, and so the joint distribution of X and Y is assumed to be known. In the beginning of the year, the expected value $\mathbf{E}[X]$ serves as a forecast of the actual sales X . In the middle of the year, the first semester sales have been realized and the value of the random variable Y is now known. This places us in a new “universe,” where everything is conditioned on the realized value of Y . Based on the knowledge of Y , the company constructs a revised forecast of yearly sales, which is $\mathbf{E}[X | Y]$.

We view $\mathbf{E}[X | Y] - \mathbf{E}[X]$ as the forecast revision, in light of the mid-year information. The law of iterated expectations implies that

$$\mathbf{E}[\mathbf{E}[X | Y] - \mathbf{E}[X]] = \mathbf{E}[\mathbf{E}[X | Y]] - \mathbf{E}[X] = \mathbf{E}[X] - \mathbf{E}[X] = 0.$$

This means that in the beginning of the year, we do not expect our forecast to be revised in any specific direction. Of course, the actual revision will usually be nonzero, but the probabilities are such that it is zero on the average. This is quite intuitive. Indeed, if a positive revision was expected, the original forecast should have been higher in the first place.

The Conditional Variance

The conditional variance of X , given that $Y = y$, is defined by the same formula as the unconditional variance, except that everything is conditioned on $Y = y$:

$$\text{var}(X | Y = y) = \mathbf{E}\left[\left(X - \mathbf{E}[X | Y = y]\right)^2 \mid Y = y\right].$$

Note that the conditional variance is a function of the value y of the random variable Y . Hence, it is a function of a random variable, and is itself a random variable that will be denoted by $\text{var}(X | Y)$.

Arguing by analogy to the law of iterated expectations, we may conjecture that the expectation of the conditional variance $\text{var}(X | Y)$ is related to the unconditional variance $\text{var}(X)$. This is indeed the case, but the relation is more complex.

Law of Total Variance: $\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y]).$

To verify the law of total variance, we start with the formula

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2,$$

and use the law of iterated expectations for each of the terms $\mathbf{E}[X^2]$ and $\mathbf{E}[X]$. We have

$$\begin{aligned}\text{var}(X) &= \mathbf{E}[\mathbf{E}[X^2 | Y]] - (\mathbf{E}[\mathbf{E}[X | Y]])^2 \\ &= \mathbf{E}[\text{var}(X | Y) + (\mathbf{E}[X | Y])^2] - (\mathbf{E}[\mathbf{E}[X | Y]])^2 \\ &= \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y]).\end{aligned}$$

The last step is justified by defining $Z = \mathbf{E}[X | Y]$ and observing that

$$\mathbf{E}[(\mathbf{E}[X | Y])^2] - (\mathbf{E}[\mathbf{E}[X | Y]])^2 = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2 = \text{var}(Z) = \text{var}(\mathbf{E}[X | Y]).$$

Example 4.16 (continued). Consider again the problem where we break twice a stick of length ℓ , at randomly chosen points, with Y being the length of the piece left after the first break and X being the length after the second break. We calculated the mean of X as $\ell/4$. We will now use the law of total variance to calculate $\text{var}(X)$.

Since X is uniformly distributed between 0 and Y , we have

$$\text{var}(X | Y) = \frac{Y^2}{12}.$$

Thus, since Y is uniformly distributed between 0 and ℓ , we have

$$\mathbf{E}[\text{var}(X | Y)] = \frac{1}{12} \int_0^\ell \frac{1}{\ell} y^2 dy = \frac{1}{12} \cdot \frac{1}{3\ell} y^3 \Big|_0^\ell = \frac{\ell^2}{36}.$$

We also have $\mathbf{E}[X | Y] = Y/2$, so

$$\text{var}(\mathbf{E}[X | Y]) = \text{var}(Y/2) = \frac{1}{4} \text{var}(Y) = \frac{1}{4} \cdot \frac{\ell^2}{12} = \frac{\ell^2}{48}.$$

Using now the law of total variance, we obtain

$$\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y]) = \frac{\ell^2}{36} + \frac{\ell^2}{48} = \frac{7\ell^2}{144}.$$

Example 4.19. Averaging Quiz Scores by Section – Variance. The setting is the same as in Example 4.17 and we consider again the random variables

X = quiz score of a student,

Y = section of a student, $(Y \in \{1, \dots, k\})$.

Let n_s be the number of students in section s , and let n be the total number of students. We interpret the different quantities in the formula

$$\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y]).$$

In this context, $\text{var}(X | Y = s)$ is the variance of the quiz scores within section s . Thus,

$$\mathbf{E}[\text{var}(X | Y)] = \sum_{s=1}^k \mathbf{P}(Y = s) \text{var}(X | Y = s) = \sum_{s=1}^k \frac{n_s}{n} \text{var}(X | Y = s),$$

so that $\mathbf{E}[\text{var}(X | Y)]$ is the weighted average of the section variances, where each section is weighted in proportion to its size.

Recall that $\mathbf{E}[X | Y = s]$ is the average score in section s . Then, $\text{var}(\mathbf{E}[X | Y])$ is a measure of the variability of the averages of the different sections. The law of total variance states that the total quiz score variance can be broken into two parts:

- (a) The average score variability $\mathbf{E}[\text{var}(X | Y)]$ *within* individual sections.
- (b) The variability $\text{var}(\mathbf{E}[X | Y])$ *between* sections.

We have seen earlier that the law of iterated expectations (in the form of the total expectation theorem) can be used to break down complicated expectation calculations, by considering different cases. A similar method applies to variance calculations.

Example 4.20. Computing Variances by Conditioning. Consider a continuous random variable X with the PDF given in Fig. 4.6. We define an auxiliary random variable Y as follows:

$$Y = \begin{cases} 1, & \text{if } x < 1, \\ 2, & \text{if } x \geq 1. \end{cases}$$

Here, $\mathbf{E}[X | Y]$ takes the values $1/2$ and 2 , each with probability $1/2$. Thus, the mean of $\mathbf{E}[X | Y]$ is $5/4$. It follows that

$$\text{var}(\mathbf{E}[X | Y]) = \frac{1}{2} \left(\frac{1}{2} - \frac{5}{4} \right)^2 + \frac{1}{2} \left(2 - \frac{5}{4} \right)^2 = \frac{9}{16}.$$

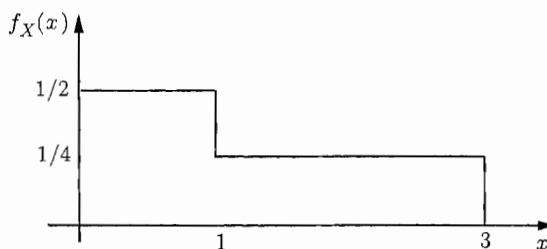


Figure 4.6: The PDF in Example 4.20.

Conditioned on $Y = 1$ or $Y = 2$, X is uniformly distributed on an interval of length 1 or 2, respectively. Therefore,

$$\text{var}(X | Y = 1) = \frac{1}{12}, \quad \text{var}(X | Y = 2) = \frac{4}{12},$$

and

$$\mathbf{E}[\text{var}(X | Y)] = \frac{1}{2} \cdot \frac{1}{12} + \frac{1}{2} \cdot \frac{4}{12} = \frac{5}{24}.$$

Putting everything together, we obtain

$$\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y]) = \frac{5}{24} + \frac{9}{16} = \frac{37}{48}.$$

We summarize the main points in this section.

Properties of the Conditional Expectation and Variance

- $\mathbf{E}[X | Y = y]$ is a number whose value depends on y .
- $\mathbf{E}[X | Y]$ is a function of the random variable Y , hence a random variable. Its value is $\mathbf{E}[X | Y = y]$ whenever the value of Y is y .
- $\mathbf{E}[\mathbf{E}[X | Y]] = \mathbf{E}[X]$ (law of iterated expectations).
- $\text{var}(X | Y)$ is a random variable whose value is $\text{var}(X | Y = y)$ whenever the value of Y is y .
- $\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y])$ (law of total variance).

4.4 SUM OF A RANDOM NUMBER OF INDEPENDENT RANDOM VARIABLES

In our discussion so far of sums of random variables, we have always assumed that the number of variables in the sum is known and fixed. In this section, we will consider the case where the number of random variables being added is itself random. In particular, we consider the sum

$$Y = X_1 + \cdots + X_N,$$

where N is a random variable that takes positive integer values, and X_1, X_2, \dots are identically distributed random variables. We assume that N, X_1, X_2, \dots are independent, meaning that any finite subcollection of these random variables are independent.

Let us denote by μ and σ^2 the common mean and variance, respectively, of the X_i . We wish to derive formulas for the mean, variance, and the transform of Y . The method that we follow is to first condition on the event $N = n$, which brings us to the more familiar case of a *fixed* number of random variables.

Fix some number n . The random variable $X_1 + \dots + X_n$ is independent of N and, therefore, independent of the event $\{N = n\}$. Hence,

$$\begin{aligned}\mathbf{E}[Y | N = n] &= \mathbf{E}[X_1 + \dots + X_N | N = n] \\ &= \mathbf{E}[X_1 + \dots + X_n | N = n] \\ &= \mathbf{E}[X_1 + \dots + X_n] \\ &= n\mu.\end{aligned}$$

This is true for every positive integer n and, therefore,

$$\mathbf{E}[Y | N] = N\mu.$$

Using the law of iterated expectations, we obtain

$$\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y | N]] = \mathbf{E}[\mu N] = \mu \mathbf{E}[N].$$

Similarly,

$$\begin{aligned}\text{var}(Y | N = n) &= \text{var}(X_1 + \dots + X_N | N = n) \\ &= \text{var}(X_1 + \dots + X_n) \\ &= n\sigma^2.\end{aligned}$$

Since this is true for every positive integer n , the random variable $\text{var}(Y | N)$ is equal to $N\sigma^2$. We now use the law of total variance to obtain

$$\begin{aligned}\text{var}(Y) &= \mathbf{E}[\text{var}(Y | N)] + \text{var}(\mathbf{E}[Y | N]) \\ &= \mathbf{E}[N\sigma^2] + \text{var}(N\mu) \\ &= \mathbf{E}[N]\sigma^2 + \mu^2\text{var}(N).\end{aligned}$$

The calculation of the transform proceeds along similar lines. The transform associated with Y , conditional on $N = n$, is $\mathbf{E}[e^{sY} | N = n]$. However, conditioned on $N = n$, Y is the sum of the independent random variables X_1, \dots, X_n , and

$$\begin{aligned}\mathbf{E}[e^{sY} | N = n] &= \mathbf{E}[e^{sX_1} \dots e^{sX_N} | N = n] \\ &= \mathbf{E}[e^{sX_1} \dots e^{sX_n}] \\ &= \mathbf{E}[e^{sX_1}] \dots \mathbf{E}[e^{sX_n}] \\ &= (M_X(s))^n,\end{aligned}$$

where $M_X(s)$ is the transform associated with X_i , for each i . Using the law of iterated expectations, the (unconditional) transform associated with Y is

$$M_Y(s) = \mathbf{E}[e^{sY}] = \mathbf{E}[\mathbf{E}[e^{sY} | N]] = \mathbf{E}[(M_X(s))^N] = \sum_{n=1}^{\infty} (M_X(s))^n p_N(n).$$

Comparing with the formula

$$M_N(s) = \mathbf{E}[e^{sN}] = \sum_{n=1}^{\infty} (e^s)^n p_N(n),$$

we see that $M_Y(s)$ is essentially the same as $M_N(s)$, except that e^s is replaced by $M_X(s)$.

Let us summarize the properties derived so far.

Properties of the Sum of a Random Number of Independent Random Variables

Let X_1, X_2, \dots be identically distributed random variables with mean μ and variance σ^2 . Let N be a random variable that takes positive integer values. We assume that all of these random variables are independent, and we consider the sum

$$Y = X_1 + \dots + X_N.$$

Then:

- $\mathbf{E}[Y] = \mu \mathbf{E}[N]$.
- $\text{var}(Y) = \sigma^2 \mathbf{E}[N] + \mu^2 \text{var}(N)$.
- The transform $M_Y(s)$ is found by starting with the transform $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$.

Example 4.21. A remote village has three gas stations. Each gas station is open on any given day with probability $1/2$, independently of the others. The amount of gas available in each gas station is unknown and is uniformly distributed between 0 and 1000 gallons. We wish to characterize the probability law of the total amount of gas available at the gas stations that are open.

The number N of open gas stations is a binomial random variable with $p = 1/2$ and the corresponding transform is

$$M_N(s) = (1 - p + pe^s)^3 = \frac{1}{8}(1 + e^s)^3.$$

The transform $M_X(s)$ associated with the amount of gas available in an open gas station is

$$M_X(s) = \frac{e^{1000s} - 1}{1000s}.$$

The transform associated with the total amount Y available is the same as $M_N(s)$, except that each occurrence of e^s is replaced with $M_X(s)$, i.e.,

$$M_Y(s) = \frac{1}{8} \left(1 + \left(\frac{e^{1000s} - 1}{1000s} \right) \right)^3.$$

Example 4.22. Sum of a Geometric Number of Independent Exponential Random Variables. Jane visits a number of bookstores, looking for *Great Expectations*. Any given bookstore carries the book with probability p , independently of the others. In a typical bookstore visited, Jane spends a random amount of time, exponentially distributed with parameter λ , until she either finds the book or she determines that the bookstore does not carry it. Assuming that Jane will keep visiting bookstores until she buys the book and that the time spent in each is independent of everything else, we wish to find the mean, variance, and PDF of the total time spent in bookstores.

The total number N of bookstores visited is geometrically distributed with parameter p . Hence, the total time Y spent in bookstores is the sum of a geometrically distributed number N of independent exponential random variables X_1, X_2, \dots . We have

$$\mathbf{E}[Y] = \mathbf{E}[N]\mathbf{E}[X] = \frac{1}{p} \cdot \frac{1}{\lambda}.$$

Using the formulas for the variance of geometric and exponential random variables, we also obtain

$$\text{var}(Y) = \mathbf{E}[N]\text{var}(X) + (\mathbf{E}[X])^2\text{var}(N) = \frac{1}{p} \cdot \frac{1}{\lambda^2} + \frac{1}{\lambda^2} \cdot \frac{1-p}{p^2} = \frac{1}{\lambda^2 p^2}.$$

In order to find the transform $M_Y(s)$, let us recall that

$$M_X(s) = \frac{\lambda}{\lambda - s}, \quad M_N(s) = \frac{pe^s}{1 - (1-p)e^s}.$$

Then, $M_Y(s)$ is found by starting with $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$. This yields

$$M_Y(s) = \frac{pM_X(s)}{1 - (1-p)M_X(s)} = \frac{\frac{p\lambda}{\lambda - s}}{1 - (1-p)\frac{\lambda}{\lambda - s}},$$

which simplifies to

$$M_Y(s) = \frac{p\lambda}{p\lambda - s}.$$

We recognize this as the transform associated with an exponentially distributed random variable with parameter $p\lambda$, and therefore,

$$f_Y(y) = p\lambda e^{-p\lambda y}, \quad y \geq 0.$$

This result can be surprising because the sum of a *fixed* number n of independent exponential random variables is not exponentially distributed. For example, if $n = 2$, the transform associated with the sum is $(\lambda/(\lambda - s))^2$, which does not correspond to the exponential distribution.

Example 4.23. Sum of a Geometric Number of Independent Geometric Random Variables. This example is a discrete counterpart of the preceding one.

We let N be geometrically distributed with parameter p . We also let each random variable X_i be geometrically distributed with parameter q . We assume that all of these random variables are independent. Let $Y = X_1 + \dots + X_N$. We have

$$M_N(s) = \frac{pe^s}{1 - (1-p)e^s}, \quad M_X(s) = \frac{qe^s}{1 - (1-q)e^s}.$$

To determine $M_Y(s)$, we start with the formula for $M_N(s)$ and replace each occurrence of e^s with $M_X(s)$. This yields

$$M_Y(s) = \frac{pM_X(s)}{1 - (1-p)M_X(s)},$$

and, after some algebra,

$$M_Y(s) = \frac{pqe^s}{1 - (1-pq)e^s}.$$

We conclude that Y is geometrically distributed, with parameter pq .

4.5 COVARIANCE AND CORRELATION

The **covariance** of two random variables X and Y is denoted by $\text{cov}(X, Y)$, and is defined by

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])].$$

When $\text{cov}(X, Y) = 0$, we say that X and Y are **uncorrelated**.

Roughly speaking, a positive or negative covariance indicates that the values of $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ obtained in a single experiment “tend” to have the same or the opposite sign, respectively (see Fig. 4.7). Thus the sign of the covariance provides an important qualitative indicator of the relation between X and Y .

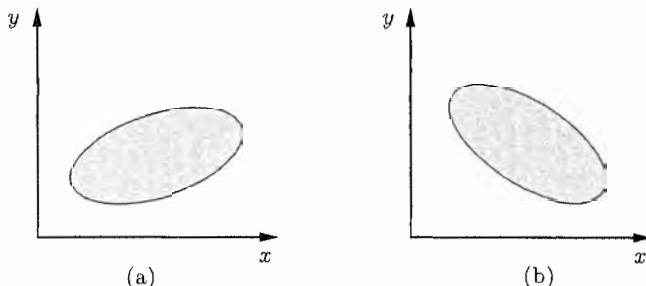


Figure 4.7: Examples of positively and negatively correlated random variables. Here, X and Y are uniformly distributed over the ellipses shown in the figure. In case (a) the covariance $\text{cov}(X, Y)$ is positive, while in case (b) it is negative.

An alternative formula for the covariance is

$$\text{cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y],$$

as can be verified by a simple calculation. Note that if X and Y are independent, we have $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$, which implies that $\text{cov}(X, Y) = 0$. Thus, if X and Y are independent, they are also uncorrelated. However, the reverse is not true, as illustrated by the following example.

Example 4.24. The pair of random variables (X, Y) takes the values $(1, 0)$, $(0, 1)$, $(-1, 0)$, and $(0, -1)$, each with probability $1/4$ (see Fig. 4.8). Thus, the marginal PMFs of X and Y are symmetric around 0, and $\mathbf{E}[X] = \mathbf{E}[Y] = 0$. Furthermore, for all possible value pairs (x, y) , either x or y is equal to 0, which implies that $XY = 0$ and $\mathbf{E}[XY] = 0$. Therefore,

$$\text{cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = 0,$$

and X and Y are uncorrelated. However, X and Y are not independent since, for example, a nonzero value of X fixes the value of Y to zero.

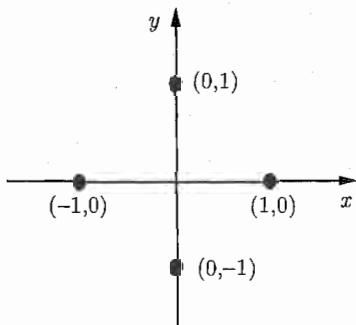


Figure 4.8: Joint PMF of X and Y for Example 4.24. Each of the four points shown has probability $1/4$. Here X and Y are uncorrelated but not independent.

The **correlation coefficient** $\rho(X, Y)$ of two random variables X and Y that have nonzero variances is defined as

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.$$

(The simpler notation ρ will also be used when X and Y are clear from the context.) It may be viewed as a normalized version of the covariance $\text{cov}(X, Y)$, and in fact, it can be shown that ρ ranges from -1 to 1 (see the end-of-chapter problems).

If $\rho > 0$ (or $\rho < 0$), then the values of $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ "tend" to have the same (or opposite, respectively) sign, and the size of $|\rho|$ provides a

normalized measure of the extent to which this is true. In fact, always assuming that X and Y have positive variances, it can be shown that $\rho = 1$ (or $\rho = -1$) if and only if there exists a positive (or negative, respectively) constant c such that

$$Y - \mathbf{E}[Y] = c(X - \mathbf{E}[X])$$

(see the end-of-chapter problems). The following example illustrates in part this property.

Example 4.25. Consider n independent tosses of a coin with probability of a head equal to p . Let X and Y be the numbers of heads and of tails, respectively, and let us look at the correlation coefficient of X and Y . Here, we have $X + Y = n$, and also $\mathbf{E}[X] + \mathbf{E}[Y] = n$. Thus,

$$X - \mathbf{E}[X] = - (Y - \mathbf{E}[Y]).$$

We will calculate the correlation coefficient of X and Y , and verify that it is indeed equal to -1 .

We have

$$\begin{aligned}\text{cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \\ &= -\mathbf{E}[(X - \mathbf{E}[X])^2] \\ &= -\text{var}(X).\end{aligned}$$

Hence, the correlation coefficient is

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{-\text{var}(X)}{\sqrt{\text{var}(X)\text{var}(X)}} = -1.$$

Covariance and Correlation

- The **covariance** of X and Y is given by

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

- If $\text{cov}(X, Y) = 0$, we say that X and Y are **uncorrelated**.
- If X and Y are independent, they are uncorrelated. The converse is not always true.

- We have

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y).$$

- The **correlation coefficient** $\rho(X, Y)$ of two random variables X and Y with positive variances is defined by

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}},$$

and satisfies

$$-1 \leq \rho(X, Y) \leq 1.$$

Variance of the Sum of Random Variables

The covariance can be used to obtain a formula for the variance of the sum of several (not necessarily independent) random variables. In particular, if X_1, X_2, \dots, X_n are random variables with finite variance, we have

$$\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2) + 2\text{cov}(X_1, X_2),$$

and, more generally,

$$\text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(X_i, X_j).$$

This can be seen from the following calculation, where for brevity, we denote $\tilde{X}_i = X_i - \mathbf{E}[X_i]$:

$$\begin{aligned} \text{var}\left(\sum_{i=1}^n X_i\right) &= \mathbf{E}\left[\left(\sum_{i=1}^n \tilde{X}_i\right)^2\right] \\ &= \mathbf{E}\left[\sum_{i=1}^n \sum_{j=1}^n \tilde{X}_i \tilde{X}_j\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbf{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1}^n \mathbf{E}[\tilde{X}_i^2] + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbf{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(X_i, X_j). \end{aligned}$$

The following example illustrates the use of this formula.

Example 4.26. Consider the hat problem discussed in Section 2.5, where n people throw their hats in a box and then pick a hat at random. Let us find the variance of X , the number of people who pick their own hat. We have

$$X = X_1 + \cdots + X_n,$$

where X_i is the random variable that takes the value 1 if the i th person selects his/her own hat, and takes the value 0 otherwise. Noting that X_i is Bernoulli with parameter $p = \mathbf{P}(X_i = 1) = 1/n$, we obtain

$$\text{var}(X_i) = \frac{1}{n} \left(1 - \frac{1}{n}\right).$$

For $i \neq j$, we have

$$\begin{aligned} \text{cov}(X_i, X_j) &= \mathbf{E} \left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j]) \right] \\ &= \mathbf{E}[X_i X_j] - \mathbf{E}[X_i] \mathbf{E}[X_j] \\ &= \mathbf{P}(X_i = 1 \text{ and } X_j = 1) - \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1) \\ &= \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1 | X_i = 1) - \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1) \\ &= \frac{1}{n} \cdot \frac{1}{n-1} - \frac{1}{n^2} \\ &= \frac{1}{n^2(n-1)}. \end{aligned}$$

Therefore

$$\begin{aligned} \text{var}(X) &= \text{var} \left(\sum_{i=1}^n X_i \right) \\ &= \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(X_i, X_j) \\ &= n \cdot \frac{1}{n} \left(1 - \frac{1}{n}\right) + 2 \cdot \frac{n(n-1)}{2} \cdot \frac{1}{n^2(n-1)} \\ &= 1. \end{aligned}$$

4.6 LEAST SQUARES ESTIMATION

In many practical contexts, we want to form an estimate of the value of a random variable X given the value of a related random variable Y , which may be viewed as some form of “measurement” of X . For example, X may be the range of

an aircraft and Y may be a noise-corrupted measurement of that range. In this section we discuss a popular formulation of the estimation problem, which is based on finding the estimate c that minimizes the expected value of the squared error $(X - c)^2$ (hence the name “least squares”).

We start by considering the simpler problem of estimating X before learning the value of Y . The estimation error $X - c$ is random (because X is random), but the mean squared error $\mathbf{E}[(X - c)^2]$ is a number that depends on c and can be minimized over c . With respect to this criterion, it turns out that the best possible estimate is to set c equal to $\mathbf{E}[X]$, as we proceed to verify.

For any estimate c , we have

$$\mathbf{E}[(X - c)^2] = \text{var}(X - c) + (\mathbf{E}[X - c])^2 = \text{var}(X) + (\mathbf{E}[X] - c)^2,$$

where the first equality makes use of the formula $\mathbf{E}[Z^2] = \text{var}(Z) + (\mathbf{E}[Z])^2$, and the second holds because when the constant c is subtracted from the random variable X , the variance is unaffected while the mean is reduced by c . We now note that the term $\text{var}(X)$ does not depend on our choice of c . Therefore, we should choose c to minimize the term $(\mathbf{E}[X] - c)^2$, which leads to $c = \mathbf{E}[X]$ (see Fig. 4.9).

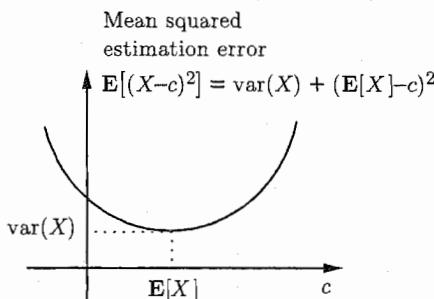


Figure 4.9: The mean squared error $\mathbf{E}[(X - c)^2]$, as a function of the estimate c , is a quadratic in c and is minimized when $c = \mathbf{E}[X]$. The minimum value of the mean squared error is $\text{var}(X)$.

Suppose now that we observe the value y of some related random variable Y , before forming an estimate of X . How can we exploit this additional information? Once we know the value y of Y , the situation is identical to the one considered earlier, except that we are now in a new “universe,” where everything is conditioned on $Y = y$. We can therefore adapt our earlier conclusion and assert that $c = \mathbf{E}[X | Y = y]$ minimizes the *conditional* mean squared error $\mathbf{E}[(c - X)^2 | Y = y]$. Note that the resulting estimate c depends on the value y of Y (as it should). Thus, we call $\mathbf{E}[X | Y = y]$ the **least-squares estimate** of X given the value y of Y .

Example 4.27. Let X be uniformly distributed in the interval $[4, 10]$ and suppose that we observe X with some random error W , that is, we observe the value of the random variable

$$Y = X + W.$$

We assume that W is uniformly distributed in the interval $[-1, 1]$, and independent of X . What is the least squares estimate of X given the value y of Y ?

We have $f_X(x) = 1/6$ for $4 \leq x \leq 10$, and $f_X(x) = 0$, elsewhere. Conditioned on X being equal to some x , Y is the same as $x + W$, and is uniform over the interval $[x - 1, x + 1]$. Thus, the joint PDF is given by

$$f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y|x) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12},$$

if $4 \leq x \leq 10$ and $x - 1 \leq y \leq x + 1$, and is zero for all other values of (x, y) . The slanted rectangle in the right-hand side of Fig. 4.10 is the set of pairs (x, y) for which $f_{X,Y}(x, y)$ is nonzero.

Given a value y of Y , the conditional PDF $f_{X|Y}$ of X is uniform on the corresponding vertical section of the slanted rectangle. The optimal estimate $E[X|Y = y]$ is the midpoint of that section. In the special case of the present example, it happens to be a piecewise linear function of y .

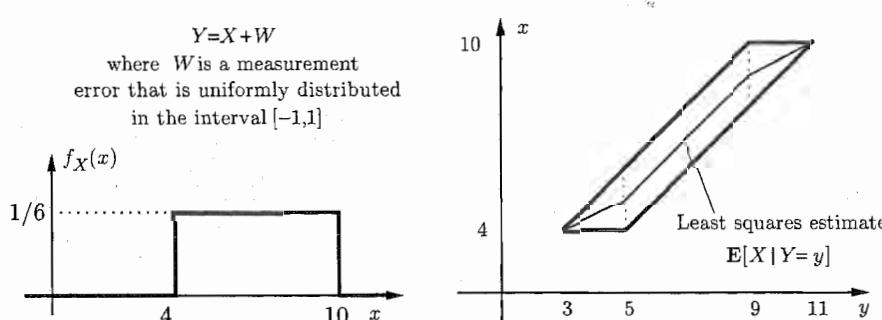


Figure 4.10: The PDFs in Example 4.27. The joint PDF of X and Y is uniform over the parallelogram shown on the right. The least squares estimate of X , given the value y of the random variable $Y = X + W$, depends on y and is represented by the piecewise linear function shown on the right.

As Example 4.27 illustrates, the estimate $E[X|Y = y]$ depends on the observed value y and should be viewed as a function of y ; see Fig. 4.11. To amplify this point, we refer to any function $g(Y)$ of the available information Y as an **estimator**. Given a value y of Y , such an estimator provides an estimate $g(y)$ (which is a number). However, if y is left unspecified, then the estimator is the random variable $g(Y)$. The expected value of the squared estimation error associated with an estimator $g(Y)$ is

$$E[(X - g(Y))^2].$$

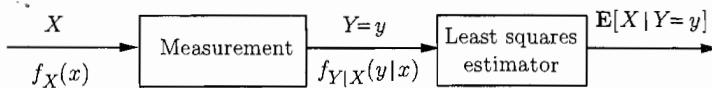


Figure 4.11: The least squares estimator.

Out of all possible estimators, it turns out that the mean squared estimation error is minimized when $g(Y) = \mathbf{E}[X | Y]$. To see this, note that if c is any number, we have

$$\mathbf{E}\left[\left(X - \mathbf{E}[X | Y = y]\right)^2 | Y = y\right] \leq \mathbf{E}\left[\left(X - c\right)^2 | Y = y\right],$$

since $\mathbf{E}[X | Y = y]$ is the least squares estimate of X , given the information $Y = y$. Consider now an arbitrary estimator $g(Y)$. For a given value y of Y , $g(y)$ is a number and, therefore,

$$\mathbf{E}\left[\left(X - \mathbf{E}[X | Y = y]\right)^2 | Y = y\right] \leq \mathbf{E}\left[\left(X - g(y)\right)^2 | Y = y\right].$$

This inequality is true for *every* possible value y of Y . Thus,

$$\mathbf{E}\left[\left(X - \mathbf{E}[X | Y]\right)^2 | Y\right] \leq \mathbf{E}\left[\left(X - g(Y)\right)^2 | Y\right],$$

which is now an inequality between random variables (functions of Y). We take expectations of both sides, and use the law of iterated expectations, to conclude that

$$\mathbf{E}\left[\left(X - \mathbf{E}[X | Y]\right)^2\right] \leq \mathbf{E}\left[\left(X - g(Y)\right)^2\right]$$

for all functions $g(Y)$.

Key Facts About Least Squares Estimation

- $\mathbf{E}\left[\left(X - c\right)^2\right]$ is minimized when $c = \mathbf{E}[X]$:

$$\mathbf{E}\left[\left(X - \mathbf{E}[X]\right)^2\right] \leq \mathbf{E}\left[\left(X - c\right)^2\right], \quad \text{for all } c.$$

- $\mathbf{E}\left[\left(X - c\right)^2 | Y = y\right]$ is minimized when $c = \mathbf{E}[X | Y = y]$:

$$\mathbf{E}\left[\left(X - \mathbf{E}[X | Y = y]\right)^2 | Y = y\right] \leq \mathbf{E}\left[\left(X - c\right)^2 | Y = y\right], \quad \text{for all } c.$$

- Out of all estimators $g(Y)$ of X based on Y , the mean squared estimation error $\mathbf{E}\left[\left(X - g(Y)\right)^2\right]$ is minimized when $g(Y) = \mathbf{E}[X | Y]$:

$$\mathbf{E}\left[\left(X - \mathbf{E}[X | Y]\right)^2\right] \leq \mathbf{E}\left[\left(X - g(Y)\right)^2\right], \quad \text{for all functions } g(Y).$$

Some Properties of the Estimation Error

Let us use the notation

$$\hat{X} = \mathbf{E}[X | Y], \quad \tilde{X} = X - \hat{X},$$

for the (optimal) estimator and the associated estimation error, respectively. The random variables \hat{X} and \tilde{X} have a number of useful properties, which are stated below.

Properties of the Estimation Error

- The estimation error \tilde{X} has zero unconditional and conditional mean:

$$\mathbf{E}[\tilde{X}] = 0, \quad \mathbf{E}[\tilde{X} | Y = y] = 0, \quad \text{for all } y.$$

- The estimation error \tilde{X} is uncorrelated with the estimate \hat{X} :

$$\text{cov}(\hat{X}, \tilde{X}) = 0.$$

- The variance of X can be decomposed as

$$\text{var}(X) = \text{var}(\hat{X}) + \text{var}(\tilde{X}).$$

To verify the first property, note that

$$\mathbf{E}[\tilde{X} | Y] = \mathbf{E}[X - \hat{X} | Y] = \mathbf{E}[X | Y] - \mathbf{E}[\hat{X} | Y] = \hat{X} - \hat{X} = 0.$$

We have used here the fact that \hat{X} is completely determined by Y and therefore $\mathbf{E}[\hat{X} | Y] = \hat{X}$. Using the law of iterated expectations, we also have

$$\mathbf{E}[\tilde{X}] = \mathbf{E}[\mathbf{E}[\tilde{X} | Y]] = 0.$$

For similar reasons, since $\hat{X} - \mathbf{E}[\hat{X}]$ is completely determined by Y ,

$$\mathbf{E}[(\hat{X} - \mathbf{E}[\hat{X}])\tilde{X} | Y] = (\hat{X} - \mathbf{E}[\hat{X}])\mathbf{E}[\tilde{X} | Y] = 0.$$

Taking expectations and using the law of iterated expectations, we obtain

$$\text{cov}(\hat{X}, \tilde{X}) = \mathbf{E}[(\hat{X} - \mathbf{E}[\hat{X}])\tilde{X}] = \mathbf{E}[\mathbf{E}[(\hat{X} - \mathbf{E}[\hat{X}])\tilde{X} | Y]] = 0.$$

We finally, note that $X = \hat{X} + \tilde{X}$. Since \hat{X} and \tilde{X} are uncorrelated, we obtain

$$\text{var}(X) = \text{var}(\hat{X}) + \text{var}(\tilde{X}) + 2\text{cov}(\hat{X}, \tilde{X}) = \text{var}(\hat{X}) + \text{var}(\tilde{X}).$$

Example 4.28. Let us say that the observed random variable Y is *uninformative* if the mean squared estimation error $\mathbf{E}[\tilde{X}^2] = \text{var}(\tilde{X})$ is the same as the unconditional variance $\text{var}(X)$ of X . When is this the case?

Using the formula

$$\text{var}(X) = \text{var}(\hat{X}) + \text{var}(\tilde{X}),$$

we see that Y is uninformative if and only if $\text{var}(\tilde{X}) = 0$. The variance of a random variable is zero if and only if that random variable is a constant, equal to its mean. We conclude that Y is uninformative if and only if the estimate $\hat{X} = \mathbf{E}[X | Y]$ is equal to $\mathbf{E}[X]$, for every value of Y .

If X and Y are independent, we have $\mathbf{E}[X | Y = y] = \mathbf{E}[X]$ for all y , and Y is indeed uninformative, which is quite intuitive. The converse, however, is not true: it is possible for $\mathbf{E}[X | Y]$ to be always equal to the constant $\mathbf{E}[X]$, without X and Y being independent. (Can you construct an example?)

Estimation Based on Several Measurements

So far, we have discussed the estimation of one random variable X on the basis of another random variable Y . In practice, one often has access to the values of several random variables Y_1, \dots, Y_n , that can be used to estimate X . Generalizing our earlier discussion, and using essentially the same argument, it can be shown that the mean squared estimation error is minimized if we use $\mathbf{E}[X | Y_1, \dots, Y_n]$ as our estimator, i.e.,

$$\mathbf{E}[(X - \mathbf{E}[X | Y_1, \dots, Y_n])^2] \leq \mathbf{E}[(X - g(Y_1, \dots, Y_n))^2],$$

for all functions $g(Y_1, \dots, Y_n)$.

This provides a complete solution to the general problem of least squares estimation, but is sometimes difficult to implement, for the following reasons.

- In order to compute the conditional expectation $\mathbf{E}[X | Y_1, \dots, Y_n]$, we need a complete probabilistic model, that is, the joint PDF f_{X, Y_1, \dots, Y_n} .
- Even if this joint PDF is available, $\mathbf{E}[X | Y_1, \dots, Y_n]$ can be a very complicated function of Y_1, \dots, Y_n .

As a consequence, practitioners often resort to approximations of the conditional expectation or focus on estimators that are not optimal but are simple and easy to implement. The most common approach involves **linear estimators**, of the form

$$a_1 Y_1 + \dots + a_n Y_n + b.$$

Given a particular choice of a_1, \dots, a_n, b , the corresponding mean squared error is

$$\mathbf{E}[(X - a_1 Y_1 - \dots - a_n Y_n - b)^2],$$

and it is meaningful to choose the coefficients a_1, \dots, a_n, b in a way that minimizes the above expression. This problem is relatively easy to solve and only requires knowledge of the means, variances, and covariances of the different random variables. We develop the solution for the case where $n = 1$, and we refer to the problem section for the case where $n > 1$.

Linear Least Squares Estimation Based on a Single Measurement

We are interested in finding a and b that minimize the mean squared estimation error $\mathbf{E}[(X - aY - b)^2]$, associated with a linear estimator $aY + b$ of X . Suppose that a has already been chosen. How should we choose b ? This is the same as having to choose a constant b to estimate the random variable $X - aY$ and, by our earlier results, the best choice is to let $b = \mathbf{E}[X - aY] = \mathbf{E}[X] - a\mathbf{E}[Y]$.

It now remains to minimize, with respect to a , the expression

$$\mathbf{E}[(X - aY - \mathbf{E}[X] + a\mathbf{E}[Y])^2],$$

which is the same as

$$\text{var}(X - aY) = \sigma_X^2 + a^2\sigma_Y^2 + 2\text{cov}(X, -aY) = \sigma_X^2 + a^2\sigma_Y^2 - 2a \cdot \text{cov}(X, Y),$$

where σ_X and σ_Y are the standard deviations of X and Y , respectively, and

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

is the covariance of X and Y . The squared error is a quadratic function of a . It is minimized at the point where its derivative is zero, which is

$$a = \frac{\text{cov}(X, Y)}{\sigma_Y^2} = \frac{\rho\sigma_X\sigma_Y}{\sigma_Y^2} = \rho\frac{\sigma_X}{\sigma_Y},$$

where

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X\sigma_Y}$$

is the correlation coefficient. With this choice of a , the mean squared estimation error is given by

$$\begin{aligned} \sigma_X^2 + a^2\sigma_Y^2 - 2a \cdot \text{cov}(X, Y) &= \sigma_X^2 + \rho^2\frac{\sigma_X^2}{\sigma_Y^2}\sigma_Y^2 - 2\rho\frac{\sigma_X}{\sigma_Y}\rho\sigma_X\sigma_Y \\ &= (1 - \rho^2)\sigma_X^2. \end{aligned}$$

Linear Least Squares Estimation Formulas

- The linear least squares estimator of X based on Y is given by

$$\mathbf{E}[X] + \rho \frac{\sigma_X}{\sigma_Y} (Y - \mathbf{E}[Y]),$$

where

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

is the correlation coefficient.

- The resulting mean squared estimation error is equal to

$$(1 - \rho^2) \sigma_X^2.$$

The formula for the linear least squares estimator has an intuitive interpretation. Suppose, for concreteness, that the correlation coefficient ρ is positive. The estimator starts with the baseline estimate $\mathbf{E}[X]$ for X , which it then adjusts by taking into account the value of $Y - \mathbf{E}[Y]$. For example, when Y is larger than its mean, the positive correlation between X and Y suggests that X is expected to be larger than its mean, and the resulting estimate is set to a value larger than X . The value of ρ also affects the quality of the estimate. When $|\rho|$ is close to 1, the two random variables are highly correlated and knowing Y allows us to accurately estimate X .

4.7 THE BIVARIATE NORMAL DISTRIBUTION

Let U and V be two independent normal random variables, and consider two new random variables X and Y of the form

$$\begin{aligned} X &= aU + bV, \\ Y &= cU + dV, \end{aligned}$$

where a, b, c, d , are some scalars. Each one of the random variables X and Y is normal, since it is a linear function of independent normal random variables (see Example 4.12).[†] Furthermore, because X and Y are linear functions of the same

[†] For the purposes of this section, we adopt the following convention. A random variable which is always equal to a constant will also be called normal, with zero variance, even though it does not have a PDF. With this convention, the family of normal random variables is closed under linear operations. That is, if X is normal, then $aX + b$ is also normal, even if $a = 0$.

two independent normal random variables, their joint PDF takes a special form, known as the **bivariate normal** PDF. The bivariate normal PDF has several useful and elegant properties and, for this reason, it is a commonly employed model. In this section, we derive many such properties, both qualitative and analytical, culminating in a closed-form expression for the joint PDF. To keep the discussion simple, we restrict ourselves to the case where X and Y have zero mean.

Jointly Normal Random Variables

Two random variables X and Y are said to be **jointly normal** if they can be expressed in the form

$$\begin{aligned} X &= aU + bV, \\ Y &= cU + dV, \end{aligned}$$

where U and V are independent normal random variables.

Note that if X and Y are jointly normal, then any linear combination

$$Z = s_1X + s_2Y$$

has a normal distribution. The reason is that if we have $X = aU + bV$ and $Y = cU + dV$ for some independent normal random variables U and V , then

$$Z = s_1(aU + bV) + s_2(cU + dV) = (as_1 + cs_2)U + (bs_1 + ds_2)V.$$

Thus, Z is the sum of the independent normal random variables $(as_1 + cs_2)U$ and $(bs_1 + ds_2)V$, and is therefore normal.

A very important property of jointly normal random variables, and which will be the starting point for our development, is that zero correlation implies independence.

Zero Correlation Implies Independence

If two random variables X and Y are jointly normal and are uncorrelated, then they are independent.

This property can be verified using multivariate transforms, as follows. Suppose that U and V are independent zero-mean normal random variables, and that $X = aU + bV$ and $Y = cU + dV$, so that X and Y are jointly normal. We assume that X and Y are uncorrelated, and we wish to show that they are independent. Our first step is to derive a formula for the multivariate transform $M_{X,Y}(s_1, s_2)$ associated with X and Y . Recall that if Z is a zero-mean normal

random variable with variance σ_Z^2 , the associated transform is

$$\mathbf{E}[e^{sZ}] = M_Z(s) = e^{\sigma_Z^2 s^2/2},$$

which implies that

$$\mathbf{E}[e^Z] = M_Z(1) = e^{\sigma_Z^2/2}.$$

Let us fix some scalars s_1, s_2 , and let $Z = s_1 X + s_2 Y$. The random variable Z is normal, by our earlier discussion, with variance

$$\sigma_Z^2 = s_1^2 \sigma_X^2 + s_2^2 \sigma_Y^2.$$

This leads to the following formula for the multivariate transform associated with the uncorrelated pair X and Y :

$$\begin{aligned} M_{X,Y}(s_1, s_2) &= \mathbf{E}[e^{s_1 X + s_2 Y}] \\ &= \mathbf{E}[e^Z] \\ &= e^{(s_1^2 \sigma_X^2 + s_2^2 \sigma_Y^2)/2}. \end{aligned}$$

Let now \bar{X} and \bar{Y} be *independent* zero-mean normal random variables with the same variances σ_X^2 and σ_Y^2 as X and Y , respectively. Since \bar{X} and \bar{Y} are independent, they are also uncorrelated, and the preceding argument yields

$$M_{\bar{X}, \bar{Y}}(s_1, s_2) = e^{(s_1^2 \sigma_X^2 + s_2^2 \sigma_Y^2)/2}.$$

Thus, the two pairs of random variables (X, Y) and (\bar{X}, \bar{Y}) are associated with the same multivariate transform. Since the multivariate transform completely determines the joint PDF, it follows that the pair (X, Y) has the same joint PDF as the pair (\bar{X}, \bar{Y}) . Since \bar{X} and \bar{Y} are independent, X and Y must also be independent, which establishes our claim.

The Conditional Distribution of X Given Y

We now turn to the problem of estimating X given the value of Y . To avoid uninteresting degenerate cases, we assume that both X and Y have positive variance. Let us define[†]

$$\hat{X} = \rho \frac{\sigma_X}{\sigma_Y} Y, \quad \tilde{X} = X - \hat{X},$$

where

$$\rho = \frac{\mathbf{E}[XY]}{\sigma_X \sigma_Y}$$

[†] Comparing with the formulas in the preceding section, it is seen that \hat{X} is defined to be the linear least squares estimator of X , and \tilde{X} is the corresponding estimation error, although these facts are not needed for the argument that follows.

is the correlation coefficient of X and Y . Since X and Y are linear combinations of independent normal random variables U and V , it follows that Y and \tilde{X} are also linear combinations of U and V . In particular, Y and \tilde{X} are jointly normal. Furthermore,

$$\mathbf{E}[Y\tilde{X}] = \mathbf{E}[YX] - \mathbf{E}[Y\hat{X}] = \rho\sigma_X\sigma_Y - \rho\frac{\sigma_X}{\sigma_Y}\sigma_Y^2 = 0.$$

Thus, Y and \tilde{X} are uncorrelated and, therefore, independent. Since \hat{X} is a scalar multiple of Y , it follows that \hat{X} and \tilde{X} are independent.

We have so far decomposed X into a sum of two independent normal random variables, namely,

$$X = \hat{X} + \tilde{X} = \rho\frac{\sigma_X}{\sigma_Y}Y + \tilde{X}.$$

We take conditional expectations of both sides, given Y , to obtain

$$\mathbf{E}[X|Y] = \rho\frac{\sigma_X}{\sigma_Y}\mathbf{E}[Y|Y] + \mathbf{E}[\tilde{X}|Y] = \rho\frac{\sigma_X}{\sigma_Y}Y = \hat{X},$$

where we have made use of the independence of Y and \tilde{X} to set $\mathbf{E}[\tilde{X}|Y] = 0$. We have therefore reached the important conclusion that the conditional expectation $\mathbf{E}[X|Y]$ is a linear function of the random variable Y .

Using the above decomposition, it is now easy to determine the conditional PDF of X . Given a value of Y , the random variable $\hat{X} = \rho\sigma_X Y / \sigma_Y$ becomes a known constant, but the normal distribution of the random variable \tilde{X} is unaffected, since \tilde{X} is independent of Y . Therefore, the conditional distribution of X given Y is the same as the unconditional distribution of \tilde{X} , shifted by \hat{X} . Since \tilde{X} is normal with mean zero and some variance $\sigma_{\tilde{X}}^2$, we conclude that the conditional distribution of X is also normal with mean \hat{X} and the same variance $\sigma_{\tilde{X}}^2$. The variance of \tilde{X} can be found with the following calculation:

$$\begin{aligned}\sigma_{\tilde{X}}^2 &= \mathbf{E}\left[\left(X - \rho\frac{\sigma_X}{\sigma_Y}Y\right)^2\right] \\ &= \sigma_X^2 - 2\rho\frac{\sigma_X}{\sigma_Y}\rho\sigma_X\sigma_Y + \rho^2\frac{\sigma_X^2}{\sigma_Y^2}\sigma_Y^2 \\ &= (1 - \rho^2)\sigma_X^2,\end{aligned}$$

where we have made use of the property $\mathbf{E}[XY] = \rho\sigma_X\sigma_Y$.

We summarize our conclusions below. Although our discussion used the zero-mean assumption, these conclusions also hold for the non-zero mean case and we state them with this added generality; see the end-of-chapter problems.

Properties of Jointly Normal Random Variables

Let X and Y be jointly normal random variables.

- X and Y are independent if and only if they are uncorrelated.
- The conditional expectation of X given Y satisfies

$$\mathbf{E}[X | Y] = \mathbf{E}[X] + \rho \frac{\sigma_X}{\sigma_Y} (Y - \mathbf{E}[Y]).$$

It is a linear function of Y and has a normal PDF.

- The estimation error $\tilde{X} = X - \mathbf{E}[X | Y]$ is zero-mean, normal, and independent of Y , with variance

$$\sigma_{\tilde{X}}^2 = (1 - \rho^2)\sigma_X^2.$$

- The conditional distribution of X given Y is normal with mean $\mathbf{E}[X | Y]$ and variance $\sigma_{\tilde{X}}^2$.

The Form of the Bivariate Normal PDF

Having determined the parameters of the PDF of \tilde{X} and of the conditional PDF of X , we can give explicit formulas for these PDFs. We keep assuming that X and Y have zero means and positive variances. Furthermore, to avoid the degenerate where \tilde{X} is identically zero, we assume that $|\rho| < 1$. We have

$$f_{\tilde{X}}(\tilde{x}) = f_{\tilde{X}|Y}(\tilde{x} | y) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_X} e^{-\tilde{x}^2/2\sigma_{\tilde{X}}^2},$$

and

$$f_{X|Y}(x | y) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_X} e^{-\left(x - \rho \frac{\sigma_X}{\sigma_Y} y\right)^2/2\sigma_{\tilde{X}}^2},$$

where

$$\sigma_{\tilde{X}}^2 = (1 - \rho^2)\sigma_X^2.$$

Using also the formula for the PDF of Y ,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-y^2/2\sigma_Y^2},$$

and the multiplication rule $f_{X,Y}(x, y) = f_Y(y)f_{X|Y}(x | y)$, we can obtain the joint PDF of X and Y . This PDF is of the form

$$f_{X,Y}(x, y) = ce^{-q(x,y)},$$

where the normalizing constant is

$$c = \frac{1}{2\pi\sqrt{1 - \rho^2}\sigma_X\sigma_Y}.$$

The exponent term $q(x, y)$ is a quadratic function of x and y ,

$$q(x, y) = \frac{y^2}{2\sigma_Y^2} + \frac{\left(x - \rho\frac{\sigma_X}{\sigma_Y}y\right)^2}{2(1 - \rho^2)\sigma_X^2},$$

which after some straightforward algebra simplifies to

$$q(x, y) = \frac{\frac{x^2}{\sigma_X^2} - 2\rho\frac{xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}}{2(1 - \rho^2)}.$$

An important observation here is that **the joint PDF is completely determined by σ_X , σ_Y , and ρ** .

In the special case where X and Y are uncorrelated ($\rho = 0$), the joint PDF takes the simple form

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{x^2}{2\sigma_X^2} - \frac{y^2}{2\sigma_Y^2}},$$

which is just the product of two independent normal PDFs. We can get some insight into the form of this PDF by considering its contours, i.e., sets of points at which the PDF takes a constant value. These contours are described by an equation of the form

$$\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} = \text{constant},$$

and are ellipses whose two axes are horizontal and vertical.

In the more general case where X and Y are dependent, a typical contour is described by

$$\frac{x^2}{\sigma_X^2} - 2\rho\frac{xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2} = \text{constant},$$

and is again an ellipse, but its axes are no longer horizontal and vertical. Figure 4.12 illustrates the contours for two cases, one in which ρ is positive and one in which ρ is negative.

Example 4.29. Suppose that X and Z are zero-mean jointly normal random variables, such that $\sigma_X^2 = 4$, $\sigma_Z^2 = 17/9$, and $\mathbf{E}[XZ] = 2$. We define a new random variable $Y = 2X - 3Z$. We wish to determine the PDF of Y , the conditional PDF of X given Y , and the joint PDF of X and Y .

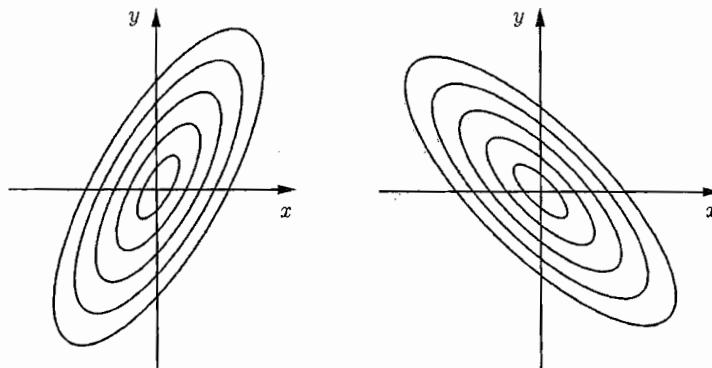


Figure 4.12: Contours of the bivariate normal PDF. The diagram on the left (respectively, right) corresponds to a case of positive (respectively, negative) correlation coefficient ρ .

As noted earlier, a linear function of two jointly normal random variables is also normal. Thus, Y is normal with variance

$$\sigma_Y^2 = \mathbf{E}[(2X - 3Z)^2] = 4\mathbf{E}[X^2] + 9\mathbf{E}[Z^2] - 12\mathbf{E}[XZ] = 4 \cdot 4 + 9 \cdot \frac{17}{9} - 12 \cdot 2 = 9.$$

Hence, Y has the normal PDF

$$f_Y(y) = \frac{1}{\sqrt{2\pi} \cdot 3} e^{-y^2/18}.$$

We next note that X and Y are jointly normal. The reason is that X and Z are linear functions of two independent normal random variables (by the definition of joint normality), so that X and Y are also linear functions of the same independent normal random variables. The covariance of X and Y is equal to

$$\begin{aligned} \mathbf{E}[XY] &= \mathbf{E}[X(2X - 3Z)] \\ &= 2\mathbf{E}[X^2] - 3\mathbf{E}[XZ] \\ &= 2 \cdot 4 - 3 \cdot 2 \\ &= 2. \end{aligned}$$

Hence, the correlation coefficient of X and Y , denoted by ρ , is equal to

$$\rho = \frac{\mathbf{E}[XY]}{\sigma_X \sigma_Y} = \frac{2}{2 \cdot 3} = \frac{1}{3}.$$

The conditional expectation of X given Y is

$$\mathbf{E}[X | Y] = \rho \frac{\sigma_X}{\sigma_Y} Y = \frac{1}{3} \cdot \frac{2}{3} Y = \frac{2}{9} Y.$$

The conditional variance of X given Y (which is the same as the variance of $\tilde{X} = X - \mathbf{E}[X|Y]$) is

$$\sigma_{\tilde{X}}^2 = (1 - \rho^2)\sigma_X^2 = \left(1 - \frac{1}{9}\right)4 = \frac{32}{9},$$

so that $\sigma_{\tilde{X}} = \sqrt{32}/3$. Hence, the conditional PDF of X given Y is

$$f_{X|Y}(x|y) = \frac{3}{\sqrt{2\pi}\sqrt{32}}e^{-\frac{(x - (2y/9))^2}{2 \cdot 32/9}}.$$

Finally, the joint PDF of X and Y is obtained using either the multiplication rule $f_{X,Y}(x,y) = f_X(x)f_{X|Y}(x|y)$, or by using the earlier developed formula for the exponent $q(x,y)$, and is equal to

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{32}}e^{-\frac{\frac{y^2}{9} + \frac{x^2}{4} - \frac{2}{3} \cdot \frac{xy}{2 \cdot 3}}{2(1 - (1/9))}}.$$

We end with a cautionary note. If X and Y are jointly normal, then each random variable X and Y is normal. However, the converse is not true. Namely, if each of the random variables X and Y is normal, it does not follow that they are jointly normal, even if they are uncorrelated. This is illustrated in the following example.

Example 4.30. Let X have a normal distribution with zero mean and unit variance. Let Z be independent of X , with $\mathbf{P}(Z = 1) = \mathbf{P}(Z = -1) = 1/2$. Let $Y = ZX$, which is also normal with zero mean. The reason is that conditioned on either value of Z , Y has the same normal distribution, hence its unconditional distribution is also normal. Furthermore,

$$\mathbf{E}[XY] = \mathbf{E}[ZX^2] = \mathbf{E}[Z]\mathbf{E}[X^2] = 0 \cdot 1 = 0,$$

so X and Y are uncorrelated. On the other hand X and Y are clearly dependent. (For example, if $X = 1$, then Y must be either -1 or 1 .) If X and Y were jointly normal, we would have a contradiction to our earlier conclusion that zero correlation implies independence. It follows that X and Y are *not* jointly normal, even though both marginal distributions are normal.

The Multivariate Normal PDF

The development in this section generalizes to the case of more than two random variables. For example, we can say that the random variables X_1, \dots, X_n are jointly normal if all of them are linear functions of a set U_1, \dots, U_n of independent

normal random variables. We can then establish the natural extensions of the results derived in this section. For example, it is still true that zero correlation implies independence, that the conditional expectation of one random variable given some of the others is a linear function of the conditioning random variables, and that the conditional PDF of X_1, \dots, X_k given X_{k+1}, \dots, X_n is multivariate normal. Finally, there is a closed-form expression for the joint PDF. Assuming that none of the random variables is a deterministic function of the others, we have

$$f_{X_1, \dots, X_n} = ce^{-q(x_1, \dots, x_n)},$$

where c is a normalizing constant and where $q(x_1, \dots, x_n)$ is a quadratic function of x_1, \dots, x_n that increases to infinity as the magnitude of the vector (x_1, \dots, x_n) tends to infinity.

Multivariate normal models are very common in statistics, econometrics, signal processing, feedback control, and many other fields. However, a full development falls outside the scope of this text.

4.8 SUMMARY AND DISCUSSION

In this chapter, we have studied a number of advanced topics. We discuss here some of the highlights.

In Section 4.1, we introduced the transform associated with a random variable, and saw how such a transform can be computed. Conversely, we indicated that given a transform, the distribution of an associated random variable is uniquely determined and can be found, for example, using tables of commonly occurring transforms. We have found transforms useful for a variety of purposes, such as the following.

- (a) Knowledge of the transform associated with a random variable provides a shortcut for calculating the moments of the random variable.
- (b) The transform associated with the sum of two independent random variables is equal to the product of the transforms associated with each one of them. This property was used to show that the sum of two independent normal (respectively, Poisson) random variables is normal (respectively, Poisson).
- (c) Transforms can be used to characterize the distribution of the sum of a random number of random variables (Section 4.4), something which is often impossible by any other means.
- (d) The inversion property of multivariate transforms was used in Section 4.7 to establish that two uncorrelated jointly normal random variables are independent.

Sums of independent random variables can also be handled without using transforms. In Section 4.2, we gave the convolution formula for the PMF or

PDF of the sum of two independent random variables, in terms of the individual PMFs or PDFs. Furthermore, in Section 4.4, we derived formulas for the mean and variance of the sum of a random number of random variables, using a divide-and-conquer strategy.

One of the most useful divide-and-conquer strategies relies on conditioning and the use of conditional expectations. In Section 4.3, we took a closer look at the conditional expectation and indicated that it can be viewed as a random variable, with an expectation and variance of its own. We derived some related properties, including the law of iterated expectations, and the law of total variance.

In Section 4.6 we discussed problems of estimation and prediction of one random variable X , based on knowledge of another random variable Y . These problems are related to the problems of inference that we discussed in Section 3.5 using Bayes' rule. In this chapter, we discussed the least squares approach. The highlights of this approach are:

- (a) The conditional expectation $\mathbf{E}[X | Y]$ is the optimal estimator under the least squares criterion.
- (b) When the conditional expectation is hard to compute, one may be interested in an optimal estimator within a restricted class of estimators that depend linearly on the available information Y . The formula for the best linear estimator involves the correlation coefficient between X and Y , which was introduced in Section 4.5, and which provides a numerical indicator of the relation between X and Y .

Finally, in Section 4.7, we introduced the concept of jointly normal random variables. Such random variables have several remarkable properties. For example, the sum of jointly normal random variables is normal. Furthermore, the conditional expectation $\mathbf{E}[X | Y]$ is a linear function of Y and is therefore normal. These properties imply that one can keep on forming linear combinations of jointly normal random variables, and taking conditional expectations, while remaining within the analytically tractable class of normal random variables. Furthermore, the somewhat cumbersome formula for the bivariate normal PDF is completely determined in terms of a few parameters: the means, variances, and the correlation coefficient of the random variables of interest. Thus, in order to determine the particular form of a bivariate normal PDF, one only needs to calculate the values of these parameters. For all these reasons, jointly normal models have found a tremendous range of applications, in statistics, signal processing, communication theory, and many other contexts.

 P R O B L E M S

SECTION 4.1. Transforms

Problem 1. Let X be a random variable that takes the values 1, 2, and 3, with the following probabilities:

$$\mathbf{P}(X = 1) = \frac{1}{2}, \quad \mathbf{P}(X = 2) = \frac{1}{4}, \quad \mathbf{P}(X = 3) = \frac{1}{4}.$$

Find the transform associated with X and use it to obtain the first three moments, $\mathbf{E}[X]$, $\mathbf{E}[X^2]$, $\mathbf{E}[X^3]$.

Problem 2. A nonnegative integer-valued random variable X has one of the following two expressions as its transform:

$$1. \quad M(s) = e^{2(e^{e^s} - 1)}.$$

$$2. \quad M(s) = e^{2(e^{e^s} - 1)}.$$

- (a) Explain why one of the two cannot possibly be the transform.
- (b) Use the true transform to find $\mathbf{P}(X = 0)$.

Problem 3. Find the PDF of the continuous random variable X associated with the transform

$$M(s) = \frac{1}{3} \cdot \frac{2}{2-s} + \frac{2}{3} \cdot \frac{3}{3-s}.$$

Problem 4. Let X be a random variable that takes nonnegative integer values, and is associated with a transform of the form

$$M_X(s) = c \cdot \frac{3 + 4e^{2s} + 2e^{3s}}{3 - e^s},$$

where c is some scalar. Find $\mathbf{E}[X]$, $p_X(1)$, and $\mathbf{E}[X \mid X \neq 0]$.

Problem 5. Let X , Y , and Z be independent random variables, where X is Bernoulli with parameter $1/3$, Y is exponential with parameter 2, and Z is Poisson with parameter 3.

- (a) Consider the new random variable $U = XY + (1 - X)Z$. Find the transform associated with U .
- (b) Find the transform associated with $2Z + 3$.
- (c) Find the transform associated with $Y + Z$.

Problem 6.* Let X be a discrete random variable taking nonnegative integer values. Let $M(s)$ be the transform associated with X .

- (a) Show that

$$\mathbf{P}(X = 0) = \lim_{s \rightarrow -\infty} M(s).$$

- (b) Use part (a) to verify that if X is a binomial random variable with parameters n and p , we have $\mathbf{P}(X = 0) = (1 - p)^n$. Furthermore, if X is a Poisson random variable with parameter λ , we have $\mathbf{P}(X = 0) = e^{-\lambda}$.
- (c) Suppose that X is instead known to take only integer values that are greater than or equal to a given integer \bar{k} . How can we calculate $P(X = \bar{k})$ using the transform associated with X ?

Solution. (a) We have

$$M(s) = \sum_{k=0}^{\infty} \mathbf{P}(X = k) e^{ks}.$$

As $s \rightarrow -\infty$, all the terms e^{ks} with $k > 0$ tend to 0, so we obtain $\lim_{s \rightarrow -\infty} M(s) = \mathbf{P}(X = 0)$.

(b) In the case of the binomial, we have from the transform tables

$$M(s) = (1 - p + pe^s)^n,$$

so that $\lim_{s \rightarrow -\infty} M(s) = (1 - p)^n$. In the case of the Poisson, we have

$$M(s) = e^{\lambda(e^s - 1)},$$

so that $\lim_{s \rightarrow -\infty} M(s) = e^{-\lambda}$.

(c) The random variable $Y = X - \bar{k}$ takes only nonnegative integer values and the associated transform is $M_Y(s) = e^{-s\bar{k}} M(s)$ (cf. Example 4.4). Since $\mathbf{P}(Y = 0) = \mathbf{P}(X = \bar{k})$, we have from part (a),

$$\mathbf{P}(X = \bar{k}) = \lim_{s \rightarrow -\infty} e^{-s\bar{k}} M(s).$$

Problem 7.* Transforms associated with uniform random variables.

- (a) Find the transform associated with an integer-valued random variable X that is uniformly distributed in the range $\{a, a + 1, \dots, b\}$.
- (b) Find the transform associated with a continuous random variable X that is uniformly distributed in the range $[a, b]$.

Solution. (a) The PMF of X is

$$p_X(k) = \begin{cases} \frac{1}{b - a + 1}, & \text{if } k = a, a + 1, \dots, b, \\ 0, & \text{otherwise.} \end{cases}$$

The transform is

$$\begin{aligned}
 M(s) &= \sum_{k=-\infty}^{\infty} e^{sk} \mathbf{P}(X = k) \\
 &= \sum_{k=a}^b \frac{1}{b-a+1} e^{sk} \\
 &= \frac{e^{sa}}{b-a+1} \sum_{k=0}^{b-a} e^{sk} \\
 &= \frac{e^{sa}}{b-a+1} \cdot \frac{1 - e^{s(b-a+1)}}{1 - e^s}.
 \end{aligned}$$

(b) We have

$$M(s) = \mathbf{E}[e^{sX}] = \int_a^b \frac{e^{sx}}{b-a} dx = \frac{e^{sb} - e^{sa}}{s(b-a)}.$$

Problem 8.* Find the third, fourth, and fifth moments of an exponential random variable with parameter λ .

Solution. The transform is

$$M(s) = \frac{\lambda}{\lambda - s}.$$

Thus,

$$\begin{aligned}
 \frac{d}{ds} M(s) &= \frac{\lambda}{(\lambda - s)^2}, & \frac{d^2}{ds^2} M(s) &= \frac{2\lambda}{(\lambda - s)^3}, & \frac{d^3}{ds^3} M(s) &= \frac{6\lambda}{(\lambda - s)^4}, \\
 \frac{d^4}{ds^4} M(s) &= \frac{24\lambda}{(\lambda - s)^5}, & \frac{d^5}{ds^5} M(s) &= \frac{120\lambda}{(\lambda - s)^6}.
 \end{aligned}$$

By setting $s = 0$, we obtain

$$\mathbf{E}[X^3] = \frac{6}{\lambda^3}, \quad \mathbf{E}[X^4] = \frac{24}{\lambda^4}, \quad \mathbf{E}[X^5] = \frac{120}{\lambda^5}.$$

Problem 9.* Suppose that the transform associated with a discrete random variable X has the form

$$M(s) = \frac{A(e^s)}{B(e^s)},$$

where $A(t)$ and $B(t)$ are polynomials of the generic variable t . Assume that $A(t)$ and $B(t)$ have no common roots and that the degree of $A(t)$ is smaller than the degree of $B(t)$. Assume also that $B(t)$ has distinct, real, and nonzero roots that have absolute value greater than 1. Then it can be seen that $M(s)$ can be written in the form

$$M(s) = \frac{a_1}{1 - r_1 e^s} + \cdots + \frac{a_m}{1 - r_m e^s},$$

where $1/r_1, \dots, 1/r_m$ are the roots of $B(t)$ and the a_i are constants that are equal to $\lim_{e^s \rightarrow 1} (1 - r_i e^s) M(s)$, $i = 1, \dots, m$.

(a) Show that the PMF of X has the form

$$\mathbf{P}(X = k) = \begin{cases} \sum_{i=1}^m a_i r_i^k, & \text{if } k = 0, 1, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

Note: For large k , the PMF of X can be approximated by $a_{\bar{i}} r_{\bar{i}}^k$, where \bar{i} is the index corresponding to the largest $|r_i|$ (assuming \bar{i} is unique).

(b) Extend the result of part (a) to the case where $M(s) = e^{bs} A(t)/B(t)$ and b is an integer.

Solution. (a) We have for all s such that $|r_i|e^s < 1$

$$\frac{1}{1 - r_i e^s} = 1 + r_i e^s + r_i^2 e^{2s} + \dots$$

Therefore,

$$M(s) = \sum_{i=1}^m a_i + \left(\sum_{i=1}^m a_i r_i \right) e^s + \left(\sum_{i=1}^m a_i r_i^2 \right) e^{2s} + \dots,$$

and by inverting this transform, we see that

$$\mathbf{P}(X = k) = \sum_{i=1}^m a_i r_i^k$$

for $k \geq 0$, and $\mathbf{P}(X = k) = 0$ for $k < 0$. Note that this PMF is a mixture of geometric PMFs.

(b) In this case, $M(s)$ corresponds to the translation by b of a random variable whose transform is $A(t)/B(t)$ (cf. Example 4.4), so we have

$$\mathbf{P}(X = k) = \begin{cases} \sum_{i=1}^m a_i r_i^{(k-b)}, & \text{if } k = b, b+1, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

SECTION 4.2. Sums of Independent Random Variables - Convolutions

Problem 10. A soccer team has three designated players who take turns striking penalty shots. The i th player has probability of success p_i , independently of the successes of the other players. Let X be the number of successful penalty shots after each player has had one turn. Use convolution to calculate the PMF of X . Confirm your

answer by first calculating the transform associated with X and then obtaining the PMF from the transform.

Problem 11. The random variables X , Y , and Z are independent and uniformly distributed between zero and one. Find the PDF of $X + Y + Z$.

Problem 12. Consider a PDF that is positive only within an interval $[a, b]$ and is symmetric around the mean $(a + b)/2$. Let X and Y be independent random variables that both have this PDF. Suppose that you have calculated the PDF and the transform associated with $X + Y$. How can you easily obtain the PDF and the transform associated with $X - Y$?

SECTION 4.3. More on Conditional Expectation and Variance

Problem 13. Pat and Nat are dating, and all of their dates are scheduled to start at 9 p.m. Nat always arrives promptly at 9 p.m. Pat is highly disorganized and arrives at a time that is uniformly distributed between 8 p.m. and 10 p.m. Let X be the time in hours between 8 p.m. and the time when Pat arrives. If Pat arrives before 9 p.m., their date will last exactly 3 hours. If Pat arrives after 9 p.m., their date will last for a time that is uniformly distributed between 0 and $3 - X$ hours. The date starts at the time they meet. Nat gets irritated when Pat is late and will end the relationship after the second date on which Pat is late by more than 45 minutes. All dates are independent of any other dates.

- What is the expected number of hours Nat waits for Pat to arrive?
- What is the expected duration of any particular date?
- What is the expected number of dates they will have before breaking up?

Problem 14. A retired professor comes to the office at a time which is uniformly distributed between 9 a.m. and 1 p.m., performs a single task, and leaves when the task is completed. The duration of the task is exponentially distributed with parameter $\lambda(y) = 1/(5 - y)$, where y is the length of the time interval between 9 a.m. and the time of his arrival.

- What is the expected amount of time that the professor devotes to the task?
- What is the expected time at which the task is completed?
- The professor has a Ph.D. student who on a given day comes to see him at a time that is uniformly distributed between 9 a.m. and 5 p.m. If the student does not find the professor, he leaves and does not return. If he finds the professor, he spends an amount of time that is uniformly distributed between 0 and 1 hour. The professor will spend the same total amount of time on his task regardless of whether he is interrupted by the student. What is the expected amount of time that the professor will spend with the student and what is the expected time at which he will leave his office?

Problem 15. Consider a gambler who at each gamble either wins or loses his bet with probabilities p and $1 - p$, independently of earlier gambles. When $p > 1/2$, a popular gambling system, known as the Kelly strategy, is to always bet the fraction

$2p - 1$ of the current fortune. Assuming $p > 1/2$, compute the expected fortune after n gambles, starting with x units and employing the Kelly strategy.

Problem 16. Let X and Y be two random variables that are uniformly distributed over the triangle formed by the points $(0, 0)$, $(1, 0)$, and $(0, 2)$ (this is an asymmetric version of the PDF of Example 4.15). Calculate $\mathbf{E}[X]$ and $\mathbf{E}[Y]$ using the law of iterated expectations and a method similar to the one of Example 4.15.

Problem 17.* Let X and Y be independent random variables. Use the law of total variance to show that

$$\text{var}(XY) = (\mathbf{E}[X])^2 \text{var}(Y) + (\mathbf{E}[Y])^2 \text{var}(X) + \text{var}(X)\text{var}(Y).$$

Solution. Let $Z = XY$. The law of total variance yields

$$\text{var}(Z) = \text{var}(\mathbf{E}[Z | X]) + \mathbf{E}[\text{var}(Z | X)].$$

We have

$$\mathbf{E}[Z | X] = \mathbf{E}[XY | X] = X\mathbf{E}[Y],$$

so that

$$\text{var}(\mathbf{E}[Z | X]) = \text{var}(X\mathbf{E}[Y]) = (\mathbf{E}[Y])^2 \text{var}(X).$$

Furthermore,

$$\text{var}(Z | X) = \text{var}(XY | X) = X^2 \text{var}(Y | X) = X^2 \text{var}(Y),$$

so that

$$\mathbf{E}[\text{var}(Z | X)] = \mathbf{E}[X^2] \text{var}(Y) = (\mathbf{E}[X])^2 \text{var}(Y) + \text{var}(X)\text{var}(Y).$$

Combining the preceding relations, we obtain

$$\text{var}(XY) = (\mathbf{E}[X])^2 \text{var}(Y) + (\mathbf{E}[Y])^2 \text{var}(X) + \text{var}(X)\text{var}(Y).$$

SECTION 4.4. Sum of a Random Number of Independent Random Variables

Problem 18. At a certain time, the number of people that enter an elevator is a Poisson random variable with parameter λ . The weight of each person is independent of every other person's weight, and is uniformly distributed between 100 and 200 lbs. Let X_i be the fraction of 100 by which the i th person exceeds 100 lbs, e.g., if the 7th person weighs 175 lbs., then $X_7 = 0.75$. Let Y be the sum of the X_i .

- Find the transform associated with Y .
- Use the transform to compute the expected value of Y .
- Verify your answer to part (b) by using the law of iterated expectations.

Problem 19. Construct an example to show that the sum of a random number of independent normal random variables is not normal (even though a fixed sum is).

Problem 20. A motorist goes through 4 lights, each of which is found to be red with probability 1/2. The waiting times at each light are modeled as independent normal random variables with mean 1 minute and standard deviation 1/2 minute. Let X be the total waiting time at the red lights.

- Use the total probability theorem to find the PDF and the transform associated with X , and the probability that X exceeds 4 minutes. Is X normal?
- Find the transform associated with X by viewing X as a sum of a random number of random variables.

Problem 21.* Use transforms to show that the sum of a Poisson-distributed number of independent, identically distributed Bernoulli random variables is Poisson.

Solution. Let N be a Poisson-distributed random variable with parameter λ . Let X_i , $i = 1, \dots, N$, be independent Bernoulli random variables with success probability p , and let

$$L = X_1 + \dots + X_N$$

be the corresponding sum. The transform associated with L is found by starting with the transform associated with N , which is

$$M_N(s) = e^{\lambda(e^s - 1)},$$

and replacing each occurrence of e^s by the transform associated with X_i , which is

$$M_{X_i}(s) = 1 - p + pe^s.$$

We obtain

$$M_L(s) = e^{\lambda(1 - p + pe^s - 1)} = e^{\lambda p(e^s - 1)}.$$

This is the transform associated with a Poisson random variable with parameter λp .

SECTION 4.5. Covariance and Correlation

Problem 22. Consider four random variables, W , X , Y , Z , with

$$\mathbf{E}[W] = \mathbf{E}[X] = \mathbf{E}[Y] = \mathbf{E}[Z] = 0,$$

$$\text{var}(W) = \text{var}(X) = \text{var}(Y) = \text{var}(Z) = 1,$$

and assume that W , X , Y , Z are pairwise uncorrelated. Find the correlation coefficients $\rho(A, B)$ and $\rho(A, C)$, where $A = W + X$, $B = X + Y$, and $C = Y + Z$.

Problem 23. Suppose that X is a standard normal random variable.

- Calculate $\mathbf{E}[X^3]$ and $\mathbf{E}[X^4]$.
- Define a new random variable Y such that

$$Y = a + bX + cX^2.$$

Find the correlation coefficient $\rho(X, Y)$.

Problem 24.* Schwarz inequality. Show that if X and Y are random variables, we have

$$(\mathbf{E}[XY])^2 \leq \mathbf{E}[X^2] \mathbf{E}[Y^2].$$

Solution. We may assume that $\mathbf{E}[Y^2] \neq 0$; otherwise, we have $Y = 0$ with probability 1, and hence $\mathbf{E}[XY] = 0$, so the inequality holds. We have

$$\begin{aligned} 0 &\leq \mathbf{E} \left[\left(X - \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]} Y \right)^2 \right] \\ &= \mathbf{E} \left[X^2 - 2 \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]} XY + \frac{(\mathbf{E}[XY])^2}{(\mathbf{E}[Y^2])^2} Y^2 \right] \\ &= \mathbf{E}[X^2] - 2 \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]} \mathbf{E}[XY] + \frac{(\mathbf{E}[XY])^2}{(\mathbf{E}[Y^2])^2} \mathbf{E}[Y^2] \\ &= \mathbf{E}[X^2] - \frac{(\mathbf{E}[XY])^2}{\mathbf{E}[Y^2]}, \end{aligned}$$

i.e., $(\mathbf{E}[XY])^2 \leq \mathbf{E}[X^2] \mathbf{E}[Y^2]$.

Problem 25.* Correlation coefficient. Consider the correlation coefficient

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}$$

of two random variables X and Y that have positive variances. Show that:

- $|\rho(X, Y)| \leq 1$. *Hint:* Use the Schwarz inequality from the preceding problem.
- If $Y - \mathbf{E}[Y]$ is a positive (or negative) multiple of $X - \mathbf{E}[X]$, then $\rho(X, Y) = 1$ [or $\rho(X, Y) = -1$, respectively].
- If $\rho(X, Y) = 1$ [or $\rho(X, Y) = -1$], then, with probability 1, $Y - \mathbf{E}[Y]$ is a positive (or negative, respectively) multiple of $X - \mathbf{E}[X]$.

Solution. (a) Let $\tilde{X} = X - \mathbf{E}[X]$ and $\tilde{Y} = Y - \mathbf{E}[Y]$. Using the Schwarz inequality, we get

$$(\rho(X, Y))^2 = \frac{(\mathbf{E}[\tilde{X}\tilde{Y}])^2}{\mathbf{E}[\tilde{X}^2] \mathbf{E}[\tilde{Y}^2]} \leq 1,$$

and hence $|\rho(X, Y)| \leq 1$.

(b) If $\tilde{Y} = a\tilde{X}$, then

$$\rho(X, Y) = \frac{\mathbf{E}[\tilde{X}a\tilde{X}]}{\sqrt{\mathbf{E}[\tilde{X}^2] \mathbf{E}[(a\tilde{X})^2]}} = \frac{a}{|a|}.$$

(c) If $(\rho(X, Y))^2 = 1$, then

$$\begin{aligned} \mathbf{E} \left[\left(\tilde{X} - \frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]} \tilde{Y} \right)^2 \right] &= \mathbf{E} \left[\tilde{X}^2 - 2 \frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]} \tilde{X}\tilde{Y} + \frac{(\mathbf{E}[\tilde{X}\tilde{Y}])^2}{(\mathbf{E}[\tilde{Y}^2])^2} \tilde{Y}^2 \right] \\ &= \mathbf{E}[\tilde{X}^2] - 2 \frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]} \mathbf{E}[\tilde{X}\tilde{Y}] + \frac{(\mathbf{E}[\tilde{X}\tilde{Y}])^2}{(\mathbf{E}[\tilde{Y}^2])^2} \mathbf{E}[\tilde{Y}^2] \\ &= \mathbf{E}[\tilde{X}^2] - \frac{(\mathbf{E}[\tilde{X}\tilde{Y}])^2}{\mathbf{E}[\tilde{Y}^2]} \\ &= \mathbf{E}[\tilde{X}^2] (1 - (\rho(X, Y))^2) \\ &= 0. \end{aligned}$$

Thus, with probability 1, the random variable

$$\tilde{X} - \frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]} \tilde{Y}$$

is equal to zero. It follows that, with probability 1,

$$\tilde{X} = \frac{\mathbf{E}[\tilde{X}\tilde{Y}]}{\mathbf{E}[\tilde{Y}^2]} \tilde{Y} = \sqrt{\frac{\mathbf{E}[\tilde{X}^2]}{\mathbf{E}[\tilde{Y}^2]}} \rho(X, Y) \tilde{Y},$$

i.e., the sign of the constant ratio of \tilde{X} and \tilde{Y} is determined by the sign of $\rho(X, Y)$.

SECTION 4.6. Least Squares Estimation

Problem 26. A police radar always overestimates the speed of incoming cars by an amount that is uniformly distributed between 0 and 5 miles/hour. Assume that car speeds are uniformly distributed from 55 to 75 miles/hour. What is the least squares estimate of the car speed based on the radar's measurement?

Problem 27. Consider the random variables X and Y of Example 4.27. Find the linear least squares estimator of X given Y .

Problem 28.* Linear least squares estimate based on several measurements. Let X be a random variable with mean μ and variance v , and let Y_1, \dots, Y_n be measurements of the form

$$Y_i = X + W_i,$$

where the W_i are random variables with mean 0 and variance v_i , which represent measurement errors. We assume that the random variables X, W_1, \dots, W_n are independent. Show that the linear least squares estimator of X based on Y_1, \dots, Y_n is

$$\frac{(\mu/v) + \sum_{i=1}^n (Y_i/v_i)}{(1/v) + \sum_{i=1}^n (1/v_i)}.$$

Note: If either v is very large, or if the number of measurements is very large so that $1/v$ is negligible relative to $\sum_{i=1}^n (1/v_i)$, this estimator can be well approximated by

$$\frac{\sum_{i=1}^n (Y_i/v_i)}{\sum_{i=1}^n (1/v_i)},$$

which does not require knowledge of μ and v . If in addition all the v_i are equal (all the measurements are of equal “quality”), the estimator can be well approximated by $\sum_{i=1}^n Y_i/n$, the sample mean of the measurements.

Solution. We want to show that the choices of a_1, \dots, a_n, b that minimize the function

$$f(a_1, \dots, a_n, b) = \frac{1}{2} \mathbf{E}[(X - a_1 Y_1 - \dots - a_n Y_n - b)^2],$$

are

$$\begin{aligned} b^* &= \frac{\mu/v}{(1/v) + \sum_{i=1}^n (1/v_i)}, \\ a_j^* &= \frac{1/v_j}{(1/v) + \sum_{i=1}^n (1/v_i)}, \quad j = 1, \dots, n. \end{aligned}$$

To this end, it is sufficient to show that the partial derivatives of f , with respect to a_1, \dots, a_n, b , are all equal to 0 when evaluated at a_1^*, \dots, a_n^*, b^* .

By differentiating f , we have

$$\begin{aligned} \frac{\partial f}{\partial b} \bigg|_{a_i^*, b^*} &= \mathbf{E} \left[\left(\sum_{i=1}^n a_i^* - 1 \right) X + \sum_{i=1}^n a_i^* W_i + b^* \right], \\ \frac{\partial f}{\partial a_i} \bigg|_{a_i^*, b^*} &= \mathbf{E} \left[Y_i \left(\left(\sum_{i=1}^n a_i^* - 1 \right) X + \sum_{i=1}^n a_i^* W_i + b^* \right) \right]. \end{aligned}$$

From the expressions for b^* and a_i^* , we see that

$$\sum_{i=1}^n a_i^* - 1 = -\frac{b^*}{\mu}.$$

Using this equation and the facts

$$\mathbf{E}[X] = \mu, \quad \mathbf{E}[W_i] = 0,$$

it follows that

$$\frac{\partial f}{\partial b} \bigg|_{a_i^*, b^*} = \mathbf{E} \left[\left(-\frac{b^*}{\mu} \right) X + \sum_{i=1}^n a_i^* W_i + b^* \right] = 0.$$

Using, in addition, the equations

$$\mathbf{E}[Y_i(\mu - X)] = \mathbf{E}[(X - \mu + W_i + \mu)(\mu - X)] = -v,$$

$$\mathbf{E}[Y_i W_i] = \mathbf{E}[(X + W_i)W_i] = v_i, \quad \text{for all } i,$$

$$\mathbf{E}[Y_j W_i] = \mathbf{E}[(X + W_j)W_i] = 0, \quad \text{for all } i \text{ and } j \text{ with } i \neq j,$$

we obtain

$$\begin{aligned} \frac{\partial f}{\partial a_i} \Bigg|_{a_i^*, b^*} &= \mathbf{E} \left[Y_i \left(\left(-\frac{b^*}{\mu} \right) X + \sum_{i=1}^n a_i^* W_i + b^* \right) \right] \\ &= \mathbf{E} \left[Y_i \left((\mu - X) \frac{b^*}{\mu} + \sum_{i=1}^n a_i^* W_i \right) \right] \\ &= -v \frac{b^*}{\mu} + a_i^* v_i \\ &= 0, \end{aligned}$$

where the last equality holds in view of the definitions of a_i^* and b^* .

Problem 29.* Let X and Y be two random variables with positive variances.

(a) Let \hat{X}_L be the linear least squares estimator of X based on Y . Show that

$$\mathbf{E}[(X - \hat{X}_L)Y] = 0$$

and that the estimation error $X - \hat{X}_L$ is uncorrelated with Y .

(b) Let $\hat{X} = \mathbf{E}[X | Y]$ be the least squares estimator of X given Y . Show that

$$\mathbf{E}[(X - \hat{X})h(Y)] = 0,$$

for any function h .

(c) Is it true that the estimation error $X - \mathbf{E}[X | Y]$ is independent of Y ?

Solution. (a) We have

$$\hat{X}_L = \mathbf{E}[X] + \rho(X, Y) \frac{\sigma_X}{\sigma_Y} (Y - \mathbf{E}[Y]) = \mathbf{E}[X] + \frac{\text{cov}(X, Y)}{\sigma_Y^2} (Y - \mathbf{E}[Y]),$$

so

$$\begin{aligned} \mathbf{E}[(X - \hat{X}_L)Y] &= \mathbf{E} \left[XY - \left(\mathbf{E}[X] + \frac{\text{cov}(X, Y)}{\sigma_Y^2} (Y - \mathbf{E}[Y]) \right) Y \right] \\ &= \mathbf{E} \left[XY - \mathbf{E}[X] Y - \frac{\text{cov}(X, Y)}{\sigma_Y^2} (Y^2 - Y \mathbf{E}[Y]) \right] \\ &= \mathbf{E}[XY] - \mathbf{E}[X] \mathbf{E}[Y] - \frac{\text{cov}(X, Y) \mathbf{E}[Y^2]}{\sigma_Y^2} + \frac{\text{cov}(X, Y) (\mathbf{E}[Y])^2}{\sigma_Y^2} \\ &= \text{cov}(X, Y) \left(1 - \frac{\mathbf{E}[Y^2]}{\sigma_Y^2} + \frac{(\mathbf{E}[Y])^2}{\sigma_Y^2} \right) \\ &= \text{cov}(X, Y) \left(1 - \frac{\sigma_Y^2}{\sigma_Y^2} \right) \\ &= 0. \end{aligned}$$

To show that $\text{cov}(X - \hat{X}_L, Y) = 0$, we write

$$\text{cov}(X - \hat{X}_L, Y) = \mathbf{E}[(X - \hat{X}_L)Y] - \mathbf{E}[X - \hat{X}_L]\mathbf{E}[Y] = -\mathbf{E}[X - \hat{X}_L]\mathbf{E}[Y],$$

and

$$\begin{aligned}\mathbf{E}[X - \hat{X}_L] &= \mathbf{E}\left[X - \mathbf{E}[X] - \frac{\text{cov}(X, Y)}{\sigma_Y^2}(Y - \mathbf{E}[Y])\right] \\ &= \mathbf{E}[X] - \mathbf{E}[X] - \frac{\text{cov}(X, Y)}{\sigma_Y^2}\mathbf{E}[Y - \mathbf{E}[Y]] \\ &= 0,\end{aligned}$$

as desired.

(b) We have

$$\begin{aligned}\mathbf{E}[(X - \hat{X})h(Y)] &= \mathbf{E}[(X - \mathbf{E}[X|Y])h(Y)] \\ &= \mathbf{E}[Xh(Y)] - \mathbf{E}[\mathbf{E}[X|Y]h(Y)] \\ &= \mathbf{E}[Xh(Y)] - \mathbf{E}[\mathbf{E}[Xh(Y)|Y]] \\ &= \mathbf{E}[Xh(Y)] - \mathbf{E}[Xh(Y)] \\ &= 0.\end{aligned}$$

(c) The answer is no. For a counterexample, let X and Y be discrete random variables with the joint PMF

$$p_{X,Y}(x,y) = \begin{cases} 1/4, & \text{for } (x,y) = (1,0), (0,1), (-1,0), (0,-1) \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\mathbf{E}[X|Y=y] = 0$ for all possible values y , so $\mathbf{E}[X|Y] = 0$. Thus, we have $X - \mathbf{E}[X|Y] = X$. Since X and Y are not independent, $X - \mathbf{E}[X|Y]$ and Y are also not independent.

SECTION 4.7. The Bivariate Normal Distribution

Problem 30. Let X_1 and X_2 be independent standard normal random variables. Define the random variables Y_1 and Y_2 by

$$Y_1 = 2X_1 + X_2, \quad Y_2 = X_1 - X_2.$$

Find $\mathbf{E}[Y_1]$, $\mathbf{E}[Y_2]$, $\text{cov}(Y_1, Y_2)$, and the joint PDF f_{Y_1, Y_2} .

Problem 31. The random variables X and Y are described by a joint PDF of the form

$$f_{X,Y}(x,y) = ce^{-8x^2 - 6xy - 18y^2}.$$

Find the means, variances, and the correlation coefficient of X and Y . Also, find the value of the constant c .

Problem 32. Suppose that X and Y are independent normal random variables with the same variance. Show that $X - Y$ and $X + Y$ are independent.

Problem 33. The coordinates X and Y of a point are independent zero-mean normal random variables with common variance σ^2 . Given that the point is at a distance of at least c from the origin, find the conditional joint PDF of X and Y .

Problem 34.* Suppose that X and Y are jointly normal random variables. Show that

$$\mathbf{E}[X | Y] = \mathbf{E}[X] + \rho \frac{\sigma_X}{\sigma_Y} (Y - \mathbf{E}[Y]).$$

Hint: Consider the random variables $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ and use the result established in the text for the zero-mean case.

Solution. Let $\tilde{X} = X - \mathbf{E}[X]$ and $\tilde{Y} = Y - \mathbf{E}[Y]$. The random variables \tilde{X} and \tilde{Y} are jointly normal. This is because if X and Y are linear functions of two independent normal random variables U and V , then \tilde{X} and \tilde{Y} are also linear functions of U and V . Therefore, as established in the text,

$$\mathbf{E}[\tilde{X} | \tilde{Y}] = \rho(\tilde{X}, \tilde{Y}) \frac{\sigma_{\tilde{X}}}{\sigma_{\tilde{Y}}} \tilde{Y}.$$

Note that conditioning on \tilde{Y} is the same as conditioning on Y . Therefore,

$$\mathbf{E}[\tilde{X} | \tilde{Y}] = \mathbf{E}[\tilde{X} | Y] = \mathbf{E}[X | Y] - \mathbf{E}[X].$$

Since X and \tilde{X} only differ by a constant, we have $\sigma_{\tilde{X}} = \sigma_X$ and, similarly, $\sigma_{\tilde{Y}} = \sigma_Y$. Finally,

$$\text{cov}(\tilde{X}, \tilde{Y}) = \mathbf{E}[\tilde{X}\tilde{Y}] = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \text{cov}(X, Y),$$

from which it follows that $\rho(\tilde{X}, \tilde{Y}) = \rho(X, Y)$. The desired formula follows by substituting the above relations in the formula for $\mathbf{E}[\tilde{X} | \tilde{Y}]$.

Problem 35.*

- (a) Let X_1, X_2, \dots, X_n be independent identically distributed random variables and let $Y = X_1 + X_2 + \dots + X_n$. Show that

$$\mathbf{E}[X_1 | Y] = \frac{Y}{n}.$$

- (b) Let X and W be independent zero-mean normal random variables, with positive integer variances k and m , respectively. Use the result of part (a) to find $\mathbf{E}[X | X + W]$, and verify that this agrees with the conditional expectation formula for jointly normal random variables given in the text. *Hint:* Think of X and W as sums of independent random variables.

Solution. (a) By symmetry, we see that $\mathbf{E}[X_i | Y]$ is the same for all i . Furthermore,

$$\mathbf{E}[X_1 + \dots + X_n | Y] = \mathbf{E}[Y | Y] = Y.$$

Therefore, $\mathbf{E}[X_1 | Y] = Y/n$.

- (b) We can think of X and W as sums of independent standard normal random variables:

$$X = X_1 + \dots + X_k, \quad W = W_1 + \dots + W_m.$$

We identify Y with $X + W$ and use the result from part (a), to obtain

$$\mathbf{E}[X_i | X + W] = \frac{X + W}{k + m}.$$

Thus,

$$\mathbf{E}[X | X + W] = \mathbf{E}[X_1 + \cdots + X_k | X + W] = \frac{k}{k + m}(X + W).$$

This formula agrees with the formula derived in the text because

$$\rho(X, X + W) \frac{\sigma_X}{\sigma_{X+W}} = \frac{\text{cov}(X, X + W)}{\sigma_{X+W}^2} = \frac{k}{k + m}.$$

We have used here the property

$$\text{cov}(X, X + W) = \mathbf{E}[X(X + W)] = \mathbf{E}[X^2] = k.$$

*Stochastic Processes***Contents**

5.1. The Bernoulli Process	p. 273
5.2. The Poisson Process	p. 285
5.3. Summary and Discussion	p. 299
Problems	p. 301

A stochastic process is a mathematical model of a probabilistic experiment that evolves in time and generates a sequence of numerical values. For example, a stochastic process can be used to model:

- (a) the sequence of daily prices of a stock;
- (b) the sequence of scores in a football game;
- (c) the sequence of failure times of a machine;
- (d) the sequence of hourly traffic loads at a node of a communication network;
- (e) the sequence of radar measurements of the position of an airplane.

Each numerical value in the sequence is modeled by a random variable, so a stochastic process is simply a (finite or infinite) sequence of random variables and does not represent a major conceptual departure from our basic framework. We are still dealing with a single basic experiment that involves outcomes governed by a probability law, and random variables that inherit their probabilistic properties from that law.[†]

However, stochastic processes involve some change in emphasis over our earlier models. In particular:

- (a) We tend to focus on the **dependencies** in the sequence of values generated by the process. For example, how do future prices of a stock depend on past values?
- (b) We are often interested in **long-term averages**, involving the entire sequence of generated values. For example, what is the fraction of time that a machine is idle?
- (c) We sometimes wish to characterize the likelihood or frequency of certain **boundary events**. For example, what is the probability that within a given hour all circuits of some telephone system become simultaneously busy, or what is the frequency with which some buffer in a computer network overflows with data?

There is a wide variety of stochastic processes, but in this book, we will only discuss two major categories.

- (i) *Arrival-Type Processes*: Here, we are interested in occurrences that have the character of an “arrival,” such as message receptions at a receiver, job completions in a manufacturing cell, customer purchases at a store, etc. We will focus on models in which the interarrival times (the times between successive arrivals) are independent random variables. In Section 5.1, we consider the case where arrivals occur in discrete time and the interarrival

[†] Let us emphasize that all of the random variables arising in a stochastic process refer to a single and common experiment, and are therefore defined on a common sample space. The corresponding probability law can be specified explicitly or implicitly (in terms of its properties), provided that it determines unambiguously the joint CDF of any subset of the random variables involved.

times are geometrically distributed – this is the *Bernoulli process*. In Section 5.2, we consider the case where arrivals occur in continuous time and the interarrival times are exponentially distributed – this is the *Poisson process*.

- (ii) *Markov Processes*: Here, we are looking at experiments that evolve in time and in which the future evolution exhibits a probabilistic dependence on the past. As an example, the future daily prices of a stock are typically dependent on past prices. However, in a Markov process, we assume a very special type of dependence: the next value depends on past values only through the current value. There is a rich methodology that applies to such processes, and is the subject of Chapter 6.

5.1 THE BERNOULLI PROCESS

The Bernoulli process can be visualized as a sequence of independent coin tosses, where the probability of heads in each toss is a fixed number p in the range $0 < p < 1$. In general, the Bernoulli process consists of a sequence of Bernoulli trials, where each trial produces a 1 (a success) with probability p , and a 0 (a failure) with probability $1 - p$, independently of what happens in other trials.

Of course, coin tossing is just a paradigm for a broad range of contexts involving a sequence of independent binary outcomes. For example, a Bernoulli process is often used to model systems involving arrivals of customers or jobs at service centers. Here, time is discretized into periods, and a “success” at the k th trial is associated with the arrival of at least one customer at the service center during the k th period. We will often use the term “arrival” in place of “success” when this is justified by the context.

In a more formal description, we define the Bernoulli process as a sequence X_1, X_2, \dots of **independent** Bernoulli random variables X_i with

$$\begin{aligned}\mathbf{P}(X_i = 1) &= \mathbf{P}(\text{success at the } i\text{th trial}) = p, \\ \mathbf{P}(X_i = 0) &= \mathbf{P}(\text{failure at the } i\text{th trial}) = 1 - p,\end{aligned}$$

for each i .[†]

Given an arrival process, one is often interested in random variables such as the number of arrivals within a certain time period, or the time until the first

[†] Generalizing from the case of a finite number of random variables, the independence of an *infinite* sequence of random variables X_i is defined by the requirement that the random variables X_1, \dots, X_n be independent for any finite n . Intuitively, knowing the values of any finite subset of the random variables does not provide any new probabilistic information on the remaining random variables, and the conditional distribution of the latter is the same as the unconditional one.

arrival. For the case of a Bernoulli process, some answers are already available from earlier chapters. Here is a summary of the main facts.

Some Random Variables Associated with the Bernoulli Process and their Properties

- **The binomial with parameters p and n .** This is the number S of successes in n independent trials. Its PMF, mean, and variance are

$$p_S(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

$$\mathbf{E}[S] = np, \quad \text{var}(S) = np(1-p).$$

- **The geometric with parameter p .** This is the number T of trials up to (and including) the first success. Its PMF, mean, and variance are

$$p_T(t) = (1-p)^{t-1}p, \quad t = 1, 2, \dots,$$

$$\mathbf{E}[T] = \frac{1}{p}, \quad \text{var}(T) = \frac{1-p}{p^2}.$$

Independence and Memorylessness

The independence assumption underlying the Bernoulli process has important implications, including a memorylessness property (whatever has happened in past trials provides no information on the outcomes of future trials). An appreciation and intuitive understanding of such properties is very useful, and allows the quick solution of many problems that would be difficult with a more formal approach. In this subsection, we aim at developing the necessary intuition.

Let us start by considering random variables that are defined in terms of what happened in a certain set of trials. For example, the random variable $Z = (X_1 + X_3)X_6X_7$ is defined in terms of the first, third, sixth, and seventh trial. If we have two random variables of this type and if the two sets of trials that define them have no common element, then these random variables are independent. This is a generalization of a fact first seen in Chapter 2: if two random variables U and V are independent, then any two functions of them, $g(U)$ and $h(V)$, are also independent.

Example 5.1.

- (a) Let U be the number of successes in trials 1 to 5. Let V be the number of successes in trials 6 to 10. Then, U and V are independent. This is because

$U = X_1 + \dots + X_5$, $V = X_6 + \dots + X_{10}$, and the two collections $\{X_1, \dots, X_5\}$, $\{X_6, \dots, X_{10}\}$ have no common elements.

- (b) Let U (respectively, V) be the first odd (respectively, even) time i in which we have a success. Then, U is determined by the odd-time sequence X_1, X_3, \dots , whereas V is determined by the even-time sequence X_2, X_4, \dots . Since these two sequences have no common elements, U and V are independent.

Suppose now that a Bernoulli process has been running for n time steps, and that we have observed the values of X_1, X_2, \dots, X_n . We notice that the sequence of future trials X_{n+1}, X_{n+2}, \dots are independent Bernoulli trials and therefore form a Bernoulli process. In addition, these future trials are independent from the past ones. We conclude that starting from any given point in time, the future is also modeled by a Bernoulli process, which is independent of the past. We refer to this as the **fresh-start** property of the Bernoulli process.

Let us now recall that the time T until the first success is a geometric random variable. Suppose that we have been watching the process for n time steps and no success has been recorded. What can we say about the number $T-n$ of remaining trials until the first success? Since the future of the process (after time n) is independent of the past and constitutes a fresh-starting Bernoulli process, the number of future trials until the first success is described by the same geometric PMF. Mathematically, we have

$$\mathbf{P}(T - n = t \mid T > n) = (1 - p)^{t-1}p = \mathbf{P}(T = t), \quad t = 1, 2, \dots$$

This **memorylessness** property can also be derived algebraically, using the definition of conditional probabilities, but the argument given here is certainly more intuitive.

Memorylessness and the Fresh-Start Property of the Bernoulli Process

- For any given time n , the sequence of random variables X_{n+1}, X_{n+2}, \dots (the future of the process) is also a Bernoulli process, and is independent from X_1, \dots, X_n (the past of the process).
- Let n be a given time and let \bar{T} be the time of the first success after time n . Then, $\bar{T} - n$ has a geometric distribution with parameter p , and is independent of the random variables X_1, \dots, X_n .

According to the fresh-start property, if we start watching a Bernoulli process at a certain time n , what we see is indistinguishable from a Bernoulli process that has just started. It turns out that the same is true if we start watching the process at some *random* time N , as long as N is determined only by the past history of the process and does not convey any information on the future. As an example, consider a roulette wheel with each occurrence of red viewed

as a success. The sequence generated starting at some fixed spin (say, the 25th spin) is probabilistically indistinguishable from the sequence generated starting immediately after red occurs in five consecutive spins. In either case, the process starts fresh (although one can certainly find gamblers with alternative theories). The example that follows makes a similar argument, but more mathematically.

Example 5.2. Fresh-Start at a Random Time. Let N be the first time in which we have a success immediately following a previous success. (That is, N is the first i for which $X_{i-1} = X_i = 1$.) What is the probability $\mathbf{P}(X_{N+1} = X_{N+2} = 0)$ that there are no successes in the two trials that follow?

Intuitively, once the condition $X_{N-1} = X_N = 1$ is satisfied, and from then on, the future of the process consists of independent Bernoulli trials. Therefore, the probability of an event that refers to the future of the process is the same as in a fresh-starting Bernoulli process, so that $\mathbf{P}(X_{N+1} = X_{N+2} = 0) = (1 - p)^2$.

To provide a rigorous justification of the above argument, we note that the time N is a random variable, and by conditioning on the possible values of N , we have

$$\begin{aligned}\mathbf{P}(X_{N+1} = X_{N+2} = 0) &= \sum_{n=1}^{\infty} \mathbf{P}(N = n) \mathbf{P}(X_{N+1} = X_{N+2} = 0 \mid N = n) \\ &= \sum_{n=1}^{\infty} \mathbf{P}(N = n) \mathbf{P}(X_{n+1} = X_{n+2} = 0 \mid N = n).\end{aligned}$$

Because of the way that N was defined, the event $\{N = n\}$ occurs if and only if the values of X_1, \dots, X_n satisfy a certain condition. But these random variables are independent of X_{n+1} and X_{n+2} . Therefore,

$$\mathbf{P}(X_{n+1} = X_{n+2} = 0 \mid N = n) = \mathbf{P}(X_{n+1} = X_{n+2} = 0) = (1 - p)^2,$$

which leads to

$$\mathbf{P}(X_{N+1} = X_{N+2} = 0) = \sum_{n=1}^{\infty} \mathbf{P}(N = n) (1 - p)^2 = (1 - p)^2.$$

The next example illustrates the use of the fresh-start property to derive the distribution of certain random variables.

Example 5.3. A computer executes two types of tasks, priority and nonpriority, and operates in discrete time units (*slots*). A priority task arrives with probability p at the beginning of each slot, independently of other slots, and requires one full slot to complete. A nonpriority task is always available and is executed at a given slot if no priority task is available. In this context, it may be important to know the probabilistic properties of the time intervals available for nonpriority tasks.

With this in mind, let us call a slot *busy* if within this slot, the computer executes a priority task, and otherwise let us call it *idle*. We call a string of idle

(or busy) slots, flanked by busy (or idle, respectively) slots, an *idle period* (or *busy period*, respectively). Let us derive the PMF, mean, and variance of the following random variables (cf. Fig. 5.1):

- (a) T = the time index of the first idle slot;
- (b) B = the length (number of slots) of the first busy period;
- (c) I = the length of the first idle period.
- (d) Z = the number of slots after the first slot of the first busy period up to and including the first subsequent idle slot.

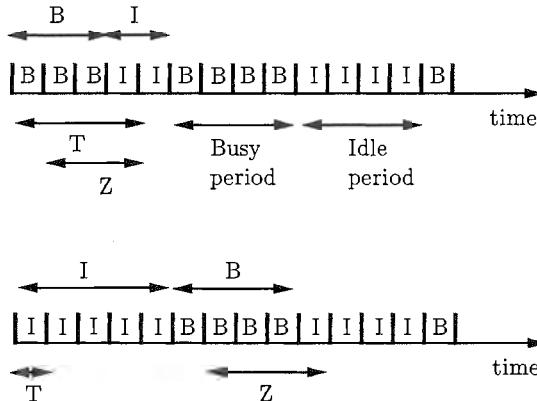


Figure 5.1: Illustration of random variables, and busy and idle periods in Example 5.3. In the top diagram, $T = 4$, $B = 3$, $I = 2$, and $Z = 3$. In the bottom diagram, $T = 1$, $I = 5$, $B = 4$, and $Z = 4$.

We recognize T as a geometrically distributed random variable with parameter $1 - p$. Its PMF is

$$p_T(k) = p^{k-1}(1-p), \quad k = 1, 2, \dots$$

Its mean and variance are

$$\mathbf{E}[T] = \frac{1}{1-p}, \quad \text{var}(T) = \frac{p}{(1-p)^2}.$$

Let us now consider the first busy period. It starts with the first busy slot, call it slot L . (In the top diagram in Fig. 5.1, $L = 1$; in the bottom diagram, $L = 6$.) The number Z of subsequent slots until (and including) the first subsequent idle slot has the same distribution as T , because the Bernoulli process starts fresh at time $L + 1$. We then notice that $Z = B$ and conclude that B has the same PMF as T .

If we reverse the roles of idle and busy slots, and interchange p with $1 - p$, we see that the length I of the first idle period has the same PMF as the time index of the first busy slot, so that

$$p_I(k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots, \quad \mathbf{E}[I] = \frac{1}{p}, \quad \text{var}(I) = \frac{1-p}{p^2}.$$

We finally note that the argument given here also works for the second, third, etc. busy (or idle) period. Thus, the PMFs calculated above apply to the i th busy and idle period, for any i .

Interarrival Times

An important random variable associated with the Bernoulli process is the time of the k th success (or arrival), which we denote by Y_k . A related random variable is the k th interarrival time, denoted by T_k . It is defined by

$$T_1 = Y_1, \quad T_k = Y_k - Y_{k-1}, \quad k = 2, 3, \dots,$$

and represents the number of trials following the $(k-1)$ st success until the next success. See Fig. 5.2 for an illustration, and also note that

$$Y_k = T_1 + T_2 + \dots + T_k.$$

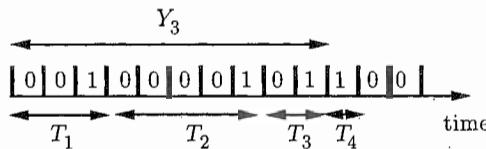


Figure 5.2: Illustration of interarrival times, where a 1 represents an arrival. In this example, $T_1 = 3$, $T_2 = 5$, $T_3 = 2$, $T_4 = 1$. Furthermore, $Y_1 = 3$, $Y_2 = 8$, $Y_3 = 10$, $Y_4 = 11$.

We have already seen that the time T_1 until the first success is a geometric random variable with parameter p . Having had a success at time T_1 , the future is a fresh-starting Bernoulli process. Thus, the number of trials T_2 until the next success has the same geometric PMF. Furthermore, past trials (up to and including time T_1) are independent of future trials (from time $T_1 + 1$ onward). Since T_2 is determined exclusively by what happens in these future trials, we see that T_2 is independent of T_1 . Continuing similarly, we conclude that the random variables T_1, T_2, T_3, \dots are independent and all have the same geometric distribution.

This important observation leads to an alternative, but equivalent way of describing the Bernoulli process, which is sometimes more convenient to work with.

Alternative Description of the Bernoulli Process

1. Start with a sequence of independent geometric random variables T_1, T_2, \dots , with common parameter p , and let these stand for the interarrival times.
2. Record a success (or arrival) at times $T_1, T_1 + T_2, T_1 + T_2 + T_3$, etc.

Example 5.4. It has been observed that after a rainy day, the number of days until it rains again is geometrically distributed with parameter p , independently of the past. Find the probability that it rains on both the 5th and the 8th day of the month.

If we attempt to approach this problem by manipulating the geometric PMFs in the problem statement, the solution is quite tedious. However, if we view rainy days as “arrivals,” we notice that the description of the weather conforms to the alternative description of the Bernoulli process given above. Therefore, any given day is rainy with probability p , independent of other days. In particular, the probability that days 5 and 8 are rainy is equal to p^2 .

The k th Arrival Time

The time Y_k of the k th success (or arrival) is equal to the sum $Y_k = T_1 + T_2 + \dots + T_k$ of k independent identically distributed geometric random variables. This allows us to derive formulas for the mean, variance, and PMF of Y_k , which are given in the table that follows.

Properties of the k th Arrival Time

- The k th arrival time is equal to the sum of the first k interarrival times

$$Y_k = T_1 + T_2 + \dots + T_k,$$

and the latter are independent geometric random variables with common parameter p .

- The mean and variance of Y_k are given by

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1] + \dots + \mathbf{E}[T_k] = \frac{k}{p},$$

$$\text{var}(Y_k) = \text{var}(T_1) + \dots + \text{var}(T_k) = \frac{k(1-p)}{p^2}.$$

- The PMF of Y_k is given by

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \quad t = k, k+1, \dots,$$

and is known as the **Pascal PMF of order k** .

To verify the formula for the PMF of Y_k , we first note that Y_k cannot be smaller than k . For $t \geq k$, we observe that the event $\{Y_k = t\}$ (the k th success comes at time t) will occur if and only if both of the following two events A and B occur:

- event A : trial t is a success;
- event B : exactly $k-1$ successes occur in the first $t-1$ trials.

The probabilities of these two events are

$$\mathbf{P}(A) = p,$$

and

$$\mathbf{P}(B) = \binom{t-1}{k-1} p^{k-1} (1-p)^{t-k},$$

respectively. In addition, these two events are independent (whether trial t is a success or not is independent of what happened in the first $t-1$ trials). Therefore,

$$p_{Y_k}(t) = \mathbf{P}(Y_k = t) = \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) = \binom{t-1}{k-1} p^k (1-p)^{t-k},$$

as claimed.

Example 5.5. In each minute of basketball play, Alicia commits a single foul with probability p and no foul with probability $1-p$. The number of fouls in different minutes are assumed to be independent. Alicia will foul out of the game once she commits her sixth foul, and will play 30 minutes if she does not foul out. What is the PMF of Alicia's playing time?

We model fouls as a Bernoulli process with parameter p . Alicia's playing time Z is equal to Y_6 , the time until the sixth foul, except if Y_6 is larger than 30, in which case, her playing time is 30; that is, $Z = \min\{Y_6, 30\}$. The random variable Y_6 has a Pascal PMF of order 6, which is given by

$$p_{Y_6}(t) = \binom{t-1}{5} p^6 (1-p)^{t-6}, \quad t = 6, 7, \dots$$

To determine the PMF $p_Z(z)$ of Z , we first consider the case where z is between 6 and 29. For z in this range, we have

$$p_Z(z) = \mathbf{P}(Z = z) = \mathbf{P}(Y_6 = z) = \binom{z-1}{5} p^6 (1-p)^{z-6}, \quad z = 6, 7, \dots, 29.$$

The probability that $Z = 30$ is then determined from

$$p_Z(30) = 1 - \sum_{z=6}^{29} p_Z(z).$$

Splitting and Merging of Bernoulli Processes

Starting with a Bernoulli process in which there is a probability p of an arrival at each time, consider **splitting** it as follows. Whenever there is an arrival, we choose to either keep it (with probability q), or to discard it (with probability $1 - q$); see Fig. 5.3. Assume that the decisions to keep or discard are independent for different arrivals. If we focus on the process of arrivals that are kept, we see that it is a Bernoulli process: in each time slot, there is a probability pq of a kept arrival, independently of what happens in other slots. For the same reason, the process of discarded arrivals is also a Bernoulli process, with a probability of a discarded arrival at each time slot equal to $p(1 - q)$.

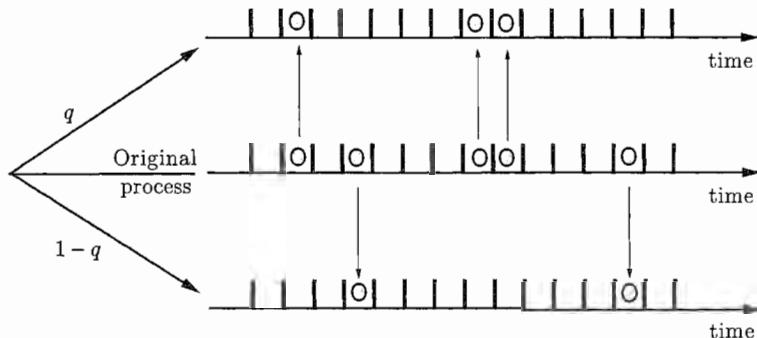


Figure 5.3: Splitting of a Bernoulli process.

In a reverse situation, we start with two *independent* Bernoulli processes (with parameters p and q , respectively) and **merge** them into a single process, as follows. An arrival is recorded in the merged process if and only if there is an arrival in at least one of the two original processes, which happens with probability $p + q - pq$ [one minus the probability $(1 - p)(1 - q)$ of no arrival in either process]. Since different time slots in either of the original processes are independent, different slots in the merged process are also independent. Thus, the merged process is Bernoulli, with success probability $p + q - pq$ at each time step; see Fig. 5.4.

Splitting and merging of Bernoulli (or other) arrival processes arises in many contexts. For example, a two-machine work center may see a stream of arriving parts to be processed and split them by sending each part to a randomly

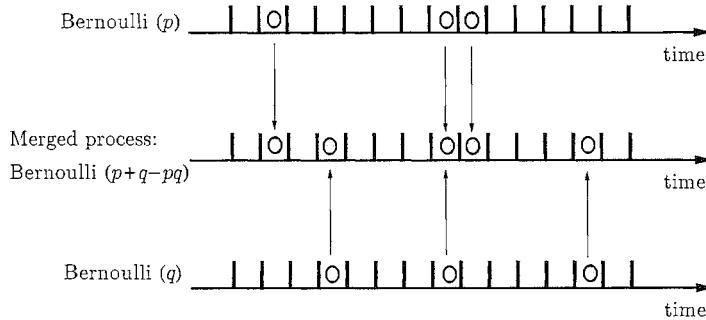


Figure 5.4: Merging of independent Bernoulli processes.

chosen machine. Conversely, a machine may be faced with arrivals of different types that can be merged into a single arrival stream.

The Poisson Approximation to the Binomial

The number of successes in n independent Bernoulli trials is a binomial random variable with parameters n and p , and its mean is np . In this subsection, we concentrate on the special case where n is large but p is small, so that the mean np has a moderate value. A situation of this type arises when one passes from discrete to continuous time, a theme to be picked up in the next section. For some examples, think of the number of airplane accidents on any given day: there is a large number n of trials (airplane flights), but each one has a very small probability p of being involved in an accident. Or think of counting the number of typos in a book: there is a large number of words, but a very small probability of misspelling any single one.

Mathematically, we can address situations of this kind, by letting n grow while simultaneously decreasing p , in a manner that keeps the product np at a constant value λ . In the limit, it turns out that the formula for the binomial PMF simplifies to the Poisson PMF. A precise statement is provided next, together with a reminder of some of the properties of the Poisson PMF that were derived in earlier chapters.

Poisson Approximation to the Binomial

- A Poisson random variable Z with parameter λ takes nonnegative integer values and is described by the PMF

$$p_Z(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

Its mean and variance are given by

$$\mathbf{E}[Z] = \lambda, \quad \text{var}(Z) = \lambda.$$

- For any fixed nonnegative integer k , the binomial probability

$$p_S(k) = \frac{n!}{(n-k)!k!} \cdot p^k(1-p)^{n-k}$$

converges to $p_Z(k)$, when we take the limit as $n \rightarrow \infty$ and $p = \lambda/n$, while keeping λ constant.

- In general, the Poisson PMF is a good approximation to the binomial as long as $\lambda = np$, n is very large, and p is very small.

To verify the validity of the Poisson approximation, we let $p = \lambda/n$ and note that

$$\begin{aligned} p_S(k) &= \frac{n!}{(n-k)!k!} \cdot p^k(1-p)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \frac{\lambda^k}{n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n}{n} \cdot \frac{(n-1)}{n} \cdots \frac{(n-k+1)}{n} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}. \end{aligned}$$

Let us focus on a fixed k and let $n \rightarrow \infty$. Each one of the ratios $(n-1)/n$, $(n-2)/n, \dots, (n-k+1)/n$ converges to 1. Furthermore,[†]

$$\left(1 - \frac{\lambda}{n}\right)^{-k} \rightarrow 1, \quad \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}.$$

We conclude that for each fixed k , and as $n \rightarrow \infty$, we have

$$p_S(k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

Example 5.6. As a rule of thumb, the Poisson/binomial approximation

$$e^{-\lambda} \frac{\lambda^k}{k!} \approx \frac{n!}{(n-k)!k!} \cdot p^k(1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

[†] We are using here, the well known formula $\lim_{x \rightarrow \infty} (1 - \frac{1}{x})^x = e^{-1}$. Letting $x = n/\lambda$, we have $\lim_{n \rightarrow \infty} (1 - \frac{\lambda}{n})^{n/\lambda} = e^{-1}$, from which $\lim_{n \rightarrow \infty} (1 - \frac{\lambda}{n})^n = e^{-\lambda}$.

is valid to several decimal places if $n \geq 100$, $p \leq 0.01$, and $\lambda = np$. To check this, consider the following.

Gary Kasparov, a world chess champion, plays against 100 amateurs in a large simultaneous exhibition. It has been estimated from past experience that Kasparov wins in such exhibitions 99% of his games on the average (in precise probabilistic terms, we assume that he wins each game with probability 0.99, independently of other games). What are the probabilities that he will win 100 games, 98 games, 95 games, and 90 games?

We model the number of games X that Kasparov does *not* win as a binomial random variable with parameters $n = 100$ and $p = 0.01$. Thus the probabilities that he will win 100 games, 98, 95 games, and 90 games are

$$p_X(0) = (1 - 0.01)^{100} = 0.366,$$

$$p_X(2) = \frac{100!}{98! 2!} \cdot 0.01^2 (1 - 0.01)^{98} = 0.185,$$

$$p_X(5) = \frac{100!}{95! 5!} \cdot 0.01^5 (1 - 0.01)^{95} = 0.00290,$$

$$p_X(10) = \frac{100!}{90! 10!} \cdot 0.01^{10} (1 - 0.01)^{90} = 7.006 \cdot 10^{-8},$$

respectively. Now let us check the corresponding Poisson approximations with $\lambda = 100 \cdot 0.01 = 1$. They are:

$$p_Z(0) = e^{-1} \frac{1}{0!} = 0.368,$$

$$p_Z(2) = e^{-1} \frac{1}{2!} = 0.184,$$

$$p_Z(5) = e^{-1} \frac{1}{5!} = 0.00306,$$

$$p_Z(10) = e^{-1} \frac{1}{10!} = 1.001 \cdot 10^{-8}.$$

By comparing the binomial PMF values $p_X(k)$ with their Poisson approximations $p_Z(k)$, we see that there is close agreement.

Suppose now that Kasparov plays simultaneously against just 5 opponents, who are, however, stronger so that his probability of a win per game is 0.9. Here are the binomial probabilities $p_X(k)$ for $n = 5$ and $p = 0.1$, and the corresponding Poisson approximations $p_Z(k)$ for $\lambda = np = 0.5$:

k	0	1	2	3	4	5
$p_X(k)$	0.590	0.328	0.0729	0.0081	0.00045	0.00001
$p_Z(k)$	0.605	0.303	0.0758	0.0126	0.0016	0.00016

We see that the approximation, while not poor, is considerably less accurate than in the case where $n = 100$ and $p = 0.01$.

Example 5.7. A packet consisting of a string of n symbols is transmitted over a noisy channel. Each symbol has probability $p = 0.0001$ of being transmitted in error, independently of errors in the other symbols. How small should n be in order for the probability of incorrect transmission (at least one symbol in error) to be less than 0.001?

Each symbol transmission is viewed as an independent Bernoulli trial. Thus, the probability of a positive number S of errors in the packet is

$$1 - \mathbf{P}(S = 0) = 1 - (1 - p)^n.$$

For this probability to be less than 0.001, we must have $1 - (1 - 0.0001)^n < 0.001$ or

$$n < \frac{\ln 0.999}{\ln 0.9999} = 10.0045.$$

We can also use the Poisson approximation for $\mathbf{P}(S = 0)$, which is $e^{-\lambda}$ with $\lambda = np = 0.0001 \cdot n$, and obtain the condition $1 - e^{-0.0001 \cdot n} < 0.001$, which leads to

$$n < \frac{-\ln 0.999}{0.0001} = 10.005.$$

Given that n must be integer, both methods lead to the same conclusion that n can be at most 10.

5.2 THE POISSON PROCESS

The Poisson process is a continuous-time analog of the Bernoulli process and applies to situations where there is no natural way of dividing time into discrete periods.

To see the need for a continuous-time version of the Bernoulli process, let us consider a possible model of traffic accidents within a city. We can start by discretizing time into one-minute periods and record a “success” during every minute in which there is at least one traffic accident. Assuming the traffic intensity to be constant over time, the probability of an accident should be the same during each period. Under the additional (and quite plausible) assumption that different time periods are independent, the sequence of successes becomes a Bernoulli process. Note that in real life, two or more accidents during the same one-minute interval are certainly possible, but the Bernoulli process model does not keep track of the exact number of accidents. In particular, it does not allow us to calculate the expected number of accidents within a given period.

One way around this difficulty is to choose the length of a time period to be very small, so that the probability of two or more accidents becomes negligible. But how small should it be? A second? A millisecond? Instead of answering this question, it is preferable to consider a limiting situation where the length of the time period becomes zero, and work with a continuous-time model.

We consider an arrival process that evolves in continuous time, in the sense that any real number t is a possible arrival time. We define

$$P(k, \tau) = \mathbf{P}(\text{there are exactly } k \text{ arrivals during an interval of length } \tau),$$

and assume that this probability is the same for all intervals of the same length τ . We also introduce a positive parameter λ , called the **arrival rate** or **intensity** of the process, for reasons that will soon become apparent.

Definition of the Poisson Process

An arrival process is called a Poisson process with rate λ if it has the following properties:

- (a) **(Time-homogeneity)** The probability $P(k, \tau)$ of k arrivals is the same for all intervals of the same length τ .
- (b) **(Independence)** The number of arrivals during a particular interval is independent of the history of arrivals outside this interval.
- (c) **(Small interval probabilities)** The probabilities $P(k, \tau)$ satisfy

$$P(0, \tau) = 1 - \lambda\tau + o(\tau),$$

$$P(1, \tau) = \lambda\tau + o_1(\tau),$$

$$P(k, \tau) = o_k(\tau), \quad \text{for } k = 2, 3, \dots$$

Here, $o(\tau)$ and $o_k(\tau)$ are functions of τ that satisfy

$$\lim_{\tau \rightarrow 0} \frac{o(\tau)}{\tau} = 0, \quad \lim_{\tau \rightarrow 0} \frac{o_k(\tau)}{\tau} = 0.$$

The first property states that arrivals are “equally likely” at all times. The arrivals during any time interval of length τ are statistically the same, i.e., they obey the same probability law. This is a counterpart to the assumption that the success probability p in a Bernoulli process is the same for all trials.

To interpret the second property, consider a particular interval $[t, t']$, of length $t' - t$. The unconditional probability of k arrivals during that interval is $P(k, t' - t)$. Suppose now that we are given complete or partial information on the arrivals outside this interval. Property (b) states that this information is irrelevant: the conditional probability of k arrivals during $[t, t']$ remains equal to the unconditional probability $P(k, t' - t)$. This property is analogous to the independence of trials in a Bernoulli process.

The third property is critical. The $o(\tau)$ and $o_k(\tau)$ terms are meant to be negligible in comparison to τ , when the interval length τ is very small. They can be thought of as the $O(\tau^2)$ terms in a Taylor series expansion of $P(k, \tau)$. Thus,

for small τ , the probability of a single arrival is roughly $\lambda\tau$, plus a negligible term. Similarly, for small τ , the probability of zero arrivals is roughly $1 - \lambda\tau$. Finally, the probability of two or more arrivals is negligible in comparison to $P(1, \tau)$, as τ becomes smaller.

Number of periods:	Probability of success per period:	Expected number of arrivals:
$n = \tau/\delta$	$p = \lambda\delta$	$np = \lambda\tau$

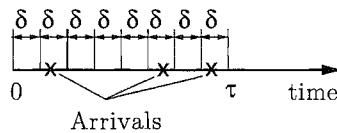


Figure 5.5: Bernoulli approximation of the Poisson process over an interval of length τ .

Number of Arrivals in an Interval

We will now derive some probability distributions associated with the arrivals in a Poisson process. We first use the connection with the Bernoulli process to obtain the PMF of the number of arrivals in a given time interval.

Let us consider a fixed time interval of length τ , and partition it into τ/δ periods of length δ , where δ is a very small number; see Fig. 5.5. The probability of more than two arrivals during any period can be neglected, because of property (c) and the preceding discussion. Different periods are independent, by property (b). Furthermore, each period has one arrival with probability approximately equal to $\lambda\delta$, or zero arrivals with probability approximately equal to $1 - \lambda\delta$. Therefore, the process being studied can be approximated by a Bernoulli process, with the approximation becoming more and more accurate as δ becomes smaller.

The probability $P(k, \tau)$ of k arrivals in time τ is approximately the same as the (binomial) probability of k successes in $n = \tau/\delta$ independent Bernoulli trials with success probability $p = \lambda\delta$ at each trial. While keeping the length τ of the interval fixed, we let the period length δ decrease to zero. We then note that the number n of periods goes to infinity, while the product np remains constant and equal to $\lambda\tau$. Under these circumstances, we saw in the previous section that the binomial PMF converges to a Poisson PMF with parameter $\lambda\tau$. We are then led to the important conclusion that

$$P(k, \tau) = e^{-\lambda\tau} \frac{(\lambda\tau)^k}{k!}, \quad k = 0, 1, \dots$$

Note that a Taylor series expansion of $e^{-\lambda\tau}$ yields

$$P(0, \tau) = e^{-\lambda\tau} = 1 - \lambda\tau + o(\tau),$$

$$P(1, \tau) = \lambda\tau e^{-\lambda\tau} = \lambda\tau - \lambda^2\tau^2 + O(\tau^3) = \lambda\tau + o_1(\tau),$$

consistently with property (c).

Using our earlier formulas for the mean and variance of the Poisson PMF, as presented in Chapter 2, we obtain

$$\mathbf{E}[N_\tau] = \lambda\tau, \quad \text{var}(N_\tau) = \lambda\tau,$$

where N_τ denotes the number of arrivals during a time interval of length τ . These formulas are hardly surprising, since we are dealing with the limit of a binomial PMF with parameters $n = \tau/\delta$, $p = \lambda\delta$, mean $np = \lambda\tau$, and variance $np(1-p) \approx np = \lambda\tau$.

Let us now derive the probability law for the time T of the first arrival, assuming that the process starts at time zero. Note that we have $T > t$ if and only if there are no arrivals during the interval $[0, t]$. Therefore,

$$F_T(t) = \mathbf{P}(T \leq t) = 1 - \mathbf{P}(T > t) = 1 - P(0, t) = 1 - e^{-\lambda t}, \quad t \geq 0.$$

We then differentiate the CDF $F_T(t)$ of T , and obtain the PDF formula

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0,$$

which shows that the time until the first arrival is exponentially distributed with parameter λ . We summarize this discussion in the table that follows. See also Fig. 5.6.

Random Variables Associated with the Poisson Process and their Properties

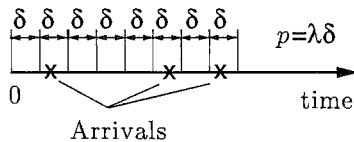
- **The Poisson with parameter $\lambda\tau$.** This is the number N_τ of arrivals in a Poisson process with rate λ , over an interval of length τ . Its PMF, mean, and variance are

$$p_{N_\tau}(k) = P(k, \tau) = e^{-\lambda\tau} \frac{(\lambda\tau)^k}{k!}, \quad k = 0, 1, \dots,$$

$$\mathbf{E}[N_\tau] = \lambda\tau, \quad \text{var}(N_\tau) = \lambda\tau.$$

- **The exponential with parameter λ .** This is the time T until the first arrival. Its PDF, mean, and variance are

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \quad \mathbf{E}[T] = \frac{1}{\lambda}, \quad \text{var}(T) = \frac{1}{\lambda^2}.$$



	POISSON	BERNOULLI
Times of Arrival	Continuous	Discrete
PMF of # of Arrivals	Poisson	Binomial
Interarrival Time CDF	Exponential	Geometric
Arrival Rate	$\lambda/\text{unit time}$	$p/\text{per trial}$

Figure 5.6: View of the Bernoulli process as the discrete-time version of the Poisson process. We discretize time in small intervals δ and associate each interval with a Bernoulli trial whose parameter is $p = \lambda\delta$. The table summarizes some of the basic correspondences.

Example 5.8. You get email according to a Poisson process at a rate of $\lambda = 0.2$ messages per hour. You check your email every hour. What is the probability of finding 0 and 1 new messages?

These probabilities can be found using the Poisson PMF $e^{-\lambda\tau}(\lambda\tau)^k/k!$, with $\tau = 1$, and $k = 0$ or $k = 1$:

$$P(0, 1) = e^{-0.2} = 0.819, \quad P(1, 1) = 0.2 \cdot e^{-0.2} = 0.164.$$

Suppose that you have not checked your email for a whole day. What is the probability of finding no new messages? We use again the Poisson PMF and obtain

$$P(0, 24) = e^{-0.2 \cdot 24} = 0.0083.$$

Alternatively, we can argue that the event of no messages in a 24-hour period is the intersection of the events of no messages during each of 24 hours. The latter events are independent and the probability of each is $P(0, 1) = e^{-0.2}$, so

$$P(0, 24) = (P(0, 1))^{24} = (e^{-0.2})^{24} = 0.0083,$$

which is consistent with the preceding calculation method.

Example 5.9. The Sum of Independent Poisson Random Variables is Poisson. Arrivals of customers at the local supermarket are modeled by a Poisson

process with a rate of $\lambda = 10$ customers per minute. Let M be the number of customers arriving between 9:00 and 9:10. Also, let N be the number of customers arriving between 9:30 and 9:35. What is the distribution of $M + N$?

We notice that M is Poisson with parameter $\mu = 10 \cdot 10 = 100$ and N is Poisson with parameter $\nu = 10 \cdot 5 = 50$. Furthermore, M and N are independent. As shown in Section 4.1, using transforms, $M + N$ is Poisson with parameter $\mu + \nu = 150$. We will now proceed to derive the same result in a more direct and intuitive manner.

Let \tilde{N} be the number of customers that arrive between 9:10 and 9:15. Note that \tilde{N} has the same distribution as N (Poisson with parameter 50). Furthermore, \tilde{N} is also independent of M . Thus, the distribution of $M + N$ is the same as the distribution of $M + \tilde{N}$. But $M + \tilde{N}$ is the number of arrivals during an interval of length 15, and has therefore a Poisson distribution with parameter $10 \cdot 15 = 150$.

This example makes a point that is valid in general. The probability of k arrivals during a *set* of times of total length τ is always given by $P(k, \tau)$, even if that set is not an interval. (In this example, we dealt with the set $[9:00, 9:10] \cup [9:30, 9:35]$, of total length 15.)

Independence and Memorylessness

The Poisson process has several properties that parallel those of the Bernoulli process, including the independence of nonoverlapping time sets, a fresh-start property, and the memorylessness of the interarrival time distribution. Given that the Poisson process can be viewed as a limiting case of a Bernoulli process, the fact that it inherits the qualitative properties of the latter should be hardly surprising.

Memorylessness and the Fresh-Start Property of the Poisson Process

- For any given time $t > 0$, the history of the process after time t is also a Poisson process, and is independent from the history of the process until time t .
- Let t be a given time and let \bar{T} be the time of the first arrival after time t . Then, $\bar{T} - t$ has an exponential distribution with parameter λ , and is independent of the history of the process until time t .

The fresh-start property is established by observing that the portion of the process that starts at time t satisfies the properties required by the definition of the Poisson process. The independence of the future from the past is a direct consequence of the independence assumption in the definition of the Poisson process. Finally, the fact that $\bar{T} - t$ has the same exponential distribution can be verified by noting that

$$\mathbf{P}(\bar{T} - t > s) = \mathbf{P}(0 \text{ arrivals during } [t, t + s]) = P(0, s) = e^{-\lambda s}.$$

The following is an example of reasoning based on the memoryless property.

Example 5.10. You and your partner go to a tennis court, and have to wait until the players occupying the court finish playing. Assume (somewhat unrealistically) that their playing time has an exponential PDF. Then the PDF of your waiting time (equivalently, their remaining playing time) also has the same exponential PDF, regardless of when they started playing.

Example 5.11. When you enter the bank, you find that all three tellers are busy serving other customers, and there are no other customers in queue. Assume that the service times for you and for each of the customers being served are independent identically distributed exponential random variables. What is the probability that you will be the last to leave?

The answer is 1/3. To see this, focus at the moment when you start service with one of the tellers. Then, the remaining time of each of the other two customers being served, as well as your own remaining time, have the same PDF. Therefore, you and the other two customers have equal probability 1/3 of being the last to leave.

Interarrival Times

An important random variable associated with a Poisson process that starts at time 0, is the time of the k th arrival, which we denote by Y_k . A related random variable is the k th interarrival time, denoted by T_k . It is defined by

$$T_1 = Y_1, \quad T_k = Y_k - Y_{k-1}, \quad k = 2, 3, \dots$$

and represents the amount of time between the $(k-1)$ st and the k th arrival. Note that

$$Y_k = T_1 + T_2 + \dots + T_k.$$

We have already seen that the time T_1 until the first arrival is an exponential random variable with parameter λ . Starting from the time T_1 of the first arrival, the future is a fresh-starting Poisson process.[†] Thus, the time until the next arrival has the same exponential PDF. Furthermore, the past of the process (up to time T_1) is independent of the future (after time T_1). Since T_2 is determined exclusively by what happens in the future, we see that T_2 is independent of T_1 . Continuing similarly, we conclude that the random variables T_1, T_2, T_3, \dots are independent and all have the same exponential distribution.

[†] We are stating here that the Poisson process starts afresh at the *random* time T_1 . This statement is a bit stronger than what was claimed earlier (fresh-start at deterministic times), but is quite intuitive, and can be formally justified using an argument analogous to the one used in Example 5.2, by conditioning on all possible values of the random variable T_1 .

This important observation leads to an alternative, but equivalent, way of describing the Poisson process.†

Alternative Description of the Poisson Process

1. Start with a sequence of independent exponential random variables T_1, T_2, \dots , with common parameter λ , and let these represent the interarrival times.
2. Record an arrival at times $T_1, T_1 + T_2, T_1 + T_2 + T_3$, etc.

The k th Arrival Time

The time Y_k of the k th arrival is equal to the sum $Y_k = T_1 + T_2 + \dots + T_k$ of k independent identically distributed exponential random variables. This allows us to derive formulas for the mean, variance, and PDF of Y_k , which are given in the table that follows.

Properties of the k th Arrival Time

- The k th arrival time is equal to the sum of the first k interarrival times

$$Y_k = T_1 + T_2 + \dots + T_k,$$

and the latter are independent exponential random variables with common parameter λ .

- The mean and variance of Y_k are given by

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1] + \dots + \mathbf{E}[T_k] = \frac{k}{\lambda},$$

$$\text{var}(Y_k) = \text{var}(T_1) + \dots + \text{var}(T_k) = \frac{k}{\lambda^2}.$$

- The PDF of Y_k is given by

† In our original definition, a process was called Poisson if it possessed certain properties. The reader may have noticed that we did not establish that a process with the required properties exists. In an alternative line of development, we can use the constructive description given here: we start with a sequence of independent, exponentially distributed interarrival times, from which the arrival times are completely determined. With this definition, it is then possible to establish that the process satisfies all of the properties that were postulated in our original Poisson process definition.

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0,$$

and is known as the **Erlang PDF of order k** .

To evaluate the PDF f_{Y_k} of Y_k , we can argue that for a small δ , the product $\delta \cdot f_{Y_k}(y)$ is the probability that the k th arrival occurs between times y and $y + \delta$.[†] When δ is very small, the probability of more than one arrival during the interval $[y, y + \delta]$ is negligible. Thus, the k th arrival occurs between y and $y + \delta$ if and only if the following two events A and B occur:

- (a) event A : there is an arrival during the interval $[y, y + \delta]$;
- (b) event B : there are exactly $k - 1$ arrivals before time y .

The probabilities of these two events are

$$\mathbf{P}(A) \approx \lambda\delta, \quad \text{and} \quad \mathbf{P}(B) = P(k-1, y) = \frac{\lambda^{k-1} y^{k-1} e^{-\lambda y}}{(k-1)!}.$$

Since A and B are independent, we have

$$\delta f_{Y_k}(y) \approx \mathbf{P}(y \leq Y_k \leq y + \delta) \approx \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \approx \lambda\delta \frac{\lambda^{k-1} y^{k-1} e^{-\lambda y}}{(k-1)!},$$

from which we obtain

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0.$$

Example 5.12. You call the IRS hotline and you are told that you are the 56th person in line, excluding the person currently being served. Callers depart

[†] For an alternative derivation that does not rely on approximation arguments, note that for a given $y \geq 0$, the event $\{Y_k \leq y\}$ is the same as the event

$$\{\text{number of arrivals in the interval } [0, y] \text{ is at least } k\}.$$

Thus the CDF of Y_k is given by

$$F_{Y_k}(y) = \mathbf{P}(Y_k \leq y) = \sum_{n=k}^{\infty} P(n, y) = 1 - \sum_{n=0}^{k-1} P(n, y) = 1 - \sum_{n=0}^{k-1} \frac{(\lambda y)^n e^{-\lambda y}}{n!}.$$

The PDF of Y_k can be obtained by differentiating the above expression, which by straightforward calculation yields the Erlang PDF formula

$$f_{Y_k}(y) = \frac{d}{dy} F_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}.$$

according to a Poisson process with a rate of $\lambda = 2$ per minute. How long will you have to wait on the average until your service starts, and what is the probability you will have to wait for more than an hour?

By the memoryless property, the remaining service time of the person currently being served is exponentially distributed with parameter 2. The service times of the 55 persons ahead of you are also exponential with the same parameter, and all of these random variables are independent. Thus, your waiting time in minutes, call it Y , is Erlang of order 56, and

$$\mathbf{E}[Y] = \frac{56}{\lambda} = 28.$$

The probability that you have to wait for more than an hour is given by the formula

$$\mathbf{P}(Y \geq 60) = \int_{60}^{\infty} \frac{\lambda^{56} y^{55} e^{-\lambda y}}{55!} dy.$$

Computing this probability is quite tedious. In Chapter 7, we will discuss a much easier way to compute approximately this probability. This is done using the central limit theorem, which allows us to approximate the CDF of the sum of a large number of independent random variables with a normal CDF, and then to calculate various probabilities of interest by using the normal tables.

Splitting and Merging of Poisson Processes

Similar to the case of a Bernoulli process, we can start with a Poisson process with rate λ and split it, as follows: each arrival is kept with probability p and discarded with probability $1-p$, independently of what happens to other arrivals. In the Bernoulli case, we saw that the result of the splitting was also a Bernoulli process. In the present context, the result of the splitting turns out to be a Poisson process with rate λp .

Alternatively, we can start with two independent Poisson processes, with rates λ_1 and λ_2 , and merge them by recording an arrival whenever an arrival occurs in either process. It turns out that the merged process is also Poisson with rate $\lambda_1 + \lambda_2$. Furthermore, any particular arrival of the merged process has probability $\lambda_1/(\lambda_1 + \lambda_2)$ of originating from the first process, and probability $\lambda_2/(\lambda_1 + \lambda_2)$ of originating from the second, independently of all other arrivals and their origins.

We discuss these properties in the context of some examples, and at the same time provide the arguments that establish their validity.

Example 5.13. Splitting of Poisson Processes. A packet that arrives at a node of a data network is either a local packet that is destined for that node (this happens with probability p), or else it is a transit packet that must be relayed to another node (this happens with probability $1 - p$). Packets arrive according to a Poisson process with rate λ , and each one is a local or transit packet independently

of other packets and of the arrival times. As stated above, the process of *local* packet arrivals is Poisson with rate λp . Let us see why.

We verify that the process of local packet arrivals satisfies the defining properties of a Poisson process. Since λ and p are constant (do not change with time), the first property (time homogeneity) clearly holds. Furthermore, there is no dependence between what happens in disjoint time intervals, verifying the second property. Finally, if we focus on a small interval of length δ , the probability of a local arrival is approximately the probability that there is a packet arrival, and that this turns out to be a local one, i.e., $\lambda\delta \cdot p$. In addition, the probability of two or more local arrivals is negligible in comparison to δ , and this verifies the third property. We conclude that local packet arrivals form a Poisson process and, in particular, the number of such arrivals during an interval of length τ has a Poisson PMF with parameter $p\lambda\tau$. By a symmetrical argument, the process of transit packet arrivals is also Poisson, with rate $\lambda(1-p)$. A somewhat surprising fact in this context is that the two Poisson processes obtained by splitting an original Poisson process are independent, see the end-of-chapter problems.

Example 5.14. Merging of Poisson Processes. People with letters to mail arrive at the post office according to a Poisson process with rate λ_1 , while people with packages to mail arrive according to an independent Poisson process with rate λ_2 . As stated earlier the merged process, which includes arrivals of both types, is Poisson with rate $\lambda_1 + \lambda_2$. Let us see why.

First, it should be clear that the merged process satisfies the time-homogeneity property. Furthermore, since different intervals in each of the two arrival processes are independent, the same property holds for the merged process. Let us now focus on a small interval of length δ . Ignoring terms that are negligible compared to δ , we have

$$\mathbf{P}(0 \text{ arrivals in the merged process}) \approx (1 - \lambda_1\delta)(1 - \lambda_2\delta) \approx 1 - (\lambda_1 + \lambda_2)\delta,$$

$$\mathbf{P}(1 \text{ arrival in the merged process}) \approx \lambda_1\delta(1 - \lambda_2\delta) + (1 - \lambda_1\delta)\lambda_2\delta \approx (\lambda_1 + \lambda_2)\delta,$$

and the third property has been verified.

Given that an arrival has just been recorded, what is the probability that it is an arrival of a person with a letter to mail? We focus again on a small interval of length δ around the current time, and we seek the probability

$$\mathbf{P}(1 \text{ arrival of person with a letter} \mid 1 \text{ arrival}).$$

Using the definition of conditional probabilities, and ignoring the negligible probability of more than one arrival, this is

$$\frac{\mathbf{P}(1 \text{ arrival of person with a letter})}{\mathbf{P}(1 \text{ arrival})} \approx \frac{\lambda_1\delta}{(\lambda_1 + \lambda_2)\delta} = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Generalizing this calculation, we let L_k be the event that the k th arrival corresponds to an arrival of a person with a letter to mail, and we have

$$\mathbf{P}(L_k) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Furthermore, since distinct arrivals happen at different times, and since, for Poisson processes, events at different times are independent, it follows that the random variables L_1, L_2, \dots are independent.

Example 5.15. Competing Exponentials. Two light bulbs have independent and exponentially distributed lifetimes T_a and T_b , with parameters λ_a and λ_b , respectively. What is the distribution of $Z = \min\{T_a, T_b\}$, the first time when a bulb burns out?

We can treat this as an exercise in derived distributions. For all $z \geq 0$, we have,

$$\begin{aligned} F_Z(z) &= \mathbf{P}(\min\{T_a, T_b\} \leq z) \\ &= 1 - \mathbf{P}(\min\{T_a, T_b\} > z) \\ &= 1 - \mathbf{P}(T_a > z, T_b > z) \\ &= 1 - \mathbf{P}(T_a > z)\mathbf{P}(T_b > z) \\ &= 1 - e^{-\lambda_a z} e^{-\lambda_b z} \\ &= 1 - e^{-(\lambda_a + \lambda_b)z}. \end{aligned}$$

This is recognized as the exponential CDF with parameter $\lambda_a + \lambda_b$. Thus, the minimum of two independent exponentials with parameters λ_a and λ_b is an exponential with parameter $\lambda_a + \lambda_b$.

For a more intuitive explanation of this fact, let us think of T_a and T_b as the times of the first arrival in two independent Poisson processes with rates λ_a and λ_b , respectively. If we merge these two processes, the first arrival time will be $\min\{T_a, T_b\}$. But we already know that the merged process is Poisson with rate $\lambda_a + \lambda_b$, and it follows that the first arrival time, $\min\{T_a, T_b\}$, is exponential with parameter $\lambda_a + \lambda_b$.

The preceding discussion can be generalized to the case of more than two processes. Thus, the total arrival process obtained by merging the arrivals of n independent Poisson processes with arrival rates $\lambda_1, \dots, \lambda_n$ is Poisson with arrival rate equal to the sum $\lambda_1 + \dots + \lambda_n$.

Example 5.16. More on Competing Exponentials. Three light bulbs have independent exponentially distributed lifetimes with a common parameter λ . What is the expected value of the time until the last bulb burns out?

We think of the times when each bulb burns out as the first arrival times in independent Poisson processes. In the beginning, we have three bulbs, and the merged process has rate 3λ . Thus, the time T_1 of the first burnout is exponential with parameter 3λ , and mean $1/3\lambda$. Once a bulb burns out, and because of the memorylessness property of the exponential distribution, the remaining lifetimes of the other two lightbulbs are again independent exponential random variables with parameter λ . We thus have two Poisson processes running in parallel, and the remaining time T_2 until the first arrival in one of these two processes is now

exponential with parameter 2λ and mean $1/2\lambda$. Finally, once a second bulb burns out, we are left with a single one. Using memorylessness once more, the remaining time T_3 until the last bulb burns out is exponential with parameter λ and mean $1/\lambda$. Thus, the expected value of the total time is

$$\mathbf{E}[T_1 + T_2 + T_3] = \frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda}.$$

Note that the random variables T_1, T_2, T_3 are independent, because of memorylessness. This allows us to also compute the variance of the total time:

$$\text{var}(T_1 + T_2 + T_3) = \text{var}(T_1) + \text{var}(T_2) + \text{var}(T_3) = \frac{1}{9\lambda^2} + \frac{1}{4\lambda^2} + \frac{1}{\lambda^2}.$$

We close by noting a related and quite deep fact, namely that the sum of a *large* number of (*not* necessarily Poisson) independent arrival processes, can be approximated by a Poisson process with arrival rate equal to the sum of the individual arrival rates. The component processes must have a small rate relative to the total (so that none of them imposes its probabilistic character on the total arrival process) and they must also satisfy some technical mathematical assumptions. Further discussion of this fact is beyond our scope, but we note that it is in large measure responsible for the abundance of Poisson-like processes in practice. For example, the telephone traffic originating in a city consists of many component processes, each of which characterizes the phone calls placed by individual residents. The component processes need not be Poisson; some people for example tend to make calls in batches, and (usually) while in the process of talking, cannot initiate or receive a second call. However, the total telephone traffic is well-modeled by a Poisson process. For the same reasons, the process of auto accidents in a city, customer arrivals at a store, particle emissions from radioactive material, etc., tend to have the character of the Poisson process.

The Random Incidence Paradox

The arrivals of a Poisson process partition the time axis into a sequence of interarrival intervals; each interarrival interval starts with an arrival and ends at the time of the next arrival. We have seen that the lengths of these interarrival intervals are independent exponential random variables with parameter λ , where λ is the rate of the process. More precisely, for every k , the length of the k th interarrival interval has this exponential distribution. In this subsection, we look at these interarrival intervals from a different perspective.

Let us fix a time instant t^* and consider the length L of the interarrival interval that contains t^* . For a concrete context, think of a person who shows up at the bus station at some arbitrary time t^* and records the time from the previous bus arrival until the next bus arrival. The arrival of this person is often referred to as a “random incidence,” but the reader should be aware that the term is misleading: t^* is just a particular time instance, not a random variable.

We assume that t^* is much larger than the starting time of the Poisson process so that we can be fairly certain that there has been an arrival prior to time t^* . To avoid the issue of how large t^* should be, we assume that the Poisson process has been running forever, so that we can be certain that there has been a prior arrival, and that L is well-defined. One might superficially argue that L is the length of a “typical” interarrival interval, and is exponentially distributed, but this turns out to be false. Instead, we will establish that L has an Erlang PDF of order two.

This is known as the *random incidence phenomenon* or *paradox*, and it can be explained with the help of Fig. 5.7. Let $[U, V]$ be the interarrival interval that contains t^* , so that $L = V - U$. In particular, U is the time of the first arrival prior to t^* and V is the time of the first arrival after t^* . We split L into two parts,

$$L = (t^* - U) + (V - t^*),$$

where $t^* - U$ is the elapsed time since the last arrival, and $V - t^*$ is the remaining time until the next arrival. Note that $t^* - U$ is determined by the past history of the process (before t^*), while $V - t^*$ is determined by the future of the process (after time t^*). By the independence properties of the Poisson process, the random variables $t^* - U$ and $V - t^*$ are independent. By the memorylessness property, the Poisson process starts fresh at time t^* , and therefore $V - t^*$ is exponential with parameter λ . The random variable $t^* - U$ is also exponential with parameter λ . The easiest way to see this is to realize that if we run a Poisson process backwards in time it remains Poisson; this is because the defining properties of a Poisson process make no reference to whether time moves forward or backward. A more formal argument is obtained by noting that

$$\mathbf{P}(t^* - U > x) = \mathbf{P}(\text{no arrivals during } [t^* - x, t^*]) = P(0, x) = e^{-\lambda x}, \quad x \geq 0.$$

We have therefore established that L is the sum of two independent exponential random variables with parameter λ , i.e., Erlang of order two, with mean $2/\lambda$.

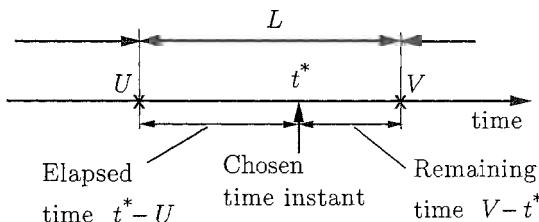


Figure 5.7: Illustration of the random incidence phenomenon. For a fixed time instant t^* , the corresponding interarrival interval $[U, V]$ consists of the elapsed time $t^* - U$ and the remaining time $V - t^*$. These two times are independent and are exponentially distributed with parameter λ , so the PDF of their sum is Erlang of order two.

Random incidence phenomena are often the source of misconceptions and errors, but these can be avoided with careful probabilistic modeling. The key issue is that even though interarrival intervals have length $1/\lambda$ on the average, an observer who arrives at an arbitrary time is more likely to fall in a large rather than a small interarrival interval. As a consequence the expected length seen by the observer is higher, $2/\lambda$ in this case. A similar situation arises in the example that follows.

Example 5.17. Random Incidence in a Non-Poisson Arrival Process.

Buses arrive at a station deterministically, on the hour, and five minutes after the hour. Thus, the interarrival times alternate between 5 and 55 minutes. The average interarrival time is 30 minutes. A person shows up at the bus station at a “random” time. We interpret “random” to mean a time that is uniformly distributed within a particular hour. Such a person falls into an interarrival interval of length 5 with probability $1/12$, and an interarrival interval of length 55 with probability $11/12$. The expected length of the chosen interarrival interval is

$$5 \cdot \frac{1}{12} + 55 \cdot \frac{11}{12} = 50.83,$$

which is considerably larger than 30, the average interarrival time.

As the preceding example indicates, random incidence is a subtle phenomenon that introduces a bias in favor of larger interarrival intervals, and can manifest itself in contexts other than the Poisson process. In general, whenever different calculations give contradictory results, the reason is that they refer to different probabilistic mechanisms. For instance, considering a fixed **nonrandom** k and the associated random value of the k th interarrival interval is a different experiment from fixing a time t and considering the **random** K such that the K th interarrival interval contains t .

For a last example with the same flavor, consider a survey of the utilization of town buses. One approach is to select a few buses at random and calculate the average number of riders in the selected buses. An alternative approach is to select a few bus riders at random, look at the buses that they rode and calculate the average number of riders in the latter set of buses. The estimates produced by the two methods have very different statistics, with the second method being biased upwards. The reason is that with the second method, it is much more likely to select a bus with a large number of riders than a bus that is near-empty.

5.3 SUMMARY AND DISCUSSION

In this chapter, we introduced and analyzed two memoryless arrival processes. The Bernoulli process evolves in discrete time, and during each discrete time step, there is a constant probability p of an arrival. The Poisson process evolves

in continuous time, and during each small interval of length $\delta > 0$, there is a probability of an arrival approximately equal to $\lambda\delta$. In both cases, the numbers of arrivals in disjoint time intervals are assumed independent. Thus the Poisson process can be viewed as a limiting case of the Bernoulli process, in which the duration of each discrete time slot is taken to be a very small number δ . This fact can be used to draw parallels between the major properties of the two processes, and to transfer insights gained from one process to the other.

Using the memorylessness property of the Bernoulli and the Poisson, we derived the following.

- (a) The PMF of the number of arrivals during a time interval of given length is binomial and Poisson, respectively.
- (b) The distribution of the time between successive arrivals is geometric and exponential, respectively.
- (c) The distribution of the time until the k th arrival, is Pascal of order k and Erlang of order k , respectively.

Furthermore, we saw that one can start with two independent Bernoulli (respectively, Poisson) processes and “merge” them to form a new Bernoulli (respectively, Poisson) process. Conversely, if one “accepts” each arrival by flipping a coin with success probability p (“splitting”), the process of accepted arrivals is a Bernoulli or Poisson process whose arrival rate is p times the original arrival rate.

We finally considered the “random incidence” phenomenon where an external observer arrives at some given time and measures the interarrival interval within which he arrives. The statistics of the measured interval are different from those of a “typical” interarrival interval, because the arriving observer is more likely to fall in a larger interarrival interval. This phenomenon indicates that when talking about a “typical” interval, one must carefully describe the mechanism by which the interval is selected. Different mechanisms will in general result in different statistical properties.

P R O B L E M S

SECTION 5.1. The Bernoulli Process

Problem 1. Each of n packages is loaded independently onto either a red truck (with probability p) or onto a green truck (with probability $1 - p$). Let R be the total number of items selected for the red truck and let G be the total number of items selected for the green truck.

- (a) Determine the PMF, expected value, and variance of the random variable R .
- (b) Evaluate the probability that the first item to be loaded ends up being the only one on its truck.
- (c) Evaluate the probability that at least one truck ends up with a total of exactly one package.
- (d) Evaluate the expected value and the variance of the difference $R - G$.
- (e) Assume that $n \geq 2$. Given that both of the first two packages to be loaded go onto the red truck, find the conditional expectation, variance, and PMF of the random variable R .

Problem 2. Dave fails quizzes with probability $1/4$, independently of other quizzes.

- (a) What is the probability that Dave fails exactly two of the next six quizzes?
- (b) What is the expected number of quizzes that Dave will pass before he has failed three times?
- (c) What is the probability that the second and third time Dave fails a quiz will occur when he takes his eighth and ninth quizzes, respectively?
- (d) What is the probability that Dave fails two quizzes in a row before he passes two quizzes in a row?

Problem 3. A computer system carries out tasks submitted by two users. Time is divided into slots. A slot can be idle, with probability $p_I = 1/6$, and busy with probability $p_B = 5/6$. During a busy slot, there is probability $p_{1|B} = 2/5$ (respectively, $p_{2|B} = 3/5$) that a task from user 1 (respectively, 2) is executed. We assume that events related to different slots are independent.

- (a) Find the probability that a task from user 1 is executed for the first time during the 4th slot.
- (b) Given that exactly 5 out of the first 10 slots were idle, find the probability that the 6th idle slot is slot 12.
- (c) Find the expected number of slots up to and including the 5th task from user 1.

- (d) Find the expected number of busy slots up to and including the 5th task from user 1.
- (e) Find the PMF, mean, and variance of the number of tasks from user 2 until the time of the 5th task from user 1.

Problem 4.* Consider a Bernoulli process with probability of success in each trial equal to p .

- (a) Relate the number of failures before the r th success (sometimes called a **negative binomial** random variable) to a Pascal random variable and derive its PMF.
- (b) Find the expected value and variance of the number of failures before the r th success.
- (c) Obtain an expression for the probability that the i th failure occurs before the r th success.

Solution. (a) Let Y be the number of trials until the r th success, which is a Pascal random variable of order r . Let X be the number of failures before the r th success, so that $X = Y - r$. Therefore, $p_X(k) = p_Y(k + r)$, and

$$p_X(k) = \binom{k+r-1}{r-1} p^r (1-p)^k, \quad k = 0, 1, \dots$$

(b) Using the notation of part (a), we have

$$\mathbf{E}[X] = \mathbf{E}[Y] - r = \frac{r}{p} - r = \frac{(1-p)r}{p}.$$

Furthermore,

$$\text{var}(X) = \text{var}(Y) = \frac{(1-p)r}{p^2}.$$

(c) Let again X be the number of failures before the r th success. The i th failure occurs before the r th success if and only if $X \geq i$. Therefore, the desired probability is equal to

$$\sum_{k=i}^{\infty} p_X(k) = \sum_{k=i}^{\infty} \binom{k+r-1}{r-1} p^r (1-p)^k, \quad i = 1, 2, \dots$$

An alternative formula is derived as follows. Consider the first $r+i-1$ trials. The number of failures in these trials is at least i if and only if the number of successes is less than r . But this is equivalent to the i th failure occurring before the r th success. Hence, the desired probability is the probability that the number of successes in $r+i-1$ trials is less than r , which is

$$\sum_{k=0}^{r-1} \binom{r+i-1}{k} p^k (1-p)^{r+i-1-k}, \quad i = 1, 2, \dots$$

Problem 5.* Random incidence in the Bernoulli process. Your cousin has been playing the same video game from time immemorial. Assume that he wins each

game with probability p , independently of the outcomes of other games. At midnight, you enter his room and witness his losing the current game. What is the PMF of the number of lost games between his most recent win and his first future win?

Solution. Let t be the number of the game when you enter the room. Let M be the number of the most recent past game that he won, and let N be the number of the first game to be won in the future. The random variable $X = N - t$ is geometrically distributed with parameter p . By symmetry and independence of the games, the random variable $Y = t - M$ is also geometrically distributed with parameter p . The games he lost between his most recent win and his first future win are all the games between M and N . Their number L is given by

$$L = N - M - 1 = X + Y - 1.$$

Thus, $L + 1$ has a Pascal PMF of order two, and

$$\mathbf{P}(L + 1 = k) = \binom{k-1}{1} p^2 (1-p)^{k-2}.$$

Hence,

$$p_L(i) = \mathbf{P}(L + 1 = i + 1) = \binom{i}{1} p^2 (1-p)^{i-1}, \quad i = 1, 2, \dots$$

Problem 6.* Sum of a geometric number of independent geometric random variables. Let $Y = X_1 + \dots + X_N$, where the random variables X_i are geometric with parameter p , and N is geometric with parameter q . Assume that the random variables N, X_1, X_2, \dots are independent. Show, without using transforms, that Y is geometric with parameter pq . *Hint:* Interpret the various random variables in terms of a split Bernoulli process.

Solution. We derived this result in Chapter 4, using transforms, but we develop a more intuitive derivation here. We interpret the random variables X_i and N as follows. We view the times $X_1, X_1 + X_2$, etc. as the arrival times in a Bernoulli process with parameter p . Each arrival is rejected with probability $1 - q$ and is accepted with probability q . We interpret N as the number of arrivals until the first acceptance. The process of accepted arrivals is obtained by splitting a Bernoulli process and is therefore itself Bernoulli with parameter pq . The random variable $Y = X_1 + \dots + X_N$ is the time of the first accepted arrival and is therefore geometric, with parameter pq .

Problem 7.* The bits in a uniform random variable form a Bernoulli process. Let X_1, X_2, \dots be a sequence of binary random variables taking values in the set $\{0, 1\}$. Let Y be a continuous random variable that takes values in the set $[0, 1]$. We relate X and Y by assuming that Y is the real number whose binary representation is $0.X_1X_2X_3\dots$. More concretely, we have

$$Y = \sum_{k=1}^{\infty} 2^{-k} X_k.$$

- (a) Suppose that the X_i form a Bernoulli process with parameter $p = 1/2$. Show that Y is uniformly distributed. *Hint:* Consider the probability of the event $(i-1)/2^k < Y < i/2^k$, where i and k are positive integers.

- (b) Suppose that Y is uniformly distributed. Show that the X_i form a Bernoulli process with parameter $p = 1/2$.

Solution. (a) We have

$$\mathbf{P}(Y \in [0, 1/2]) = \mathbf{P}(X_1 = 0) = \frac{1}{2} = \mathbf{P}(Y \in [1/2, 1]).$$

Furthermore,

$$\mathbf{P}(Y \in [0, 1/4]) = \mathbf{P}(X_1 = 0, X_2 = 0) = \frac{1}{4}.$$

Arguing similarly, we consider an interval of the form $[(i-1)/2^k, i/2^k]$, where i and k are positive integers and $i \leq 2^k$. For Y to fall in the interior of this interval, we need X_1, \dots, X_k to take on a particular set of values (namely, the binary expansion of $i-1$). Hence,

$$\mathbf{P}((i-1)/2^k < Y < i/2^k) = \frac{1}{2^k}.$$

Note also that for any $y \in [0, 1]$, we have $\mathbf{P}(Y = y) = 0$, because the event $\{Y = y\}$ can only occur if infinitely many X_i s take on particular values, a zero probability event. Therefore, the CDF of Y is continuous and satisfies

$$\mathbf{P}(Y \leq i/2^k) = i/2^k.$$

Since every number y in $[0, 1]$ can be closely approximated by a number of the form $i/2^k$, we have $\mathbf{P}(Y \leq y) = y$, for every $y \in [0, 1]$, which establishes that Y is uniform.

(b) As in part (a), we observe that every possible zero-one pattern for X_1, \dots, X_k is associated to one particular interval of the form $[(i-1)/2^k, i/2^k]$ for Y . These intervals have equal length, and therefore have the same probability $1/2^k$, since Y is uniform. This particular joint PMF for X_1, \dots, X_k , corresponds to independent Bernoulli random variables with parameter $p = 1/2$.

SECTION 5.2. The Poisson Process

Problem 8. During rush hour, from 8 a.m. to 9 a.m., traffic accidents occur according to a Poisson process with a rate of 5 accidents per hour. Between 9 a.m. and 11 a.m., they occur as an independent Poisson process with a rate of 3 accidents per hour. What is the PMF of the total number of accidents between 8 a.m. and 11 a.m.?

Problem 9. A fisherman catches fish according to a Poisson process with rate $\lambda = 0.6$ per hour. The fisherman will keep fishing for two hours. If he has caught at least one fish, he quits. Otherwise, he continues until he catches at least one fish.

- (a) Find the probability that he stays for more than two hours.
- (b) Find the probability that the total time he spends fishing is between two and five hours.
- (c) Find the probability that he catches at least two fish.
- (d) Find the expected number of fish that he catches.
- (e) Find the expected total fishing time, given that he has been fishing for four hours.

Problem 10. Customers depart from a bookstore according to a Poisson process with rate λ per hour. Each customer buys a book with probability p , independent of everything else.

- (a) Find the distribution of the time until the first sale of a book.
- (b) Find the probability that no books are sold during a particular hour.
- (c) Find the expected number of customers who buy a book during a particular hour.

Problem 11. Transmitters A and B independently send messages to a single receiver in a Poisson manner, with rates of λ_A and λ_B , respectively. All messages are so brief that we may assume that they occupy single points in time. The number of words in a message, regardless of the source that is transmitting it, is a random variable with PMF

$$p_w(w) = \begin{cases} 2/6, & \text{if } w = 1, \\ 3/6, & \text{if } w = 2, \\ 1/6, & \text{if } w = 3, \\ 0, & \text{otherwise,} \end{cases}$$

and is independent of everything else.

- (a) What is the probability that during an interval of duration t , a total of exactly nine messages will be received?
- (b) Let N be the total number of words received during an interval of duration t . Determine the expected value of N .
- (c) Determine the PDF of the time from $t = 0$ until the receiver has received exactly eight three-word messages from transmitter A.
- (d) What is the probability that exactly eight of the next twelve messages received will be from transmitter A?

Problem 12. Beginning at time $t = 0$, we start using bulbs, one at a time, to illuminate a room. Bulbs are replaced immediately upon failure. Each new bulb is selected independently by an equally likely choice between a type-A bulb and a type-B bulb. The lifetime, X , of any particular bulb of a particular type is a random variable, independent of everything else, with the following PDF:

$$\text{for type-A Bulbs: } f_X(x) = \begin{cases} e^{-x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise;} \end{cases}$$

$$\text{for type-B Bulbs: } f_X(x) = \begin{cases} 3e^{-3x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Find the expected time until the first failure.
- (b) Find the probability that there are no bulb failures before time t .
- (c) Given that there are no failures until time t , determine the conditional probability that the first bulb used is a type-A bulb.
- (d) Find the variance of the time until the first bulb failure.
- (e) Find the probability that the 12th bulb failure is also the 4th type-A bulb failure.

- (f) Up to and including the 12th bulb failure, what is the probability that a total of exactly 4 type-A bulbs have failed?
- (g) Determine either the PDF or the transform associated with the time until the 12th bulb failure.
- (h) Determine the probability that the total period of illumination provided by the first two type-B bulbs is longer than that provided by the first type-A bulb.
- (i) Suppose the process terminates as soon as a total of exactly 12 bulb failures have occurred. Determine the expected value and variance of the total period of illumination provided by type-B bulbs while the process is in operation.
- (j) Given that there are no failures until time t , find the expected value of the time until the first failure.

Problem 13. A service station handles jobs of two types, A and B. (Multiple jobs can be processed simultaneously.) Arrivals of the two job types are independent Poisson processes with parameters $\lambda_A = 3$ and $\lambda_B = 4$ per minute, respectively. Type A jobs stay in the service station for exactly one minute. Each type B job stays in the service station for a random but integer amount of time which is geometrically distributed, with mean equal to 2, and independent of everything else. The service station started operating at some time in the remote past.

- (a) What is the mean, variance, and PMF of the total number of jobs that arrive within a given three-minute interval?
- (b) We are told that during a 10-minute interval, exactly 10 new jobs arrived. What is the probability that exactly 3 of them are of type A?
- (c) At time 0, no job is present in the service station. What is the PMF of the number of type B jobs that arrive in the future, but before the first type A arrival?
- (d) At time $t = 0$, there were exactly two type A jobs in the service station. What is the PDF of the time of the last (before time 0) type A arrival?
- (e) At time 1, there was exactly one type B job in the service station. Find the distribution of the time until this type B job departs.

Problem 14. Each morning, as you pull out of your driveway, you would like to make a U-turn rather than drive around the block. Unfortunately, U-turns are illegal in your neighborhood, and police cars drive by according to a Poisson process with rate λ . You decide to make a U-turn once you see that the road has been clear of police cars for τ time units. Let N be the number of police cars you see before you make the U-turn.

- (a) Find $E[N]$.
- (b) Find the conditional expectation of the time elapsed between police cars $n - 1$ and n , given that $N \geq n$.
- (c) Find the expected time that you wait until you make the U-turn. *Hint:* Condition on N .

Problem 15. A wombat in the San Diego zoo spends the day walking from a burrow to a food tray, eating, walking back to the burrow, resting, and repeating the cycle. The amount of time to walk from the burrow to the tray (and also from the tray to the

burrow) is 20 secs. The amounts of time spent at the tray and resting are exponentially distributed with mean 30 secs. The wombat, with probability 1/3, will momentarily stand still (for a negligibly small time) during a walk to or from the tray, with all times being equally likely (and independently of what happened in the past). A photographer arrives at a random time and will take a picture at the first time the wombat will stand still. What is the expected value of the length of time the photographer has to wait to snap the wombat's picture?

Problem 16.* Consider a Poisson process. Given that a single arrival occurred in a given interval $[0, t]$, show that the PDF of the arrival time is uniform over $[0, t]$.

Solution. Consider an interval $[a, b] \subset [0, t]$ of length $l = b - a$. Let T be the time of the first arrival, and let A be the event that a single arrival occurred during $[0, t]$. We have

$$P(T \in [a, b] | A) = \frac{P(T \in [a, b] \text{ and } A)}{P(A)}.$$

The numerator is equal to the probability $P(1, l)$ that the Poisson process has exactly one arrival during the length l interval $[a, b]$, times the probability $P(0, t-l)$ that the process has zero arrivals during the set $[0, a] \cup (b, t]$, of total length $t-l$. Hence

$$P(T \in [a, b] | A) = \frac{P(1, l)P(0, t-l)}{P(1, t)} = \frac{(\lambda l)e^{-\lambda l}e^{-\lambda(t-l)}}{(\lambda t)e^{-\lambda t}} = \frac{l}{t},$$

which establishes that T is uniformly distributed.

Problem 17.*

- (a) Let X_1 and X_2 be independent and exponentially distributed, with parameters λ_1 and λ_2 , respectively. Find the expected value of $\max\{X_1, X_2\}$.
- (b) Let Y be exponentially distributed with parameter λ_1 . Let Z be Erlang of order 2 with parameter λ_2 . Assume that Y and Z are independent. Find the expected value of $\max\{Y, Z\}$.

Solution. A direct but tedious approach would be to find the PDF of the random variable of interest and then evaluate an integral to find its expectation. A much simpler solution is obtained by interpreting the random variables of interest in terms of underlying Poisson processes.

(a) Consider two independent Poisson processes with rates λ_1 and λ_2 , respectively. We interpret X_1 as the first arrival time in the first process, and X_2 as the first arrival time in the second process. Let $T = \min\{X_1, X_2\}$ be the first time when one of the processes registers an arrival. Let $S = \max\{X_1, X_2\} - T$ be the additional time until both have registered an arrival. Since the merged process is Poisson with rate $\lambda_1 + \lambda_2$, we have

$$E[T] = \frac{1}{\lambda_1 + \lambda_2}.$$

Concerning S , there are two cases to consider.

- (i) The first arrival comes from the first process; this happens with probability $\lambda_1/(\lambda_1 + \lambda_2)$. We then have to wait for an arrival from the second process, which takes $1/\lambda_2$ time on the average.

- (ii) The first arrival comes from the second process; this happens with probability $\lambda_2/(\lambda_1 + \lambda_2)$. We then have to wait for an arrival from the first process, which takes $1/\lambda_1$ time on the average.

Putting everything together, we obtain

$$\begin{aligned}\mathbf{E}[\max\{X_1, X_2\}] &= \frac{1}{\lambda_1 + \lambda_2} + \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \frac{1}{\lambda_2} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot \frac{1}{\lambda_1} \\ &= \frac{1}{\lambda_1 + \lambda_2} \left(1 + \frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1}\right).\end{aligned}$$

- (b) Consider two independent Poisson processes with rates λ_1 and λ_2 , respectively. We interpret Y as the first arrival time in the first process, and Z as the second arrival time in the second process. Let T be the first time when one of the processes registers an arrival. Since the merged process is Poisson with rate $\lambda_1 + \lambda_2$, we have $\mathbf{E}[T] = 1/(\lambda_1 + \lambda_2)$. There are two cases to consider.

- (i) The arrival at time T comes from the first process; this happens with probability $\lambda_1/(\lambda_1 + \lambda_2)$. In this case, we have to wait an additional time until the second process registers two arrivals. This additional time is Erlang of order 2, with parameter λ_2 , and its expected value is $2/\lambda_2$.
- (ii) The arrival at time T comes from the second process; this happens with probability $\lambda_2/(\lambda_1 + \lambda_2)$. In this case, the additional time S we have to wait is the time until each of the two processes registers an arrival. This is the maximum of two independent exponential random variables and, according to the result of part (a), we have

$$\mathbf{E}[S] = \frac{1}{\lambda_1 + \lambda_2} \left(1 + \frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1}\right).$$

Putting everything together, we have

$$\mathbf{E}[\max\{Y, Z\}] = \frac{1}{\lambda_1 + \lambda_2} + \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \frac{2}{\lambda_2} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot \mathbf{E}[S],$$

where $\mathbf{E}[S]$ is given by the previous formula.

Problem 18.* Let Y_k be the time of the k th arrival in a Poisson process with rate λ . Show that for all $y > 0$,

$$\sum_{k=1}^{\infty} f_{Y_k}(y) = \lambda.$$

Solution. We have

$$\begin{aligned}\sum_{k=1}^{\infty} f_{Y_k}(y) &= \sum_{k=1}^{\infty} \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!} \\ &= \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1} y^{k-1} e^{-\lambda y}}{(k-1)!} \quad (\text{let } m = k-1) \\ &= \lambda \sum_{m=0}^{\infty} \frac{\lambda^m y^m e^{-\lambda y}}{m!} \\ &= \lambda.\end{aligned}$$

The last equality holds because the $\lambda^m y^m e^{-\lambda y} / m!$ terms are the values of a Poisson PMF with parameter λy and must therefore sum to 1.

For a more intuitive derivation, let δ be a small positive number and consider the following events:

A_k : the k th arrival occurs between y and $y + \delta$; the probability of this event is $P(A_k) \approx f_{Y_k}(y)\delta$;

A : an arrival occurs between y and $y + \delta$; the probability of this event is $P(A) \approx f_Y(y)\delta$.

Suppose that δ is taken small enough so that the possibility of two or more arrivals during an interval of length δ can be ignored. With this approximation, the events A_1, A_2, \dots become disjoint, and their union is A . Therefore,

$$\begin{aligned} \sum_{k=1}^{\infty} f_{Y_k}(y) \cdot \delta &\approx \sum_{k=1}^{\infty} P(A_k) \\ &\approx P(A) \\ &\approx \lambda \delta, \end{aligned}$$

and the desired result follows by canceling δ from both sides.

Problem 19.* Consider an experiment involving two independent Poisson processes with rates λ_1 and λ_2 . Let $X_1(k)$ and $X_2(k)$ be the times of the k th arrival in the 1st and the 2nd process, respectively. Show that

$$P(X_1(n) < X_2(m)) = \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n+m-1-k}.$$

Solution. Consider the merged Poisson process, which has rate $\lambda_1 + \lambda_2$. Each time there is an arrival in the merged process, this arrival comes from the first process (“success”) with probability $\lambda_1/(\lambda_1 + \lambda_2)$, and from the second process (“failure”) with probability $\lambda_2/(\lambda_1 + \lambda_2)$. Consider the situation after $n + m - 1$ arrivals. The number of arrivals from the first process is at least n if and only if the number of arrivals from the second process is less than m , which happens if and only if the n th success occurs before the m th failure. Thus, the event $\{X_1(n) < X_2(m)\}$ is the same as the event of having at least n successes in the first $n + m - 1$ trials. Therefore, using the binomial PMF for the number of successes in a given number of trials, we have

$$P(X_1(n) < X_2(m)) = \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n+m-1-k}.$$

Problem 20.* Sum of a geometric number of independent exponential random variables. Let $Y = X_1 + \dots + X_N$, where the random variables X_i are exponential with parameter λ , and N is geometric with parameter p . Assume that the random variables N, X_1, X_2, \dots are independent. Show, without using transforms, that Y is exponential with parameter λp . *Hint:* Interpret the various random variables in terms of a split Poisson process.

Solution. We derived this result in Chapter 4, using transforms, but we develop a more intuitive derivation here. We interpret the random variables X_i and N as follows. We view the times X_1 , $X_1 + X_2$, etc. as the arrival times in a Poisson process with parameter λ . Each arrival is rejected with probability $1 - p$ and is accepted with probability p . We interpret N as the number of arrivals until the first acceptance. The process of accepted arrivals is obtained by splitting a Poisson process and is therefore itself Poisson with parameter $p\lambda$. Note that $Y = X_1 + \dots + X_N$ is the time of the first accepted arrival and is therefore exponential with parameter $p\lambda$.

Problem 21.* The number of Poisson arrivals during an exponentially distributed interval. Consider a Poisson process with parameter λ , and an independent random variable T , which is exponential with parameter ν . Find the PMF of the number of Poisson arrivals during the time interval $[0, T]$.

Solution. Let us view T as the first arrival time in a new, independent, Poisson process with parameter ν , and merge this process with the original Poisson process. Each arrival in the merged process comes from the original Poisson process with probability $\lambda/(\lambda + \nu)$, independently from other arrivals. If we view each arrival in the merged process as a trial, and an arrival from the new process as a success, we note that the number K of trials/arrivals until the first success has a geometric PMF, of the form

$$p_K(k) = \left(\frac{\nu}{\lambda + \nu} \right) \left(\frac{\lambda}{\lambda + \nu} \right)^{k-1}, \quad k = 1, 2, \dots$$

Now the number L of arrivals from the original Poisson process until the first “success” is equal to $K - 1$ and its PMF is

$$p_L(l) = p_K(l+1) = \left(\frac{\nu}{\lambda + \nu} \right) \left(\frac{\lambda}{\lambda + \nu} \right)^l, \quad l = 0, 1, \dots$$

Problem 22.* An infinite server queue. We consider a queueing system with an infinite number of servers, in which customers arrive according to a Poisson process with rate λ . The i th customer stays in the system for a random amount of time, denoted by X_i . We assume that the random variables X_i are independent identically distributed, and also independent from the arrival process. We also assume, for simplicity, that the X_i take integer values in the range $1, 2, \dots, n$, with given probabilities. Find the PMF of N_t , the number of customers in the system at time t .

Solution. Let us refer to those customers i whose service time X_i is equal to k as “type- k ” customers. We view the overall arrival process as the merging of n Poisson subprocesses; the k th subprocess corresponds to arrivals of type- k customers, is independent of the other arrival subprocesses, and has rate λp_k , where $p_k = \mathbf{P}(X_i = k)$. Let N_t^k be the number of type- k customers in the system at time t . Thus,

$$N_t = \sum_{k=1}^n N_t^k,$$

and the random variables N_t^k are independent.

We now determine the PMF of N_t^k . A type- k customer is in the system at time t if and only if that customer arrived between times $t - k$ and t . Thus, N_t^k has a Poisson

PMF with mean λkp_k . Since the sum of independent Poisson random variables is Poisson, it follows that N_t has a Poisson PMF with parameter

$$\mathbf{E}[N_t] = \lambda \sum_{k=1}^n kp_k = \lambda \mathbf{E}[X_i].$$

Problem 23.* Independence of Poisson processes obtained by splitting. Consider a Poisson process whose arrivals are split, with each arrival assigned to one of two subprocesses by flipping an independent coin with success probability p . In Example 5.13, it was established that each of the subprocesses is a Poisson process. Show that the two subprocesses are independent.

Solution. Let us start with two independent Poisson processes \mathcal{P}_1 and \mathcal{P}_2 , with rates $p\lambda$ and $(1-p)\lambda$, respectively. We merge the two processes and obtain a Poisson process \mathcal{P} with rate λ . We now split the process \mathcal{P} into two new subprocesses \mathcal{P}'_1 and \mathcal{P}'_2 , according to the following rule: an arrival is assigned to subprocess \mathcal{P}'_1 (respectively, \mathcal{P}'_2) if and only if that arrival corresponds to an arrival from subprocess \mathcal{P}_1 (respectively, \mathcal{P}_2). Clearly, the two new subprocesses \mathcal{P}'_1 and \mathcal{P}'_2 are independent, since they are identical to the original subprocesses \mathcal{P}_1 and \mathcal{P}_2 . However, \mathcal{P}'_1 and \mathcal{P}'_2 were generated by a splitting mechanism that looks different than the one in the problem statement. We will now verify that the new splitting mechanism considered here is statistically identical to the one in the problem statement. It will then follow that the subprocesses constructed in the problem statement have the same statistical properties as \mathcal{P}'_1 and \mathcal{P}'_2 , and are also independent.

So, let us consider the above described splitting mechanism. Given that \mathcal{P} had an arrival at time t , this was due to either an arrival in \mathcal{P}_1 (with probability p), or to an arrival in \mathcal{P}_2 (probability $1-p$). Therefore, the arrival to \mathcal{P} is assigned to \mathcal{P}'_1 or \mathcal{P}'_2 with probabilities p and $1-p$, respectively, exactly as in the splitting procedure described in the problem statement. Consider now the k th arrival in \mathcal{P} and let L_k be the event that this arrival originated from subprocess \mathcal{P}_1 ; this is the same as the event that the k th arrival is assigned to subprocess \mathcal{P}'_1 . As explained in the context of Example 5.14, the events L_k are independent. Thus, the assignments of arrivals to the subprocesses \mathcal{P}'_1 and \mathcal{P}'_2 are independent for different arrivals, which is the other requirement of the splitting mechanism described in the problem statement.

Problem 24.* Random incidence in an Erlang arrival process. Consider an arrival process in which the interarrival times are independent Erlang random variables of order 2, with mean $2/\lambda$. Assume that the arrival process has been ongoing for a very long time. An external observer arrives at a given time t . Find the PDF of the length of the interarrival interval that contains t .

Solution. We view the Erlang arrival process in the problem statement as part of a Poisson process with rate λ . In particular, the Erlang arrival process registers an arrival once every two arrivals of the Poisson process. For concreteness, let us say that the Erlang process arrivals correspond to even-numbered arrivals in the Poisson process. Let Y_k be the time of the k th arrival in the Poisson process.

Let K be such that $Y_K \leq t < Y_{K+1}$. By the discussion of random incidence in Poisson processes in the text, we have that $Y_{K+1} - Y_K$ is Erlang of order 2. The interarrival interval for the Erlang process considered in this problem is of the form $[Y_K, Y_{K+2}]$ or $[Y_{K-1}, Y_{K+1}]$, depending on whether K is even or odd, respectively. In

the first case, the interarrival interval in the Erlang process is of the form $(Y_{K+1} - Y_K) + (Y_{K+2} - Y_{K+1})$. We claim that $Y_{K+2} - Y_{K+1}$ is exponential with parameter λ and independent of $Y_{K+1} - Y_K$. Indeed, an observer who arrives at time t and notices that K is even, must first wait until the time Y_{K+1} of the next Poisson arrival. At that time, the Poisson process starts afresh, and the time $Y_{K+2} - Y_{K+1}$ until the next Poisson arrival is independent of the past (hence, independent of $Y_{K+1} - Y_K$) and has an exponential distribution with parameter λ , as claimed. This establishes that, conditioned on K being even, the interarrival interval length $Y_{K+2} - Y_K$ of the Erlang process is Erlang of order 3 (since it is the sum of an exponential random variable and a random variable which is Erlang of order 2). By a symmetrical argument, if we condition on K being odd, the conditional PDF of the interarrival interval length $Y_{K+1} - Y_{K-1}$ of the Erlang process is again the same. Since the conditional PDF of the length of the interarrival interval that contains t is Erlang of order 3, for every conditioning event, it follows that the unconditional PDF is also Erlang of order 3.

Markov Chains

Contents

6.1. Discrete-Time Markov Chains	p. 314
6.2. Classification of States	p. 321
6.3. Steady-State Behavior	p. 326
6.4. Absorption Probabilities and Expected Time to Absorption . .	p. 337
6.5. Continuous-Time Markov Chains	p. 344
6.6. Summary and Discussion	p. 352
Problems	p. 354

The Bernoulli and Poisson processes studied in the preceding chapter are memoryless, in the sense that the future does not depend on the past: the occurrences of new “successes” or “arrivals” do not depend on the past history of the process. In this chapter, we consider processes where the future depends on and can be predicted to some extent by what has happened in the past.

We emphasize models where the effect of the past on the future is summarized by a **state**, which changes over time according to given probabilities. We restrict ourselves to models in which the state can only take a finite number values and can change according to probabilities that do not depend on the time of the change. We want to analyze the probabilistic properties of the sequence of state values.

The range of applications of the type of models described in this chapter is truly vast. It includes just about any dynamical system whose evolution over time involves uncertainty, provided the state of the system is suitably defined. Such systems arise in a broad variety of fields, such as, for example, communications, automatic control, signal processing, manufacturing, economics, and operations research.

6.1 DISCRETE-TIME MARKOV CHAINS

We will first consider **discrete-time Markov chains**, in which the state changes at certain discrete time instants, indexed by an integer variable n . At each time step n , the state of the chain is denoted by X_n , and belongs to a **finite** set \mathcal{S} of possible states, called the **state space**. Without loss of generality, and unless there is a statement to the contrary, we will assume that $\mathcal{S} = \{1, \dots, m\}$, for some positive integer m . The Markov chain is described in terms of its **transition probabilities** p_{ij} : whenever the state happens to be i , there is probability p_{ij} that the next state is equal to j . Mathematically,

$$p_{ij} = \mathbf{P}(X_{n+1} = j \mid X_n = i), \quad i, j \in \mathcal{S}.$$

The key assumption underlying Markov chains is that the transition probabilities p_{ij} apply whenever state i is visited, no matter what happened in the past, and no matter how state i was reached. Mathematically, we assume the **Markov property**, which requires that

$$\begin{aligned} \mathbf{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) &= \mathbf{P}(X_{n+1} = j \mid X_n = i) \\ &= p_{ij}, \end{aligned}$$

for all times n , all states $i, j \in \mathcal{S}$, and all possible sequences i_0, \dots, i_{n-1} of earlier states. Thus, the probability law of the next state X_{n+1} depends on the past only through the value of the present state X_n .

The transition probabilities p_{ij} must be of course nonnegative, and sum to one:

$$\sum_{j=1}^m p_{ij} = 1, \quad \text{for all } i.$$

We will generally allow the probabilities p_{ii} to be positive, in which case it is possible for the next state to be the same as the current one. Even though the state does not change, we still view this as a state transition of a special type (a “self-transition”).

Specification of Markov Models

- A Markov chain model is specified by identifying:
 - the set of states $\mathcal{S} = \{1, \dots, m\}$,
 - the set of possible transitions, namely, those pairs (i, j) for which $p_{ij} > 0$, and,
 - the numerical values of those p_{ij} that are positive.
- The Markov chain specified by this model is a sequence of random variables X_0, X_1, X_2, \dots , that take values in \mathcal{S} , and which satisfy

$$\mathbf{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = p_{ij},$$

for all times n , all states $i, j \in \mathcal{S}$, and all possible sequences i_0, \dots, i_{n-1} of earlier states.

All of the elements of a Markov chain model can be encoded in a **transition probability matrix**, which is simply a two-dimensional array whose element at the i th row and j th column is p_{ij} :

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}.$$

It is also helpful to lay out the model in the so-called **transition probability graph**, whose nodes are the states and whose arcs are the possible transitions. By recording the numerical values of p_{ij} near the corresponding arcs, one can visualize the entire model in a way that can make some of its major properties readily apparent.

Example 6.1. Alice is taking a probability class and in each week, she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given

week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in the given week, the probability that she will be up-to-date (or behind) in the next week is 0.6 (or 0.4, respectively). We assume that these probabilities do not depend on whether she was up-to-date or behind in previous weeks, so the problem has the typical Markov chain character (the future depends on the past only through the present).

Let us introduce states 1 and 2, and identify them with being up-to-date and behind, respectively. Then, the transition probabilities are

$$p_{11} = 0.8, \quad p_{12} = 0.2, \quad p_{21} = 0.6, \quad p_{22} = 0.4,$$

and the transition probability matrix is

$$\begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix}.$$

The transition probability graph is shown in Fig. 6.1.

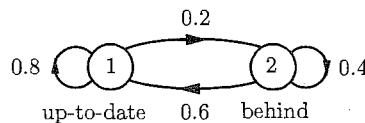


Figure 6.1: The transition probability graph in Example 6.1.

Example 6.2. Spiders and Fly. A fly moves along a straight line in unit increments. At each time period, it moves one unit to the left with probability 0.3, one unit to the right with probability 0.3, and stays in place with probability 0.4, independently of the past history of movements. Two spiders are lurking at positions 1 and m : if the fly lands there, it is captured by a spider, and the process terminates. We want to construct a Markov chain model, assuming that the fly starts in a position between 1 and m .

Let us introduce states $1, 2, \dots, m$, and identify them with the corresponding positions of the fly. The nonzero transition probabilities are

$$p_{11} = 1, \quad p_{mm} = 1,$$

$$p_{ij} = \begin{cases} 0.3, & \text{if } j = i - 1 \text{ or } j = i + 1, \\ 0.4, & \text{if } j = i, \end{cases} \quad \text{for } i = 2, \dots, m - 1.$$

The transition probability graph and matrix are shown in Fig. 6.2.

Example 6.3. Machine Failure, Repair, and Replacement. A machine can be either working or broken down on a given day. If it is working, it will break down in the next day with probability b , and will continue working with probability

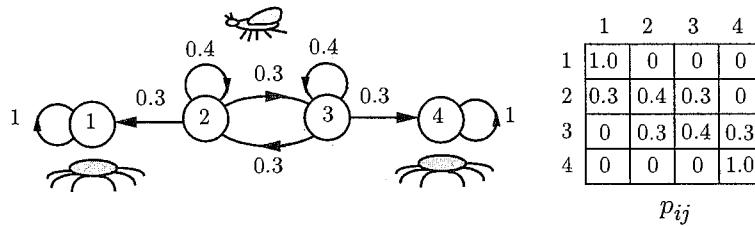


Figure 6.2: The transition probability graph and the transition probability matrix in Example 6.2, for the case where $m = 4$.

$1 - b$. If it breaks down on a given day, it will be repaired and be working in the next day with probability r , and will continue to be broken down with probability $1 - r$.

We model the machine by a Markov chain with the following two states:

State 1: Machine is working, State 2: Machine is broken down.

The transition probability graph of the chain is given in Fig. 6.3. The transition probability matrix is

$$\begin{bmatrix} 1 - b & b \\ r & 1 - r \end{bmatrix}.$$

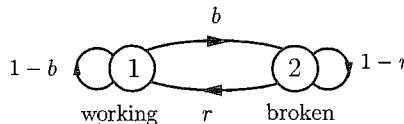


Figure 6.3: Transition probability graph for Example 6.3.

The situation considered here evidently has the Markov property: the state of the machine at the next day depends explicitly only on its state at the present day. However, it is possible to use a Markov chain model even if there is a dependence on the states at several past days. The general idea is to introduce some additional states which encode relevant information from preceding periods, as in the variation that we consider next.

Suppose that whenever the machine remains broken for a given number of ℓ days, despite the repair efforts, it is replaced by a new working machine. To model this as a Markov chain, we replace the single state 2, corresponding to a broken down machine, with several states that indicate the number of days that the machine is broken. These states are

State $(2, i)$: The machine has been broken for i days, $i = 1, 2, \dots, \ell$.

The transition probability graph is given in Fig. 6.4 for the case where $\ell = 4$.

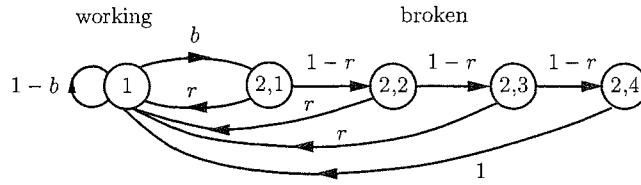


Figure 6.4: Transition probability graph for the second part of Example 6.3. A machine that has remained broken for $\ell = 4$ days is replaced by a new, working machine.

The second half of the preceding example illustrates that in order to construct a Markov model, there is often a need to introduce new states that capture the dependence of the future on the model's past history. We note that there is some freedom in selecting these additional states, but their number should be generally kept small, for reasons of analytical or computational tractability.

The Probability of a Path

Given a Markov chain model, we can compute the probability of any particular sequence of future states. This is analogous to the use of the multiplication rule in sequential (tree) probability models. In particular, we have

$$\mathbf{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \mathbf{P}(X_0 = i_0)p_{i_0 i_1}p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

To verify this property, note that

$$\begin{aligned} \mathbf{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) &= \mathbf{P}(X_n = i_n | X_0 = i_0, \dots, X_{n-1} = i_{n-1})\mathbf{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1}) \\ &= p_{i_{n-1} i_n} \mathbf{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1}), \end{aligned}$$

where the last equality made use of the Markov property. We then apply the same argument to the term $\mathbf{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1})$ and continue similarly, until we eventually obtain the desired expression. If the initial state X_0 is given and is known to be equal to some i_0 , a similar argument yields

$$\mathbf{P}(X_1 = i_1, \dots, X_n = i_n | X_0 = i_0) = p_{i_0 i_1}p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

Graphically, a state sequence can be identified with a sequence of arcs in the transition probability graph, and the probability of such a path (given the initial state) is given by the product of the probabilities associated with the arcs traversed by the path.

Example 6.4. For the spider and fly example (Example 6.2), we have

$$\mathbf{P}(X_1 = 2, X_2 = 2, X_3 = 3, X_4 = 4 | X_0 = 2) = p_{22}p_{22}p_{23}p_{34} = (0.4)^2(0.3)^2.$$

We also have

$$\begin{aligned}\mathbf{P}(X_0 = 2, X_1 = 2, X_2 = 2, X_3 = 3, X_4 = 4) &= \mathbf{P}(X_0 = 2)p_{22}p_{22}p_{23}p_{34} \\ &= \mathbf{P}(X_0 = 2)(0.4)^2(0.3)^2.\end{aligned}$$

Note that in order to calculate a probability of this form, in which there is no conditioning on a fixed initial state, we need to specify a probability law for the initial state X_0 .

***n*-Step Transition Probabilities**

Many Markov chain problems require the calculation of the probability law of the state at some future time, conditioned on the current state. This probability law is captured by the ***n*-step transition probabilities**, defined by

$$r_{ij}(n) = \mathbf{P}(X_n = j \mid X_0 = i).$$

In words, $r_{ij}(n)$ is the probability that the state after n time periods will be j , given that the current state is i . It can be calculated using the following basic recursion, known as the **Chapman-Kolmogorov equation**.

Chapman-Kolmogorov Equation for the *n*-Step Transition Probabilities

The n -step transition probabilities can be generated by the recursive formula

$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1)p_{kj}, \quad \text{for } n > 1, \text{ and all } i, j,$$

starting with

$$r_{ij}(1) = p_{ij}.$$

To verify the formula, we apply the total probability theorem as follows:

$$\begin{aligned}\mathbf{P}(X_n = j \mid X_0 = i) &= \sum_{k=1}^m \mathbf{P}(X_{n-1} = k \mid X_0 = i) \mathbf{P}(X_n = j \mid X_{n-1} = k, X_0 = i) \\ &= \sum_{k=1}^m r_{ik}(n-1)p_{kj};\end{aligned}$$

see Fig. 6.5 for an illustration. We have used here the Markov property: once we condition on $X_{n-1} = k$, the conditioning on $X_0 = i$ does not affect the probability p_{kj} of reaching j at the next step.

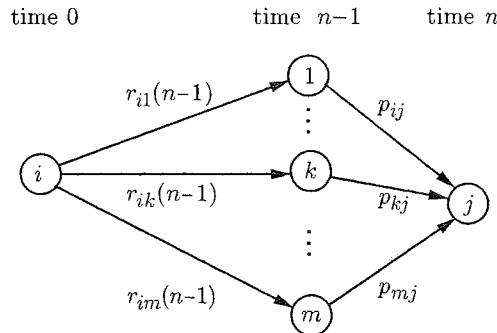
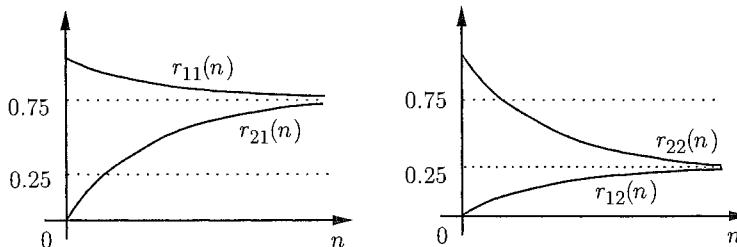


Figure 6.5: Derivation of the Chapman-Kolmogorov equation. The probability of being at state j at time n is the sum of the probabilities $r_{ik}(n-1)p_{kj}$ of the different ways of reaching j .

We can view $r_{ij}(n)$ as the element at the i th row and j th column of a two-dimensional array, called the **n -step transition probability matrix**.[†] Figures 6.6 and 6.7 give the n -step transition probabilities $r_{ij}(n)$ for the cases of Examples 6.1 and 6.2, respectively. There are some interesting observations about the limiting behavior of $r_{ij}(n)$ in these two examples. In Fig. 6.6, we see that each $r_{ij}(n)$ converges to a limit, as $n \rightarrow \infty$, and this limit does not depend on the initial state i . Thus, each state has a positive “steady-state” probability of being occupied at times far into the future. Furthermore, the probability $r_{ij}(n)$ depends on the initial state i when n is small, but over time this dependence diminishes. Many (but by no means all) probabilistic models that evolve over time have such a character: after a sufficiently long time, the effect of their initial condition becomes negligible.

In Fig. 6.7, we see a qualitatively different behavior: $r_{ij}(n)$ again converges, but the limit depends on the initial state, and can be zero for selected states. Here, we have two states that are “absorbing,” in the sense that they are infinitely repeated, once reached. These are the states 1 and 4 that correspond to the capture of the fly by one of the two spiders. Given enough time, it is certain that some absorbing state will be reached. Accordingly, the probability of being at the non-absorbing states 2 and 3 diminishes to zero as time increases. Furthermore, the probability that a particular absorbing state will be reached depends on how “close” we start to that state.

[†] Those readers familiar with matrix multiplication, may recognize that the Chapman-Kolmogorov equation can be expressed as follows: the matrix of n -step transition probabilities $r_{ij}(n)$ is obtained by multiplying the matrix of $(n-1)$ -step transition probabilities $r_{ik}(n-1)$, with the one-step transition probability matrix. Thus, the n -step transition probability matrix is the n th power of the transition probability matrix.

n-step transition probabilities as a function of the number n of transitions

	U	B				
U	0.8	0.2	.76	.24	.752	.248
B	0.6	0.4	.72	.28	.744	.256
	$r_{ij}(1)$		$r_{ij}(2)$		$r_{ij}(3)$	$r_{ij}(4)$
						$r_{ij}(5)$

Sequence of n -step transition probability matrices

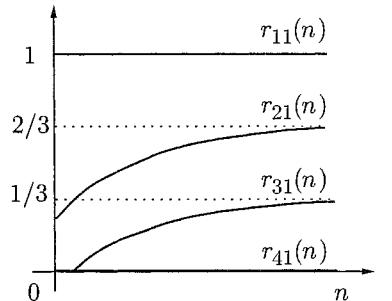
Figure 6.6: n -step transition probabilities for the “up-to-date/behind” Example 6.1. Observe that as n increases, $r_{ij}(n)$ converges to a limit that does not depend on the initial state i .

These examples illustrate that there is a variety of types of states and asymptotic occupancy behavior in Markov chains. We are thus motivated to classify and analyze the various possibilities, and this is the subject of the next three sections.

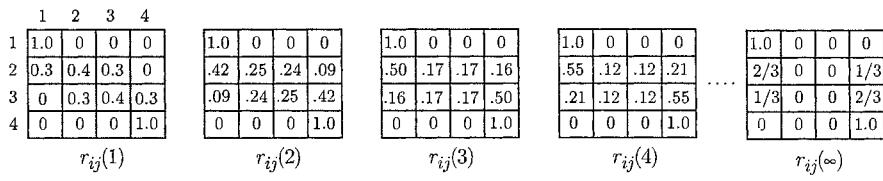
6.2 CLASSIFICATION OF STATES

In the preceding section, we saw some examples indicating that the various states of a Markov chain can have qualitatively different characteristics. In particular, some states, after being visited once, are certain to be visited again, while for some other states this may not be the case. In this section, we focus on the mechanism by which this occurs. In particular, we wish to classify the states of a Markov chain with a focus on the long-term frequency with which they are visited.

As a first step, we make the notion of revisiting a state precise. Let us say that a state j is **accessible** from a state i if for some n , the n -step transition probability $r_{ij}(n)$ is positive, i.e., if there is positive probability of reaching j , starting from i , after some number of time periods. An equivalent definition is that there is a possible state sequence $i, i_1, \dots, i_{n-1}, j$, that starts at i and ends



n-step transition probabilities into state 1



Sequence of transition probability matrices

Figure 6.7: The top part of the figure shows the n -step transition probabilities $r_{i1}(n)$ for the “spiders-and-fly” Example 6.2. These are the probabilities of reaching state 1 by time n , starting from state i . We observe that these probabilities converge to a limit, but the limit depends on the starting state. In this example, note that the probabilities $r_{i2}(n)$ and $r_{i3}(n)$ of being at the non-absorbing states 2 or 3, go to zero, as n increases.

at j , in which the transitions $(i, i_1), (i_1, i_2), \dots, (i_{n-2}, i_{n-1}), (i_{n-1}, j)$ all have positive probability. Let $A(i)$ be the set of states that are accessible from i . We say that i is **recurrent** if for every j that is accessible from i , i is also accessible from j ; that is, for all j that belong to $A(i)$ we have that i belongs to $A(j)$.

When we start at a recurrent state i , we can only visit states $j \in A(i)$ from which i is accessible. Thus, from any future state, there is always some probability of returning to i and, given enough time, this is certain to happen. By repeating this argument, if a recurrent state is visited once, it is certain to be revisited an infinite number of times. (See the end-of-chapter problems for a more rigorous version of this argument.)

A state is called **transient** if it is not recurrent. Thus, a state i is transient if there is a state $j \in A(i)$ such that i is not accessible from j . After each visit to state i , there is positive probability that the state enters such a j . Given enough time, this will happen, and state i cannot be visited after that. Thus, a transient state will only be visited a finite number of times; see again the end-of-chapter problems.

Note that transience or recurrence is determined by the arcs of the tran-

sition probability graph [those pairs (i, j) for which $p_{ij} > 0$] and not by the numerical values of the p_{ij} . Figure 6.8 provides an example of a transition probability graph, and the corresponding recurrent and transient states.

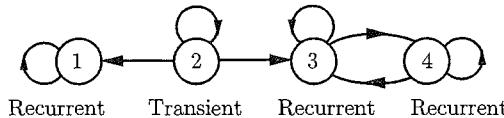


Figure 6.8: Classification of states given the transition probability graph. Starting from state 1, the only accessible state is itself, and so 1 is a recurrent state. States 1, 3, and 4 are accessible from 2, but 2 is not accessible from any of them, so state 2 is transient. States 3 and 4 are accessible from each other, and they are both recurrent.

If i is a recurrent state, the set of states $A(i)$ that are accessible from i form a **recurrent class** (or simply **class**), meaning that states in $A(i)$ are all accessible from each other, and no state outside $A(i)$ is accessible from them. Mathematically, for a recurrent state i , we have $A(i) = A(j)$ for all j that belong to $A(i)$, as can be seen from the definition of recurrence. For example, in the graph of Fig. 6.8, states 3 and 4 form a class, and state 1 by itself also forms a class.

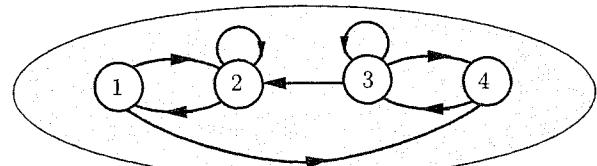
It can be seen that at least one recurrent state must be accessible from any given transient state. This is intuitively evident and is left as an end-of-chapter problem. It follows that there must exist at least one recurrent state, and hence at least one class. Thus, we reach the following conclusion.

Markov Chain Decomposition

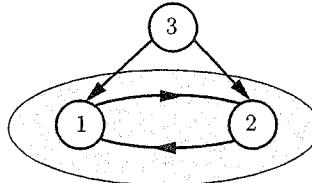
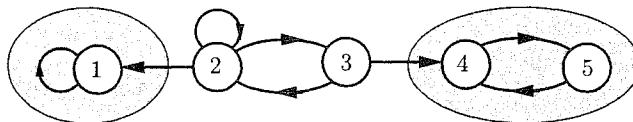
- A Markov chain can be decomposed into one or more recurrent classes, plus possibly some transient states.
- A recurrent state is accessible from all states in its class, but is not accessible from recurrent states in other classes.
- A transient state is not accessible from any recurrent state.
- At least one, possibly more, recurrent states are accessible from a given transient state.

Figure 6.9 provides examples of Markov chain decompositions. Decomposition provides a powerful conceptual tool for reasoning about Markov chains and visualizing the evolution of their state. In particular, we see that:

- (a) once the state enters (or starts in) a class of recurrent states, it stays within



Single class of recurrent states

Single class of recurrent states (1 and 2)
and one transient state (3)Two classes of recurrent states
(class of state 1 and class of states 4 and 5)
and two transient states (2 and 3)**Figure 6.9:** Examples of Markov chain decompositions into recurrent classes and transient states.

that class; since all states in the class are accessible from each other, all states in the class will be visited an infinite number of times;

- (b) if the initial state is transient, then the state trajectory contains an initial portion consisting of transient states and a final portion consisting of recurrent states from the same class.

For the purpose of understanding long-term behavior of Markov chains, it is important to analyze chains that consist of a single recurrent class. For the purpose of understanding short-term behavior, it is also important to analyze the mechanism by which any particular class of recurrent states is entered starting from a given transient state. These two issues, long-term and short-term behavior, are the focus of Sections 6.3 and 6.4, respectively.

Periodicity

There is another important characterization of a recurrent class, which relates to the presence or absence of a certain periodic pattern in the times that a state can be visited. In particular, a recurrent class is said to be **periodic** if its states can be grouped in $d > 1$ disjoint subsets S_1, \dots, S_d so that all transitions from one subset lead to the next subset; see Fig. 6.10. More precisely,

$$\text{if } i \in S_k \text{ and } p_{ij} > 0, \text{ then } \begin{cases} j \in S_{k+1}, & \text{if } k = 1, \dots, d-1, \\ j \in S_1, & \text{if } k = d. \end{cases}$$

A recurrent class that is not periodic, is said to be **aperiodic**.

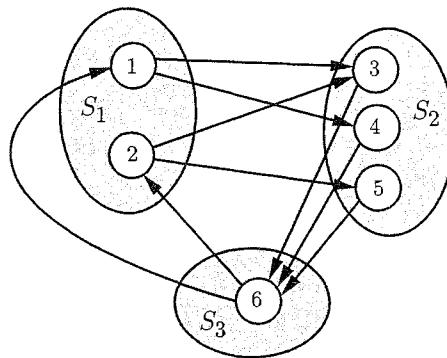


Figure 6.10: Structure of a periodic recurrent class. In this example, $d = 3$.

Thus, in a periodic recurrent class, we move through the sequence of subsets in order, and after d steps, we end up in the same subset. As an example, the recurrent class in the second chain of Fig. 6.9 (states 1 and 2) is periodic, and the same is true of the class consisting of states 4 and 5 in the third chain of Fig. 6.9. All other recurrent classes in the chains of this figure are aperiodic.

Note that given a periodic recurrent class, a positive time n , and a state i in the class, there must exist one or more states j for which $r_{ij}(n) = 0$. The reason is that starting from i , only one of the sets S_k is possible at time n . Thus, a way to verify aperiodicity of a given recurrent class R , is to check whether there is a special time $n \geq 1$ and a special state $i \in R$ from which all states in R can be reached in n steps, i.e., $r_{ij}(n) > 0$ for all $j \in R$. As an example, consider the first chain in Fig. 6.9. Starting from state 1, every state is possible at time $n = 3$, so the unique recurrent class of that chain is aperiodic.

A converse statement, which we do not prove, also turns out to be true: if a recurrent class R is aperiodic, then there exists a time n such that $r_{ij}(n) > 0$ for every i and j in R ; see the end-of-chapter problems.

Periodicity

Consider a recurrent class R .

- The class is called **periodic** if its states can be grouped in $d > 1$ disjoint subsets S_1, \dots, S_d , so that all transitions from S_k lead to S_{k+1} (or to S_1 if $k = d$).
- The class is **aperiodic** (not periodic) if and only if there exists a time n such that $r_{ij}(n) > 0$, for all $i, j \in R$.

6.3 STEADY-STATE BEHAVIOR

In Markov chain models, we are often interested in long-term state occupancy behavior, that is, in the n -step transition probabilities $r_{ij}(n)$ when n is very large. We have seen in the example of Fig. 6.6 that the $r_{ij}(n)$ may converge to steady-state values that are independent of the initial state. We wish to understand the extent to which this behavior is typical.

If there are two or more classes of recurrent states, it is clear that the limiting values of the $r_{ij}(n)$ must depend on the initial state (the possibility of visiting j far into the future depends on whether j is in the same class as the initial state i). We will, therefore, restrict attention to chains involving a single recurrent class, plus possibly some transient states. This is not as restrictive as it may seem, since we know that once the state enters a particular recurrent class, it will stay within that class. Thus, the asymptotic behavior of a multiclass chain can be understood in terms of the asymptotic behavior of a single-class chain.

Even for chains with a single recurrent class, the $r_{ij}(n)$ may fail to converge. To see this, consider a recurrent class with two states, 1 and 2, such that from state 1 we can only go to 2, and from 2 we can only go to 1 ($p_{12} = p_{21} = 1$). Then, starting at some state, we will be in that same state after any even number of transitions, and in the other state after any odd number of transitions. Formally,

$$r_{ii}(n) = \begin{cases} 1, & n \text{ even,} \\ 0, & n \text{ odd.} \end{cases}$$

What is happening here is that the recurrent class is periodic, and for such a class, it can be seen that the $r_{ij}(n)$ generically oscillate.

We now assert that for every state j , the probability $r_{ij}(n)$ of being at state j approaches a limiting value that is independent of the initial state i , provided we exclude the two situations discussed above (multiple recurrent classes and/or a periodic class). This limiting value, denoted by π_j , has the interpretation

$$\pi_j \approx \mathbf{P}(X_n = j), \quad \text{when } n \text{ is large,}$$

and is called the **steady-state probability** of j . The following is an important theorem. Its proof is quite complicated and is outlined together with several other proofs in the end-of-chapter problems.

Steady-State Convergence Theorem

Consider a Markov chain with a single recurrent class, which is aperiodic. Then, the states j are associated with steady-state probabilities π_j that have the following properties.

- (a) For each j , we have

$$\lim_{n \rightarrow \infty} r_{ij}(n) = \pi_j, \quad \text{for all } i.$$

- (b) The π_j are the unique solution to the system of equations below:

$$\begin{aligned} \pi_j &= \sum_{k=1}^m \pi_k p_{kj}, \quad j = 1, \dots, m, \\ 1 &= \sum_{k=1}^m \pi_k. \end{aligned}$$

- (c) We have

$$\begin{aligned} \pi_j &= 0, & \text{for all transient states } j, \\ \pi_j &> 0, & \text{for all recurrent states } j. \end{aligned}$$

The steady-state probabilities π_j sum to 1 and form a probability distribution on the state space, called the **stationary distribution** of the chain. The reason for the qualification “stationary” is that if the initial state is chosen according to this distribution, i.e., if

$$\mathbf{P}(X_0 = j) = \pi_j, \quad j = 1, \dots, m,$$

then, using the total probability theorem, we have

$$\mathbf{P}(X_1 = j) = \sum_{k=1}^m \mathbf{P}(X_0 = k) p_{kj} = \sum_{k=1}^m \pi_k p_{kj} = \pi_j,$$

where the last equality follows from part (b) of the steady-state convergence theorem. Similarly, we obtain $\mathbf{P}(X_n = j) = \pi_j$, for all n and j . Thus, if the initial state is chosen according to the stationary distribution, the state at any future time will have the same distribution.

The equations

$$\pi_j = \sum_{k=1}^m \pi_k p_{kj}, \quad j = 1, \dots, m,$$

are called the **balance equations**. They are a simple consequence of part (a) of the theorem and the Chapman-Kolmogorov equation. Indeed, once the convergence of $r_{ij}(n)$ to some π_j is taken for granted, we can consider the equation,

$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1) p_{kj},$$

take the limit of both sides as $n \rightarrow \infty$, and recover the balance equations.[†] Together with the **normalization equation**

$$\sum_{k=1}^m \pi_k = 1,$$

the balance equations can be solved to obtain the π_j . The following examples illustrate the solution process.

Example 6.5. Consider a two-state Markov chain with transition probabilities

$$\begin{aligned} p_{11} &= 0.8, & p_{12} &= 0.2, \\ p_{21} &= 0.6, & p_{22} &= 0.4. \end{aligned}$$

(This is the same as the chain of Example 6.1 and Fig. 6.1.) The balance equations take the form

$$\pi_1 = \pi_1 p_{11} + \pi_2 p_{21}, \quad \pi_2 = \pi_1 p_{12} + \pi_2 p_{22},$$

or

$$\pi_1 = 0.8 \cdot \pi_1 + 0.6 \cdot \pi_2, \quad \pi_2 = 0.2 \cdot \pi_1 + 0.4 \cdot \pi_2.$$

Note that the above two equations are dependent, since they are both equivalent to

$$\pi_1 = 3\pi_2.$$

[†] According to a famous and important theorem from linear algebra (called the Perron-Frobenius theorem), the balance equations always have a nonnegative solution, for any Markov chain. What is special about a chain that has a single recurrent class, which is aperiodic, is that given also the normalization equation, the solution is unique and is equal to the limit of the n -step transition probabilities $r_{ij}(n)$.

This is a generic property, and in fact it can be shown that any one of the balance equations can always be derived from the remaining equations. However, we also know that the π_j satisfy the normalization equation

$$\pi_1 + \pi_2 = 1,$$

which supplements the balance equations and suffices to determine the π_j uniquely. Indeed, by substituting the equation $\pi_1 = 3\pi_2$ into the equation $\pi_1 + \pi_2 = 1$, we obtain $3\pi_2 + \pi_2 = 1$, or

$$\pi_2 = 0.25,$$

which using the equation $\pi_1 + \pi_2 = 1$, yields

$$\pi_1 = 0.75.$$

This is consistent with what we found earlier by iterating the Chapman-Kolmogorov equation (cf. Fig. 6.6).

Example 6.6. An absent-minded professor has two umbrellas that she uses when commuting from home to office and back. If it rains and an umbrella is available in her location, she takes it. If it is not raining, she always forgets to take an umbrella. Suppose that it rains with probability p each time she commutes, independently of other times. What is the steady-state probability that she gets wet during a commute?

We model this problem using a Markov chain with the following states:

State i : i umbrellas are available in her current location, $i = 0, 1, 2$.

The transition probability graph is given in Fig. 6.11, and the transition probability matrix is

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1-p & p \\ 1-p & p & 0 \end{bmatrix}.$$

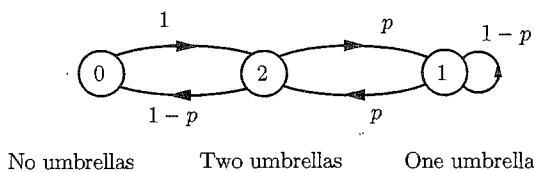


Figure 6.11: Transition probability graph for Example 6.6.

The chain has a single recurrent class that is aperiodic (assuming $0 < p < 1$), so the steady-state convergence theorem applies. The balance equations are

$$\pi_0 = (1-p)\pi_2, \quad \pi_1 = (1-p)\pi_1 + p\pi_2, \quad \pi_2 = \pi_0 + p\pi_1.$$

From the second equation, we obtain $\pi_1 = \pi_2$, which together with the first equation $\pi_0 = (1 - p)\pi_2$ and the normalization equation $\pi_0 + \pi_1 + \pi_2 = 1$, yields

$$\pi_0 = \frac{1 - p}{3 - p}, \quad \pi_1 = \frac{1}{3 - p}, \quad \pi_2 = \frac{1}{3 - p}.$$

According to the steady-state convergence theorem, the steady-state probability that the professor finds herself in a place without an umbrella is π_0 . The steady-state probability that she gets wet is π_0 times the probability of rain p .

Example 6.7. A superstitious professor works in a circular building with m doors, where m is odd, and never uses the same door twice in a row. Instead he uses with probability p (or probability $1 - p$) the door that is adjacent in the clockwise direction (or the counterclockwise direction, respectively) to the door he used last. What is the probability that a given door will be used on some particular day far into the future?

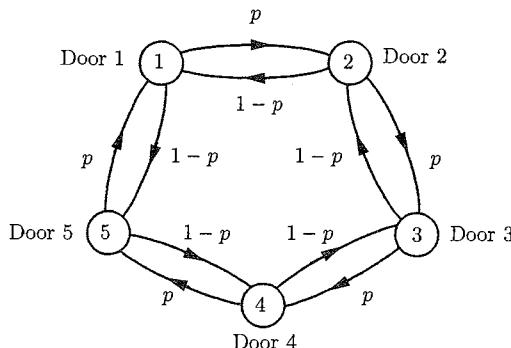


Figure 6.12: Transition probability graph in Example 6.7, for the case of $m = 5$ doors. Assuming that $0 < p < 1$, it is not hard to see that given an initial state j can be reached in exactly 5 steps, and therefore the chain is aperiodic.

We introduce a Márkov chain with the following m states:

State i : last door used is door i , $i = 1, \dots, m$.

The transition probability graph of the chain is given in Fig. 6.12, for the case $m = 5$. The transition probability matrix is

$$\begin{bmatrix} 0 & p & 0 & 0 & \dots & 0 & 1 - p \\ 1 - p & 0 & p & 0 & \dots & 0 & 0 \\ 0 & 1 - p & 0 & p & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p & 0 & 0 & 0 & \dots & 1 - p & 0 \end{bmatrix}.$$

Assuming that $0 < p < 1$, the chain has a single recurrent class that is aperiodic. (To verify aperiodicity, we leave it to the reader to verify that given an initial state, every state j can be reached in exactly m steps, and the criterion for aperiodicity given at the end of the preceding section is satisfied.) The balance equations are

$$\begin{aligned}\pi_1 &= (1-p)\pi_2 + p\pi_m, \\ \pi_i &= p\pi_{i-1} + (1-p)\pi_{i+1}, \quad i = 2, \dots, m-1, \\ \pi_m &= (1-p)\pi_1 + p\pi_{m-1}.\end{aligned}$$

These equations are easily solved once we observe that by symmetry, all doors should have the same steady-state probability. This suggests the solution

$$\pi_j = \frac{1}{m}, \quad j = 1, \dots, m.$$

Indeed, we see that these π_j satisfy the balance equations as well as the normalization equation, so they must be the desired steady-state probabilities (by the uniqueness part of the steady-state convergence theorem).

Note that if either $p = 0$ or $p = 1$, the chain still has a single recurrent class but is periodic. In this case, the n -step transition probabilities $r_{ij}(n)$ do not converge to a limit, because the doors are used in a cyclic order. Similarly, if m is even, the recurrent class of the chain is periodic, since the states can be grouped into two subsets, the even and the odd numbered states, such that from each subset one can only go to the other subset.

Long-Term Frequency Interpretations

Probabilities are often interpreted as relative frequencies in an infinitely long string of independent trials. The steady-state probabilities of a Markov chain admit a similar interpretation, despite the absence of independence.

Consider, for example, a Markov chain involving a machine, which at the end of any day can be in one of two states, working or broken down. Each time it breaks down, it is immediately repaired at a cost of \$1. How are we to model the long-term expected cost of repair per day? One possibility is to view it as the expected value of the repair cost on a randomly chosen day far into the future; this is just the steady-state probability of the broken down state. Alternatively, we can calculate the total expected repair cost in n days, where n is very large, and divide it by n . Intuition suggests that these two methods of calculation should give the same result. Theory supports this intuition, and in general we have the following interpretation of steady-state probabilities (a justification is given in the end-of-chapter problems).

Steady-State Probabilities as Expected State Frequencies

For a Markov chain with a single class which is aperiodic, the steady-state probabilities π_j satisfy

$$\pi_j = \lim_{n \rightarrow \infty} \frac{v_{ij}(n)}{n},$$

where $v_{ij}(n)$ is the expected value of the number of visits to state j within the first n transitions, starting from state i .

Based on this interpretation, π_j is the long-term expected fraction of time that the state is equal to j . Each time that state j is visited, there is probability p_{jk} that the next transition takes us to state k . We conclude that $\pi_j p_{jk}$ can be viewed as the long-term expected fraction of transitions that move the state from j to k .[†]

Expected Frequency of a Particular Transition

Consider n transitions of a Markov chain with a single class which is aperiodic, starting from a given initial state. Let $q_{jk}(n)$ be the expected number of such transitions that take the state from j to k . Then, regardless of the initial state, we have

$$\lim_{n \rightarrow \infty} \frac{q_{jk}(n)}{n} = \pi_j p_{jk}.$$

The frequency interpretation of π_j and $\pi_j p_{jk}$ allows for a simple interpretation of the balance equations. The state is equal to j if and only if there is a transition that brings the state to j . Thus, the expected frequency π_j of visits to j is equal to the sum of the expected frequencies $\pi_k p_{kj}$ of transitions that lead to j , and

$$\pi_j = \sum_{k=1}^m \pi_k p_{kj};$$

see Fig. 6.13.

[†] In fact, some stronger statements are also true, such as the following. Whenever we carry out a probabilistic experiment and generate a trajectory of the Markov chain over an infinite time horizon, the observed long-term frequency with which state j is visited will be exactly equal to π_j , and the observed long-term frequency of transitions from j to k will be exactly equal to $\pi_j p_{jk}$. Even though the trajectory is random, these equalities hold with essential certainty, that is, with probability 1. The exact meaning of such a statement will become more apparent in the next chapter, when we discuss concepts related to the limiting behavior of random processes.

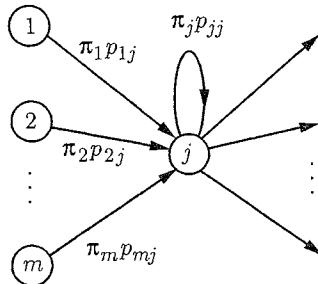


Figure 6.13: Interpretation of the balance equations in terms of frequencies. In a very large number of transitions, we expect a fraction $\pi_k p_{kj}$ that bring the state from k to j . (This also applies to transitions from j to itself, which occur with frequency $\pi_j p_{jj}$.) The sum of the expected frequencies of such transitions is the expected frequency π_j of being at state j .

Birth-Death Processes

A **birth-death** process is a Markov chain in which the states are linearly arranged and transitions can only occur to a neighboring state, or else leave the state unchanged. They arise in many contexts, especially in queueing theory. Figure 6.14 shows the general structure of a birth-death process and also introduces some generic notation for the transition probabilities. In particular,

$$\begin{aligned} b_i &= \mathbf{P}(X_{n+1} = i+1 \mid X_n = i), && \text{("birth" probability at state } i\text{),} \\ d_i &= \mathbf{P}(X_{n+1} = i-1 \mid X_n = i), && \text{("death" probability at state } i\text{).} \end{aligned}$$

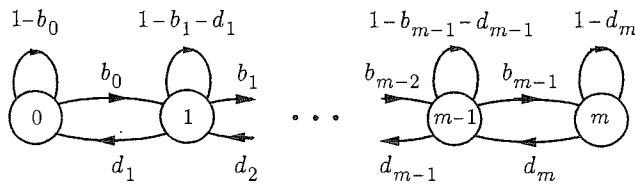


Figure 6.14: Transition probability graph for a birth-death process.

For a birth-death process, the balance equations can be substantially simplified. Let us focus on two neighboring states, say, i and $i+1$. In any trajectory of the Markov chain, a transition from i to $i+1$ has to be followed by a transition from $i+1$ to i , before another transition from i to $i+1$ can occur. Therefore, the expected frequency of transitions from i to $i+1$, which is $\pi_i b_i$, must be equal

to the expected frequency of transitions from $i + 1$ to i , which is $\pi_{i+1}d_{i+1}$. This leads to the **local balance** equations[†]

$$\pi_i b_i = \pi_{i+1} d_{i+1}, \quad i = 0, 1, \dots, m - 1.$$

Using the local balance equations, we obtain

$$\pi_i = \pi_0 \frac{b_0 b_1 \cdots b_{i-1}}{d_1 d_2 \cdots d_i}, \quad i = 1, \dots, m,$$

from which, using also the normalization equation $\sum_i \pi_i = 1$, the steady-state probabilities π_i are easily computed.

Example 6.8. Random Walk with Reflecting Barriers. A person walks along a straight line and, at each time period, takes a step to the right with probability b , and a step to the left with probability $1 - b$. The person starts in one of the positions $1, 2, \dots, m$, but if he reaches position 0 (or position $m + 1$), his step is instantly reflected back to position 1 (or position m , respectively). Equivalently, we may assume that when the person is in positions 1 or m , he will stay in that position with corresponding probability $1 - b$ and b , respectively. We introduce a Markov chain model whose states are the positions $1, \dots, m$. The transition probability graph of the chain is given in Fig. 6.15.

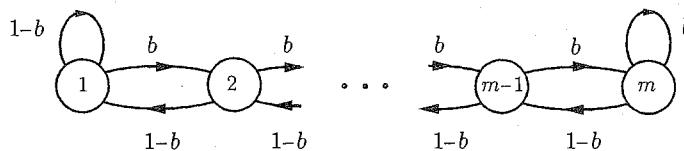


Figure 6.15: Transition probability graph for the random walk Example 6.8.

The local balance equations are

$$\pi_i b = \pi_{i+1} (1 - b), \quad i = 1, \dots, m - 1.$$

Thus, $\pi_{i+1} = \rho \pi_i$, where

$$\rho = \frac{b}{1 - b},$$

[†] A more formal derivation that does not rely on the frequency interpretation proceeds as follows. The balance equation at state 0 is $\pi_0(1 - b_0) + \pi_1 d_1 = \pi_0$, which yields the first local balance equation $\pi_0 b_0 = \pi_1 d_1$.

The balance equation at state 1 is $\pi_0 b_0 + \pi_1 (1 - b_1 - d_1) + \pi_2 d_2 = \pi_1$. Using the local balance equation $\pi_0 b_0 = \pi_1 d_1$ at the previous state, this is rewritten as $\pi_1 d_1 + \pi_1 (1 - b_1 - d_1) + \pi_2 d_2 = \pi_1$, which simplifies to $\pi_1 b_1 = \pi_2 d_2$. We can then continue similarly to obtain the local balance equations at all other states.

and we can express all the π_j in terms of π_1 , as

$$\pi_i = \rho^{i-1} \pi_1, \quad i = 1, \dots, m.$$

Using the normalization equation $1 = \pi_1 + \dots + \pi_m$, we obtain

$$1 = \pi_1(1 + \rho + \dots + \rho^{m-1})$$

which leads to

$$\pi_i = \frac{\rho^{i-1}}{1 + \rho + \dots + \rho^{m-1}}, \quad i = 1, \dots, m.$$

Note that if $\rho = 1$ (left and right steps are equally likely), then $\pi_i = 1/m$ for all i .

Example 6.9. Queueing. Packets arrive at a node of a communication network, where they are stored in a buffer and then transmitted. The storage capacity of the buffer is m : if m packets are already present, any newly arriving packets are discarded. We discretize time in very small periods, and we assume that in each period, at most one event can happen that can change the number of packets stored in the node (an arrival of a new packet or a completion of the transmission of an existing packet). In particular, we assume that at each period, exactly one of the following occurs:

- (a) one new packet arrives; this happens with a given probability $b > 0$;
- (b) one existing packet completes transmission; this happens with a given probability $d > 0$ if there is at least one packet in the node, and with probability 0 otherwise;
- (c) no new packet arrives and no existing packet completes transmission; this happens with probability $1 - b - d$ if there is at least one packet in the node, and with probability $1 - b$ otherwise.

We introduce a Markov chain with states $0, 1, \dots, m$, corresponding to the number of packets in the buffer. The transition probability graph is given in Fig. 6.16.

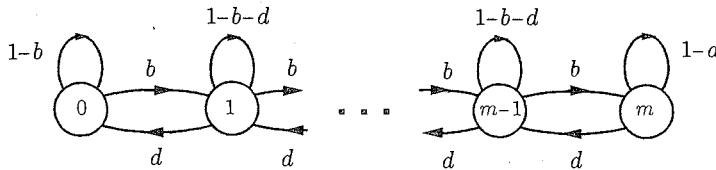


Figure 6.16: Transition probability graph in Example 6.9.

The local balance equations are

$$\pi_i b = \pi_{i+1} d, \quad i = 0, 1, \dots, m-1.$$

We define

$$\rho = \frac{b}{d},$$

and obtain $\pi_{i+1} = \rho\pi_i$, which leads to

$$\pi_i = \rho^i \pi_0, \quad i = 0, 1, \dots, m.$$

By using the normalization equation $1 = \pi_0 + \pi_1 + \dots + \pi_m$, we obtain

$$1 = \pi_0(1 + \rho + \dots + \rho^m),$$

and

$$\pi_0 = \begin{cases} \frac{1 - \rho}{1 - \rho^{m+1}}, & \text{if } \rho \neq 1, \\ \frac{1}{m+1}, & \text{if } \rho = 1. \end{cases}$$

Using the equation $\pi_i = \rho^i \pi_0$, the steady-state probabilities are

$$\pi_i = \begin{cases} \frac{1 - \rho}{1 - \rho^{m+1}} \rho^i, & \text{if } \rho \neq 1, \\ \frac{1}{m+1}, & \text{if } \rho = 1, \end{cases} \quad i = 0, 1, \dots, m.$$

It is interesting to consider what happens when the buffer size m is so large that it can be considered as practically infinite. We distinguish two cases.

- (a) Suppose that $b < d$, or $\rho < 1$. In this case, arrivals of new packets are less likely than departures of existing packets. This prevents the number of packets in the buffer from growing, and the steady-state probabilities π_i decrease with i , as in a (truncated) geometric PMF. We observe that as $m \rightarrow \infty$, we have $1 - \rho^{m+1} \rightarrow 1$, and

$$\pi_i \rightarrow \rho^i(1 - \rho), \quad \text{for all } i.$$

We can view these as the steady-state probabilities in a system with an infinite buffer. [As a check, note that we have $\sum_{i=0}^{\infty} \rho^i(1 - \rho) = 1$.]

- (b) Suppose that $b > d$, or $\rho > 1$. In this case, arrivals of new packets are more likely than departures of existing packets. The number of packets in the buffer tends to increase, and the steady-state probabilities π_i increase with i . As we consider larger and larger buffer sizes m , the steady-state probability of any fixed state i decreases to zero:

$$\pi_i \rightarrow 0, \quad \text{for all } i.$$

Were we to consider a system with an infinite buffer, we would have a Markov chain with a countably infinite number of states. Although we do not have the machinery to study such chains, the preceding calculation suggests that every state will have zero steady-state probability and will be “transient.” The number of packets in queue will generally grow to infinity, and any particular state will be visited only a finite number of times.

The preceding analysis provides a glimpse into the character of Markov chains with an infinite number of states. In such chains, even if there is a single and aperiodic recurrent class, the chain may never reach steady-state and a steady-state distribution may not exist.

6.4 ABSORPTION PROBABILITIES AND EXPECTED TIME TO ABSORPTION

In this section, we study the short-term behavior of Markov chains. We first consider the case where the Markov chain starts at a transient state. We are interested in the first recurrent state to be entered, as well as in the time until this happens.

When addressing such questions, the subsequent behavior of the Markov chain (after a recurrent state is encountered) is immaterial. We can therefore focus on the case where every recurrent state k is **absorbing**, i.e.,

$$p_{kk} = 1, \quad p_{kj} = 0 \text{ for all } j \neq k.$$

If there is a unique absorbing state k , its steady-state probability is 1 (because all other states are transient and have zero steady-state probability), and will be reached with probability 1, starting from any initial state. If there are multiple absorbing states, the probability that one of them will be eventually reached is still 1, but the identity of the absorbing state to be entered is random and the associated probabilities may depend on the starting state. In the sequel, we fix a particular absorbing state, denoted by s , and consider the absorption probability a_i that s is eventually reached, starting from i :

$$a_i = \mathbf{P}(X_n \text{ eventually becomes equal to the absorbing state } s \mid X_0 = i).$$

Absorption probabilities can be obtained by solving a system of linear equations, as indicated below.

Absorption Probability Equations

Consider a Markov chain where each state is either transient or absorbing, and fix a particular absorbing state s . Then, the probabilities a_i of eventually reaching state s , starting from i , are the unique solution to the equations

$$\begin{aligned} a_s &= 1, \\ a_i &= 0, \quad \text{for all absorbing } i \neq s, \\ a_i &= \sum_{j=1}^m p_{ij} a_j, \quad \text{for all transient } i. \end{aligned}$$

The equations $a_s = 1$, and $a_i = 0$, for all absorbing $i \neq s$, are evident from the definitions. To verify the remaining equations, we argue as follows. Let us consider a transient state i and let A be the event that state s is eventually

reached. We have

$$\begin{aligned}
 a_i &= \mathbf{P}(A \mid X_0 = i) \\
 &= \sum_{j=1}^m \mathbf{P}(A \mid X_0 = i, X_1 = j) \mathbf{P}(X_1 = j \mid X_0 = i) \quad (\text{total probability thm.}) \\
 &= \sum_{j=1}^m \mathbf{P}(A \mid X_1 = j) p_{ij} \quad (\text{Markov property}) \\
 &= \sum_{j=1}^m a_j p_{ij}.
 \end{aligned}$$

The uniqueness property of the solution to the absorption probability equations requires a separate argument, which is given in the end-of-chapter problems.

The next example illustrates how we can use the preceding method to calculate the probability of entering a given recurrent class (rather than a given absorbing state).

Example 6.10. Consider the Markov chain shown in Fig. 6.17(a). Note that there are two recurrent classes, namely $\{1\}$ and $\{4, 5\}$. We would like to calculate the probability that the state eventually enters the recurrent class $\{4, 5\}$ starting from one of the transient states. For the purposes of this problem, the possible transitions within the recurrent class $\{4, 5\}$ are immaterial. We can therefore lump the states in this recurrent class and treat them as a single absorbing state (call it state 6), as in Fig. 6.17(b). It then suffices to compute the probability of eventually entering state 6 in this new chain.

The probabilities of eventually reaching state 6, starting from the transient states 2 and 3, satisfy the following equations:

$$\begin{aligned}
 a_2 &= 0.2a_1 + 0.3a_2 + 0.4a_3 + 0.1a_6, \\
 a_3 &= 0.2a_2 + 0.8a_6.
 \end{aligned}$$

Using the facts $a_1 = 0$ and $a_6 = 1$, we obtain

$$\begin{aligned}
 a_2 &= 0.3a_2 + 0.4a_3 + 0.1, \\
 a_3 &= 0.2a_2 + 0.8.
 \end{aligned}$$

This is a system of two equations in the two unknowns a_2 and a_3 , which can be readily solved to yield $a_2 = 21/31$ and $a_3 = 29/31$.

Example 6.11. Gambler's Ruin. A gambler wins \$1 at each round, with probability p , and loses \$1, with probability $1 - p$. Different rounds are assumed independent. The gambler plays continuously until he either accumulates a target amount of $\$m$, or loses all his money. What is the probability of eventually accumulating the target amount (winning) or of losing his fortune?

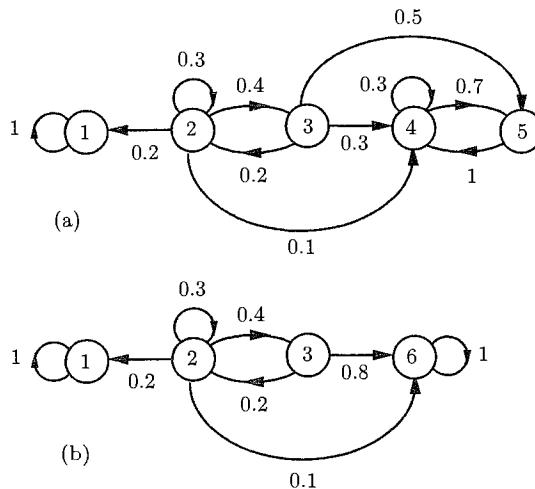


Figure 6.17: (a) Transition probability graph in Example 6.10. (b) A new graph in which states 4 and 5 have been lumped into the absorbing state 6.

We introduce the Markov chain shown in Fig. 6.18 whose state i represents the gambler's wealth at the beginning of a round. The states $i = 0$ and $i = m$ correspond to losing and winning, respectively.

All states are transient, except for the winning and losing states which are absorbing. Thus, the problem amounts to finding the probabilities of absorption at each one of these two absorbing states. Of course, these absorption probabilities depend on the initial state i .

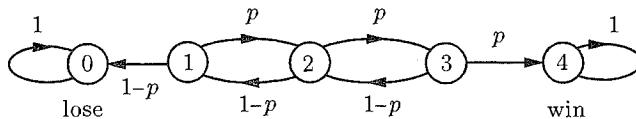


Figure 6.18: Transition probability graph for the gambler's ruin problem (Example 6.11). Here $m = 4$.

Let us set $s = m$ in which case the absorption probability a_i is the probability of winning, starting from state i . These probabilities satisfy

$$a_0 = 0,$$

$$a_i = (1 - p)a_{i-1} + pa_{i+1}, \quad i = 1, \dots, m - 1,$$

$$a_m = 1.$$

These equations can be solved in a variety of ways. It turns out there is an elegant method that leads to a nice closed form solution.

Let us write the equations for the a_i as

$$(1-p)(a_i - a_{i-1}) = p(a_{i+1} - a_i), \quad i = 1, \dots, m-1.$$

Then, by denoting

$$\delta_i = a_{i+1} - a_i, \quad i = 0, \dots, m-1,$$

and

$$\rho = \frac{1-p}{p},$$

the equations are written as

$$\delta_i = \rho \delta_{i-1}, \quad i = 1, \dots, m-1,$$

from which we obtain

$$\delta_i = \rho^i \delta_0, \quad i = 1, \dots, m-1.$$

This, together with the equation $\delta_0 + \delta_1 + \dots + \delta_{m-1} = a_m - a_0 = 1$, implies that

$$(1 + \rho + \dots + \rho^{m-1}) \delta_0 = 1,$$

and

$$\delta_0 = \frac{1}{1 + \rho + \dots + \rho^{m-1}}.$$

Since $a_0 = 0$ and $a_{i+1} = a_i + \delta_i$, the probability a_i of winning starting from a fortune i is equal to

$$\begin{aligned} a_i &= \delta_0 + \delta_1 + \dots + \delta_{i-1} \\ &= (1 + \rho + \dots + \rho^{i-1}) \delta_0 \\ &= \frac{1 + \rho + \dots + \rho^{i-1}}{1 + \rho + \dots + \rho^{m-1}}, \end{aligned}$$

which simplifies to

$$a_i = \begin{cases} \frac{1 - \rho^i}{1 - \rho^m}, & \text{if } \rho \neq 1, \\ \frac{i}{m}, & \text{if } \rho = 1. \end{cases}$$

The solution reveals that if $\rho > 1$, which corresponds to $p < 1/2$ and unfavorable odds for the gambler, the probability of winning approaches 0 as $m \rightarrow \infty$, for any fixed initial fortune. This suggests that if you aim for a large profit under unfavorable odds, financial ruin is almost certain.

Expected Time to Absorption

We now turn our attention to the expected number of steps until a recurrent state is entered (an event that we refer to as “absorption”), starting from a particular transient state. For any state i , we denote

$$\begin{aligned}\mu_i &= \mathbf{E}[\text{number of transitions until absorption, starting from } i] \\ &= \mathbf{E}[\min\{n \geq 0 \mid X_n \text{ is recurrent}\} \mid X_0 = i].\end{aligned}$$

Note that if i is recurrent, then $\mu_i = 0$ according to this definition.

We can derive equations for the μ_i by using the total expectation theorem. We argue that the time to absorption starting from a transient state i is equal to 1 plus the expected time to absorption starting from the next state, which is j with probability p_{ij} . We then obtain a system of linear equations, stated below, which has a unique solution (see the end-of-chapter problems).

Equations for the Expected Time to Absorption

The expected times to absorption, μ_1, \dots, μ_m , are the unique solution to the equations

$$\begin{aligned}\mu_i &= 0, && \text{for all recurrent states } i, \\ \mu_i &= 1 + \sum_{j=1}^m p_{ij} \mu_j, && \text{for all transient states } i.\end{aligned}$$

Example 6.12. Spiders and Fly. Consider the spiders-and-fly model of Example 6.2. This corresponds to the Markov chain shown in Fig. 6.19. The states correspond to possible fly positions, and the absorbing states 1 and m correspond to capture by a spider.

Let us calculate the expected number of steps until the fly is captured. We have

$$\mu_1 = \mu_m = 0,$$

and

$$\mu_i = 1 + 0.3\mu_{i-1} + 0.4\mu_i + 0.3\mu_{i+1}, \quad \text{for } i = 2, \dots, m-1.$$

We can solve these equations in a variety of ways, such as for example by successive substitution. As an illustration, let $m = 4$, in which case, the equations reduce to

$$\mu_2 = 1 + 0.4\mu_2 + 0.3\mu_3, \quad \mu_3 = 1 + 0.3\mu_2 + 0.4\mu_3.$$

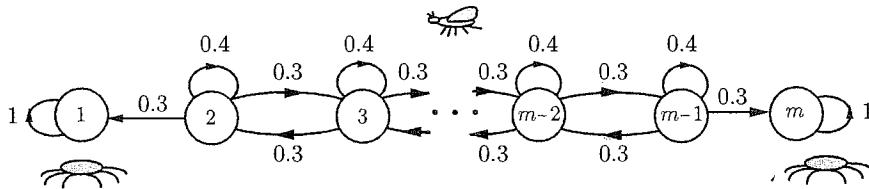


Figure 6.19: Transition probability graph in Example 6.12.

The first equation yields $\mu_2 = (1/0.6) + (1/2)\mu_3$, which we can substitute in the second equation and solve for μ_3 . We obtain $\mu_3 = 10/3$ and by substitution again, $\mu_2 = 10/3$.

Mean First Passage and Recurrence Times

The idea used to calculate the expected time to absorption can also be used to calculate the expected time to reach a particular recurrent state, starting from any other state. For simplicity, we consider a Markov chain with a single recurrent class. We focus on a special recurrent state s , and we denote by t_i the **mean first passage time from state i to state s** , defined by

$$\begin{aligned} t_i &= \mathbf{E}[\text{number of transitions to reach } s \text{ for the first time, starting from } i] \\ &= \mathbf{E}[\min\{n \geq 0 \mid X_n = s\} \mid X_0 = i]. \end{aligned}$$

The transitions out of state s are irrelevant to the calculation of the mean first passage times. We may thus consider a new Markov chain which is identical to the original, except that the special state s is converted into an absorbing state (by setting $p_{ss} = 1$, and $p_{sj} = 0$ for all $j \neq s$). With this transformation, all states other than s become transient. We then compute t_i as the expected number of steps to absorption starting from i , using the formulas given earlier in this section. We have

$$\begin{aligned} t_i &= 1 + \sum_{j=1}^m p_{ij} t_j, \quad \text{for all } i \neq s, \\ t_s &= 0. \end{aligned}$$

This system of linear equations can be solved for the unknowns t_i , and has a unique solution (see the end-of-chapter problems).

The above equations give the expected time to reach the special state s starting from any other state. We may also want to calculate the **mean recurrence time** of the special state s , which is defined as

$$\begin{aligned} t_s^* &= \mathbf{E}[\text{number of transitions up to the first return to } s, \text{ starting from } s] \\ &= \mathbf{E}[\min\{n \geq 1 \mid X_n = s\} \mid X_0 = s]. \end{aligned}$$

We can obtain t_s^* , once we have the first passage times t_i , by using the equation

$$t_s^* = 1 + \sum_{j=1}^m p_{sj} t_j.$$

To justify this equation, we argue that the time to return to s , starting from s , is equal to 1 plus the expected time to reach s from the next state, which is j with probability p_{sj} . We then apply the total expectation theorem.

Example 6.13. Consider the “up-to-date”–“behind” model of Example 6.1. States 1 and 2 correspond to being up-to-date and being behind, respectively, and the transition probabilities are

$$\begin{aligned} p_{11} &= 0.8, & p_{12} &= 0.2, \\ p_{21} &= 0.6, & p_{22} &= 0.4. \end{aligned}$$

Let us focus on state $s = 1$ and calculate the mean first passage time to state 1, starting from state 2. We have $t_1 = 0$ and

$$t_2 = 1 + p_{21} t_1 + p_{22} t_2 = 1 + 0.4 t_2,$$

from which

$$t_2 = \frac{1}{0.6} = \frac{5}{3}.$$

The mean recurrence time to state 1 is given by

$$t_1^* = 1 + p_{11} t_1 + p_{12} t_2 = 1 + 0 + 0.2 \cdot \frac{5}{3} = \frac{4}{3}.$$

Equations for Mean First Passage and Recurrence Times

Consider a Markov chain with a single recurrent class, and let s be a particular recurrent state.

- The mean first passage times t_i to reach state s starting from i , are the unique solution to the system of equations

$$t_s = 0, \quad t_i = 1 + \sum_{j=1}^m p_{ij} t_j, \quad \text{for all } i \neq s.$$

- The mean recurrence time t_s^* of state s is given by

$$t_s^* = 1 + \sum_{j=1}^m p_{sj} t_j.$$

6.5 CONTINUOUS-TIME MARKOV CHAINS

In the Markov chain models that we have considered so far, we have assumed that the transitions between states take unit time. In this section, we consider a related class of models that evolve in continuous time and can be used to study systems involving continuous-time arrival processes. Examples are distribution centers or nodes in communication networks where some events of interest (e.g., arrivals of orders or of new calls) are described in terms of Poisson processes.

As before, we will consider a process that involves transitions from one state to the next, according to given transition probabilities, but we will model the times spent between transitions as continuous random variables. We will still assume that the number of states is finite and, in the absence of a statement to the contrary, we will let the state space be the set $\mathcal{S} = \{1, \dots, m\}$.

To describe the process, we introduce certain random variables of interest:

X_n : the state right after the n th transition;

Y_n : the time of the n th transition;

T_n : the time elapsed between the $(n - 1)$ st and the n th transition.

For completeness, we denote by X_0 the initial state, and we let $Y_0 = 0$. We also introduce some assumptions.

Continuous-Time Markov Chain Assumptions

- If the current state is i , the time until the next transition is exponentially distributed with a given parameter ν_i , independently of the past history of the process and of the next state.
- If the current state is i , the next state will be j with a given probability p_{ij} , independently of the past history of the process and of the time until the next transition.

The above assumptions are a complete description of the process and provide an unambiguous method for simulating it: given that we just entered state i , we remain at state i for a time that is exponentially distributed with parameter ν_i , and then move to a next state j according to the transition probabilities p_{ij} . As an immediate consequence, the sequence of states X_n obtained after successive transitions is a discrete-time Markov chain, with transition probabilities p_{ij} , called the **embedded** Markov chain.

In mathematical terms, our assumptions can be formulated as follows. Let

$$A = \{T_1 = t_1, \dots, T_n = t_n, X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i\}$$

be an event that captures the history of the process until the n th transition. We then have

$$\begin{aligned}\mathbf{P}(X_{n+1} = j, T_{n+1} \geq t \mid A) &= \mathbf{P}(X_{n+1} = j, T_{n+1} \geq t \mid X_n = i) \\ &= \mathbf{P}(X_{n+1} = j \mid X_n = i) \mathbf{P}(T_{n+1} \geq t \mid X_n = i) \\ &= p_{ij} e^{-\nu_i t}, \quad \text{for all } t \geq 0.\end{aligned}$$

The expected time to the next transition is

$$\mathbf{E}[T_{n+1} \mid X_n = i] = \int_0^\infty \tau \nu_i e^{-\nu_i \tau} d\tau = \frac{1}{\nu_i},$$

so we can interpret ν_i as the average number of transitions out of state i , per unit time spent at state i . Consequently, the parameter ν_i is called the **transition rate out of state i** . Since only a fraction p_{ij} of the transitions out of state i will lead to state j , we may also view

$$q_{ij} = \nu_i p_{ij}$$

as the average number of transitions from i to j , per unit time spent at i . Accordingly, we call q_{ij} the **transition rate from i to j** . Note that given the transition rates q_{ij} , one can obtain the transition rates ν_i using the formula

$$\nu_i = \sum_{j=1}^m q_{ij},$$

and the transition probabilities using the formula

$$p_{ij} = \frac{q_{ij}}{\nu_i}.$$

Note that the model allows for self-transitions, from a state back to itself, which can indeed happen if a self-transition probability p_{ii} is nonzero. However, such self-transitions have no observable effects: because of the memorylessness of the exponential distribution, the remaining time until the next transition is the same, irrespective of whether a self-transition just occurred or not. For this reason, we can ignore self-transitions and we will henceforth assume that

$$p_{ii} = q_{ii} = 0, \quad \text{for all } i.$$

Example 6.14. A machine, once in production mode, operates continuously until an alarm signal is generated. The time up to the alarm signal is an exponential random variable with parameter 1. Subsequent to the alarm signal, the machine is tested for an exponentially distributed amount of time with parameter 5. The test results are positive, with probability $1/2$, in which case the machine returns to production mode, or negative, with probability $1/2$, in which case the machine is taken

for repair. The duration of the repair is exponentially distributed with parameter 3. We assume that the above mentioned random variables are all independent and also independent of the test results.

Let states 1, 2, and 3, correspond to production mode, testing, and repair, respectively. The transition rates are $\nu_1 = 1$, $\nu_2 = 5$, and $\nu_3 = 3$. The transition probabilities and the transition rates are given by the following two matrices:

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & 1 & 0 \\ 5/2 & 0 & 5/2 \\ 3 & 0 & 0 \end{bmatrix}.$$

See Fig. 6.20 for an illustration.

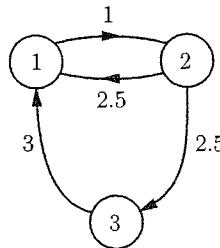


Figure 6.20: Illustration of the Markov chain in Example 6.14. The quantities indicated next to each arc are the transition rates q_{ij} .

We finally note that the continuous-time process we have described has a Markov property similar to its discrete-time counterpart: the future is independent of the past, given the present. To see this, denote by $X(t)$ the state of a continuous-time Markov chain at time $t \geq 0$, and note that it stays constant between transitions.[†] Let us recall the memorylessness property of the exponential distribution, which in our context implies that for any time t between the n th and $(n + 1)$ st transition times Y_n and Y_{n+1} , the additional time $Y_{n+1} - t$ until the next transition is independent of the time $t - Y_n$ that the system has been in the current state. It follows that for any time t , and given the present state $X(t)$, the future of the process [the random variables $X(\tau)$ for $\tau > t$], is independent of the past [the random variables $X(\tau)$ for $\tau < t$].

Approximation by a Discrete-Time Markov Chain

We now elaborate on the relation between a continuous-time Markov chain and a corresponding discrete-time version. This relation will lead to an alternative

[†] If a transition takes place at time t , the notation $X(t)$ is ambiguous. A common convention is to let $X(t)$ refer to the state right after the transition, so that $X(Y_n)$ is the same as X_n .

description of a continuous-time Markov chain, and to a set of balance equations characterizing the steady-state behavior.

Let us fix a small positive number δ and consider the discrete-time Markov chain Z_n that is obtained by observing $X(t)$ every δ time units:

$$Z_n = X(n\delta), \quad n = 0, 1, \dots$$

The fact that Z_n is a Markov chain (the future is independent from the past, given the present) follows from the Markov property of $X(t)$. We will use the notation \bar{p}_{ij} to describe the transition probabilities of Z_n .

Given that $Z_n = i$, there is a probability approximately equal to $\nu_i \delta$ that there is a transition between times $n\delta$ and $(n+1)\delta$, and in that case there is a further probability p_{ij} that the next state is j . Therefore,

$$\bar{p}_{ij} = \mathbf{P}(Z_{n+1} = j \mid Z_n = i) = \nu_i p_{ij} \delta + o(\delta) = q_{ij} \delta + o(\delta), \quad \text{if } j \neq i,$$

where $o(\delta)$ is a term that is negligible compared to δ , as δ gets smaller. The probability of remaining at i [i.e., no transition occurs between times $n\delta$ and $(n+1)\delta$] is

$$\bar{p}_{ii} = \mathbf{P}(Z_{n+1} = i \mid Z_n = i) = 1 - \sum_{j \neq i} \bar{p}_{ij}.$$

This gives rise to the following alternative description.[†]

Alternative Description of a Continuous-Time Markov Chain

Given the current state i of a continuous-time Markov chain, and for any $j \neq i$, the state δ time units later is equal to j with probability

$$q_{ij} \delta + o(\delta),$$

independent of the past history of the process.

Example 6.14 (continued). Neglecting $o(\delta)$ terms, the transition probability matrix for the corresponding discrete-time Markov chain Z_n is

$$\begin{bmatrix} 1 - \delta & \delta & 0 \\ 5\delta/2 & 1 - 5\delta & 5\delta/2 \\ 3\delta & 0 & 1 - 3\delta \end{bmatrix}.$$

[†] Our argument so far shows that a continuous-time Markov chain satisfies this alternative description. Conversely, it can be shown that if we start with this alternative description, the time until a transition out of state i is an exponential random variable with parameter $\nu_i = \sum_j q_{ij}$. Furthermore, given that such a transition has just occurred, the next state is j with probability $q_{ij}/\nu_i = p_{ij}$. This establishes that the alternative description is equivalent to the original one.

Example 6.15. Queueing. Packets arrive at a node of a communication network according to a Poisson process with rate λ . The packets are stored at a buffer with room for up to m packets, and are then transmitted one at a time. However, if a packet finds a full buffer upon arrival, it is discarded. The time required to transmit a packet is exponentially distributed with parameter μ . The transmission times of different packets are independent and are also independent from all the interarrival times.

We will model this system using a continuous-time Markov chain with state $X(t)$ equal to the number of packets in the system at time t [if $X(t) > 0$, then $X(t) - 1$ packets are waiting in the queue and one packet is under transmission]. The state increases by one when a new packet arrives and decreases by one when an existing packet departs. To show that $X(t)$ is indeed a Markov chain, we verify that we have the property specified in the above alternative description, and at the same time identify the transition rates q_{ij} .

Consider first the case where the system is empty, i.e., the state $X(t)$ is equal to 0. A transition out of state 0 can only occur if there is a new arrival, in which case the state becomes equal to 1. Since arrivals are Poisson, we have

$$\mathbf{P}(X(t + \delta) = 1 | X(t) = 0) = \lambda\delta + o(\delta),$$

and

$$q_{0j} = \begin{cases} \lambda, & \text{if } j = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Consider next the case where the system is full, i.e., the state $X(t)$ is equal to m . A transition out of state m will occur upon the completion of the current packet transmission, at which point the state will become $m - 1$. Since the duration of a transmission is exponential (and in particular, memoryless), we have

$$\mathbf{P}(X(t + \delta) = m - 1 | X(t) = m) = \mu\delta + o(\delta),$$

and

$$q_{mj} = \begin{cases} \mu, & \text{if } j = m - 1, \\ 0, & \text{otherwise.} \end{cases}$$

Consider finally the case where $X(t)$ is equal to some intermediate state i , with $0 < i < m$. During the next δ time units, there is a probability $\lambda\delta + o(\delta)$ of a new packet arrival, which will bring the state to $i + 1$, and a probability $\mu\delta + o(\delta)$ that a packet transmission is completed, which will bring the state to $i - 1$. [The probability of both an arrival and a departure within an interval of length δ is of the order of δ^2 and can be neglected, as is the case with other $o(\delta)$ terms.] Hence,

$$\mathbf{P}(X(t + \delta) = i - 1 | X(t) = i) = \mu\delta + o(\delta),$$

$$\mathbf{P}(X(t + \delta) = i + 1 | X(t) = i) = \lambda\delta + o(\delta),$$

and

$$q_{ij} = \begin{cases} \lambda, & \text{if } j = i + 1, \\ \mu, & \text{if } j = i - 1, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } i = 1, 2, \dots, m - 1;$$

see Fig. 6.21.

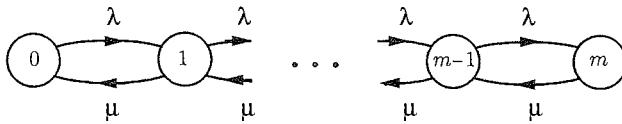


Figure 6.21: Transition graph in Example 6.15.

Steady-State Behavior

We now turn our attention to the long-term behavior of a continuous-time Markov chain and focus on the state occupancy probabilities $\mathbf{P}(X(t) = i)$, in the limit as t gets large. We approach this problem by studying the steady-state probabilities of the corresponding discrete-time chain Z_n .

Since $Z_n = X(n\delta)$, it is clear that the limit π_j of $\mathbf{P}(Z_n = j | Z_0 = i)$, if it exists, is the same as the limit of $\mathbf{P}(X(t) = j | X(0) = i)$. It therefore suffices to consider the steady-state probabilities associated with Z_n . Reasoning as in the discrete-time case, we see that for the steady-state probabilities to be independent of the initial state, we need the chain Z_n to have a single recurrent class, which we will henceforth assume. We also note that the Markov chain Z_n is automatically aperiodic. This is because the self-transition probabilities are of the form

$$\bar{p}_{ii} = 1 - \delta \sum_{j \neq i} q_{ij} + o(\delta),$$

which is positive when δ is small, and because chains with nonzero self-transition probabilities are always aperiodic.

The balance equations for the chain Z_n are of the form

$$\pi_j = \sum_{k=1}^m \pi_k \bar{p}_{kj} \quad \text{for all } j,$$

or

$$\begin{aligned} \pi_j &= \pi_j \bar{p}_{jj} + \sum_{k \neq j} \pi_k \bar{p}_{kj} \\ &= \pi_j \left(1 - \delta \sum_{k \neq j} q_{jk} + o(\delta) \right) + \sum_{k \neq j} \pi_k (q_{kj} \delta + o(\delta)). \end{aligned}$$

We cancel out the term π_j that appears on both sides of the equation, divide by δ , and take the limit as δ decreases to zero, to obtain the **balance equations**

$$\pi_j \sum_{k \neq j} q_{jk} = \sum_{k \neq j} \pi_k q_{kj}.$$

We can now invoke the Steady-State Convergence Theorem for the chain Z_n to obtain the following.

Steady-State Convergence Theorem

Consider a continuous-time Markov chain with a single recurrent class. Then, the states j are associated with steady-state probabilities π_j that have the following properties.

(a) For each j , we have

$$\lim_{t \rightarrow \infty} \mathbf{P}(X(t) = j \mid X(0) = i) = \pi_j, \quad \text{for all } i.$$

(b) The π_j are the unique solution to the system of equations below:

$$\begin{aligned} \pi_j \sum_{k \neq j} q_{jk} &= \sum_{k \neq j} \pi_k q_{kj}, \quad j = 1, \dots, m, \\ 1 &= \sum_{k=1}^m \pi_k. \end{aligned}$$

(c) We have

$$\begin{aligned} \pi_j &= 0, & \text{for all transient states } j, \\ \pi_j &> 0, & \text{for all recurrent states } j. \end{aligned}$$

To interpret the balance equations, we view π_j as the expected long-term fraction of time the process spends in state j . It follows that $\pi_k q_{kj}$ can be viewed as the expected frequency of transitions from k to j (expected number of transitions from k to j per unit time). It is seen therefore that the balance equations express the intuitive fact that the frequency of transitions out of state j (the left-hand side term $\pi_j \sum_{k \neq j} q_{jk}$) is equal to the frequency of transitions into state j (the right-hand side term $\sum_{k \neq j} \pi_k q_{kj}$).

Example 6.14 (continued). The balance and normalization equations for this example are

$$\begin{aligned} \pi_1 &= \frac{5}{2}\pi_2 + 3\pi_3, & 5\pi_2 &= \pi_1, & 3\pi_3 &= \frac{5}{2}\pi_2, \\ 1 &= \pi_1 + \pi_2 + \pi_3. \end{aligned}$$

As in the discrete-time case, one of these equations is redundant, e.g., the third equation can be obtained from the first two. Still, there is a unique solution:

$$\pi_1 = \frac{30}{41}, \quad \pi_2 = \frac{6}{41}, \quad \pi_3 = \frac{5}{41}.$$

Thus, for example, if we let the process run for a long time, $X(t)$ will be at state 1 with probability $30/41$, independent of the initial state.

The steady-state probabilities π_j are to be distinguished from the steady-state probabilities $\bar{\pi}_j$ of the embedded Markov chain X_n . Indeed, the balance and normalization equations for the embedded Markov chain are

$$\begin{aligned}\bar{\pi}_1 &= \frac{1}{2}\bar{\pi}_2 + \bar{\pi}_3, & \bar{\pi}_2 &= \bar{\pi}_1, & \bar{\pi}_3 &= \frac{1}{2}\bar{\pi}_2, \\ 1 &= \bar{\pi}_1 + \bar{\pi}_2 + \bar{\pi}_3,\end{aligned}$$

yielding the solution

$$\bar{\pi}_1 = \frac{2}{5}, \quad \bar{\pi}_2 = \frac{2}{5}, \quad \bar{\pi}_3 = \frac{1}{5}.$$

To interpret the probabilities $\bar{\pi}_j$, we can say, for example, that if we let the process run for a long time, the expected fraction of transitions that lead to state 1 is equal to $2/5$.

Note that even though $\bar{\pi}_1 = \bar{\pi}_2$ (that is, there are about as many transitions into state 1 as there are transitions into state 2), we have $\pi_1 > \pi_2$. The reason is that the process tends to spend more time during a typical visit to state 1 than during a typical visit to state 2. Hence, at a given time t , the process $X(t)$ is more likely to be found at state 1. This situation is typical, and the two sets of steady-state probabilities (π_j and $\bar{\pi}_j$) are generically different. The main exception arises in the special case where the transition rates ν_i are the same for all i ; see the end-of-chapter problems.

Birth-Death Processes

As in the discrete-time case, the states in a **birth-death process** are linearly arranged and transitions can only occur to a neighboring state, or else leave the state unchanged; formally, we have

$$q_{ij} = 0, \quad \text{for } |i - j| > 1.$$

In a birth-death process, the long-term expected frequencies of transitions from i to j and of transitions from j to i must be the same, leading to the **local balance equations**

$$\pi_j q_{ji} = \pi_i q_{ij}, \quad \text{for all } i, j.$$

The local balance equations have the same structure as in the discrete-time case, leading to closed-form formulas for the steady-state probabilities.

Example 6.15 (continued). The local balance equations take the form

$$\pi_i \lambda = \pi_{i+1} \mu, \quad i = 0, 1, \dots, m-1,$$

and we obtain $\pi_{i+1} = \rho \pi_i$, where $\rho = \lambda/\mu$. Thus, we have $\pi_i = \rho^i \pi_0$ for all i . The normalization equation $1 = \sum_{i=0}^m \pi_i$ yields

$$1 = \pi_0 \sum_{i=0}^m \rho^i,$$

and the steady-state probabilities are

$$\pi_i = \frac{\rho^i}{1 + \rho + \dots + \rho^m}, \quad i = 0, 1, \dots, m.$$

6.6 SUMMARY AND DISCUSSION

In this chapter, we have introduced Markov chain models with a finite number of states. In a discrete-time Markov chain, transitions occur at integer times according to given transition probabilities p_{ij} . The crucial property that distinguishes Markov chains from general random processes is that the transition probabilities p_{ij} apply each time that the state is equal to i , independently of the previous values of the state. Thus, given the present, the future of the process is independent of the past.

Coming up with a suitable Markov chain model of a given physical situation is to some extent an art. In general, we need to introduce a rich enough set of states so that the current state summarizes whatever information from the history of the process is relevant to its future evolution. Subject to this requirement, we usually aim at a model that does not involve more states than necessary.

Given a Markov chain model, there are several questions of interest.

- (a) Questions referring to the statistics of the process over a finite time horizon.
We have seen that we can calculate the probability that the process follows a particular path by multiplying the transition probabilities along the path. The probability of a more general event can be obtained by adding the probabilities of the various paths that lead to the occurrence of the event. In some cases, we can exploit the Markov property to avoid listing each and every path that corresponds to a particular event. A prominent example is the recursive calculation of the n -step transition probabilities, using the Chapman-Kolmogorov equations.
- (b) Questions referring to the steady-state behavior of the Markov chain. To address such questions, we classified the states of a Markov chain as transient and recurrent. We discussed how the recurrent states can be divided into disjoint recurrent classes, so that each state in a recurrent class is accessible from every other state in the same class. We also distinguished between periodic and aperiodic recurrent classes. The central result of Markov chain theory is that if a chain consists of a single aperiodic recurrent class, plus possibly some transient states, the probability $r_{ij}(n)$ that the state is equal to some j converges, as time goes to infinity, to a steady-state probability π_j , which does not depend on the initial state i . In other words, the identity of the initial state has no bearing on the statistics of X_n when n is very large. The steady-state probabilities can be found by solving a system of linear equations, consisting of the balance equations and the normalization equation $\sum_j \pi_j = 1$.
- (c) Questions referring to the transient behavior of a Markov chain. We discussed the absorption probabilities (the probability that the state eventually enters a given recurrent class, given that it starts at a given transient state), and the mean first passage times (the expected time until a particular recurrent state is entered, assuming that the chain has a single recurrent

class). In both cases, we showed that the quantities of interest can be found by considering the unique solution to a system of linear equations.

We finally considered continuous-time Markov chains. In such models, given the current state, the next state is determined by the same mechanism as in discrete-time Markov chains. However, the time until the next transition is an exponentially distributed random variable, whose parameter depends only on the current state. Continuous-time Markov chains are in many ways similar to their discrete-time counterparts. They have the same Markov property (the future is independent from the past, given the present). In fact, we can visualize a continuous-time Markov chain in terms of a related discrete-time Markov chain obtained by a fine discretization of the time axis. Because of this correspondence, the steady-state behaviors of continuous-time and discrete-time Markov chains are similar: assuming that there is a single recurrent class, the occupancy probability of any particular state converges to a steady-state probability that does not depend on the initial state. These steady-state probabilities can be found by solving a suitable set of balance and normalization equations.

P R O B L E M S

SECTION 6.1. Discrete-Time Markov Chains

Problem 1. The times between successive customer arrivals at a facility are independent and identically distributed random variables with the following PMF:

$$p(k) = \begin{cases} 0.2, & k=1, \\ 0.3, & k=3, \\ 0.5, & k=4, \\ 0, & \text{otherwise.} \end{cases}$$

Construct a four-state Markov chain model that describes the arrival process. In this model, one of the states should correspond to the times when an arrival occurs.

Problem 2. Consider the Markov chain in Example 6.2, for the case where $m = 4$, as in Fig. 6.2, and assume that the process starts at any of the four states, with equal probability. Let $Y_n = 1$ whenever the Markov chain is at state 1 or 2, and $Y_n = 2$ whenever the Markov chain is at state 3 or 4. Is the process Y_n a Markov chain?

SECTION 6.2. Classification of States

Problem 3. A spider and a fly move along a straight line in unit increments. The spider always moves towards the fly by one unit. The fly moves towards the spider by one unit with probability 0.3, moves away from the spider by one unit with probability 0.3, and stays in place with probability 0.4. The initial distance between the spider and the fly is integer. When the spider and the fly land in the same position, the spider captures the fly.

- (a) Construct a Markov chain that describes the relative location of the spider and fly.
- (b) Identify the transient and recurrent states.

Problem 4.* Existence of a recurrent state. Show that in a Markov chain at least one recurrent state must be accessible from any given state, i.e., for any i , there is at least one recurrent j in the set $A(i)$ of accessible states from i .

Solution. Fix a state i . If i is recurrent, then every $j \in A(i)$ is also recurrent so we are done. Assume that i is transient. Then, there exists a state $i_1 \in A(i)$ such that $i \notin A(i_1)$. If i_1 is recurrent, then we have found a recurrent state that is accessible from i , and we are done. Suppose now that i_1 is transient. Then, $i_1 \neq i$ because otherwise the assumptions $i_1 \in A(i)$ and $i \notin A(i_1)$ would yield $i \in A(i)$ and $i \notin A(i)$, which is a contradiction. Since i_1 is transient, there exists some i_2 such that $i_2 \in A(i_1)$ and $i_2 \notin A(i_2)$. In particular, $i_2 \in A(i)$. If i_2 is recurrent, we are done. So, suppose that i_2 is transient. The same argument as before shows that $i_2 \neq i_1$. Furthermore, we must

also have $i_2 \neq i$. This is because if we had $i_2 = i$, we would have $i_1 \in A(i) = A(i_2)$, contradicting the assumption $i_1 \notin A(i_2)$. Continuing this process, at the k th step, we will either obtain a recurrent state i_k which is accessible from i , or else we will obtain a transient state i_k which is different than all the preceding states i, i_1, \dots, i_{k-1} . Since there is only a finite number of states, a recurrent state must ultimately be obtained.

Problem 5.* Consider a Markov chain with some transient and some recurrent states.

(a) Show that for some numbers c and γ , with $c > 0$ and $0 < \gamma < 1$, we have

$$\mathbf{P}(X_n \text{ is transient} \mid X_0 = i) \leq c\gamma^n, \quad \text{for all } i \text{ and } n \geq 1.$$

(b) Let T be the first time n at which X_n is recurrent. Show that such a time is certain to exist (i.e., the probability of the event that there exists a time n at which X_n is recurrent is equal to 1) and that $\mathbf{E}[T] < \infty$.

Solution. (a) For notational convenience, let

$$q_i(n) = \mathbf{P}(X_n \text{ transient} \mid X_0 = i).$$

A recurrent state that is accessible from state i can be reached in at most m steps, where m is the number of states. Therefore, $q_i(m) < 1$. Let

$$\beta = \max_{i=1, \dots, m} q_i(m)$$

and note that for all i , we have $q_i(m) \leq \beta < 1$. If a recurrent state has not been reached by time m , which happens with probability at most β , the conditional probability that a recurrent state is not reached in the next m steps is at most β as well, which suggests that $q_i(2m) \leq \beta^2$. Indeed, conditioning on the possible values of X_m , we obtain

$$\begin{aligned} q_i(2m) &= \mathbf{P}(X_{2m} \text{ transient} \mid X_0 = i) \\ &= \sum_{j \text{ transient}} \mathbf{P}(X_{2m} \text{ transient} \mid X_m = j, X_0 = i) \mathbf{P}(X_m = j \mid X_0 = i) \\ &= \sum_{j \text{ transient}} \mathbf{P}(X_{2m} \text{ transient} \mid X_m = j) \mathbf{P}(X_m = j \mid X_0 = i) \\ &= \sum_{j \text{ transient}} \mathbf{P}(X_m \text{ transient} \mid X_0 = j) \mathbf{P}(X_m = j \mid X_0 = i) \\ &\leq \beta \sum_{j \text{ transient}} \mathbf{P}(X_m = j \mid X_0 = i) \\ &= \beta \mathbf{P}(X_m \text{ transient} \mid X_0 = i) \\ &\leq \beta^2. \end{aligned}$$

Continuing similarly, we obtain

$$q_i(km) \leq \beta^k, \quad \text{for all } i \text{ and } k \geq 1.$$

Let n be any positive integer, and let k be the integer such that $km \leq n < (k+1)m$. Then, we have

$$q_i(n) \leq q_i(km) \leq \beta^k = \beta^{-1} (\beta^{1/m})^{(k+1)m} \leq \beta^{-1} (\beta^{1/m})^n.$$

Thus, the desired relation holds with $c = \beta^{-1}$ and $\gamma = \beta^{1/m}$.

(b) Let A be the event that the state never enters the set of recurrent states. Using the result from part (a), we have

$$\mathbf{P}(A) \leq \mathbf{P}(X_n \text{ transient}) \leq c\gamma^n.$$

Since this is true for every n and since $\gamma < 1$, we must have $\mathbf{P}(A) = 0$. This establishes that there is certainty (probability equal to 1) that there is a finite time T that a recurrent state is first entered. We then have

$$\begin{aligned} \mathbf{E}[T] &= \sum_{n=1}^{\infty} n \mathbf{P}(X_{n-1} \text{ transient}, X_n \text{ recurrent}) \\ &\leq \sum_{n=1}^{\infty} n \mathbf{P}(X_{n-1} \text{ transient}) \\ &\leq \sum_{n=1}^{\infty} nc\gamma^{n-1} \\ &= \frac{c}{1-\gamma} \sum_{n=1}^{\infty} n(1-\gamma)\gamma^{n-1} \\ &= \frac{c}{(1-\gamma)^2}, \end{aligned}$$

where the last relation is obtained using the expression for the mean of the geometric distribution.

Problem 6.* Recurrent states. Show that if a recurrent state is visited once, the probability that it will be visited again in the future is equal to 1 (and, therefore, the probability that it will be visited an infinite number of times is equal to 1). *Hint:* Modify the chain to make the recurrent state of interest the only recurrent state, and use the result from Problem 5(b).

Solution. Let s be a recurrent state, and suppose that s has been visited once. From then on, the only possible states are those in the same recurrence class as s . Therefore, without loss of generality, we can assume that there is a single recurrent class. Suppose that the current state is some $i \neq s$. We want to show that s is guaranteed to be visited some time in the future.

Consider a new Markov chain in which the transitions out of state s are disabled, so that $p_{ss} = 1$. The transitions out of states i , for $i \neq s$ are unaffected. Clearly, s is recurrent in the new chain. Furthermore, for any state $i \neq s$, there is a positive probability path from i to s in the original chain (since s is recurrent in the original chain), and the same holds true in the new chain. Since i is not accessible from s in the new chain, it follows that every $i \neq s$ in the new chain is transient. By the result

of Problem 5(b), state s will be eventually visited by the new chain (with probability 1). But the original chain is identical to the new one until the time that s is first visited. Hence, state s is guaranteed to be eventually visited by the original chain s . By repeating this argument, we see that s is guaranteed to be visited an infinite number of times (with probability 1).

Problem 7.* Periodic classes. Consider a recurrent class R . Show that exactly one of the following two alternatives must hold:

- (i) The states in R can be grouped in $d > 1$ disjoint subsets S_1, \dots, S_d , so that all transitions from S_k lead to S_{k+1} , or to S_1 if $k = d$. (In this case, R is periodic.)
- (ii) There exists a time n such that $r_{ij}(n) > 0$ for all $i, j \in R$. (In this case R is aperiodic.)

Hint: Fix a state i and let d be the greatest common divisor of the elements of the set $Q = \{n \mid r_{ii}(n) > 0\}$. If $d = 1$, use the following fact from elementary number theory: if the positive integers $\alpha_1, \alpha_2, \dots$ have no common divisor other than 1, then every positive integer n outside a finite set can be expressed in the form $n = k_1\alpha_1 + k_2\alpha_2 + \dots + k_t\alpha_t$ for some nonnegative integers k_1, \dots, k_t , and some $t \geq 1$.

Solution. Fix a state i and consider the set $Q = \{n \mid r_{ii}(n) > 0\}$. Let d be the greatest common divisor of the elements of Q . Consider first the case where $d \neq 1$. For $k = 1, \dots, d$, let S_k be the set of all states that are reachable from i in $ld + k$ steps for some nonnegative integer l . Suppose that $s \in S_k$ and $p_{ss'} > 0$. Since $s \in S_k$, we can reach s from i in $ld + k$ steps for some l , which implies that we can reach s' from i in $ld + k + 1$ steps. This shows that $s' \in S_{k+1}$ if $k < d$, and that $s' \in S_1$ if $k = d$. It only remains to show that the sets S_1, \dots, S_d are disjoint. Suppose, to derive a contradiction, that $s \in S_k$ and $s \in S_{k'}$ for some $k \neq k'$. Let q be the length of some positive probability path from s to i . Starting from i , we can get to s in $ld + k$ steps, and then return to i in q steps. Hence $ld + k + q$ belongs to Q , which implies that d divides $k + q$. By the same argument, d must also divide $k' + q$. Hence d divides $k - k'$, which is a contradiction because $1 \leq |k - k'| \leq d - 1$.

Consider next the case where $d = 1$. Let $Q = \{\alpha_1, \alpha_2, \dots\}$. Since these are the possible lengths of positive probability paths that start and end at i , it follows that any integer n of the form $n = k_1\alpha_1 + k_2\alpha_2 + \dots + k_t\alpha_t$ is also in Q . (To see this, use k_1 times a path of length α_1 , followed by using k_2 times a path of length α_2 , etc.) By the number-theoretic fact given in the hint, the set Q contains all but finitely many positive integers. Let n_i be such that

$$r_{ii}(n) > 0, \quad \text{for all } n > n_i.$$

Fix some $j \neq i$ and let q be the length of a shortest positive probability path from i to j , so that $q < m$, where m is the number of states. Consider some n that satisfies $n > n_i + m$, and note that $n - q > n_i + m - q > n_i$. Thus, we can go from i to itself in $n - q$ steps, and then from i to j in q steps. Therefore, there is a positive probability path from i to j , of length n , so that $r_{ij}(n) > 0$.

We have so far established that at least one of the alternatives given in the problem statement must hold. To establish that they cannot hold simultaneously, note that the first alternative implies that $r_{ii}(n) = 0$ whenever n is not an integer multiple of d , which is incompatible with the second alternative.

For completeness, we now provide a proof of the number-theoretic fact that was used in this problem. We start with the set of positive integers $\alpha_1, \alpha_2, \dots$, and assume

that they have no common divisor other than 1. We define M as the set of all positive integers the form $\sum_{i=1}^t k_i \alpha_i$, where the k_i are nonnegative integers. Note that this set is closed under addition (the sum of two elements of M is of the same form and must also belong to M). Let g be the smallest difference between two distinct elements of M . Then, $g \geq 1$ and $g \leq \alpha_i$ for all i , since α_i and $2\alpha_i$ both belong to M .

Suppose that $g > 1$. Since the greatest common divisor of the α_i is 1, there exists some α_{i^*} which is not divisible by g . We then have

$$\alpha_{i^*} = \ell g + r,$$

for some positive integer ℓ , where the remainder r satisfies $0 < r < g$. Furthermore, in view of the definition of g , there exist nonnegative integers $k_1, k'_1, \dots, k_t, k'_t$ such that

$$\sum_{i=1}^t k_i \alpha_i = \sum_{i=1}^t k'_i \alpha_i + g.$$

Multiplying this equation by ℓ and using the equation $\alpha_{i^*} = \ell g + r$, we obtain

$$\sum_{i=1}^t (\ell k_i) \alpha_i = \sum_{i=1}^t (\ell k'_i) \alpha_i + \ell g = \sum_{i=1}^t (\ell k'_i) \alpha_i + \alpha_{i^*} - r.$$

This shows that there exist two numbers in the set M , whose difference is equal to r . Since $0 < r < g$, this contradicts our definition of g as the smallest possible difference. This contradiction establishes that g must be equal to 1.

Since $g = 1$, there exists some positive integer x such that $x \in M$ and $x + 1 \in M$. We will now show that every integer n larger than $\alpha_1 x$ belongs to M . Indeed, by dividing n by α_1 , we obtain $n = k\alpha_1 + r$, where $k \geq x$ and where the remainder r satisfies $0 \leq r < \alpha_1$. We rewrite n in the form

$$n = x(\alpha_1 - r) + (x + 1)r + (k - x)\alpha_1.$$

Since x , $x + 1$, and α_1 all belong to M , this shows that n is the sum of elements of M and must also belong to M , as desired.

SECTION 6.3. Steady-State Behavior

Problem 8. Consider the two models of machine failure and repair in Example 6.3. Find conditions on b and r for the chain to have a single recurrent class which is aperiodic and, under those conditions, find closed form expressions for the steady-state probabilities.

Problem 9. A professor gives tests that are hard, medium, or easy. If she gives a hard test, her next test will be either medium or easy, with equal probability. However, if she gives a medium or easy test, there is a 0.5 probability that her next test will be of the same difficulty, and a 0.25 probability for each of the other two levels of difficulty. Construct an appropriate Markov chain and find the steady-state probabilities.

Problem 10. Alvin likes to sail each Saturday to his cottage on a nearby island off the coast. Alvin is an avid fisherman, and enjoys fishing off his boat on the way to and

from the island, as long as the weather is good. Unfortunately, the weather is good on the way to or from the island with probability p , independently of what the weather was on any past trip (so the weather could be nice on the way to the island, but poor on the way back). Now, if the weather is nice, Alvin will take one of his n fishing rods for the trip, but if the weather is bad, he will not bring a fishing rod with him. We want to find the probability that on a given leg of the trip to or from the island the weather will be nice, but Alvin will not fish because all his fishing rods are at his other home.

- (a) Formulate an appropriate Markov chain model with $n + 1$ states and find the steady-state probabilities.
- (b) What is the steady-state probability that on a given trip, Alvin sails with nice weather but without a fishing rod?

Problem 11. Consider the Markov chain in Fig. 6.22. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.

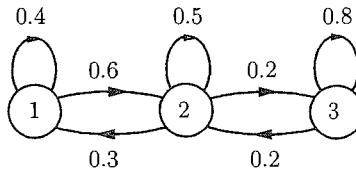


Figure 6.22: Transition probability graph for Problem 11.

- (a) For each state i , the probability that the current state is i .
- (b) The probability that the first transition we observe is a birth.
- (c) The probability that the first change of state we observe is a birth.
- (d) The conditional probability that the process was in state 2 before the first transition that we observe, given that this transition was a birth.
- (e) The conditional probability that the process was in state 2 before the first change of state that we observe, given that this change of state was a birth.
- (f) The conditional probability that the first observed transition is a birth given that it resulted in a change of state.
- (g) The conditional probability that the first observed transition leads to state 2, given that it resulted in a change of state.

Problem 12. Consider a Markov chain with given transition probabilities and with a single recurrent class that is aperiodic. Assume that for $n \geq 500$, the n -step transition probabilities are very close to the steady-state probabilities.

- (a) Find an approximate formula for $\mathbf{P}(X_{1000} = j, X_{1001} = k, X_{2000} = l \mid X_0 = i)$.
 (b) Find an approximate formula for $\mathbf{P}(X_{1000} = i \mid X_{1001} = j)$.

Problem 13. Ehrenfest model of diffusion. We have a total of n balls, some of them black, some white. At each time step, we either do nothing, which happens with probability ϵ , where $0 < \epsilon < 1$, or we select a ball at random, so that each ball has probability $(1 - \epsilon)/n > 0$ of being selected. In the latter case, we change the color of the selected ball (if white it becomes black, and vice versa), and the process is repeated indefinitely. What is the steady-state distribution of the number of white balls?

Problem 14. Bernoulli-Laplace model of diffusion. Each of two urns contains m balls. Out of the total of the $2m$ balls, m are white and m are black. A ball is simultaneously selected from each urn and moved to the other urn, and the process is indefinitely repeated. What is the steady-state distribution of the number of white balls in each urn?

Problem 15. Consider a Markov chain with two states denoted 1 and 2, and transition probabilities

$$\begin{aligned} p_{11} &= 1 - \alpha, & p_{12} &= \alpha, \\ p_{21} &= \beta, & p_{22} &= 1 - \beta, \end{aligned}$$

where α and β are such that $0 < \alpha < 1$ and $0 < \beta < 1$.

- (a) Show that the two states of the chain form a recurrent and aperiodic class.
 (b) Use induction to show that for all n , we have

$$\begin{aligned} r_{11}(n) &= \frac{\beta}{\alpha + \beta} + \frac{\alpha(1 - \alpha - \beta)^n}{\alpha + \beta}, & r_{12}(n) &= \frac{\alpha}{\alpha + \beta} - \frac{\alpha(1 - \alpha - \beta)^n}{\alpha + \beta}, \\ r_{21}(n) &= \frac{\beta}{\alpha + \beta} - \frac{\beta(1 - \alpha - \beta)^n}{\alpha + \beta}, & r_{22}(n) &= \frac{\alpha}{\alpha + \beta} + \frac{\beta(1 - \alpha - \beta)^n}{\alpha + \beta}. \end{aligned}$$

- (c) What are the steady-state probabilities π_1 and π_2 ?

Problem 16. The parking garage at MIT has installed a card-operated gate, which, unfortunately, is vulnerable to absent-minded faculty and staff. In particular, in each day, a car crashes the gate with probability p , in which case a new gate must be installed. Also a gate that has survived for m days must be replaced as a matter of periodic maintenance. What is the long-term expected frequency of gate replacements?

Problem 17.* Steady-state convergence. Consider a Markov chain with a single recurrent class, and assume that there exists a time \bar{n} such that

$$r_{ij}(\bar{n}) > 0,$$

for all i and all recurrent j . (This is equivalent to assuming that the class is aperiodic.) We wish to show that for any i and j , the limit

$$\lim_{n \rightarrow \infty} r_{ij}(n)$$

exists and does not depend on i . To derive this result, we need to show that the choice of the initial state has no long-term effect. To quantify this effect, we consider two different initial states i and k , and consider two independent Markov chains, X_n and Y_n , with the same transition probabilities and with $X_0 = i$, $Y_0 = k$. Let $T = \min\{n \mid X_n = Y_n\}$ be the first time that the two chains enter the same state.

- (a) Show that there exist positive constants c and $\gamma < 1$ such that

$$\mathbf{P}(T \geq n) \leq c\gamma^n.$$

- (b) Show that if the states of the two chains became equal by time n , their occupancy probabilities at time n are the same, that is,

$$\mathbf{P}(X_n = j \mid T \leq n) = \mathbf{P}(Y_n = j \mid T \leq n).$$

- (c) Show that $|r_{ij}(n) - r_{kj}(n)| \leq c\gamma^n$, for all i, j, k , and n . *Hint:* Condition on the two events $\{T > n\}$ and $\{T \leq n\}$.

- (d) Let $q_j^+(n) = \max_i r_{ij}(n)$ and $q_j^-(n) = \min_i r_{ij}(n)$. Show that

$$q_j^-(n) \leq q_j^-(n+1) \leq q_j^+(n+1) \leq q_j^+(n), \quad \text{for all } n.$$

- (e) Show that the sequence $r_{ij}(n)$ converges to a limit that does not depend on i . *Hint:* Combine the results of parts (c) and (d) to show that the two sequences $q_j^-(n)$ and $q_j^+(n)$ converge and have the same limit.

Solution. (a) The argument is similar to the one used to bound the PMF of the time until a recurrent state is entered (Problem 5). Let l be some recurrent state and let $\beta = \min_i r_{il}(\bar{n}) > 0$. No matter what is the current state of X_n and Y_n , there is probability of at least β^2 that both chains are at state l after \bar{n} time steps. Thus,

$$\mathbf{P}(T > \bar{n}) \leq 1 - \beta^2.$$

Similarly,

$$\mathbf{P}(T > 2\bar{n}) = \mathbf{P}(T > \bar{n}) \mathbf{P}(T > 2\bar{n} \mid T > \bar{n}) \leq (1 - \beta^2)^2,$$

and

$$\mathbf{P}(T > k\bar{n}) \leq (1 - \beta^2)^k.$$

This implies that

$$\mathbf{P}(T \geq n) \leq c\gamma^n,$$

where $\gamma = (1 - \beta^2)^{1/\bar{n}}$, and $c = 1/(1 - \beta^2)^{\bar{n}}$.

- (b) We condition on the possible values of T and on the common state l of the two chains at time T , and use the total probability theorem. We have

$$\begin{aligned} \mathbf{P}(X_n = j \mid T \leq n) &= \sum_{t=0}^n \sum_{l=1}^m \mathbf{P}(X_n = j \mid T = t, X_t = l) \mathbf{P}(T = t, X_t = l \mid T \leq n) \\ &= \sum_{t=0}^n \sum_{l=1}^m \mathbf{P}(X_n = j \mid X_t = l) \mathbf{P}(T = t, X_t = l \mid T \leq n) \\ &= \sum_{t=0}^n \sum_{l=1}^m r_{lj}(n-t) \mathbf{P}(T = t, X_t = l \mid T \leq n). \end{aligned}$$

Similarly,

$$\mathbf{P}(Y_n = j \mid T \leq n) = \sum_{t=0}^n \sum_{l=1}^m r_{lj}(n-t) \mathbf{P}(T = t, Y_t = l \mid T \leq n).$$

But the events $\{T = t, X_t = l\}$ and $\{T = t, Y_t = l\}$ are identical, and therefore have the same probability, which implies that $\mathbf{P}(X_n = j \mid T \leq n) = \mathbf{P}(Y_n = j \mid T \leq n)$.

(c) We have

$$r_{ij}(n) = \mathbf{P}(X_n = j) = \mathbf{P}(X_n = j \mid T \leq n) \mathbf{P}(T \leq n) + \mathbf{P}(X_n = j \mid T > n) \mathbf{P}(T > n)$$

and

$$r_{kj}(n) = \mathbf{P}(Y_n = j) = \mathbf{P}(Y_n = j \mid T \leq n) \mathbf{P}(T \leq n) + \mathbf{P}(Y_n = j \mid T > n) \mathbf{P}(T > n).$$

By subtracting these two equations, using the result of part (b) to eliminate the first terms in their right-hand sides, and by taking the absolute value of both sides, we obtain

$$\begin{aligned} |r_{ij}(n) - r_{kj}(n)| &\leq |\mathbf{P}(X_n = j \mid T > n) \mathbf{P}(T > n) - \mathbf{P}(Y_n = j \mid T > n) \mathbf{P}(T > n)| \\ &\leq \mathbf{P}(T > n) \\ &\leq c\gamma^n. \end{aligned}$$

(d) By conditioning on the state after the first transition, and using the total probability theorem, we have the following variant of the Chapman-Kolmogorov equation:

$$r_{ij}(n+1) = \sum_{k=1}^m p_{ik} r_{kj}(n).$$

Using this equation, we obtain

$$q_j^+(n+1) = \max_i r_{ij}(n+1) = \max_i \sum_{k=1}^m p_{ik} r_{kj}(n) \leq \max_i \sum_{k=1}^m p_{ik} q_j^+(n) = q_j^+(n).$$

The inequality $q_j^-(n) \leq q_j^-(n+1)$ is established by a symmetrical argument. The inequality $q_j^-(n+1) \leq q_j^+(n+1)$ is a consequence of the definitions.

(e) The sequences $q_j^-(n)$ and $q_j^+(n)$ converge because they are monotonic. The inequality $|r_{ij}(n) - r_{kj}(n)| \leq c\gamma^n$, for all i and k , implies that $q_j^+(n) - q_j^-(n) \leq c\gamma^n$. Taking the limit as $n \rightarrow \infty$, we obtain that the limits of $q_j^+(n)$ and $q_j^-(n)$ are the same. Let π_j denote this common limit. Since $q_j^-(n) \leq r_{ij}(n) \leq q_j^+(n)$, it follows that $r_{ij}(n)$ also converges to π_j , and the limit is independent of i .

Problem 18.* Uniqueness of solutions to the balance equations. Consider a Markov chain with a single recurrent class, plus possibly some transient states.

- (a) Assuming that the recurrent class is aperiodic, show that the balance equations together with the normalization equation have a unique nonnegative solution.

Hint: Given a solution different from the steady-state probabilities, let it be the PMF of X_0 and consider what happens as time goes to infinity.

- (b) Show that the uniqueness result of part (a) is also true when the recurrent class is periodic. *Hint:* Introduce self-transitions in the Markov chain, in a manner that results in an equivalent set of balance equations, and use the result of part (a).

Solution. (a) Let π_1, \dots, π_m be the steady-state probabilities, that is, the limits of the $r_{ij}(n)$. These satisfy the balance and normalization equations. Suppose that there is another nonnegative solution $\bar{\pi}_1, \dots, \bar{\pi}_m$. Let us initialize the Markov chain according to these probabilities, so that $\mathbf{P}(X_0 = j) = \bar{\pi}_j$ for all j . Using the argument given in the text, we obtain $\mathbf{P}(X_n = j) = \bar{\pi}_j$, for all times. Thus,

$$\begin{aligned}\bar{\pi}_j &= \lim_{n \rightarrow \infty} \mathbf{P}(X_n = j) \\ &= \lim_{n \rightarrow \infty} \sum_{k=1}^m \bar{\pi}_k r_{kj}(n) \\ &= \sum_{k=1}^m \bar{\pi}_k \pi_j \\ &= \pi_j.\end{aligned}$$

- (b) Consider a new Markov chain, whose transition probabilities \bar{p}_{ij} are given by

$$\bar{p}_{ii} = (1 - \alpha)p_{ii} + \alpha, \quad \bar{p}_{ij} = (1 - \alpha)p_{ij}, \quad j \neq i.$$

Here, α is a number satisfying $0 < \alpha < 1$. The balance equations for the new Markov chain take the form

$$\pi_j = \pi_j \left((1 - \alpha)p_{jj} + \alpha \right) + \sum_{i \neq j} \pi_i (1 - \alpha)p_{ij},$$

or

$$(1 - \alpha)\pi_j = (1 - \alpha) \sum_{i=1}^m \pi_i p_{ij}.$$

These equations are equivalent to the balance equations for the original chain. Notice that the new chain is aperiodic, because self-transitions have positive probability. This establishes uniqueness of solutions for the new chain, and implies the same for the original chain.

Problem 19.* Expected long-term frequency interpretation. Consider a Markov chain with a single recurrent class which is aperiodic. Show that

$$\pi_j = \lim_{n \rightarrow \infty} \frac{v_{ij}(n)}{n}, \quad \text{for all } i, j = 1, \dots, m,$$

where the π_j are the steady-state probabilities, and $v_{ij}(n)$ is the expected value of the number of visits to state j within the first n transitions, starting from state i . *Hint:* Use the following fact from analysis. If a sequence a_n converges to a number a , the sequence b_n defined by $b_n = (1/n) \sum_{k=1}^n a_k$ also converges to a .

Solution. We first assert that for all n , i , and j , we have

$$v_{ij}(n) = \sum_{k=1}^n r_{ij}(k).$$

To see this, note that

$$v_{ij}(n) = \mathbf{E} \left[\sum_{k=1}^n I_k \mid X_0 = i \right],$$

where I_k is the random variable that takes the value 1 if $X_k = j$, and the value 0 otherwise, so that

$$\mathbf{E}[I_k \mid X_0 = i] = r_{ij}(k).$$

Since

$$\frac{v_{ij}(n)}{n} = \frac{1}{n} \sum_{k=1}^n r_{ij}(k),$$

and $r_{ij}(k)$ converges to π_j , it follows that $v_{ij}(n)/n$ also converges to π_j , which is the desired result.

For completeness, we also provide the proof of the fact given in the hint, and which was used in the last step of the above argument. Consider a sequence a_n that converges to some a , and let $b_n = (1/n) \sum_{k=1}^n a_k$. Fix some $\epsilon > 0$. Since a_n converges to a , there exists some n_0 such that $a_k \leq a + (\epsilon/2)$, for all $k > n_0$. Let also $c = \max_k a_k$. We then have

$$b_n = \frac{1}{n} \sum_{k=1}^{n_0} a_k + \frac{1}{n} \sum_{k=n_0+1}^n a_k \leq \frac{n_0}{n} c + \frac{n - n_0}{n} \left(a + \frac{\epsilon}{2} \right).$$

The limit of the right-hand side, as n tends to infinity, is $a + (\epsilon/2)$. Therefore, there exists some n_1 such that $b_n \leq a + \epsilon$, for every $n \geq n_1$. By a symmetrical argument, there exists some n_2 such that $b_n \geq a - \epsilon$, for every $n \geq n_2$. We have shown that for every $\epsilon > 0$, there exists some n_3 (namely, $n_3 = \max\{n_1, n_2\}$) such that $|b_n - a| \leq \epsilon$, for all $n \geq n_3$. This means that b_n converges to a .

Problem 20.* Doubly stochastic matrices. Consider a Markov chain with a single recurrent class which is aperiodic, and whose transition probability matrix is **doubly stochastic**, i.e., it has the property that the entries in any column (as well as in any row) add to unity, so that

$$\sum_{i=1}^m p_{ij} = 1, \quad j = 1, \dots, m.$$

- (a) Show that the transition probability matrix of the chain in Example 6.7 is doubly stochastic.
- (b) Show that the steady-state probabilities are

$$\pi_j = \frac{1}{m}, \quad j = 1, \dots, m.$$

- (c) Suppose that the recurrent class of the chain is instead periodic. Show that $\pi_1 = \dots = \pi_m = 1/m$ is the unique solution to the balance and normalization equations. Discuss your answer in the context of Example 6.7 for the case where m is even.

Solution. (a) Indeed the rows and the columns of the transition probability matrix in this example all add to 1.

(b) We have

$$\sum_{i=1}^m \frac{1}{m} p_{ij} = \frac{1}{m}.$$

Thus, the given probabilities $\pi_j = 1/m$ satisfy the balance equations and must therefore be the steady-state probabilities.

(c) Let π_j be a solution to the balance and normalization equations. Consider a particular j such that $\pi_j \geq \pi_i$ for all i , and let $q = \pi_j$. The balance equation for state j yields

$$q = \pi_j = \sum_{i=1}^m \pi_i p_{ij} \leq q \sum_{i=1}^m p_{ij} = q,$$

where the last step follows because the transition probability matrix is doubly stochastic. It follows that the above inequality is actually an equality and

$$\sum_{i=1}^m \pi_i p_{ij} = \sum_{i=1}^m q p_{ij}.$$

Since $\pi_i \leq q$ for all i , we must have $\pi_i p_{ij} = q p_{ij}$ for every i . Thus, $\pi_i = q$ for every state i from which a transition to j is possible. By repeating this argument, we see that $\pi_i = q$ for every state i such that there is a positive probability path from i to j . Since all states are recurrent and belong to the same class, all states i have this property, and therefore π_i is the same for all i . Since the π_i add to 1, we obtain $\pi_1 = 1/m$ for all i .

If m is even in Example 6.7, the chain is periodic with period 2. Despite this fact, the result we have just established shows that $\pi_j = 1/m$ is the unique solution to the balance and normalization equations.

Problem 21.* Queueing. Consider the queueing Example 6.9, but assume that the probabilities of a packet arrival and a packet transmission depend on the state of the queue. In particular, in each period where there are i packets in the node, one of the following occurs:

- (i) one new packet arrives; this happens with a given probability b_i . We assume that $b_i > 0$ for $i < m$, and $b_m = 0$.
- (ii) one existing packet completes transmission; this happens with a given probability $d_i > 0$ if $i \geq 1$, and with probability 0 otherwise;
- (iii) no new packet arrives and no existing packet completes transmission; this happens with probability $1 - b_i - d_i$ if $i \geq 1$, and with probability $1 - b_i$ if $i = 0$.

Calculate the steady-state probabilities of the corresponding Markov chain.

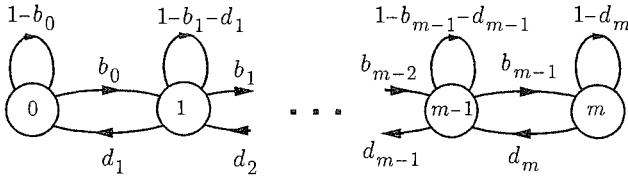


Figure 6.23: Transition probability graph for Problem 21.

Solution. We introduce a Markov chain where the states are $0, 1, \dots, m$, and correspond to the number of packets currently stored at the node. The transition probability graph is given in Fig. 6.23.

Similar to Example 6.9, we write down the local balance equations, which take the form

$$\pi_i b_i = \pi_{i+1} d_{i+1}, \quad i = 0, 1, \dots, m-1.$$

Thus we have $\pi_{i+1} = \rho_i \pi_i$, where

$$\rho_i = \frac{b_i}{d_{i+1}}.$$

Hence $\pi_i = (\rho_0 \cdots \rho_{i-1}) \pi_0$ for $i = 1, \dots, m$. By using the normalization equation $1 = \pi_0 + \pi_1 + \cdots + \pi_m$, we obtain

$$1 = \pi_0(1 + \rho_0 + \rho_0 \rho_1 + \cdots + \rho_0 \cdots \rho_{m-1}),$$

from which

$$\pi_0 = \frac{1}{1 + \rho_0 + \rho_0 \rho_1 + \cdots + \rho_0 \cdots \rho_{m-1}}.$$

The remaining steady-state probabilities are

$$\pi_i = \frac{\rho_0 \cdots \rho_{i-1}}{1 + \rho_0 + \rho_0 \rho_1 + \cdots + \rho_0 \cdots \rho_{m-1}}, \quad i = 1, \dots, m.$$

Problem 22.* Dependence of the balance equations. Show that if we add the first $m-1$ balance equations $\pi_j = \sum_{k=1}^m \pi_k p_{kj}$, for $j = 1, \dots, m-1$, we obtain the last equation $\pi_m = \sum_{k=1}^m \pi_k p_{km}$.

Solution. By adding the first $m-1$ balance equations, we obtain

$$\begin{aligned} \sum_{j=1}^{m-1} \pi_j &= \sum_{j=1}^{m-1} \sum_{k=1}^m \pi_k p_{kj} \\ &= \sum_{k=1}^m \pi_k \sum_{j=1}^{m-1} p_{kj} \\ &= \sum_{k=1}^m \pi_k (1 - p_{km}) \\ &= \pi_m + \sum_{k=1}^{m-1} \pi_k - \sum_{k=1}^m \pi_k p_{km}. \end{aligned}$$

This equation is equivalent to the last balance equation $\pi_m = \sum_{k=1}^m \pi_k p_{km}$.

Problem 23.* Local balance equations. We are given a Markov chain that has a single recurrent class which is aperiodic. Suppose that we have found a solution π_1, \dots, π_m to the following system of local balance and normalization equations:

$$\begin{aligned}\pi_i p_{ij} &= \pi_j p_{ji}, & i, j = 1, \dots, m, \\ \sum_{i=1}^m \pi_i &= 1, & i = 1, \dots, m.\end{aligned}$$

- (a) Show that the π_j are the steady-state probabilities.
- (b) What is the interpretation of the equations $\pi_i p_{ij} = \pi_j p_{ji}$ in terms of expected long-term frequencies of transitions between i and j ?
- (c) Construct an example where the local balance equations are not satisfied by the steady-state probabilities.

Solution. (a) By adding the local balance equations $\pi_i p_{ij} = \pi_j p_{ji}$ over i , we obtain

$$\sum_{i=1}^m \pi_i p_{ij} = \sum_{i=1}^m \pi_j p_{ji} = \pi_j,$$

so the π_j also satisfy the balance equations. Therefore, they are equal to the steady-state probabilities.

(b) We know that $\pi_i p_{ij}$ can be interpreted as the expected long-term frequency of transitions from i to j , so the local balance equations imply that the expected long-term frequency of any transition is equal to the expected long-term frequency of the reverse transition. (This property is also known as *time reversibility* of the chain.)

(c) We need a minimum of three states for such an example. Let the states be 1, 2, 3, and let $p_{12} > 0$, $p_{13} > 0$, $p_{21} > 0$, $p_{32} > 0$, with all other transition probabilities being 0. The chain has a single recurrent aperiodic class. The local balance equations do not hold because the expected frequency of transitions from 1 to 3 is positive, but the expected frequency of reverse transitions is 0.

Problem 24.* Sampled Markov chains. Consider a Markov chain X_n with transition probabilities p_{ij} , and let $r_{ij}(n)$ be the n -step transition probabilities.

- (a) Show that for all $n \geq 1$ and $l \geq 1$, we have

$$r_{ij}(n+l) = \sum_{k=1}^m r_{ik}(n) r_{kj}(l).$$

- (b) Suppose that there is a single recurrent class, which is aperiodic. We sample the Markov chain every l transitions, thus generating a process Y_n , where $Y_n = X_{ln}$. Show that the sampled process can be modeled by a Markov chain with a single aperiodic recurrent class and transition probabilities $r_{ij}(l)$.
- (c) Show that the Markov chain of part (b) has the same steady-state probabilities as the original process.

Solution. (a) We condition on X_n and use the total probability theorem. We have

$$\begin{aligned}
 r_{ij}(n+l) &= \mathbf{P}(X_{n+l} = j \mid X_0 = i) \\
 &= \sum_{k=1}^m \mathbf{P}(X_n = k \mid X_0 = i) \mathbf{P}(X_{n+l} = j \mid X_n = k, X_0 = i) \\
 &= \sum_{k=1}^m \mathbf{P}(X_n = k \mid X_0 = i) \mathbf{P}(X_{n+l} = j \mid X_n = k) \\
 &= \sum_{k=1}^m r_{ik}(n) r_{kj}(l),
 \end{aligned}$$

where in the third equality we used the Markov property.

(b) Since X_n is Markov, once we condition on X_{ln} , the past of the process (the states X_k for $k < ln$) becomes independent of the future (the states X_k for $k > ln$). This implies that given Y_n , the past (the states Y_k for $k < n$) is independent of the future (the states Y_k for $k > n$). Thus, Y_n has the Markov property. Because of our assumptions on X_n , there is a time \bar{n} such that

$$\mathbf{P}(X_n = j \mid X_0 = i) > 0,$$

for every $n \geq \bar{n}$, every state i , and every state j in the single recurrent class R of the process X_n . This implies that

$$\mathbf{P}(Y_n = j \mid Y_0 = i) > 0,$$

for every $n \geq \bar{n}$, every i , and every $j \in R$. Therefore, the process Y_n has a single recurrent class, which is aperiodic.

(c) The n -step transition probabilities $r_{ij}(n)$ of the process X_n converge to the steady-state probabilities π_j . The n -step transition probabilities of the process Y_n are of the form $r_{ij}(ln)$, and therefore also converge to the same limits π_j . This establishes that the π_j are the steady-state probabilities of the process Y_n .

Problem 25.* Given a Markov chain X_n with a single recurrent class which is aperiodic, consider the Markov chain whose state at time n is (X_{n-1}, X_n) . Thus, the state in the new chain can be associated with the last transition in the original chain.

(a) Show that the steady-state probabilities of the new chain are

$$\eta_{ij} = \pi_i p_{ij},$$

where the π_i are the steady-state probabilities of the original chain.

(b) Generalize part (a) to the case of the Markov chain $(X_{n-k}, X_{n-k+1}, \dots, X_n)$, whose state can be associated with the last k transitions of the original chain.

Solution. (a) For every state (i, j) of the new Markov chain, we have

$$\mathbf{P}((X_{n-1}, X_n) = (i, j)) = \mathbf{P}(X_{n-1} = i) \mathbf{P}(X_n = j \mid X_{n-1} = i) = \mathbf{P}(X_{n-1} = i) p_{ij}.$$

Since the Markov chain X_n has a single recurrent class which is aperiodic, $\mathbf{P}(X_{n-1} = i)$ converges to the steady-state probability π_i , for every i . It follows that $\mathbf{P}((X_{n-1}, X_n) = (i, j))$ converges to $\pi_i p_{ij}$, which is therefore the steady-state probability of (i, j) .

(b) Using the multiplication rule, we have

$$\mathbf{P}((X_{n-k}, \dots, X_n) = (i_0, \dots, i_k)) = \mathbf{P}(X_{n-k} = i_0) p_{i_0 i_1} \cdots p_{i_{k-1} i_k}.$$

Therefore, by an argument similar to the one in part (a), the steady-state probability of state (i_0, \dots, i_k) is equal to $\pi_{i_0} p_{i_0 i_1} \cdots p_{i_{k-1} i_k}$.

SECTION 6.4. Absorption Probabilities and Expected Time to Absorption

Problem 26. There are m classes offered by a particular department, and each year, the students rank each class from 1 to m , in order of difficulty, with rank m being the highest. Unfortunately, the ranking is completely arbitrary. In fact, any given class is equally likely to receive any given rank on a given year (two classes may not receive the same rank). A certain professor chooses to remember only the highest ranking his class has ever gotten.

- Find the transition probabilities of the Markov chain that models the ranking that the professor remembers.
- Find the recurrent and the transient states.
- Find the expected number of years for the professor to achieve the highest ranking given that in the first year he achieved the i th ranking.

Problem 27. Consider the Markov chain specified in Fig. 6.24. The steady-state probabilities are known to be:

$$\pi_1 = \frac{6}{31}, \quad \pi_2 = \frac{9}{31}, \quad \pi_3 = \frac{6}{31}, \quad \pi_4 = \frac{10}{31}.$$

Assume that the process is in state 1 just before the first transition.

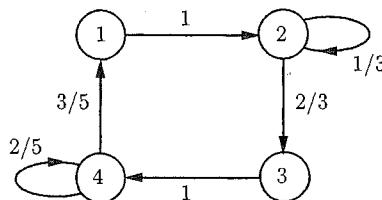


Figure 6.24: Transition probability graph for Problem 27.

- What is the probability that the process will be in state 1 just after the sixth transition?

- (b) Determine the expected value and variance of the number of transitions up to and including the next transition during which the process returns to state 1.
- (c) What is (approximately) the probability that the state of the system resulting from transition 1000 is neither the same as the state resulting from transition 999 nor the same as the state resulting from transition 1001?

Problem 28. Consider the Markov chain specified in Fig. 6.25.

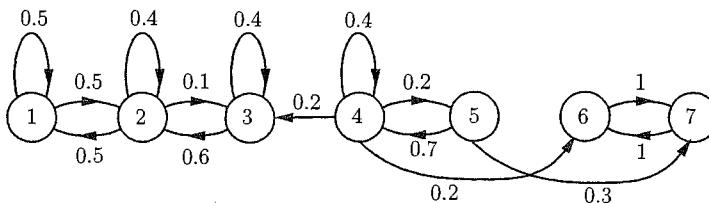


Figure 6.25: Transition probability graph for Problem 28.

- (a) Identify the transient and recurrent states. Also, determine the recurrent classes and indicate which ones, if any are periodic.
- (b) Do there exist steady-state probabilities given that the process starts in state 1? If so, what are they?
- (c) Do there exist steady-state probabilities given that the process starts in state 6? If so, what are they?
- (d) Assume that the process starts in state 1 but we begin observing it after it reaches steady-state.
- Find the probability that the state increases by one during the first transition we observe.
 - Find the conditional probability that the process was in state 2 when we started observing it, given that the state increased by one during the first transition that we observed.
 - Find the probability that the state increased by one during the first change of state that we observed.
- (e) Assume that the process starts in state 4.
- For each recurrent class, determine the probability that we eventually reach that class.
 - What is the expected number of transitions up to and including the transition at which we reach a recurrent state for the first time?

Problem 29.* Absorption probabilities. Consider a Markov chain where each state is either transient or absorbing. Fix an absorbing state s . Show that the probabilities a_i of eventually reaching s starting from a state i are the unique solution to

the equations

$$\begin{aligned} a_s &= 1, \\ a_i &= 0, && \text{for all absorbing } i \neq s, \\ a_i &= \sum_{j=1}^m p_{ij} a_j, && \text{for all transient } i. \end{aligned}$$

Hint: If there are two solutions, find a system of equations that is satisfied by their difference, and look for its solutions.

Solution. The fact that the a_i satisfy these equations was established in the text, using the total probability theorem. To show uniqueness, let \bar{a}_i be another solution, and let $\delta_i = \bar{a}_i - a_i$. Denoting by A the set of absorbing states and using the fact $\delta_j = 0$ for all $j \in A$, we have

$$\delta_i = \sum_{j=1}^m p_{ij} \delta_j = \sum_{j \notin A} p_{ij} \delta_j, \quad \text{for all transient } i.$$

Applying this relation m successive times, we obtain

$$\delta_i = \sum_{j_1 \notin A} p_{ij_1} \sum_{j_2 \notin A} p_{j_1 j_2} \cdots \sum_{j_m \notin A} p_{j_{m-1} j_m} \cdot \delta_{j_m}.$$

Hence

$$\begin{aligned} |\delta_i| &\leq \sum_{j_1 \notin A} p_{ij_1} \sum_{j_2 \notin A} p_{j_1 j_2} \cdots \sum_{j_m \notin A} p_{j_{m-1} j_m} \cdot |\delta_{j_m}| \\ &= \mathbf{P}(X_1 \notin A, \dots, X_m \notin A \mid X_0 = i) \cdot |\delta_{j_m}| \\ &\leq \mathbf{P}(X_1 \notin A, \dots, X_m \notin A \mid X_0 = i) \cdot \max_{j \notin A} |\delta_j|. \end{aligned}$$

The above relation holds for all transient i , so we obtain

$$\max_{j \notin A} |\delta_j| \leq \beta \cdot \max_{j \notin A} |\delta_j|,$$

where

$$\beta = \mathbf{P}(X_1 \notin A, \dots, X_m \notin A \mid X_0 = i).$$

Note that $\beta < 1$, because there is positive probability that X_m is absorbing, regardless of the initial state. It follows that $\max_{j \notin A} |\delta_j| = 0$, or $a_i = \bar{a}_i$ for all i that are not absorbing. We also have $a_j = \bar{a}_j$ for all absorbing j , so $a_i = \bar{a}_i$ for all i .

Problem 30.* **Multiple recurrent classes.** Consider a Markov chain that has more than one recurrent class, as well as some transient states. Assume that all the recurrent classes are aperiodic.

- For any transient state i , let $a_i(k)$ be the probability that starting from i we will reach a state in the k th recurrent class. Derive a system of equations whose solution are the $a_i(k)$.
- Show that each of the n -step transition probabilities $r_{ij}(n)$ converges to a limit, and discuss how these limits can be calculated.

Solution. (a) We introduce a new Markov chain that has only transient and absorbing states. The transient states correspond to the transient states of the original, while the absorbing states correspond to the recurrent classes of the original. The transition probabilities \hat{p}_{ij} of the new chain are as follows: if i and j are transient, $\hat{p}_{ij} = p_{ij}$; if i is a transient state and k corresponds to a recurrent class, \hat{p}_{ik} is the sum of the transition probabilities from i to states in the recurrent class in the original Markov chain.

The desired probabilities $a_i(k)$ are the absorption probabilities in the new Markov chain and are given by the corresponding formulas:

$$a_i(k) = \hat{p}_{ik} + \sum_{j: \text{ transient}} \hat{p}_{ij} a_j(k), \quad \text{for all transient } i.$$

(b) If i and j are recurrent but belong to different classes, $r_{ij}(n)$ is always 0. If i and j are recurrent but belong to the same class, $r_{ij}(n)$ converges to the steady-state probability of j in a chain consisting of only this particular recurrent class. If j is transient, $r_{ij}(n)$ converges to 0. Finally, if i is transient and j is recurrent, then $r_{ij}(n)$ converges to the product of two probabilities: (1) the probability that starting from i we will reach a state in the recurrent class of j , and (2) the steady-state probability of j conditioned on the initial state being in the class of j .

Problem 31.* Mean first passage times and expected times to absorption. Consider a Markov chain with a single recurrent class, and let s be a fixed recurrent state. Show that the system of equations

$$t_s = 0, \quad t_i = 1 + \sum_{j=1}^m p_{ij} t_j, \quad \text{for all } i \neq s,$$

satisfied by the mean first passage times, has a unique solution. Show a similar result for the system of equations satisfied by the expected times to absorption μ_i . *Hint:* If there are two solutions, find a system of equations that is satisfied by their difference, and look for its solutions.

Solution. Let t_i be the mean first passage times. These satisfy the given system of equations. To show uniqueness, let \bar{t}_i be another solution. Then we have for all $i \neq s$

$$t_i = 1 + \sum_{j \neq s} p_{ij} t_j, \quad \bar{t}_i = 1 + \sum_{j \neq s} p_{ij} \bar{t}_j,$$

and by subtraction, we obtain

$$\delta_i = \sum_{j \neq s} p_{ij} \delta_j,$$

where $\delta_i = \bar{t}_i - t_i$. By applying m successive times this relation, it follows that

$$\delta_i = \sum_{j_1 \neq s} p_{ij_1} \sum_{j_2 \neq s} p_{j_1 j_2} \cdots \sum_{j_m \neq s} p_{j_{m-1} j_m} \cdot \delta_{j_m}.$$

Hence, we have for all $i \neq s$

$$\begin{aligned} |\delta_i| &\leq \sum_{j_1 \neq s} p_{ij_1} \sum_{j_2 \neq s} p_{j_1 j_2} \cdots \sum_{j_m \neq s} p_{j_{m-1} j_m} \cdot \max_j |\delta_j| \\ &= \mathbf{P}(X_1 \neq s, \dots, X_m \neq s \mid X_0 = i) \cdot \max_j |\delta_j|. \end{aligned}$$

On the other hand, we have $\mathbf{P}(X_1 \neq s, \dots, X_m \neq s | X_0 = i) < 1$. This is because starting from any state there is positive probability that s is reached in m steps. It follows that all the δ_i must be equal to zero. The proof for the case of the system of equations satisfied by the expected times to absorption is nearly identical.

Problem 32.* Balance equations and mean recurrence times. Consider a Markov chain with a single recurrent class, and let s be a fixed recurrent state. For any state i , let

$$\rho_i = \mathbf{E}[\text{Number of visits to } i \text{ between two successive visits to } s],$$

where by convention, $\rho_s = 1$.

- (a) Show that for all i , we have

$$\rho_i = \sum_{k=1}^m \rho_k p_{ki}.$$

- (b) Show that the numbers

$$\pi_i = \frac{\rho_i}{t_s^*}, \quad i = 1, \dots, m,$$

sum to 1 and satisfy the balance equations, where t_s^* is the mean recurrence time of s (the expected number of transitions up to the first return to s , starting from s).

- (c) Show that if π_1, \dots, π_m are nonnegative, satisfy the balance equations, and sum to 1, then

$$\pi_i = \begin{cases} \frac{1}{t_i^*}, & \text{if } i \text{ is recurrent,} \\ 0, & \text{if } i \text{ is transient,} \end{cases}$$

where t_i^* is the mean recurrence time of i .

- (d) Show that the distribution of part (b) is the unique probability distribution that satisfies the balance equations.

Note: This problem not only provides an alternative proof of the existence and uniqueness of probability distributions that satisfy the balance equations, but also makes an intuitive connection between steady-state probabilities and mean recurrence times. The main idea is to break the process into “cycles,” with a new cycle starting each time that the recurrent state s is visited. The steady-state probability of state s can be interpreted as the long-term expected frequency of visits to state s , which is inversely proportional to the average time between consecutive visits (the mean recurrence time); cf. part (c). Furthermore, if a state i is expected to be visited, say, twice as often as some other state j during a typical cycle, it is plausible that the long-term expected frequency π_i of state i will be twice as large as π_j . Thus, the steady-state probabilities π_i should be proportional to the expected number of visits ρ_i during a cycle; cf. part (b).

Solution. (a) Let us consider the Markov chain X_n , initialized with $X_0 = s$. We first assert that for all i ,

$$\rho_i = \sum_{n=1}^{\infty} \mathbf{P}(X_1 \neq s, \dots, X_{n-1} \neq s, X_n = i).$$

To see this, we first consider the case $i \neq s$, and let I_n be the random variable that takes the value 1 if $X_1 \neq s, \dots, X_{n-1} \neq s$, and $X_n = i$, and the value 0 otherwise. Then, the number of visits to state i before the next visit to state s is equal to $\sum_{n=1}^{\infty} I_n$. Thus,[†]

$$\begin{aligned}\rho_i &= \mathbf{E} \left[\sum_{n=1}^{\infty} I_n \right] \\ &= \sum_{n=1}^{\infty} \mathbf{E}[I_n] \\ &= \sum_{n=1}^{\infty} \mathbf{P}(X_1 \neq s, \dots, X_{n-1} \neq s, X_n = i).\end{aligned}$$

When $i = s$, the events $\{X_1 \neq s, \dots, X_{n-1} \neq s, X_n = s\}$, for the different values of n , form a partition of the sample space, because they correspond to the different possibilities for the time of the next visit to state s . Thus,

$$\sum_{n=1}^{\infty} \mathbf{P}(X_1 \neq s, \dots, X_{n-1} \neq s, X_n = s) = 1 = \rho_s,$$

which completes the verification of our assertion.

We next use the total probability theorem to write for $n \geq 2$,

$$\mathbf{P}(X_1 \neq s, \dots, X_{n-1} \neq s, X_n = i) = \sum_{k \neq s} \mathbf{P}(X_1 \neq s, \dots, X_{n-2} \neq s, X_{n-1} = k) p_{ki}.$$

[†] The interchange of the infinite summation and the expectation in the subsequent calculation can be justified by the following argument. We have for any $k > 0$,

$$\mathbf{E} \left[\sum_{n=1}^{\infty} I_n \right] = \mathbf{E} \left[\sum_{n=1}^k I_n \right] + \mathbf{E} \left[\sum_{n=k+1}^{\infty} I_n \right] = \sum_{n=1}^k \mathbf{E}[I_n] + \mathbf{E} \left[\sum_{n=k+1}^{\infty} I_n \right].$$

Let T be the first positive time that state s is visited. Then,

$$\mathbf{E} \left[\sum_{n=k+1}^{\infty} I_n \right] = \sum_{t=k+2}^{\infty} \mathbf{P}(T = t) \mathbf{E} \left[\sum_{n=k+1}^{\infty} I_n \mid T = t \right] \leq \sum_{t=k+2}^{\infty} t \mathbf{P}(T = t).$$

Since the mean recurrence time $\sum_{t=1}^{\infty} t \mathbf{P}(T = t)$ is finite, the limit, as $k \rightarrow \infty$ of $\sum_{t=k+2}^{\infty} t \mathbf{P}(T = t)$ is equal to zero. We take the limit of both sides of the earlier equation, as $k \rightarrow \infty$, to obtain the desired relation

$$\mathbf{E} \left[\sum_{n=1}^{\infty} I_n \right] = \sum_{n=1}^{\infty} \mathbf{E}[I_n].$$

We thus obtain

$$\begin{aligned}
 \rho_i &= \sum_{n=1}^{\infty} \mathbf{P}(X_1 \neq s, \dots, X_{n-1} \neq s, X_n = i) \\
 &= p_{si} + \sum_{n=2}^{\infty} \mathbf{P}(X_1 \neq s, \dots, X_{n-1} \neq s, X_n = i) \\
 &= p_{si} + \sum_{n=2}^{\infty} \sum_{k \neq s} \mathbf{P}(X_1 \neq s, \dots, X_{n-2} \neq s, X_{n-1} = k) p_{ki} \\
 &= p_{si} + \sum_{k \neq s} p_{ki} \sum_{n=2}^{\infty} \mathbf{P}(X_1 \neq s, \dots, X_{n-2} \neq s, X_{n-1} = k) \\
 &= \rho_s p_{si} + \sum_{k \neq s} p_{ki} \rho_k \\
 &= \sum_{k=1}^m \rho_k p_{ki}.
 \end{aligned}$$

(b) Dividing both sides of the relation established in part (a) by t_s^* , we obtain

$$\pi_i = \sum_{k=1}^m \pi_k p_{ki},$$

where $\pi_i = \rho_i/t_s^*$. Thus, the π_i solve the balance equations. Furthermore, the π_i are nonnegative, and we clearly have $\sum_{i=1}^m \rho_i = t_s^*$ or $\sum_{i=1}^m \pi_i = 1$. Hence, (π_1, \dots, π_m) is a probability distribution.

(c) Consider a probability distribution (π_1, \dots, π_m) that satisfies the balance equations. Fix a recurrent state s , let t_s^* be the mean recurrence time of s , and let t_i be the mean first passage time from a state $i \neq s$ to state s . We will show that $\pi_s t_s^* = 1$. Indeed, we have

$$\begin{aligned}
 t_s^* &= 1 + \sum_{j \neq s} p_{sj} t_j, \\
 t_i &= 1 + \sum_{j \neq s} p_{ij} t_j, \quad \text{for all } i \neq s.
 \end{aligned}$$

Multiplying these equations with π_s and π_i , respectively, and adding, we obtain

$$\pi_s t_s^* + \sum_{i \neq s} \pi_i t_i = 1 + \sum_{i=1}^m \pi_i \sum_{j \neq s} p_{ij} t_j.$$

By using the balance equations, the right-hand side is equal to

$$1 + \sum_{i=1}^m \pi_i \sum_{j \neq s} p_{ij} t_j = 1 + \sum_{j \neq s} t_j \sum_{i=1}^m \pi_i p_{ij} = 1 + \sum_{j \neq s} t_j \pi_j.$$

By combining the last two equations, we obtain $\pi_s t_s^* = 1$.

Since the probability distribution (π_1, \dots, π_m) satisfies the balance equations, if the initial state X_0 is chosen according to this distribution, all subsequent states X_n have the same distribution. If we start at a transient state i , the probability of being at that state at time n diminishes to 0 as $n \rightarrow \infty$. It follows that we must have $\pi_i = 0$.

(d) Part (b) shows that there exists at least one probability distribution that satisfies the balance equations. Part (c) shows that there can be only one such probability distribution.

SECTION 6.5. Continuous-Time Markov Chains

Problem 33. A facility of m identical machines is sharing a single repairperson. The time to repair a failed machine is exponentially distributed with mean $1/\lambda$. A machine once operational, fails after a time that is exponentially distributed with mean $1/\mu$. All failure and repair times are independent.

- (a) Find the steady-state probability that there is no operational machine.
- (b) Find the expected number of operational machines, in steady-state.

Problem 34. An athletic facility has 5 tennis courts. Pairs of players arrive at the courts according to a Poisson process with rate of one pair per 10 minutes, and use a court for an exponentially distributed time with mean 40 minutes. Suppose a pair of players arrives and finds all courts busy and k other pairs waiting in queue. What is the expected waiting time to get a court?

Problem 35. Empty taxis pass by a street corner at a Poisson rate of two per minute and pick up a passenger if one is waiting there. Passengers arrive at the street corner at a Poisson rate of one per minute and wait for a taxi only if there are less than four persons waiting; otherwise they leave and never return. Penelope arrives at the street corner at a given time. Find her expected waiting time, given that she joins the queue. Assume that the process is in steady-state.

Problem 36. There are m users who share a computer system. Each user alternates between “thinking” intervals whose durations are independent exponentially distributed with parameter λ , and an “active” mode that starts by submitting a service request. The server can only serve one request at a time, and will serve a request completely before serving other requests. The service times of different requests are independent exponentially distributed random variables with parameter μ , and also independent of the thinking times of the users. Construct a Markov chain model and derive the steady-state distribution of the number of pending requests, including the one presently served, if any.

Problem 37.* Consider a continuous-time Markov chain in which the transition rates ν_i are the same for all i . Assume that the process has a single recurrent class.

- (a) Explain why the sequence Y_n of transition times form a Poisson process.
- (b) Show that the steady-state probabilities of the Markov chain $X(t)$ are the same as the steady-state probabilities of the embedded Markov chain X_n .

Solution. (a) Denote by ν the common value of the transition rates ν_i . The sequence $\{Y_n\}$ is a sequence of independent exponentially distributed time intervals with parameter ν . Therefore they can be associated with the arrival times of a Poisson process with rate ν .

(b) The balance and normalization equations for the continuous-time chain are

$$\begin{aligned}\pi_j \sum_{k \neq j} q_{jk} &= \sum_{k \neq j} \pi_k q_{kj}, \quad j = 1, \dots, m, \\ 1 &= \sum_{k=1}^m \pi_k.\end{aligned}$$

By using the relation $q_{jk} = \nu p_{jk}$, and by canceling the common factor ν , these equations are written as

$$\begin{aligned}\pi_j \sum_{k \neq j} p_{jk} &= \sum_{k \neq j} \pi_k p_{kj}, \quad j = 1, \dots, m, \\ 1 &= \sum_{k=1}^m \pi_k.\end{aligned}$$

We have $\sum_{k \neq j} p_{jk} = 1 - p_{jj}$, so the first of these two equations is written as

$$\pi_j (1 - p_{jj}) = \sum_{k \neq j} \pi_k p_{kj},$$

or

$$\pi_j = \sum_{k=1}^m \pi_k p_{kj}, \quad j = 1, \dots, m.$$

These are the balance equations for the embedded Markov chain, which have a unique solution since the embedded Markov chain has a single recurrent class, which is aperiodic. Hence the π_j are the steady-state probabilities for the embedded Markov chain.

Limit Theorems

Contents

7.1. Markov and Chebyshev Inequalities	p. 381
7.2. The Weak Law of Large Numbers	p. 383
7.3. Convergence in Probability	p. 386
7.4. The Central Limit Theorem	p. 388
7.5. The Strong Law of Large Numbers	p. 395
7.6. Summary and Discussion	p. 397
Problems	p. 399

In this chapter, we discuss some fundamental issues related to the asymptotic behavior of sequences of random variables. Our principal context involves a sequence X_1, X_2, \dots of independent identically distributed random variables with mean μ and variance σ^2 . Let

$$S_n = X_1 + \dots + X_n$$

be the sum of the first n of them. Limit theorems are mostly concerned with the properties of S_n and related random variables as n becomes very large.

Because of independence, we have

$$\text{var}(S_n) = \text{var}(X_1) + \dots + \text{var}(X_n) = n\sigma^2.$$

Thus, the distribution of S_n spreads out as n increases, and cannot have a meaningful limit. The situation is different if we consider the **sample mean**

$$M_n = \frac{X_1 + \dots + X_n}{n} = \frac{S_n}{n}.$$

A quick calculation yields

$$\mathbf{E}[M_n] = \mu, \quad \text{var}(M_n) = \frac{\sigma^2}{n}.$$

In particular, the variance of M_n decreases to zero as n increases, and the bulk of the distribution of M_n must be very close to the mean μ . This phenomenon is the subject of certain laws of large numbers, which generally assert that the sample mean M_n (a random variable) converges to the true mean μ (a number), in a precise sense. These laws provide a mathematical basis for the loose interpretation of an expectation $\mathbf{E}[X] = \mu$ as the average of a large number of independent samples drawn from the distribution of X .

We will also consider a quantity which is intermediate between S_n and M_n . We first subtract $n\mu$ from S_n , to obtain the zero-mean random variable $S_n - n\mu$ and then divide by $\sigma\sqrt{n}$, to form the random variable

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

It can be seen that

$$\mathbf{E}[Z_n] = 0, \quad \text{var}(Z_n) = 1.$$

Since the mean and the variance of Z_n remain unchanged as n increases, its distribution neither spreads, nor shrinks to a point. The **central limit theorem** is concerned with the asymptotic shape of the distribution of Z_n and asserts that it becomes the standard normal distribution.

Limit theorems are useful for several reasons:

- Conceptually, they provide an interpretation of expectations (as well as probabilities) in terms of a long sequence of identical independent experiments.
- They allow for an approximate analysis of the properties of random variables such as S_n . This is to be contrasted with an exact analysis which would require a formula for the PMF or PDF of S_n , a complicated and tedious task when n is large.

7.1 MARKOV AND CHEBYSHEV INEQUALITIES

In this section, we derive some important inequalities. These inequalities use the mean, and possibly the variance, of a random variable to draw conclusions on the probabilities of certain events. They are primarily useful in situations where the mean and variance of a random variable X are easily computable, but the distribution of X is either unavailable or hard to calculate.

We first present the **Markov inequality**. Loosely speaking, it asserts that if a *nonnegative* random variable has a small mean, then the probability that it takes a large value must also be small.

Markov Inequality

If a random variable X can only take nonnegative values, then

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}, \quad \text{for all } a > 0.$$

To justify the Markov inequality, let us fix a positive number a and consider the random variable Y_a defined by

$$Y_a = \begin{cases} 0, & \text{if } X < a, \\ a, & \text{if } X \geq a. \end{cases}$$

It is seen that the relation

$$Y_a \leq X$$

always holds and therefore,

$$\mathbf{E}[Y_a] \leq \mathbf{E}[X].$$

On the other hand,

$$\mathbf{E}[Y_a] = a\mathbf{P}(Y_a = a) = a\mathbf{P}(X \geq a),$$

from which we obtain

$$a\mathbf{P}(X \geq a) \leq \mathbf{E}[X];$$

see Fig. 7.1 for an illustration.

Example 7.1. Let X be uniformly distributed in the interval $[0, 4]$ and note that $\mathbf{E}[X] = 2$. Then, the Markov inequality asserts that

$$\mathbf{P}(X \geq 2) \leq \frac{2}{2} = 1, \quad \mathbf{P}(X \geq 3) \leq \frac{2}{3} = 0.67, \quad \mathbf{P}(X \geq 4) \leq \frac{2}{4} = 0.5.$$

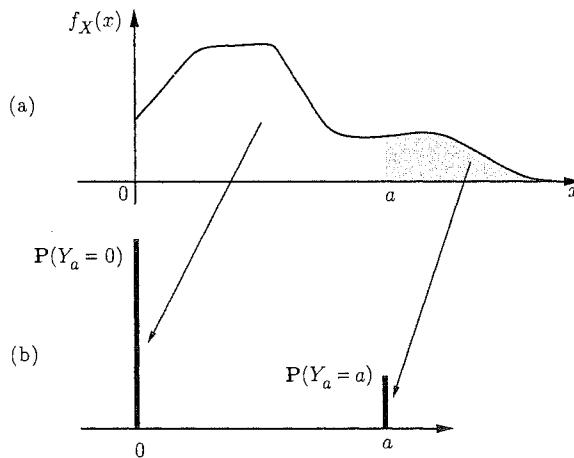


Figure 7.1: Illustration of the derivation of the Markov inequality. Part (a) of the figure shows the PDF of a nonnegative random variable X . Part (b) shows the PMF of a related random variable Y_a , which is constructed as follows. All of the probability mass in the PDF of X that lies between 0 and a is assigned to 0, and all of the mass that lies above a is assigned to a . Since mass is shifted to the left, the expectation can only decrease and, therefore,

$$\mathbf{E}[X] \geq \mathbf{E}[Y_a] = a\mathbf{P}(Y_a = a) = a\mathbf{P}(X \geq a).$$

By comparing with the exact probabilities

$$\mathbf{P}(X \geq 2) = 0.5, \quad \mathbf{P}(X \geq 3) = 0.25, \quad \mathbf{P}(X \geq 4) = 0,$$

we see that the bounds provided by the Markov inequality can be quite loose.

We continue with the **Chebyshev inequality**. Loosely speaking, it asserts that if the variance of a random variable is small, then the probability that it takes a value far from its mean is also small. Note that the Chebyshev inequality does not require the random variable to be nonnegative.

Chebyshev Inequality

If X is a random variable with mean μ and variance σ^2 , then

$$\mathbf{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}, \quad \text{for all } c > 0.$$

To justify the Chebyshev inequality, we consider the nonnegative random variable $(X - \mu)^2$ and apply the Markov inequality with $a = c^2$. We obtain

$$\mathbf{P}((X - \mu)^2 \geq c^2) \leq \frac{\mathbf{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}.$$

We complete the derivation by observing that the event $(X - \mu)^2 \geq c^2$ is identical to the event $|X - \mu| \geq c$, so that

$$\mathbf{P}(|X - \mu| \geq c) = \mathbf{P}((X - \mu)^2 \geq c^2) \leq \frac{\sigma^2}{c^2}.$$

An alternative form of the Chebyshev inequality is obtained by letting $c = k\sigma$, where k is positive, which yields

$$\mathbf{P}(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Thus, the probability that a random variable takes a value more than k standard deviations away from its mean is at most $1/k^2$.

The Chebyshev inequality tends to be more powerful than the Markov inequality (the bounds that it provides are more accurate), because it also uses information on the variance of X . Still, the mean and the variance of a random variable are only a rough summary of its properties, and we cannot expect the bounds to be close approximations of the exact probabilities.

Example 7.2. As in Example 7.1, let X be uniformly distributed in $[0, 4]$. Let us use the Chebyshev inequality to bound the probability that $|X - 2| \geq 1$. We have $\sigma^2 = 16/12 = 4/3$, and

$$\mathbf{P}(|X - 2| \geq 1) \leq \frac{4}{3},$$

which is uninformative.

For another example, let X be exponentially distributed with parameter $\lambda = 1$, so that $\mathbf{E}[X] = \text{var}(X) = 1$. For $c > 1$, using the Chebyshev inequality, we obtain

$$\mathbf{P}(X \geq c) = \mathbf{P}(X - 1 \geq c - 1) \leq \mathbf{P}(|X - 1| \geq c - 1) \leq \frac{1}{(c - 1)^2}.$$

This is again conservative compared to the exact answer $\mathbf{P}(X \geq c) = e^{-c}$.

7.2 THE WEAK LAW OF LARGE NUMBERS

The weak law of large numbers asserts that the sample mean of a large number of independent identically distributed random variables is very close to the true mean, with high probability.

As in the introduction to this chapter, we consider a sequence X_1, X_2, \dots of independent identically distributed random variables with mean μ and variance σ^2 , and define the sample mean by

$$M_n = \frac{X_1 + \dots + X_n}{n}.$$

We have

$$\mathbf{E}[M_n] = \frac{\mathbf{E}[X_1] + \dots + \mathbf{E}[X_n]}{n} = \frac{n\mu}{n} = \mu,$$

and, using independence,

$$\text{var}(M_n) = \frac{\text{var}(X_1 + \dots + X_n)}{n^2} = \frac{\text{var}(X_1) + \dots + \text{var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

We apply the Chebyshev inequality and obtain

$$\mathbf{P}(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}, \quad \text{for any } \epsilon > 0.$$

We observe that for any fixed $\epsilon > 0$, the right-hand side of this inequality goes to zero as n increases. As a consequence, we obtain the weak law of large numbers, which is stated below. It turns out that this law remains true even if the X_i have infinite variance, but a much more elaborate argument is needed, which we omit. The only assumption needed is that $\mathbf{E}[X_i]$ is finite.

The Weak Law of Large Numbers

Let X_1, X_2, \dots be independent identically distributed random variables with mean μ . For every $\epsilon > 0$, we have

$$\mathbf{P}(|M_n - \mu| \geq \epsilon) = \mathbf{P}\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

The weak law of large numbers states that for large n , the bulk of the distribution of M_n is concentrated near μ . That is, if we consider a positive length interval $[\mu - \epsilon, \mu + \epsilon]$ around μ , then there is high probability that M_n will fall in that interval; as $n \rightarrow \infty$, this probability converges to 1. Of course, if ϵ is very small, we may have to wait longer (i.e., need a larger value of n) before we can assert that M_n is highly likely to fall in that interval.

Example 7.3. Probabilities and Frequencies. Consider an event A defined in the context of some probabilistic experiment. Let $p = \mathbf{P}(A)$ be the probability

of that event. We consider n independent repetitions of the experiment, and let M_n be the fraction of time that event A occurs; in this context, M_n is often called the **empirical frequency** of A . Note that

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

where X_i is 1 whenever A occurs, and 0 otherwise; in particular, $\mathbf{E}[X_i] = p$. The weak law applies and shows that when n is large, the empirical frequency is most likely to be within ϵ of p . Loosely speaking, this allows us to say that empirical frequencies are faithful estimates of p . Alternatively, this is a step towards interpreting the probability p as the frequency of occurrence of A .

Example 7.4. Polling. Let p be the fraction of voters who support a particular candidate for office. We interview n “randomly selected” voters and record M_n , the fraction of them that support the candidate. We view M_n as our estimate of p and would like to investigate its properties.

We interpret “randomly selected” to mean that the n voters are chosen independently and uniformly from the given population. Thus, the reply of each person interviewed can be viewed as an independent Bernoulli random variable X_i with success probability p and variance $\sigma^2 = p(1 - p)$. The Chebyshev inequality yields

$$\mathbf{P}(|M_n - p| \geq \epsilon) \leq \frac{p(1 - p)}{n\epsilon^2}.$$

The true value of the parameter p is assumed to be unknown. On the other hand, it is easily verified that $p(1 - p) \leq 1/4$, which yields

$$\mathbf{P}(|M_n - p| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}.$$

For example, if $\epsilon = 0.1$ and $n = 100$, we obtain

$$\mathbf{P}(|M_{100} - p| \geq 0.1) \leq \frac{1}{4 \cdot 100 \cdot (0.1)^2} = 0.25.$$

In words, with a sample size of $n = 100$, the probability that our estimate is incorrect by more than 0.1 is no larger than 0.25.

Suppose now that we impose some tight specifications on our poll. We would like to have high confidence (probability at least 95%) that our estimate will be very accurate (within .01 of p). How many voters should be sampled?

The only guarantee that we have at this point is the inequality

$$\mathbf{P}(|M_n - p| \geq 0.01) \leq \frac{1}{4n(0.01)^2}.$$

We will be sure to satisfy the above specifications if we choose n large enough so that

$$\frac{1}{4n(0.01)^2} \leq 1 - 0.95 = 0.05,$$

which yields $n \geq 50,000$. This choice of n satisfies our specifications, but turns out to be fairly conservative, because it is based on the rather loose Chebyshev inequality. A refinement will be considered in Section 7.4.

7.3 CONVERGENCE IN PROBABILITY

We can interpret the weak law of large numbers as stating that “ M_n converges to μ .” However, since M_1, M_2, \dots is a sequence of random variables, not a sequence of numbers, the meaning of convergence has to be made precise. A particular definition is provided below. To facilitate the comparison with the ordinary notion of convergence, we also include the definition of the latter.

Convergence of a Deterministic Sequence

Let a_1, a_2, \dots be a sequence of real numbers, and let a be another real number. We say that the sequence a_n converges to a , or $\lim_{n \rightarrow \infty} a_n = a$, if for every $\epsilon > 0$ there exists some n_0 such that

$$|a_n - a| \leq \epsilon, \quad \text{for all } n \geq n_0.$$

Intuitively, if $\lim_{n \rightarrow \infty} a_n = a$, then for any given accuracy level ϵ , a_n must be within ϵ of a , when n is large enough.

Convergence in Probability

Let Y_1, Y_2, \dots be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence Y_n **converges to a in probability**, if for every $\epsilon > 0$, we have

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Y_n - a| \geq \epsilon) = 0.$$

Given this definition, the weak law of large numbers simply says that the sample mean converges in probability to the true mean μ .

If the random variables Y_1, Y_2, \dots have a PMF or a PDF and converge in probability to a , then according to the above definition, “almost all” of the PMF or PDF of Y_n is concentrated in an ϵ -interval around a for large values of n . It is also instructive to rephrase the above definition as follows: for every $\epsilon > 0$, and for every $\delta > 0$, there exists some n_0 such that

$$\mathbf{P}(|Y_n - a| \geq \epsilon) \leq \delta, \quad \text{for all } n \geq n_0.$$

If we refer to ϵ as the *accuracy* level, and δ as the *confidence* level, the definition takes the following intuitive form: for any given level of accuracy and confidence, Y_n will be equal to a , within these levels of accuracy and confidence, provided that n is large enough.

Example 7.5. Consider a sequence of independent random variables X_n that are uniformly distributed in the interval $[0, 1]$, and let

$$Y_n = \min\{X_1, \dots, X_n\}.$$

The sequence of values of Y_n cannot increase as n increases, and it will occasionally decrease (when a value of X_n that is smaller than the preceding values is obtained). Thus, we intuitively expect that Y_n converges to zero. Indeed, for $\epsilon > 0$, we have using the independence of the X_n ,

$$\begin{aligned}\mathbf{P}(|Y_n - 0| \geq \epsilon) &= \mathbf{P}(X_1 \geq \epsilon, \dots, X_n \geq \epsilon) \\ &= \mathbf{P}(X_1 \geq \epsilon) \cdots \mathbf{P}(X_n \geq \epsilon) \\ &= (1 - \epsilon)^n.\end{aligned}$$

In particular,

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Y_n - 0| \geq \epsilon) = \lim_{n \rightarrow \infty} (1 - \epsilon)^n = 0.$$

Since this is true for every $\epsilon > 0$, we conclude that Y_n converges to zero, in probability.

Example 7.6. Let Y be an exponentially distributed random variable with parameter $\lambda = 1$. For any positive integer n , let $Y_n = Y/n$. (Note that these random variables are dependent.) We wish to investigate whether the sequence Y_n converges to zero.

For $\epsilon > 0$, we have

$$\mathbf{P}(|Y_n - 0| \geq \epsilon) = \mathbf{P}(Y_n \geq \epsilon) = \mathbf{P}(Y \geq n\epsilon) = e^{-n\epsilon}.$$

In particular,

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Y_n - 0| \geq \epsilon) = \lim_{n \rightarrow \infty} e^{-n\epsilon} = 0.$$

Since this is the case for every $\epsilon > 0$, Y_n converges to zero, in probability.

One might be tempted to believe that if a sequence Y_n converges to a number a , then $\mathbf{E}[Y_n]$ must also converge to a . The following example shows that this need not be the case, and illustrates some of the limitations of the notion of convergence in probability.

Example 7.7. Consider a sequence of discrete random variables Y_n with the following distribution:

$$\mathbf{P}(Y_n = y) = \begin{cases} 1 - \frac{1}{n}, & \text{for } y = 0, \\ \frac{1}{n}, & \text{for } y = n^2, \\ 0, & \text{elsewhere;} \end{cases}$$

see Fig. 7.2 for an illustration. For every $\epsilon > 0$, we have

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Y_n| \geq \epsilon) = \lim_{n \rightarrow \infty} \frac{1}{n} = 0,$$

and Y_n converges to zero in probability. On the other hand, $\mathbf{E}[Y_n] = n^2/n = n$, which goes to infinity as n increases.

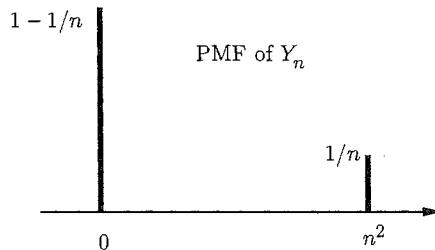


Figure 7.2: The PMF of the random variable Y_n in Example 7.7.

7.4 THE CENTRAL LIMIT THEOREM

According to the weak law of large numbers, the distribution of the sample mean M_n is increasingly concentrated in the near vicinity of the true mean μ . In particular, its variance tends to zero. On the other hand, the variance of the sum

$$S_n = X_1 + \cdots + X_n = nM_n$$

increases to infinity, and the distribution of S_n cannot be said to converge to anything meaningful. An intermediate view is obtained by considering the deviation $S_n - n\mu$ of S_n from its mean $n\mu$, and scaling it by a factor proportional to $1/\sqrt{n}$. What is special about this particular scaling is that it keeps the variance at a constant level. The central limit theorem asserts that the distribution of this scaled random variable approaches a normal distribution.

More specifically, let X_1, X_2, \dots be a sequence of independent identically distributed random variables with mean μ and variance σ^2 . We define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

An easy calculation yields

$$\mathbf{E}[Z_n] = \frac{\mathbf{E}[X_1 + \cdots + X_n] - n\mu}{\sigma\sqrt{n}} = 0,$$

and

$$\text{var}(Z_n) = \frac{\text{var}(X_1 + \cdots + X_n)}{\sigma^2 n} = \frac{\text{var}(X_1) + \cdots + \text{var}(X_n)}{\sigma^2 n} = \frac{n\sigma^2}{n\sigma^2} = 1.$$

The Central Limit Theorem

Let X_1, X_2, \dots be a sequence of independent identically distributed random variables with common mean μ and variance σ^2 , and define

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Then, the CDF of Z_n converges to the standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx,$$

in the sense that

$$\lim_{n \rightarrow \infty} \mathbf{P}(Z_n \leq z) = \Phi(z), \quad \text{for every } z.$$

The central limit theorem is surprisingly general. Besides independence, and the implicit assumption that the mean and variance are finite, it places no other requirement on the distribution of the X_i , which could be discrete, continuous, or mixed; see the end-of-chapter problems for an outline of its proof.

The theorem is of tremendous importance for several reasons, both conceptual and practical. On the conceptual side, it indicates that the sum of a large number of independent random variables is approximately normal. As such, it applies to many situations in which a random effect is the sum of a large number of small but independent random factors. Noise in many natural or engineered systems has this property. In a wide array of contexts, it has been found empirically that the statistics of noise are well-described by normal distributions, and the central limit theorem provides a convincing explanation for this phenomenon.

On the practical side, the central limit theorem eliminates the need for detailed probabilistic models, and for tedious manipulations of PMFs and PDFs. Rather, it allows the calculation of certain probabilities by simply referring to the normal CDF table. Furthermore, these calculations only require the knowledge of means and variances.

Approximations Based on the Central Limit Theorem

The central limit theorem allows us to calculate probabilities related to Z_n as

if Z_n were normal. Since normality is preserved under linear transformations, this is equivalent to treating S_n as a normal random variable with mean $n\mu$ and variance $n\sigma^2$.

Normal Approximation Based on the Central Limit Theorem

Let $S_n = X_1 + \dots + X_n$, where the X_i are independent identically distributed random variables with mean μ and variance σ^2 . If n is large, the probability $\mathbf{P}(S_n \leq c)$ can be approximated by treating S_n as if it were normal, according to the following procedure.

1. Calculate the mean $n\mu$ and the variance $n\sigma^2$ of S_n .
2. Calculate the normalized value $z = (c - n\mu)/\sigma\sqrt{n}$.
3. Use the approximation

$$\mathbf{P}(S_n \leq c) \approx \Phi(z),$$

where $\Phi(z)$ is available from standard normal CDF tables.

Example 7.8. We load on a plane 100 packages whose weights are independent random variables that are uniformly distributed between 5 and 50 pounds. What is the probability that the total weight will exceed 3000 pounds? It is not easy to calculate the CDF of the total weight and the desired probability, but an approximate answer can be quickly obtained using the central limit theorem.

We want to calculate $\mathbf{P}(S_{100} > 3000)$, where S_{100} is the sum of the weights of 100 packages. The mean and the variance of the weight of a single package are

$$\mu = \frac{5 + 50}{2} = 27.5, \quad \sigma^2 = \frac{(50 - 5)^2}{12} = 168.75,$$

based on the formulas for the mean and variance of the uniform PDF. We thus calculate the normalized value

$$z = \frac{3000 - 100 \cdot 27.5}{\sqrt{168.75 \cdot 100}} = \frac{250}{129.9} = 1.92,$$

and use the standard normal tables to obtain the approximation

$$\mathbf{P}(S_{100} \leq 3000) \approx \Phi(1.92) = 0.9726.$$

Thus the desired probability is

$$\mathbf{P}(S_{100} > 3000) = 1 - \mathbf{P}(S_{100} \leq 3000) \approx 1 - 0.9726 = 0.0274.$$

Example 7.9. A machine processes parts, one at a time. The processing times of different parts are independent random variables, uniformly distributed in $[1, 5]$.

We wish to approximate the probability that the number of parts processed within 320 time units, denoted by N_{320} , is at least 100.

There is no obvious way of expressing the random variable N_{320} as a sum of independent random variables, but we can proceed differently. Let X_i be the processing time of the i th part, and let $S_{100} = X_1 + \dots + X_{100}$ be the total processing time of the first 100 parts. The event $\{N_{320} \geq 100\}$ is the same as the event $\{S_{100} \leq 320\}$, and we can now use a normal approximation to the distribution of S_{100} . Note that $\mu = \mathbf{E}[X_i] = 3$ and $\sigma^2 = \text{var}(X_i) = 16/12 = 4/3$. We calculate the normalized value

$$z = \frac{320 - n\mu}{\sigma\sqrt{n}} = \frac{320 - 300}{\sqrt{100 \cdot 4/3}} = 1.73,$$

and use the approximation

$$\mathbf{P}(S_{100} \leq 320) \approx \Phi(1.73) = 0.9582.$$

If the variance of the X_i is unknown, but an upper bound is available, the normal approximation can be used to obtain bounds on the probabilities of interest.

Example 7.10. Polling. Let us revisit the polling problem in Example 7.4. We poll n voters and record the fraction M_n of those polled who are in favor of a particular candidate. If p is the fraction of the entire voter population that supports this candidate, then

$$M_n = \frac{X_1 + \dots + X_n}{n},$$

where the X_i are independent Bernoulli random variables with parameter p . In particular, M_n has mean p and variance $p(1-p)/n$. By the normal approximation, $X_1 + \dots + X_n$ is approximately normal, and therefore M_n is also approximately normal.

We are interested in the probability $\mathbf{P}(|M_n - p| \geq \epsilon)$ that the polling error is larger than some desired accuracy ϵ . Because of the symmetry of the normal PDF around the mean, we have

$$\mathbf{P}(|M_n - p| \geq \epsilon) \approx 2\mathbf{P}(M_n - p \geq \epsilon).$$

The variance $p(1-p)/n$ of $M_n - p$ depends on p and is therefore unknown. We note that the probability of a large deviation from the mean increases with the variance. Thus, we can obtain an upper bound on $\mathbf{P}(M_n - p \geq \epsilon)$ by assuming that $M_n - p$ has the largest possible variance, namely, $1/(4n)$ which corresponds to $p = 1/2$. To calculate this upper bound, we evaluate the standardized value

$$z = \frac{\epsilon}{1/(2\sqrt{n})},$$

and use the normal approximation

$$\mathbf{P}(M_n - p \geq \epsilon) \leq 1 - \Phi(z) = 1 - \Phi(2\epsilon\sqrt{n}).$$

For instance, consider the case where $n = 100$ and $\epsilon = 0.1$. Assuming the worst-case variance, and treating M_n as if it were normal, we obtain

$$\begin{aligned}\mathbf{P}(|M_{100} - p| \geq 0.1) &\approx 2\mathbf{P}(M_n - p \geq 0.1) \\ &\leq 2 - 2\Phi(2 \cdot 0.1 \cdot \sqrt{100}) = 2 - 2\Phi(2) = 2 - 2 \cdot 0.977 = 0.046.\end{aligned}$$

This is much smaller (and more accurate) than the estimate of 0.25 that was obtained in Example 7.4 using the Chebyshev inequality.

We now consider a reverse problem. How large a sample size n is needed if we wish our estimate M_n to be within 0.01 of p with probability at least 0.95? Assuming again the worst possible variance, we are led to the condition

$$2 - 2\Phi(2 \cdot 0.01 \cdot \sqrt{n}) \leq 0.05,$$

or

$$\Phi(2 \cdot 0.01 \cdot \sqrt{n}) \geq 0.975.$$

From the normal tables, we see that $\Phi(1.96) = 0.975$, which leads to

$$2 \cdot 0.01 \cdot \sqrt{n} \geq 1.96,$$

or

$$n \geq \frac{(1.96)^2}{4 \cdot (0.01)^2} = 9604.$$

This is significantly better than the sample size of 50,000 that we found using Chebyshev's inequality.

The normal approximation is increasingly accurate as n tends to infinity, but in practice we are generally faced with specific and finite values of n . It would be useful to know how large n should be before the approximation can be trusted, but there are no simple and general guidelines. Much depends on whether the distribution of the X_i is close to normal and, in particular, whether it is symmetric. For example, if the X_i are uniform, then S_8 is already very close to normal. But if the X_i are, say, exponential, a significantly larger n will be needed before the distribution of S_n is close to a normal one. Furthermore, the normal approximation to $\mathbf{P}(S_n \leq c)$ tends to be more faithful when c is in the vicinity of the mean of S_n .

De Moivre – Laplace Approximation to the Binomial

A binomial random variable S_n with parameters n and p can be viewed as the sum of n independent Bernoulli random variables X_1, \dots, X_n , with common parameter p :

$$S_n = X_1 + \dots + X_n.$$

Recall that

$$\mu = \mathbf{E}[X_i] = p, \quad \sigma = \sqrt{\text{var}(X_i)} = \sqrt{p(1-p)},$$

We will now use the approximation suggested by the central limit theorem to provide an approximation for the probability of the event $\{k \leq S_n \leq l\}$, where k and l are given integers. We express the event of interest in terms of a standardized random variable, using the equivalence

$$k \leq S_n \leq l \iff \frac{k - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{l - np}{\sqrt{np(1-p)}}.$$

By the central limit theorem, $(S_n - np)/\sqrt{np(1-p)}$ has approximately a standard normal distribution, and we obtain

$$\begin{aligned} \mathbf{P}(k \leq S_n \leq l) &= \mathbf{P}\left(\frac{k - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{l - np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{l - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right). \end{aligned}$$

An approximation of this form is equivalent to treating S_n as a normal random variable with mean np and variance $np(1-p)$. Figure 7.3 provides an illustration and indicates that a more accurate approximation may be possible if we replace k and l by $k - \frac{1}{2}$ and $l + \frac{1}{2}$, respectively. The corresponding formula is given below.

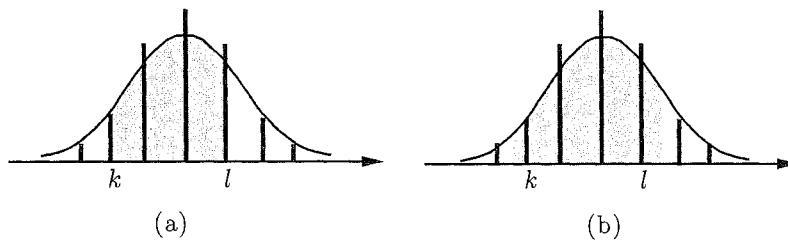


Figure 7.3: The central limit approximation treats a binomial random variable S_n as if it were normal with mean np and variance $np(1-p)$. This figure shows a binomial PMF together with the approximating normal PDF. (a) A first approximation of a binomial probability $\mathbf{P}(k \leq S_n \leq l)$ is obtained by integrating the area under the normal PDF from k to l , which is the shaded area in the figure. With this approach, if we have $k = l$, the probability $\mathbf{P}(S_n = k)$ would be approximated by zero. (b) A possible remedy is to use the normal probability between $k - \frac{1}{2}$ and $k + \frac{1}{2}$ to approximate $\mathbf{P}(S_n = k)$. By extending this idea, $\mathbf{P}(k \leq S_n \leq l)$ can be approximated by using the area under the normal PDF from $k - \frac{1}{2}$ to $l + \frac{1}{2}$, which corresponds to the shaded area.

De Moivre – Laplace Approximation to the Binomial

If S_n is a binomial random variable with parameters n and p , n is large, and k, l are nonnegative integers, then

$$\mathbf{P}(k \leq S_n \leq l) \approx \Phi\left(\frac{l + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

When p is close to $1/2$, in which case the PMF of the X_i is symmetric, the above formula yields a very good approximation for n as low as 40 or 50. When p is near 1 or near 0, the quality of the approximation drops, and a larger value of n is needed to maintain the same accuracy.

Example 7.11. Let S_n be a binomial random variable with parameters $n = 36$ and $p = 0.5$. An exact calculation yields

$$\mathbf{P}(S_n \leq 21) = \sum_{k=0}^{21} \binom{36}{k} (0.5)^{36} = 0.8785.$$

The central limit theorem approximation, without the above discussed refinement, yields

$$\mathbf{P}(S_n \leq 21) \approx \Phi\left(\frac{21 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{21 - 18}{3}\right) = \Phi(1) = 0.8413.$$

Using the proposed refinement, we have

$$\mathbf{P}(S_n \leq 21) \approx \Phi\left(\frac{21.5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{21.5 - 18}{3}\right) = \Phi(1.17) = 0.879,$$

which is much closer to the exact value.

The de Moivre – Laplace formula also allows us to approximate the probability of a single value. For example,

$$\mathbf{P}(S_n = 19) \approx \Phi\left(\frac{19.5 - 18}{3}\right) - \Phi\left(\frac{18.5 - 18}{3}\right) = 0.6915 - 0.5675 = 0.124.$$

This is very close to the exact value which is

$$\binom{36}{19} (0.5)^{36} = 0.1251.$$

7.5 THE STRONG LAW OF LARGE NUMBERS

The strong law of large numbers is similar to the weak law in that it also deals with the convergence of the sample mean to the true mean. It is different, however, because it refers to another type of convergence.

The following is a general statement of the strong law of large numbers. A proof of the strong law, under the mildly restrictive assumption that the X_i have finite fourth moments is developed in the end-of-chapter problems.

The Strong Law of Large Numbers

Let X_1, X_2, \dots be a sequence of independent identically distributed random variables with mean μ . Then, the sequence of sample means $M_n = (X_1 + \dots + X_n)/n$ converges to μ , **with probability 1**, in the sense that

$$P\left(\lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1.$$

In order to interpret the strong law of large numbers, we need to go back to our original description of probabilistic models in terms of sample spaces. The contemplated experiment is infinitely long and generates a sequence of values, one value for each one of the random variables in the sequence X_1, X_2, \dots . Thus, it is best to think of the sample space as a set of infinite sequences $\omega = (x_1, x_2, \dots)$ of real numbers: any such sequence is a possible outcome of the experiment. Let us now consider the set A consisting of those sequences (x_1, x_2, \dots) whose long-term average is μ , i.e.,

$$(x_1, x_2, \dots) \in A \iff \lim_{n \rightarrow \infty} \frac{x_1 + \dots + x_n}{n} = \mu.$$

The strong law of large numbers states that all of the probability is concentrated on this particular subset of the sample space. Equivalently, the collection of outcomes that do not belong to A (infinite sequences whose long-term average is not μ) has probability zero.

The difference between the weak and the strong law is subtle and deserves close scrutiny. The weak law states that the probability $P(|M_n - \mu| \geq \epsilon)$ of a significant deviation of M_n from μ goes to zero as $n \rightarrow \infty$. Still, for any finite n , this probability can be positive and it is conceivable that once in a while, even if infrequently, M_n deviates significantly from μ . The weak law provides no conclusive information on the number of such deviations, but the strong law does. According to the strong law, and with probability 1, M_n converges to μ . This implies that for any given $\epsilon > 0$, the difference $|M_n - \mu|$ will exceed ϵ only a finite number of times.

Example 7.12. Probabilities and Frequencies. As in Example 7.3, consider an event A defined in terms of some probabilistic experiment. We consider a sequence of independent repetitions of the same experiment, and let M_n be the fraction of the first n trials in which A occurs. The strong law of large numbers asserts that M_n converges to $\mathbf{P}(A)$, with probability 1. In contrast, the weak law of large numbers asserts that M_n converges to $\mathbf{P}(A)$ in probability (cf. Example 7.3).

We have often talked intuitively about the probability of an event A as the frequency with which it occurs in an infinitely long sequence of independent trials. The strong law backs this intuition and establishes that the long-term frequency of occurrence of A is indeed equal to $\mathbf{P}(A)$, with essential certainty (the probability of this happening is 1).

Convergence with Probability 1

The convergence concept behind the strong law is different than the notion employed in the weak law. We provide here a definition and some discussion of this new convergence concept.

Convergence with Probability 1

Let Y_1, Y_2, \dots be a sequence of random variables (not necessarily independent) associated with the same probabilistic model. Let c be a real number. We say that Y_n converges to c with probability 1 (or almost surely) if

$$\mathbf{P} \left(\lim_{n \rightarrow \infty} Y_n = c \right) = 1.$$

Similar to our earlier discussion, a proper interpretation of this type of convergence involves a sample space consisting of infinite sequences: all of the probability is concentrated on those sequences that converge to c . This does not mean that other sequences are impossible, only that they are extremely unlikely, in the sense that their total probability is zero.

Example 7.13. Let X_1, X_2, \dots be a sequence of independent random variables that are uniformly distributed in $[0, 1]$, and let $Y_n = \min\{X_1, \dots, X_n\}$. We wish to show that Y_n converges to 0, with probability 1.

In any execution of the experiment, the sequence Y_n is nonincreasing, i.e., $Y_{n+1} \leq Y_n$ for all n . Since this sequence is bounded below by zero, it must have a limit, which we denote by Y . Let us fix some $\epsilon > 0$. If $Y \geq \epsilon$, then $X_i \geq \epsilon$ for all i , which implies that

$$\mathbf{P}(Y \geq \epsilon) \leq \mathbf{P}(X_1 \geq \epsilon, \dots, X_n \geq \epsilon) = (1 - \epsilon)^n.$$

Since this is true for all n , we must have

$$\mathbf{P}(Y \geq \epsilon) \leq \lim_{n \rightarrow \infty} (1 - \epsilon)^n = 0.$$

This shows that $\mathbf{P}(Y \geq \epsilon) = 0$, for any positive ϵ . We conclude that $\mathbf{P}(Y > 0) = 0$, which implies that $\mathbf{P}(Y = 0) = 1$. Since Y is the limit of Y_n , we see that Y_n converges to zero with probability 1.

Convergence with probability 1 implies convergence in probability (see the end-of-chapter problems), but the converse is not necessarily true. Our last example illustrates the difference between convergence in probability and convergence with probability 1.

Example 7.14. Consider a discrete-time arrival process. The set of times is partitioned into consecutive intervals of the form $I_k = \{2^k, 2^k + 1, \dots, 2^{k+1} - 1\}$. Note that the length of I_k is 2^k , which increases with k . During each interval I_k , there is exactly one arrival, and all times within an interval are equally likely. The arrival times within different intervals are assumed to be independent. Let us define $Y_n = 1$ if there is an arrival at time n , and $Y_n = 0$ if there is no arrival.

We have $\mathbf{P}(Y_n \neq 0) = 1/2^k$, if $n \in I_k$. Note that as n increases, it belongs to intervals I_k with increasingly large indices k . Consequently,

$$\lim_{n \rightarrow \infty} \mathbf{P}(Y_n \neq 0) = \lim_{k \rightarrow \infty} \frac{1}{2^k} = 0,$$

and we conclude that Y_n converges to 0 in probability. However, when we carry out the experiment, the total number of arrivals is infinite (one arrival during each interval I_k). Therefore, Y_n is unity for infinitely many values of n , the event $\{\lim_{n \rightarrow \infty} Y_n = 0\}$ has zero probability, and we do not have convergence with probability 1.

Intuitively, the following is happening. At any given time, there is only a small, and diminishing with n , probability of a substantial deviation from 0, which implies convergence in probability. On the other hand, given enough time, a substantial deviation from 0 is certain to occur and for this reason, we do not have convergence with probability 1.

7.6 SUMMARY AND DISCUSSION

In this chapter, we explored some fundamental aspects of probability theory that have major conceptual and practical implications. On the conceptual side, they put on a firm ground the interpretation of probability as relative frequency in a large number of independent trials. On the practical side, they allow the approximate calculation of probabilities in models that involve sums of independent random variables and that would be too hard to compute with other means.

We discussed three major laws that take the form of limit theorems.

- (a) The first one, the weak law of large numbers, indicates that the sample mean is very likely to be close to the true mean, as the sample size increases. It is based on the Chebyshev inequality, which is of independent interest and is representative of a large collection of useful inequalities that permeate probability theory.
- (b) The second one, the central limit theorem, is one of the most remarkable results of probability theory, and asserts that the sum of a large number of independent random variables is approximately normal. The central limit theorem finds many applications: it is one of the principal tools of statistical analysis and also justifies the use of normal random variables in modeling a wide array of situations.
- (c) The third one, the strong law of large numbers, makes a more emphatic connection of probabilities and relative frequencies, and is often an important tool in theoretical studies.

While developing the various limit theorems, we introduced a number of convergence concepts (convergence in probability and convergence with probability 1), which provide a precise language for discussing convergence in probabilistic models. The limit theorems and the convergence concepts discussed in this chapter underlie several more advanced topics in the study of probabilistic models and random processes.

 P R O B L E M S

SECTION 7.1. Some Useful Inequalities

Problem 1.* The Chernoff bound. The Chernoff bound is a powerful tool that relies on the transform associated with a random variable, and provides bounds on the probabilities of certain tail events.

(a) Show that the inequality

$$\mathbf{P}(X \geq a) \leq e^{-sa} M(s)$$

holds for every a and every $s \geq 0$, where $M(s) = \mathbf{E}[e^{sX}]$ is the transform associated with the random variable X .

(b) Show that the inequality

$$\mathbf{P}(X \leq a) \leq e^{-sa} M(s)$$

holds for every a and every $s \leq 0$.

(c) Show that the inequality

$$\mathbf{P}(X \geq a) \leq e^{-\phi(a)}$$

holds for every a , where

$$\phi(a) = \max_{s \geq 0} (sa - \ln M(s)).$$

(d) Show that if $a > \mathbf{E}[X]$, then $\phi(a) > 0$.

(e) Apply the result of part (c) to obtain a bound for $\mathbf{P}(X \geq a)$, for the case where X is a standard normal random variable and $a > 0$.

(f) Let X_1, X_2, \dots be independent random variables with the same distribution as X . Show that for any $a > \mathbf{E}[X]$, we have

$$\mathbf{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq a\right) \leq e^{-n\phi(a)},$$

so that the probability that the sample mean exceeds the mean by a certain amount decreases exponentially with n .

Solution. (a) Given some a and $s \geq 0$, consider the random variable Y_a defined by

$$Y_a = \begin{cases} 0, & \text{if } X < a, \\ e^{sa}, & \text{if } X \geq a. \end{cases}$$

It is seen that the relation

$$Y_a \leq e^{sX}$$

always holds and therefore,

$$\mathbf{E}[Y_a] \leq \mathbf{E}[e^{sX}] = M(s).$$

On the other hand,

$$\mathbf{E}[Y_a] = e^{sa} \mathbf{P}(Y_a = e^{sa}) = e^{sa} \mathbf{P}(X \geq a),$$

from which we obtain

$$\mathbf{P}(X \geq a) \leq e^{-sa} M(s).$$

(b) The argument is similar to the one for part (a). We define Y_a by

$$Y_a = \begin{cases} e^{sa}, & \text{if } X \leq a, \\ 0, & \text{if } X > a. \end{cases}$$

Since $s < 0$, the relation

$$Y_a \leq e^{sX}$$

always holds and therefore,

$$\mathbf{E}[Y_a] \leq \mathbf{E}[e^{sX}] = M(s).$$

On the other hand,

$$\mathbf{E}[Y_a] = e^{sa} \mathbf{P}(Y_a = e^{sa}) = e^{sa} \mathbf{P}(X \leq a),$$

from which we obtain

$$\mathbf{P}(X \leq a) \leq e^{-sa} M(s).$$

(c) Since the inequality from part (a) is valid for every $s \geq 0$, we obtain

$$\begin{aligned} \mathbf{P}(X \geq a) &\leq \min_{s \geq 0} (e^{-sa} M(s)) \\ &= \min_{s \geq 0} e^{-(sa - \ln M(s))} \\ &= e^{-\max_{s \geq 0} (sa - \ln M(s))} \\ &= e^{-\phi(a)}. \end{aligned}$$

(d) For $s = 0$, we have

$$sa - \ln M(s) = 0 - \ln 1 = 0,$$

where we have used the generic property $M(0) = 1$ of transforms. Furthermore,

$$\frac{d}{ds} (sa - \ln M(s)) \Big|_{s=0} = a - \frac{1}{M(s)} \cdot \frac{d}{ds} M(s) \Big|_{s=0} = a - 1 \cdot \mathbf{E}[X] > 0.$$

Since the function $sa - \ln M(s)$ is zero and has a positive derivative at $s = 0$, it must be positive when s is positive and small. It follows that the maximum $\phi(a)$ of the function $sa - \ln M(s)$ over all $s \geq 0$ is also positive.

(e) For a standard normal random variable X , we have $M(s) = e^{s^2/2}$. Therefore, $sa - \ln M(s) = sa - s^2/2$. To maximize this expression over all $s \geq 0$, we form the derivative, which is $a - s$, and set it to zero, resulting in $s = a$. Thus, $\phi(a) = a^2/2$, which leads to the bound

$$\mathbf{P}(X \geq a) \leq e^{-a^2/2}.$$

Note: In the case where $a < 0$, the maximizing value of s turns out to be $s = 0$, resulting in $\phi(a) = 0$ and in the uninteresting bound

$$\mathbf{P}(X \geq a) \leq 1.$$

(f) Let $Y = X_1 + \cdots + X_n$. Using the result of part (c), we have

$$\mathbf{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq a\right) = \mathbf{P}(Y \geq na) \leq e^{-\phi_Y(na)},$$

where

$$\phi_Y(na) = \max_{s \geq 0} (nsa - \ln M_Y(s)),$$

and

$$M_Y(s) = (M(s))^n$$

is the transform associated with Y . We have $\ln M_Y(s) = n \ln M(s)$, from which we obtain

$$\phi_Y(na) = n \cdot \max_{s \geq 0} (sa - \ln M(s)) = n\phi(a),$$

and

$$\mathbf{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq a\right) \leq e^{-n\phi(a)}.$$

Note that when $a > \mathbf{E}[X]$, part (d) asserts that $\phi(a) > 0$, so the probability of interest decreases exponentially with n .

Problem 2.* Jensen inequality. A twice differentiable real-valued function f of a single variable is called **convex** if its second derivative $(d^2 f / dx^2)(x)$ is nonnegative for all x in its domain of definition.

- (a) Show that the functions $f(x) = e^{\alpha x}$, $f(x) = -\ln x$, and $f(x) = x^4$ are all convex.
- (b) Show that if f is twice differentiable and convex, then the first order Taylor approximation of f is an underestimate of the function, that is,

$$f(a) + (x - a) \frac{df}{dx}(a) \leq f(x),$$

for every a and x .

(c) Show that if f has the property in part (b), and if X is a random variable, then

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)].$$

Solution. (a) We have

$$\frac{d^2}{dx^2} e^{ax} = a^2 e^{ax} > 0, \quad \frac{d^2}{dx^2} (-\ln x) = \frac{1}{x^2} > 0, \quad \frac{d^2}{dx^2} x^4 = 4 \cdot 3 \cdot x^2 \geq 0.$$

(b) Since the second derivative of f is nonnegative, its first derivative must be nondecreasing. Using the fundamental theorem of calculus, we obtain

$$f(x) = f(a) + \int_a^x \frac{df}{dt}(t) dt \geq f(a) + \int_a^x \frac{df}{dt}(a) dt = f(a) + (x-a) \frac{df}{dx}(a).$$

(c) Since the inequality from part (b) is assumed valid for every possible value x of the random variable X , we obtain

$$f(a) + (X-a) \frac{df}{dx}(a) \leq f(X).$$

We now choose $a = \mathbf{E}[X]$ and take expectations, to obtain

$$f(\mathbf{E}[X]) + (\mathbf{E}[X] - \mathbf{E}[X]) \frac{df}{dx}(\mathbf{E}[X]) \leq \mathbf{E}[f(X)],$$

or

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)].$$

SECTION 7.2. The Weak Law of Large Numbers

Problem 3. In order to estimate f , the true fraction of smokers in a large population, Alvin selects n people at random. His estimator M_n is obtained by dividing S_n , the number of smokers in his sample, by n , i.e., $M_n = S_n/n$. Alvin chooses the sample size n to be the smallest possible number for which the Chebyshev inequality yields a guarantee that

$$\mathbf{P}(|M_n - f| \geq \epsilon) \leq \delta,$$

where ϵ and δ are some prespecified tolerances. Determine how the value of n recommended by the Chebyshev inequality changes in the following cases.

- (a) The value of ϵ is reduced to half its original value.
- (b) The probability δ is reduced to $\delta/2$.

SECTION 7.3. Convergence in Probability

Problem 4.* Consider two sequences of random variables X_1, X_2, \dots and Y_1, Y_2, \dots , associated with the same experiment. Suppose that X_n converges to a and Y_n converges to b , in probability. Show that $X_n + Y_n$ converges to $a + b$, in probability.

Solution. We need to show that $\mathbf{P}(|X_n + Y_n - a - b| \geq \epsilon)$ converges to zero, for any $\epsilon > 0$. To bound this probability, we note that for $|X_n + Y_n - a - b|$ to be as large as ϵ , we need either $|X_n - a|$ or $|Y_n - b|$ (or both) to exceed $\epsilon/2$. Therefore, in terms of events, we have

$$\{|X_n + Y_n - a - b| \geq \epsilon\} \subset \{|X_n - a| \geq \epsilon/2\} \cup \{|Y_n - b| \geq \epsilon/2\}.$$

This implies that

$$\mathbf{P}(|X_n + Y_n - a - b| \geq \epsilon) \leq \mathbf{P}(|X_n - a| \geq \epsilon/2) + \mathbf{P}(|Y_n - b| \geq \epsilon/2),$$

and

$$\lim_{n \rightarrow \infty} \mathbf{P}(|X_n + Y_n - a - b| \geq \epsilon) \leq \lim_{n \rightarrow \infty} \mathbf{P}(|X_n - a| \geq \epsilon/2) + \lim_{n \rightarrow \infty} \mathbf{P}(|Y_n - b| \geq \epsilon/2) = 0,$$

where the last equality follows because of the assumption that the sequences X_n and Y_n converge, in probability, to a and b , respectively.

Problem 5.* A sequence X_n of random variables is said to converge to a number c in the mean square, if

$$\lim_{n \rightarrow \infty} \mathbf{E}[(X_n - c)^2] = 0.$$

- (a) Show that convergence in the mean square implies convergence in probability.
- (b) Give an example that shows that convergence in probability does not imply convergence in the mean square.

Solution. (a) Suppose that X_n converges to c in the mean square. Using the Markov inequality, we have

$$\mathbf{P}(|X_n - c| \geq \epsilon) = \mathbf{P}(|X_n - c|^2 \geq \epsilon^2) \leq \frac{\mathbf{E}[(X_n - c)^2]}{\epsilon^2}.$$

Taking the limit as $n \rightarrow \infty$, we obtain

$$\lim_{n \rightarrow \infty} \mathbf{P}(|X_n - c| \geq \epsilon) = 0,$$

which establishes convergence in probability.

- (b) In Example 7.7, we have convergence in probability to 0 but $\mathbf{E}[Y_n^2] = n^3$, which diverges to infinity.

SECTION 7.4. The Central Limit Theorem

Problem 6. Before starting to play the roulette in a casino, you want to look for biases that you can exploit. You therefore watch 100 rounds that result in a number between 1 and 36, and count the number of rounds for which the result is odd. If the count exceeds 55, you decide that the roulette is not fair. Assuming that the roulette is fair, find an approximation for the probability that you will make the wrong decision.

Problem 7. During each day, the probability that your computer's operating system crashes at least once is 5%, independent of every other day. You are interested in the probability of at least 45 crash-free days out of the next 50 days.

- (a) Find the probability of interest by using the normal approximation to the binomial.
- (b) Repeat part (a), this time using the Poisson approximation to the binomial.

Problem 8. A factory produces X_n gadgets on day n , where the X_n are independent and identically distributed random variables, with mean 5 and variance 9.

- (a) Find an approximation to the probability that the total number of gadgets produced in 100 days is less than 440.
- (b) Find (approximately) the largest value of n such that

$$\mathbf{P}(X_1 + \cdots + X_n \geq 200 + 5n) \leq 0.05.$$

- (c) Let N be the first day on which the total number of gadgets produced exceeds 1000. Calculate an approximation to the probability that $N \geq 220$.

Problem 9. Let $X_1, Y_1, X_2, Y_2, \dots$ be independent random variables, uniformly distributed in the unit interval $[0, 1]$. Let

$$W = \frac{(X_1 + \cdots + X_{16}) - (Y_1 + \cdots + Y_{16})}{16}.$$

Find a numerical approximation to the quantity

$$\mathbf{P}(|W - \mathbf{E}[W]| < 0.001).$$

Problem 10.* Proof of the central limit theorem. Let X_1, X_2, \dots be a sequence of independent identically distributed zero-mean random variables with common variance σ^2 , and associated transform $M_X(s)$. We assume that $M_X(s)$ is finite when $-d < s < d$, where d is some positive number. Let

$$Z_n = \frac{X_1 + \cdots + X_n}{\sigma\sqrt{n}}.$$

- (a) Show that the transform associated with Z_n satisfies

$$M_{Z_n}(s) = \left(M_X \left(\frac{s}{\sigma\sqrt{n}} \right) \right)^n.$$

- (b) Suppose that the transform $M_X(s)$ has a second order Taylor series expansion around $s = 0$, of the form

$$M_X(s) = a + bs + cs^2 + o(s^2),$$

where $o(s^2)$ is a function that satisfies $\lim_{s \rightarrow 0} o(s^2)/s^2 = 0$. Find a , b , and c in terms of σ^2 .

- (c) Combine the results of parts (a) and (b) to show that the transform $M_{Z_n}(s)$ converges to the transform associated with a standard normal random variable, that is,

$$\lim_{n \rightarrow \infty} M_{Z_n}(s) = e^{s^2/2}, \quad \text{for all } s.$$

Note: The central limit theorem follows from the result of part (c), together with the fact (whose proof lies beyond the scope of this text) that if the transforms $M_{Z_n}(s)$ converge to the transform $M_Z(s)$ of a random variable Z whose CDF is continuous, then the CDFs F_{Z_n} converge to the CDF of Z . In our case, this implies that the CDF of Z_n converges to the CDF of a standard normal.

Solution. (a) We have, using the independence of the X_i ,

$$\begin{aligned} M_{Z_n}(s) &= \mathbb{E}[e^{sZ_n}] \\ &= \mathbb{E}\left[\exp\left\{\frac{s}{\sigma\sqrt{n}} \sum_{i=1}^n X_i\right\}\right] \\ &= \prod_{i=1}^n \mathbb{E}[e^{sX_i/(\sigma\sqrt{n})}] \\ &= \left(M_X\left(\frac{s}{\sigma\sqrt{n}}\right)\right)^n. \end{aligned}$$

(b) Using the moment generating properties of the transform, we have

$$a = M_X(0) = 1, \quad b = \frac{d}{ds} M_X(s) \Big|_{s=0} = \mathbb{E}[X] = 0,$$

and

$$c = \frac{1}{2} \cdot \frac{d^2}{ds^2} M_X(s) \Big|_{s=0} = \frac{\mathbb{E}[X^2]}{2} = \frac{\sigma^2}{2}.$$

(c) We combine the results of parts (a) and (b). We have

$$M_{Z_n}(s) = \left(M_X\left(\frac{s}{\sigma\sqrt{n}}\right)\right)^n = \left(a + \frac{bs}{\sigma\sqrt{n}} + \frac{cs^2}{\sigma^2 n} + o\left(\frac{s^2}{\sigma^2 n}\right)\right)^n,$$

and using the formulas for a , b , and c from part (b), it follows that

$$M_{Z_n}(s) = \left(1 + \frac{s^2}{2n} + o\left(\frac{s^2}{\sigma^2 n}\right)\right)^n.$$

We now take the limit as $n \rightarrow \infty$, and use the identity

$$\lim_{n \rightarrow \infty} \left(1 + \frac{c}{n}\right)^n = e^c,$$

to obtain

$$\lim_{n \rightarrow \infty} M_{Z_n}(s) = e^{s^2/2}.$$

SECTION 7.5. The Strong Law of Large Numbers

Problem 11.* Consider two sequences of random variables X_1, X_2, \dots and Y_1, Y_2, \dots Suppose that X_n converges to a and Y_n converges to b , with probability 1. Show that $X_n + Y_n$ converges to $a + b$, with probability 1. Also, assuming that the random variables Y_n cannot be equal to zero, show that X_n/Y_n converges to a/b , with probability 1.

Solution. Let A (respectively, B) be the event that the sequence of values of the random variables X_n (respectively, Y_n) does not converge to a (respectively, b). Let C be the event that the sequence of values of $X_n + Y_n$ does not converge to $a + b$ and notice that $C \subset A \cup B$.

Since X_n and Y_n converge to a and b , respectively, with probability 1, we have $P(A) = 0$ and $P(B) = 0$. Hence,

$$P(C) \leq P(A \cup B) \leq P(A) + P(B) = 0.$$

Therefore, $P(C^c) = 1$, or equivalently, $X_n + Y_n$ converges to $a + b$ with probability 1. For the convergence of X_n/Y_n , the argument is similar.

Problem 12.* Let X_1, X_2, \dots be a sequence of independent identically distributed random variables. Let Y_1, Y_2, \dots be another sequence of independent identically distributed random variables. We assume that the X_i and Y_i have finite mean, and that $Y_1 + \dots + Y_n$ cannot be equal to zero. Does the sequence

$$Z_n = \frac{X_1 + \dots + X_n}{Y_1 + \dots + Y_n}$$

converge with probability 1, and if so, what is the limit?

Solution. We have

$$Z_n = \frac{(X_1 + \dots + X_n)/n}{(Y_1 + \dots + Y_n)/n}.$$

By the strong law of large numbers, the numerator and denominator converge with probability 1 to $E[X]$ and $E[Y]$, respectively. It follows that Z_n converges to $E[X]/E[Y]$, with probability 1 (cf. the preceding problem).

Problem 13.* Suppose that a sequence Y_1, Y_2, \dots of random variables converges to a real number c , with probability 1. Show that the sequence also converges to c in probability.

Solution. Let C be the event that the sequence of values of the random variables Y_n converges to c . By assumption, we have $P(C) = 1$. Fix some $\epsilon > 0$, and let A_k be the event that $|Y_n - c| < \epsilon$ for every $n \geq k$. If the sequence of values of the random variables Y_n converges to c , then there must exist some k such that for every $n \geq k$, this sequence of values is within less than ϵ from c . Therefore, every element of C belongs to A_k for some k , or

$$C \subset \bigcup_{k=1}^{\infty} A_k.$$

Note also that the sequence of events A_k is monotonically increasing, in the sense that $A_k \subset A_{k+1}$ for all k . Finally, note that the event A_k is a subset of the event $\{|Y_k - c| < \epsilon\}$. Therefore,

$$\lim_{k \rightarrow \infty} \mathbf{P}(|Y_k - c| < \epsilon) \geq \lim_{k \rightarrow \infty} \mathbf{P}(A_k) = \mathbf{P}(\cup_{k=1}^{\infty} A_k) \geq \mathbf{P}(C) = 1,$$

where the first equality uses the continuity property of probabilities (Problem 10 in Chapter 1). It follows that

$$\lim_{k \rightarrow \infty} \mathbf{P}(|Y_k - c| \geq \epsilon) = 0,$$

which establishes convergence in probability.

Problem 14.* Consider a sequence Y_n of nonnegative random variables and suppose that

$$\mathbf{E} \left[\sum_{n=1}^{\infty} Y_n \right] < \infty.$$

Show that Y_n converges to 0, with probability 1.

Note: This result provides a commonly used method for establishing convergence with probability 1. To evaluate the expectation of $\sum_{n=1}^{\infty} Y_n$, one typically uses the formula

$$\mathbf{E} \left[\sum_{n=1}^{\infty} Y_n \right] = \sum_{n=1}^{\infty} \mathbf{E}[Y_n].$$

The fact that the expectation and the infinite summation can be interchanged, for the case of nonnegative random variables, is known as the **monotone convergence theorem**, a fundamental result of probability theory, whose proof lies beyond the scope of this text.

Solution. We note that the infinite sum $\sum_{n=1}^{\infty} Y_n$ must be finite, with probability 1. Indeed, if it had a positive probability of being infinite, then its expectation would also be infinite. But if the sum of the values of the random variables Y_n is finite, the sequence of these values must converge to zero. Since the probability of this event is equal to 1, it follows that the sequence Y_n converges to zero, with probability 1.

Problem 15.* Consider a sequence of Bernoulli random variables X_n , and let $p_n = \mathbf{P}(X_n = 1)$ be the probability of success in the n th trial. Assuming that $\sum_{n=1}^{\infty} p_n < \infty$, show that the number of successes is finite, with probability 1. [Compare with Exercise 1.43(b).]

Solution. Using the monotone convergence theorem (see above note), we have

$$\mathbf{E} \left[\sum_{n=1}^{\infty} X_n \right] = \sum_{n=1}^{\infty} \mathbf{E}[X_n] = \sum_{n=1}^{\infty} p_n < \infty.$$

This implies that

$$\sum_{n=1}^{\infty} X_n < \infty,$$

with probability 1. We then note that the event $\{\sum_{n=1}^{\infty} X_n < \infty\}$ is the same as the event that there is a finite number of successes.

Problem 16.* The strong law of large numbers. Let X_1, X_2, \dots be a sequence of independent identically distributed random variables and assume that $\mathbf{E}[X_i^4] < \infty$. Prove the strong law of large numbers.

Solution. We note that the assumption $\mathbf{E}[X_i^4] < \infty$ implies that the expected value of the X_i is finite. Indeed, using the inequality $|x| \leq 1 + x^4$, we have

$$\mathbf{E}[|X_i|] \leq 1 + \mathbf{E}[X_i^4] < \infty.$$

Let us assume first that $\mathbf{E}[X_i] = 0$. We will show that

$$\mathbf{E}\left[\sum_{n=1}^{\infty} \frac{(X_1 + \dots + X_n)^4}{n^4}\right] < \infty.$$

We have

$$\mathbf{E}\left[\frac{(X_1 + \dots + X_n)^4}{n^4}\right] = \frac{1}{n^4} \sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{i_3=1}^n \sum_{i_4=1}^n \mathbf{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}].$$

Let us consider the various terms in this sum. If one of the indices is different from all of the other indices, the corresponding term is equal to zero. For example, if i_1 is different from i_2, i_3 , or i_4 , the assumption $\mathbf{E}[X_i] = 0$ yields

$$\mathbf{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}] = \mathbf{E}[X_{i_1}] \mathbf{E}[X_{i_2} X_{i_3} X_{i_4}] = 0.$$

Therefore, the nonzero terms in the above sum are either of the form $\mathbf{E}[X_i^4]$ (there are n such terms), or of the form $\mathbf{E}[X_i^2 X_j^2]$, with $i \neq j$. Let us count how many terms there are of this form. Such terms are obtained in three different ways: by setting $i_1 = i_2 \neq i_3 = i_4$, or by setting $i_1 = i_3 \neq i_2 = i_4$, or by setting $i_1 = i_4 \neq i_2 = i_3$. For each one of these three ways, we have n choices for the first pair of indices, and $n - 1$ choices for the second pair. We conclude that there are $3n(n - 1)$ terms of this type. Thus,

$$\mathbf{E}\left[\frac{(X_1 + \dots + X_n)^4}{n^4}\right] = \frac{n\mathbf{E}[X_1^4] + 3n(n - 1)\mathbf{E}[X_1^2 X_2^2]}{n^4}.$$

Using the inequality $xy \leq (x^2 + y^2)/2$, we obtain $\mathbf{E}[X_1^2 X_2^2] \leq \mathbf{E}[X_1^4]$, and

$$\mathbf{E}\left[\frac{(X_1 + \dots + X_n)^4}{n^4}\right] \leq \frac{(n + 3n(n - 1))\mathbf{E}[X_1^4]}{n^4} \leq \frac{3n^2\mathbf{E}[X_1^4]}{n^4} = \frac{3\mathbf{E}[X_1^4]}{n^2}.$$

It follows that

$$\mathbf{E}\left[\sum_{n=1}^{\infty} \frac{(X_1 + \dots + X_n)^4}{n^4}\right] = \sum_{n=1}^{\infty} \frac{1}{n^4} \overline{\mathbf{E}[(X_1 + \dots + X_n)^4]} \leq \sum_{n=1}^{\infty} \frac{3}{n^2} \mathbf{E}[X_1^4] < \infty,$$

where the last step uses the well known property $\sum_{n=1}^{\infty} n^{-2} < \infty$. This implies that $(X_1 + \dots + X_n)^4/n^4$ converges to zero with probability 1 (cf. Problem 14), and therefore, $(X_1 + \dots + X_n)/n$ also converges to zero with probability 1, which is the strong law of large numbers.

For the more general case where the mean of the random variables X_i is nonzero, the preceding argument establishes that $(X_1 + \dots + X_n - n\mathbf{E}[X_1])/n$ converges to zero, which is the same as $(X_1 + \dots + X_n)/n$ converging to $\mathbf{E}[X_1]$, with probability 1.

Problem 17.* The strong law of large numbers for Markov chains. Consider a finite-state Markov chain in which all states belong to a single recurrent class which is aperiodic. For a fixed state s , let Y_k be the time of the k th visit to state s . Let also V_n be the number of visits to state s during the first n transitions.

- (a) Show that Y_k/k converges with probability 1 to the mean recurrence time t_s^* of state s .
- (b) Show that V_n/n converges with probability 1 to $1/t_s^*$.
- (c) Can you relate the limit of V_n/n to the steady-state probability of state s ?

Solution. (a) Let us fix an initial state i , not necessarily the same as s . Thus, the random variables $Y_{k+1} - Y_k$, for $k \geq 1$, correspond to the time between successive visits to state s . Because of the Markov property (the past is independent of the future, given the present), the process “starts fresh” at each revisit to state s and, therefore, the random variables $Y_{k+1} - Y_k$ are independent and identically distributed, with mean equal to the mean recurrence time t_s^* . Using the strong law of large numbers, we obtain

$$\lim_{k \rightarrow \infty} \frac{Y_k}{k} = \lim_{k \rightarrow \infty} \frac{Y_1}{k} + \lim_{k \rightarrow \infty} \frac{(Y_2 - Y_1) + \dots + (Y_k - Y_{k-1})}{k} = 0 + t_s^*,$$

with probability 1.

(b) Let us fix an element of the sample space (a trajectory of the Markov chain). Let y_k and v_n be the values of the random variables Y_k and V_n , respectively. Furthermore, let us assume that the sequence y_k/k converges to t_s^* ; according to the result of part (a), the set of trajectories with this property has probability 1. Let us consider some n between the time of the k th visit to state s and the time just before the next visit to that state:

$$y_k \leq n < y_{k+1}.$$

For every n in this range, we have $v_n = k$, and also

$$\frac{1}{y_{k+1}} < \frac{1}{n} \leq \frac{1}{y_k},$$

from which we obtain

$$\frac{k}{y_{k+1}} \leq \frac{v_n}{n} \leq \frac{k}{y_k}.$$

Note that

$$\lim_{k \rightarrow \infty} \frac{k}{y_{k+1}} = \lim_{k \rightarrow \infty} \frac{k+1}{y_{k+1}} \cdot \lim_{k \rightarrow \infty} \frac{k}{k+1} = \lim_{k \rightarrow \infty} \frac{k}{y_k} = \frac{1}{t_s^*}.$$

If we now let n go to infinity, the corresponding values of k , chosen to satisfy $y_k \leq n < y_{k+1}$ also go to infinity. Therefore, the sequence v_n/n is between two sequences both

of which converge to $1/t_s^*$, which implies that the sequence v_n/n converges to $1/t_s^*$ as well. Since this happens for every trajectory in a set of trajectories that has probability equal to 1, we conclude that V_n/n converges to $1/t_s^*$, with probability 1.

(c) It turns out that $1/t_s^* = \pi_s$, as established in Problem 32 of Chapter 6. This implies the intuitive result that V_n/n converges to π_s , with probability 1.

It is tempting to try to establish the convergence of V_n/n to π_s by combining the facts that V_n/n converges [part (b)] together with the fact that $\mathbf{E}[V_n]/n$ converges to π_s (cf. the long-term expected frequency interpretation of steady-state probabilities in Section 6.3). However, this line of reasoning is not valid. This is because it is generally possible for a sequence Y_n of random variables to converge with probability 1 to a constant, while the expected values converge to a different constant. An example is the following. Let X be uniformly distributed in the unit interval $[0, 1]$. let

$$Y_n = \begin{cases} 0, & \text{if } X \geq 1/n, \\ n, & \text{if } X < 1/n. \end{cases}$$

As long as X is nonzero (which happens with probability 1), the sequence Y_n converges to zero. On the other hand, it can be seen that

$$\mathbf{E}[Y_n] = \mathbf{P}(X < 1/n) \mathbf{E}[Y_n \mid X < 1/n] = \frac{1}{n} \cdot \frac{n}{2} = \frac{1}{2}, \quad \text{for all } n.$$

INDEX

A

Absorbing state, 320, 337
Absorption probabilities, 337, 370
Accessible state, 321
Additivity axiom, 9
Almost surely, 396
Aperiodic recurrent class, 325
Axioms, 9, 17

B

Balance equations, 328, 349, 373
dependence, 366
local, 334, 351, 367
uniqueness of solutions, 327, 362
Bayes, 17
Bayes' rule, 28, 31
continuous, 175
Bernoulli, Daniel, 17
problem of joint lives, 127
Bernoulli-Laplace model of diffusion, 360
Bernoulli, Jacob, 17
Bernoulli process, 273
alternative description, 279
definition, 273
fresh-start at random time, 275
fresh-start property, 275
independence, 274
interarrival times, 278
memorylessness, 274
merging, 281
number of successes, 274
Poisson approximation, 282
random incidence in, 302
relation to uniform, 303
splitting, 281
time until first success, 274
time until k th success, 279
Bernoulli r.v., 75, 117
mean, 88
sum of, 96, 263
transform, 218
variance, 88
Bernoulli trial, 41
Bertrand's paradox, 16
Binomial
coefficient, 42, 47, 64

formula, 43, 48
probabilities, 42, 44
Binomial r.v., 77, 117
de Moivre-Laplace approximation, 392
mean, 96
Poisson approximation, 79, 121
recursive calculation of PMF, 120
transform, 218
variance, 114
Birth-death process, 333, 351
Birthday problem, 66
Bivariate normal, 247
conditional expectation, 250, 269
conditional PDF, 249
conditional variance, 250
PDF, 251
properties, 251
Bonferroni's inequality, 53
Borel-Cantelli lemma, 65
Buffon's needle, 166, 200

C

Cantor's diagonalization, 52
Cardano, 17
Cauchy r.v., 204
CDF, 148
joint, 174
properties, 149
Central limit theorem, 158, 380, 388
history, 17
proof, 404
Chain, *see* Markov chain
Chapman-Kolmogorov equation, 319, 328
Chebyshev, 17
Chebyshev inequality, 382
Chernoff bound, 399
Classification of states, 321
Collectively exhaustive, 7
Combination, 45, 47
Competing exponentials, 205
Conditional
expectation, *see* expectation, conditional
independence, 111
PDF, *see* conditional PDF
PMF, *see* conditional PMF
probability, 18
variance, *see* variance, conditional

- Conditioning
 continuous r.v. on an event, 158
 continuous r.v. on another, 168, 170
 discrete r.v. on an event, 98
 discrete r.v. on another, 100
- Continuity property of probabilities, 54
- Continuous random variable, *see* random variable
- Convergence
 in probability, 386, 397
 in the mean square, 403
 of a deterministic sequence, 386
 with probability one, 396
- Convolution, 221
 graphical calculation, 224
- Correlation coefficient, 237, 264
- Covariance, 236
- Countably infinite, 3, 82
- Counting, 44
 principle, 45
- Cumulative distribution function, *see* CDF
- D**
- De Méré, 61, 66
- De Moivre-Laplace approximation, 17, 392
- De Morgan's laws, 5, 52
- Decomposition, 323
- Density, *see* PDF
- Derived distributions, 179
 of function of two r.v.'s, 186
- Detection, 22, 32, 157, 177
- Diffusion
 Bernoulli-Laplace model, 360
 Ehrenfest model, 360
- Discrete probability law, 11
- Discrete uniform probability law, 11
- Discrete random variable, *see* random variable
- Discrete uniform random variable, *see* uniform
- Doubly stochastic, 364
- E**
- Ehrenfest model of diffusion, 360
- Empirical frequency, 385
- Entropy, 135
- Erlang process, 311
- Erlang r.v., 293
- Estimation
 error, 244-267
 least squares, 240
 linear least squares, 245
 w. several measurements, 245, 265
- Estimator, 242
- linear, 245
- Event, 6, 7
 independent, 34
- Expectation
 as center of gravity, 83
 conditional, 104, 160, 225
 linearity, 87, 94, 96, 98, 130, 145, 168, 179
 of continuous r.v., 144, 192
 of discrete r.v., 81, 82
 of function of r.v., *see* expected value rule
 well-defined, 82, 144
- Expected time to absorption, 341
- Expected value, *see* expectation
- Expected value rule
 for continuous r.v.'s, 145, 160, 168, 172, 179, 193
 for discrete r.v.'s, 84, 94, 96, 97, 130
- Experiment, 6
- Exponential random variable, 146, 190
 CDF, 151
 in Markov chain, 344
 in Poisson process, 288
 linear function of, 183
 maximum of, 307
 mean, 147
 memorylessness, 159
 minimum of, 205, 296
 moments, 259
 relation to geometric, 152
 sum of geometric number, 235, 309
 transform, 211
 two-sided, 190, 196
 variance, 147
- F**
- False-positive puzzle, 33
- Fermat, 17, 61
- First passage time, 342, 372
- Frequency interpretation of probabilities, 2, 9, 17, 385, 396
- Function of a r.v., 74, 80, *see also* derived distributions
 mean, *see* expected value rule
- Function of multiple r.v.'s
 continuous case, 179
 discrete case, 94
- G**
- Gambler's ruin, 61, 338
- Gauss, 17
- Gaussian r.v., *see* normal r.v.
- Geometric r.v., 77, 117, 133
 CDF, 151

- mean, 106
- relation to exponential, 152
- sum of a geometric number, 235, 303
- transform, 216
- variance, 106
- Grade of service, 43
- H**
- Hat problem, 97, 240
- History of probability, 17
- Hypergeometric probabilities, 68
- I**
- Inclusion-exclusion formula, 54, 126
- Independence
 - conditional, 36, 111
 - of continuous r.v.'s, 173, 179
 - of discrete r.v.'s, 111
 - of events, 34
 - of functions of r.v.'s, 112, 134, 174
 - of r.v. from an event, 110
 - of several events, 38
 - pairwise, 38
- Independent trials, 41
- Indistinguishable objects, 68
- Inequality
 - Chebyshev, 382
 - Chernov bound, 399
 - Jensen, 401
 - Markov, 381
 - Schwarz, 264
- Inference, 31, 175
- Inversion property, 215
- Iterated expectations, law of, 227
- J**
- Jensen inequality, 401
- Joint PDF, 164
- Joint PMF, 92
- Jointly normal, *see* bivariate normal
- K**
- Kelly strategy, 261
- King's sibling, 59
- L**
- Lagrange, 17
- Laplace, 17, 166, 192, 200
- Laplace r.v., 190
- Laplace transform, 211
- Laplace's rule of succession, 63
- Law of iterated expectations, 227
- Law of large numbers
 - weak, *see* weak law of large numbers
 - strong, *see* strong law of large numbers
- Law of total variance, 229
- Least squares estimate, 241
- Least squares estimation, 17, 240
 - with several measurements, 245
- Limit, *see* convergence
- Linear least squares estimation, 245
 - with several measurements, 265
- Legendre, 17
- Leibnitz, 17
- Linear function of r.v.'s
 - mean, 87, *see also* expectation, linearity
 - normal, 154
 - PDF, 182
 - transform, 212
 - variance, 87, 145
- Local balance equations, 334, 351, 367
- Lyapunov, 17
- M**
- Marginal
 - PDF, 165
 - PMF, 93
- Markov, 17
 - Markov chain in continuous time, 344
 - alternative description, 347
 - assumptions, 344
 - balance equations, 349
 - birth-death, 351
 - definition, 344
 - discrete-time approximation, 346
 - embedded chain, 344
 - local balance equations, 351
 - normalization equation, 350
 - steady-state, 349
 - steady-state convergence theorem, 350
 - transition rate, 345
 - Markov chain in discrete time, 314
 - absorption probabilities, 337, 370
 - balance equations, 328, 362, 373
 - birth-death, 333
 - Chapman-Kolmogorov equation, 319, 328
 - classification, 321
 - decomposition, 323
 - definition, 314
 - expected state frequencies, 332, 363
 - expected time to absorption, 341
 - expected transition frequencies, 332
 - first passage time, 342, 372
 - local balance equations, 334, 367

- multiple recurrent classes, 371
- n*-step transition probability, 319
- normalization equation, 328
- periodicity, 326, 357
- probability of a path, 318
- recurrence time, 342, 373
- sampled, 367
- stationary distribution, 327
- steady-state, 326
- steady-state convergence theorem, 327, 360
- strong law of large numbers, 409
- transition graph, 315
- transition matrix, 315, 320
- transition probability, 314
- Markov inequality, 381
- Markov model, 315
- Markov property, 314, 346
- Matchbox problem, 120
- Matrix
 - doubly stochastic, 364
 - transition, 315, 320
- Mean, *see* expectation
- Mean first passage time, 342, 372
- Mean recurrence time, 342, 373
- Merging
 - Bernoulli processes, 281
 - Poisson processes, 294
- Minimum of r.v.'s, 124
- Mixed r.v., 198
- Mixture of distributions, 216
- Moment, 83, 145
 - calculation using transforms, 213
- Moment generating function, 213
 - see also* transform
- Monotone convergence theorem, 407
- Monotonic function of a r.v., 184
- Monty Hall problem, 27
- Multinomial coefficient, 49
- Multiplication rule
 - for PDFs, 178, 200
 - for PMFs, 104, 131
 - for probabilities, 24
- Multivariate normal, 254
- Multivariate transform, 219, 249
- Mutual information, 136
- Mutually exclusive, 7
- N**
- Negative binomial r.v., 302
- Nonnegativity axiom, 9
- Normal approximation, 390
- Normal r.v.
 - CDF, 156
- central limit theorem, 388
- jointly, 248
- linear function of, 154, 183
- mean, 153
- multivariate, 254
- normalization property, 153, 195
- standard, 154
- sum of, 218
- table, 155
- transform, 212, 249
- uncorrelated, 248
- variance, 153
 - see also* bivariate normal
- Normalization axiom, 9
- Normalization equation, in Markov chains, 328
- O**
- Ordered pair, 5
- Outcome, 6
 - mutually exclusive, 7
- P**
- Pairwise independence, 38
- Paradox
 - Bertrand's, 16
 - of induction, 57
 - random incidence, 298
 - St. Petersburg, 123
 - two-envelopes, 107
- Parallel connection, 40
- Partition, 48
- Pascal, 17, 61, 66
- Pascal r.v., 280, 302
- Pascal triangle, 64
- PDF, 140
 - conditional, 159, 168, 170
 - joint, 164
 - marginal, 165
 - of function of r.v.'s, *see* derived distributions
 - of linear function of r.v.'s, 182
 - of monotonic function of r.v.'s, 184
- Periodic class, 325, 357
- Permutation, 45, 68
 - k*-permutation, 46
- Perron-Frobenius theorem, 328
- Piecewise constant PDF, 142
 - mean, 161
 - variance, 161
- PMF, 74
 - calculation of, 75
 - conditional, 98, 100
 - summary, 103

- joint, 92
- marginal, 93
- Poisson, 17
- Poisson process, 285
 - alternative description, 292
 - arrival rate, 286
 - definition, 286
 - fresh-start at random time, 291
 - fresh-start property, 290
 - independence, 286, 290
 - intensity, 286
 - interarrival times, 291
 - memorylessness, 290
 - merging, 294
 - number of arrivals, 287, 310
 - random incidence, 297
 - small interval probabilities, 286
 - splitting, 294, 311
 - time-homogeneity, 286
 - time until first arrival, 288
 - time until k th arrival, 292
- Poisson random variable, 78, 117, 288
 - approximation of binomial, 79, 121
 - mean, 90
 - splitting, 131
 - sum of, 218, 263
 - transform, 211
 - variance, 123
- Polling, 385, 391
- Prisoner's dilemma, 57
- Probabilistic model, 6
- Probability
 - conditional, 18
 - history, 17
 - steady-state, *see* steady-state probability
 - subjective, 3
- Probability density function, *see* PDF
- Probability mass function, *see* PMF
- Probability law, 6, 8
 - of random variable, 149
 - properties, 14
- Problem of joint lives, 127
- Problem of points, 61
- Process
 - arrival, 272
 - Bernoulli, 273
 - birth-death, 333, 351
 - Erlang, 311
 - Markov, 273
 - Poisson, 273, 285
- Q**
- Queueing, 310, 335, 348, 365
- Quiz problem, 91, 124
- R**
- Random incidence
 - in Bernoulli process, 302
 - in Erlang process, 311
 - in Poisson process, 297
 - in non-Poisson process, 299
- Random variable
 - continuous, 140
 - definition, 73
 - discrete, 74
 - function of, 74, 80, 94
 - independent, 111, 173, 179
 - mixed, 198
- Random walk, 334
- Rayleigh r.v., 206
- Recurrence time, 342, 373
- Recurrent
 - class, 323
 - aperiodic, 325
 - multiple, 371
 - periodic, 325
 - state, 322, 356
 - existence, 354
- Relative frequency, 9, 17, *see also* frequency interpretation
- Reliability, 40, 60
- Reversibility, 367
- S**
- Sample mean, 115, 380
 - mean, 115
 - variance, 115
- Sample space, 6
 - uncountable, 7
- Schwarz inequality, 264
- Sequential model, 8, 22
- Sequential method, 51
- Series connection, 40
- Set, 3
 - complement, 4
 - countably infinite, 3
 - disjoint, 4
 - element of, 3
 - empty, 3
 - intersection, 4
 - partition, 4
 - uncountable, 4
 - union, 4
 - universal, 4
- Shannon, 135
- Simulation, 115, 194, 200, 201

- Splitting
 Bernoulli process, 281
 Poisson r.v., 131
 Poisson process, 294, 311
 St. Petersburg paradox, 123
 Standard deviation, 83
 Standard normal random variable, 154
 Standard normal table, 155
 use for calculating normal CDF, 156
- State
 absorbing, 320, 337
 accessible, 321
 classification, 321
 of Markov chain, 314
 recurrent, 322, 354, 356
 transient, 322, 355
- State space, 314
- Stationary distribution, 327
- Statistics, 15
- Steady-state, in Markov chains, 326, 349
- Steady-state convergence theorem, 327, 350, 360
- Steady-state probability, 327, 350
- Strong law of large numbers, 395
 for Markov chains, 409
 proof, 408
- Subset, 4
- Sum of random number of r.v.'s, 232
 geometric number of exponential r.v.'s, 235, 309
 geometric number of geometric r.v.'s, 235, 303
 mean, 203, 233
 variance, 203, 233
 transform, 233
- Sum of random variables
 of Bernoulli, 96
 convolution, 221
 expectation, *see* expectation, linearity
 of normal, 218
 of Poisson, 218
 transform, 217
 variability extremes, 134
 variance, 112, 114, 174, 239
- T**
- Tabular method, 93, 94
- Total expectation theorem
 for continuous r.v.'s, 161, 172
 for discrete r.v.'s, 105
- Total probability theorem, 28
 conditional version, 57
 for PDFs, 161, 172
- for PMFs, 104
- Transient state, 322, 355
- Transition
 graph, 315
 matrix, 315, 320
 probability, 314
 rate, 345
- Transform, 210
 inversion property, 215
 multivariate, 219
 of linear function, 212
 of mixture, 216
 of sum of r.v.'s 217
 of sum of random number of r.v.'s, 233
 table, 220
- Trial
 Bernoulli, 41
 independent, 41
- Two-envelopes paradox, 107
- Two-envelopes puzzle, 57
- U**
- Uncorrelated r.v.'s, 236, 248
- Uniform r.v.
 continuous, 141, 190
 mean, 145
 relation to Bernoulli process, 303
 transform, 258
 variance, 145
- discrete, 116
 mean, 88
 transform, 258
 variance, 88
- two-dimensional continuous, 165
- Uncountable, 4, 7, 13, 52
- V**
- Value of a r.v., 72
- Variability extremes, 134
- Variance, 83, 86, 145
 conditional, 104, 160, 229
 law of total, 229
 of product of r.v.'s, 262
 of sum of r.v.'s, 113, 114, 174, 239
 in terms of moments, 87, 145
- Venn diagram, 5, 15
- W**
- weak law of large numbers, 384
- Z**
- z-transform, 211

ATHENA SCIENTIFIC BOOKS

1. Introduction to Probability, by Dimitri P. Bertsekas and John N. Tsitsiklis, 2002, ISBN 1-886529-40-X, 430 pages
2. Dynamic Programming and Optimal Control: Second Edition, Vols. I and II, by Dimitri P. Bertsekas, 2001, ISBN 1-886529-08-6, 848 pages
3. Nonlinear Programming, Second Edition, by Dimitri P. Bertsekas, 1999, ISBN 1-886529-00-0, 791 pages
4. Network Optimization: Continuous and Discrete Models, by Dimitri P. Bertsekas, 1998, ISBN 1-886529-02-7, 606 pages
5. Network Flows and Monotropic Optimization, by R. Tyrrell Rockafellar, 1998, ISBN 1-886529-06-X, 634 pages
6. Introduction to Linear Optimization, by Dimitris Bertsimas and John N. Tsitsiklis, 1997, ISBN 1-886529-19-1, 602 pages
7. Parallel and Distributed Computation: Numerical Methods, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1997, ISBN 1-886529-01-9, 718 pages
8. Neuro-Dynamic Programming, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1996, ISBN 1-886529-10-8, 512 pages
9. Constrained Optimization and Lagrange Multiplier Methods, by Dimitri P. Bertsekas, 1996, ISBN 1-886529-04-3, 410 pages
10. Stochastic Optimal Control: The Discrete-Time Case, by Dimitri P. Bertsekas and Steven E. Shreve, 1996, ISBN 1-886529-03-5, 330 pages