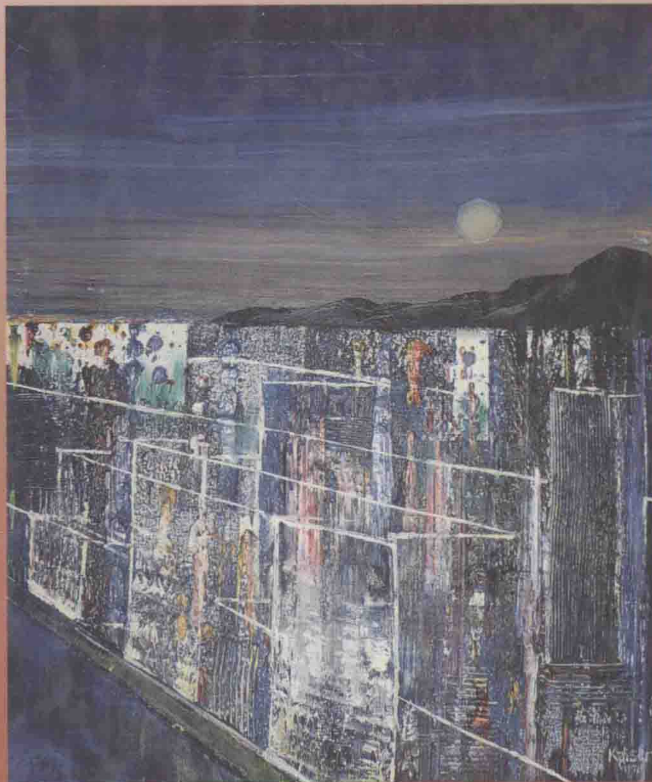
Anthony J. Hayter



Probability and Statistics

FOR ENGINEERS AND SCIENTISTS

Probability and Statistics

FOR ENGINEERS AND SCIENTISTS

Anthony J. Hayter
Georgia Institute of Technology

PWS PUBLISHING COMPANY

ITP A division of International Thomson Publishing

Boston Albany Bonn Cincinnati Detroit London Madrid
Melbourne Mexico City New York Paris San Francisco Singapore
Tokyo Toronto Washington

PWS PUBLISHING COMPANY
20 Park Plaza, Boston, Massachusetts 02116-4324

This book is printed on recycled, acid-free paper.

I(T)P International Thomson Publishing
The trademark ITP is used under license.

For more information, contact:

PWS Publishing Co.
20 Park Plaza
Boston, MA 02116

International Thomson
Publishing Europe
Berkshire House I68-I73
High Holborn
London WC1V 7AA
England

Thomas Nelson Australia
102 Dodds Street
South Melbourne, 3205
Victoria, Australia

Nelson Canada
1120 Birchmont Road
Scarborough, Ontario
Canada M1K 5G4

International Thomson
Editores
Campos Eliseos 385, Piso 7
Col. Polanco
11560 Mexico D.F., Mexico

International Thomson
Publishing GmbH
Konigswinterer Strasse 418
53227 Bonn, Germany

International Thomson
Publishing Asia
221 Henderson Road
#05-10 Henderson Building
Singapore 0315

International Thomson
Publishing Japan
Hirakawacho Kyowa
Building 31
2-2-1 Hirakawacho
Chiyoda Ku, Tokyo 102
Japan

Library of Congress Cataloging-in-Publication Data

Hayter, Anthony J.
Probability and statistics for engineers and scientists / Anthony J. Hayter.
p. cm.
Includes index.
ISBN 0–534–95610–6
1. Engineering—Statistical methods. 2. Probabilities.
I. Title.
TA340.H38 1995
519.5—dc20 95-26470
CIP

Sponsoring Editor: Jonathan Plant
Editorial Assistant: Monica Block
Developmental Editor: Mary Thomas Stone
Production Coordinator: Helen Walden
Marketing Manager: Nathan Wilbur
Manufacturing Coordinator: Wendy Kilborn
Production: Textbook Writers Associates
Interior Designer: Merry Obrecht Sawdey
Cover Designer: Elise Kaiser
Compositor: TechBooks
Cover Art: Judy Kaiser
Cover Printer: Coral Graphic Services
Text Printer: Quebecor, Hawkins

Printed and bound in the United States of America
10 9 8 7 6 5

Preface

This book provides an introduction to probability and statistics for students in scientific disciplines such as engineering, physics, chemistry, computing, biology, and management. It can typically be used for a two-semester or three-quarter undergraduate series, or for graduate-level service courses. It is intended for students with reasonable quantitative abilities, although it primarily provides an applied rather than a theoretical exposition. Basic calculus methods obtainable from an introductory course in calculus are used somewhat in the book.

A good knowledge of and familiarity with data analysis methodologies and statistical inference techniques is becoming increasingly important for today's engineers and scientists. The recent surge in interest toward quality control is one aspect of this phenomenon, since the quality control implementations and operations depend heavily on probabilistic and statistical techniques. Moreover, the development of statistical package software has literally made data analysis available at the push of a button.

The job of statistical educators is to motivate students to obtain the statistical skills that they will require. A key issue in this motivation is to illustrate to the students why statistical analysis methods are relevant and useful for engineering and the sciences. This book provides such a motivation by illustrating the ideas presented with appropriate and interesting examples. These examples are built up and developed throughout the course of the chapters as increasingly sophisticated methodologies are considered. Moreover, the relevance and importance of the statistical analysis to these problems is indicated. A list of these examples together with the page numbers where they appear in the text is provided on the inside front cover.

This book concentrates on allowing the students to obtain an understanding of the concepts behind the methodologies presented, rather than providing an unnecessary amount of theory. Students are encouraged to be able to look at a formula and understand what it is doing and why it works. When students use statistical packages in the workplace, they then will understand which analysis techniques to employ and how to interpret the computer output with which they are presented.

A considerable amount of computer output from a selection of statistical software packages is presented in the text in conjunction with the discussion of the examples. This book does not teach the reader how to use these software packages, but it does instruct the reader how to correctly interpret the output that typically is obtained. In this way the student is not limited to any one particular package but can feel confident of successfully utilizing any statistical software package that is available. Output from the Minitab package is always given, and the other packages used are SAS, JMP, Systat, BMDP, Statview, and Datadesk.

The accompanying figure illustrates the composition of this book. Chapters 1–5 provide basic material on probability theory and probability distributions. Chapter 1 introduces probability through sample spaces and events, and the basic probability laws are discussed. Chapter 2 introduces the concept of random variables, and discrete and continuous probability distributions are discussed in general terms. In Chapter 3 some specific, commonly used discrete probability distributions are presented, and Chapter 4 presents some specific, commonly used continuous probability distributions. The normal distribution is discussed in a chapter of its own, Chapter 5.

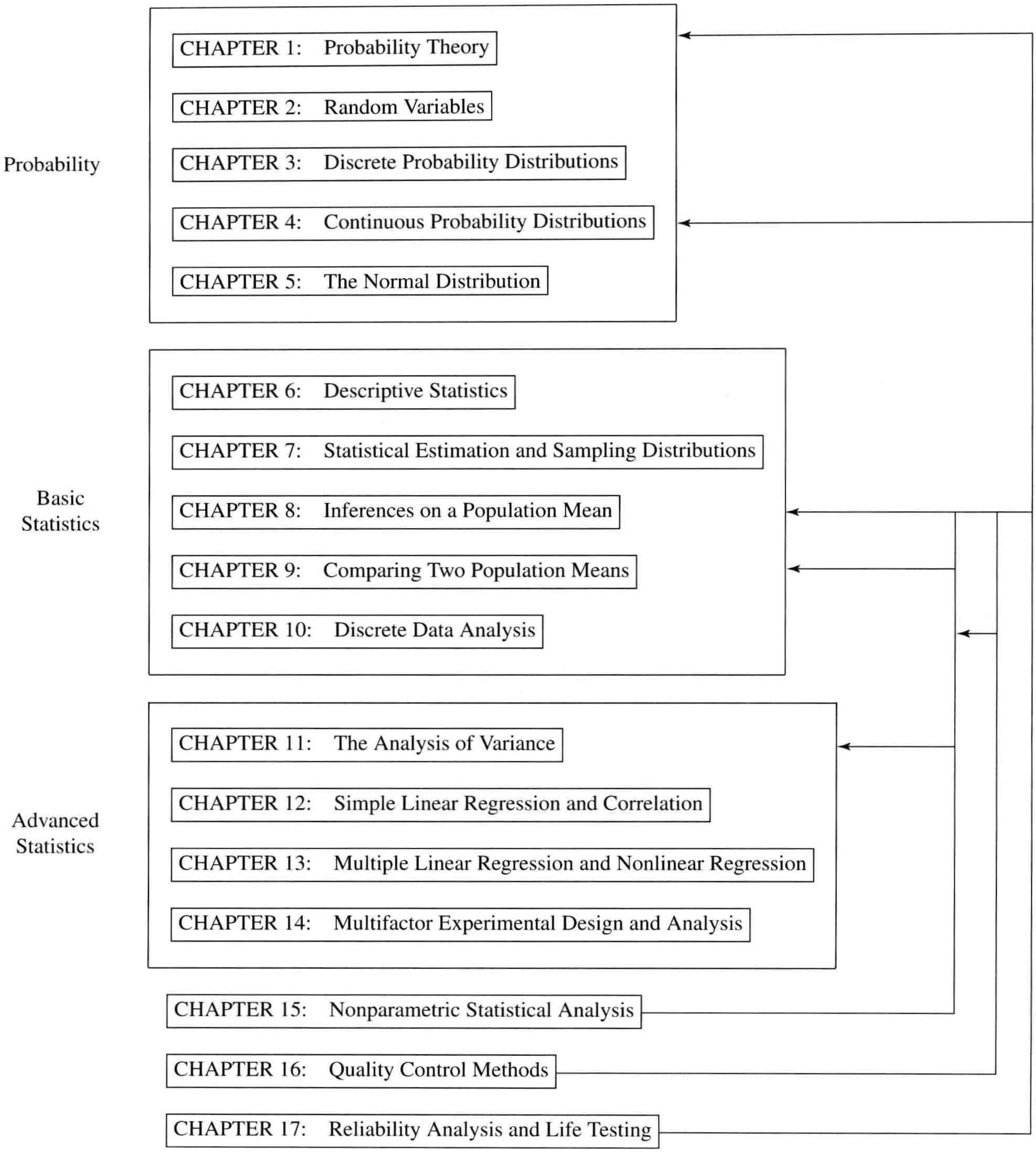
Chapters 6–10 contain the basic statistics material. Chapters 6 and 7 provide the link between the probability material discussed in Chapters 1–5 and the ensuing statistical inference methodologies. Chapter 6 discusses sampling, data presentation, and sample statistics while in Chapter 7 estimation and sampling distributions are discussed. The core statistical inference techniques, confidence interval construction and hypothesis testing, are introduced in Chapter 8 in the context of inferences on a population mean. These ideas are extended in Chapter 9 to two sample problems and to discrete data analysis, such as inferences on population proportions, in Chapter 10.

Chapters 11–14 provide more advanced statistical material on linear models. In Chapter 11 the one-factor analysis of variance is discussed together with randomized block designs. This material is extended in Chapter 14 where multifactor experimental designs are considered. Simple linear regression analysis and correlation are presented in Chapter 12, and more general modeling techniques, primarily multiple linear regression models, are addressed in Chapter 13.

The final three chapters deal with three important topics that may be considered separately or that may be interjected among the previous chapters. Chapter 15 is concerned with nonparametric or distribution-free statistical analysis, and this may be treated in isolation, or the material in its sections on single populations, two populations, and three or more populations may be presented with Chapters 8, 9, and 11, respectively. Chapter 16 contains material on quality control methods such as control charts and acceptance sampling methods. It is becoming somewhat fashionable to present this material quite early in a statistics course, and if desired this material can easily be presented along with the material on population means in Chapter 8 and population proportions in Chapter 10. Finally, Chapter 17 provides an introduction to reliability analysis and life testing and this material can actually be discussed any time after Chapter 4.

Teaching aids to accompany the book include the Student's Solutions Manual and the data diskette. The Student's Solutions Manual contains worked solutions to all of the odd-numbered problems at the ends of the chapter sections. These worked solutions demonstrate to the students how the problems are solved and provide the answers. The data diskette, attached to the inside back cover of the book, has copies of all the data sets included in this book. These data sets are all of those used in the examples discussed in the text, together with all of the data sets used in the problems. Each data set is provided in ASCII format and as a Minitab worksheet.

Finally, I would like to thank various people who have helped in the development of this book. The reviewers, who provided valuable comments on this project, include Mary R. Anderson, Arizona State University; Charles E. Antle, The Pennsylvania State University, University Park; Sant Ram Arora, University of Minnesota—Twin


Probability
CHAPTER 1: Probability Theory
CHAPTER 2: Random Variables
CHAPTER 3: Discrete Probability Distributions
CHAPTER 4: Continuous Probability Distributions
CHAPTER 5: The Normal Distribution
Basic Statistics
CHAPTER 6: Descriptive Statistics
CHAPTER 7: Statistical Estimation and Sampling Distributions
CHAPTER 8: Inferences on a Population Mean
CHAPTER 9: Comparing Two Population Means
CHAPTER 10: Discrete Data Analysis
Advanced Statistics
CHAPTER 11: The Analysis of Variance
CHAPTER 12: Simple Linear Regression and Correlation
CHAPTER 13: Multiple Linear Regression and Nonlinear Regression
CHAPTER 14: Multifactor Experimental Design and Analysis
CHAPTER 15: Nonparametric Statistical Analysis
CHAPTER 16: Quality Control Methods
CHAPTER 17: Reliability Analysis and Life Testing

Cities Campus; William R. Astle, Colorado School of Mines; Lee J. Bain, University of Missouri—Rolla; Douglas M. Bates, University of Wisconsin—Madison; Rajan Batta, The State University of New York at Buffalo; Alan C. Bovik, University of Texas at Austin; Don B. Campbell, Western Illinois University; M. Jeya Chandra, Pennsylvania State University—University Park; Yueh-Jane Chang, Idaho State University; Chung-Lung Chen, Mississippi State University; Inchan Choi, Wichita State University; John R. Cook, North Dakota State University; Rianto A. Djojosugito, South Dakota School of Mines and Technology; Lucien Duckstein, University of Arizona; Earnest W. Fant, University of Arkansas; Richard F. Feldman, Texas A & M University; Sam Gutmann, Northeastern University; Carol O'Connor Holloman, University of Louisville; Chi-Ming Ip, University of Miami; Rasul A. Khan, Cleveland State University; Stojan Kotefski, New Jersey Institute of Technology; Walter S. Kuklinski, University of Massachusetts—Lowell; S. Kumar, Rochester Institute of Technology; Gang Li, University of North Carolina at Charlotte; Jiye-Chyi Lu, North Carolina State University; Ditlev Monrad, University of Illinois at Urbana—Champaign; John Morgan, California Polytechnic State University—Pomona; Paul J. Nahin, University of New Hampshire; Larry Ringer, Texas A & M University; Paul L. Schillings, Montana State University; Ioannis Stavrakakis, The University of Vermont; and James J. Swain, University of Alabama—Huntsville.

In addition I would like to thank Tina Barr, Carolina Barcenas, Tory Chen, Helen Bush, Karen Emanuel, Russ Heikes, MAJ Tom Kastner, Paul Kvam, CPT Andy Napoli, Kurt Palmer, and Alex Shapiro at Georgia Tech for their comments and help with the solutions manual and proofreading.

Anthony J. Hayter

Contents

CHAPTER 1 • Probability Theory

1.1 Probabilities

1.1.1 Introduction

Probability theory is a branch of mathematics that has been developed to deal with **uncertainty**. Classical mathematical theory had been successful in describing the world as a series of fixed and real observable events, yet prior to the seventeenth century, it was largely inadequate in coping with processes or experiments that involved uncertain or random outcomes. Spurred on initially by the mathematician's desire to analyze gambling games, and later by the scientific analysis of mortality tables within the medical profession, the theory of probability has been developed as a scientific tool dealing with **chance**.

Today, probability theory is recognized as one of the most interesting and also one of the most useful areas of mathematics. It provides the basis for the science of statistical inference through experimentation and data analysis—an area of crucial importance in an increasingly quantitative world. Through its applications to problems such as the assessment of system reliability, the interpretation of measurement accuracy, and the maintenance of suitable quality controls, probability theory is particularly relevant to the engineering sciences today.

1.1.2 Sample Spaces

An **experiment** can in general be thought of as referring to any process or procedure for which more than one **outcome** is possible. The goal of probability theory is to provide a mathematical structure for understanding or explaining the chances or likelihoods of the various outcomes actually occurring. A first step in the development of this theory is the construction of a list of the possible experimental outcomes. This collection of outcomes is called the **sample space** or **state space** and is denoted by $\mathcal{S}$. Mathematically, the sample space $\mathcal{S}$ is defined to be a **set** consisting of all of the possible experimental outcomes.

Sample Space

The **sample space** $\mathcal{S}$ of an experiment is a set consisting of all of the possible experimental outcomes.

The following examples help illustrate the concept of a sample space.

EXAMPLE 1 •

Machine Breakdowns An engineer in charge of the maintenance of a particular machine notices that its breakdowns can be characterized as being due to either an electrical failure within the machine, a mechanical failure of some component of the machine, or operator misuse. When the machine is running, the engineer is uncertain what will be the cause of the next breakdown, and it can be thought of as an experiment with a sample space

$$\mathcal{S} = \{\text{electrical, mechanical, misuse}\}.$$

EXAMPLE 2 •

Defective Computer Chips A company sells computer chips in boxes of 500 chips, and each chip can be classified as either satisfactory or defective. The number of defective chips in a particular box is uncertain, and the sample space is

$$\mathcal{S} = \{0 \text{ defectives}, 1 \text{ defective}, 2 \text{ defectives}, 3 \text{ defectives}, 4 \text{ defectives}, \ldots, 499 \text{ defectives}, 500 \text{ defectives}\}.$$

EXAMPLE 3 •

Software Errors The control of errors in computer software products is obviously an area of great importance. The number of separate errors in a particular piece of software can be viewed as having a sample space

$$\mathcal{S} = \{0 \text{ errors}, 1 \text{ error}, 2 \text{ errors}, 3 \text{ errors}, 4 \text{ errors}, 5 \text{ errors}, \ldots\}.$$

In practice there will be an upper bound on the possible number of errors in the software, although in this case it does not hurt to allow the sample space to consist of all the positive integers.

EXAMPLE 4 •

Power Plant Operation A manager supervises the operation of three power plants, plant X, plant Y, and plant Z. At any given time, each of the three plants can be classified as either generating electricity (1) or as being idle (0). Using the notation (0,1,0) to represent the situation where machine Y is generating electricity but machines X and Z are both idle, the sample space for the status of the three plants at a particular point in time is

$$\mathcal{S} = \{(0,0,0)\ (0,0,1)\ (0,1,0)\ (0,1,1)\ (1,0,0)\ (1,0,1)\ (1,1,0)\ (1,1,1)\}.$$

	$\mathcal{S}$
(0, 0, 0)	(1, 0, 0)
(0, 0, 1)	(1, 0, 1)
(0, 1, 0)	(1, 1, 0)
(0, 1, 1)	(1, 1, 1)

FIGURE 1.1 •
Sample space for power plants example

It is often helpful to portray a sample space as a diagram. Figure 1.1 shows a diagram of the sample space for this example, where the sample space is represented by a square box containing the eight individual outcomes. Diagrams of this kind are known as **Venn diagrams**.

GAMES OF CHANCE

Games of chance commonly involve the tossing of a coin, the rolling of a die, or the use of a pack of cards. The tossing of a single coin has a sample space

$$\mathcal{S} = \{\text{head, tail}\},$$

FIGURE 1.2 • Sample space for rolling two dice

$\mathcal{S}$

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

and the tossing of two coins (or one coin twice) has a sample space

$$\mathcal{S} = \{(\text{head, head})\ (\text{head, tail})\ (\text{tail, head})\ (\text{tail, tail})\},$$

where (head, tail), say, represents the event that the first coin resulted in a head and the second coin resulted in a tail. Notice that (head, tail) and (tail, head) are two distinct outcomes since observing a head on the first coin and a tail on the second coin is different from observing a tail on the first coin and a head on the second coin.

A usual six-sided die has a sample space

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}.$$

If two dice are rolled (or, equivalently, if one die is rolled twice), then the sample space is shown in Figure 1.2, where $(1, 2)$ represents the event that the first die recorded a 1 and the second die recorded a 2. Again, notice that the events $(1, 2)$ and $(2, 1)$ are both included in the sample space because they represent two distinct events. This can be seen by considering one die to be red and the other die to be blue, and by distinguishing between obtaining a 1 on the red die and a 2 on the blue die, and between obtaining a 2 on the red die and a 1 on the blue die.

If a card is chosen from an ordinary pack of 52 playing cards, the sample space consists of the 52 individual cards as shown in Figure 1.3. If two cards are drawn, then it is necessary to consider whether they are drawn with or without **replacement**. If the drawing is performed with replacement, so that the initial card drawn is returned to the pack and the second drawing is from a full pack of 52 cards, then the sample space consists of events such as $(6\heartsuit, 8\clubsuit)$, where the first card drawn is $6\heartsuit$ and the second card drawn is $8\clubsuit$. Altogether there will be $52 \times 52 = 2704$ elements of the sample space, including events such as $(A\heartsuit, A\heartsuit)$ where the $A\heartsuit$ is drawn twice. This sample space is shown in Figure 1.4.

If two cards are drawn without replacement, so that the second card is drawn from a reduced pack of 51 cards, then the sample space will be a subset of that above, as shown in Figure 1.5. Specifically, events such as $(A\heartsuit, A\heartsuit)$ where a particular card is drawn twice will not be in the sample space. The total number of elements in this new sample space will therefore be $2704 - 52 = 2652$.

FIGURE 1.3 •
Sample space for choosing one card

$\mathcal{S}$

A♡	2♡	3♡	4♡	5♡	6♡	7♡	8♡	9♡	10♡	J♡	Q♡	K♡
A♣	2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣
A◇	2◇	3◇	4◇	5◇	6◇	7◇	8◇	9◇	10◇	J◇	Q◇	K◇
A♠	2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠

FIGURE 1.4 •
Sample space for choosing two cards with replacement

$\mathcal{S}$

(A♡, A♡)	(A♡, 2♡)	(A♡, 3♡)	···	(A♡, Q♠)	(A♡, K♠)
(2♡, A♡)	(2♡, 2♡)	(2♡, 3♡)	···	(2♡, Q♠)	(2♡, K♠)
(3♡, A♡)	(3♡, 2♡)	(3♡, 3♡)	···	(3♡, Q♠)	(3♡, K♠)
⋮	⋮	⋮		⋮	⋮
(Q♠, A♡)	(Q♠, 2♡)	(Q♠, 3♡)	···	(Q♠, Q♠)	(Q♠, K♠)
(K♠, A♡)	(K♠, 2♡)	(K♠, 3♡)	···	(K♠, Q♠)	(K♠, K♠)

FIGURE 1.5 •
Sample space for choosing two cards without replacement

$\mathcal{S}$

	(A♡, 2♡)	(A♡, 3♡)	···	(A♡, Q♠)	(A♡, K♠)
(2♡, A♡)		(2♡, 3♡)	···	(2♡, Q♠)	(2♡, K♠)
(3♡, A♡)	(3♡, 2♡)		···	(3♡, Q♠)	(3♡, K♠)
⋮	⋮	⋮		⋮	⋮
(Q♠, A♡)	(Q♠, 2♡)	(Q♠, 3♡)	···		(Q♠, K♠)
(K♠, A♡)	(K♠, 2♡)	(K♠, 3♡)	···	(K♠, Q♠)	

1.1.3 Probability Values

The characterization of the likelihoods of particular experimental outcomes actually occurring is achieved through the assignment of a set of **probability values** to each of the elements of the sample space. Specifically, each outcome in the sample space is assigned a probability value which is a number between zero and one. The probabilities are chosen so that the sum of the probability values over all of the elements in the sample space is one.

> **Probabilities**
>
> A set of **probability** values for an experiment with a sample space $\mathcal{S} = \{O_1, O_2, \ldots, O_n\}$ consists of some probabilities
>
> $$p_1, p_2, \ldots, p_n$$
>
> satisfying
>
> $$0 \leq p_1 \leq 1, 0 \leq p_2 \leq 1, \ldots, 0 \leq p_n \leq 1,$$
>
> and
>
> $$p_1 + p_2 + \cdots + p_n = 1.$$
>
> The probability of outcome O_i occurring is said to be p_i, and this is written $P(O_i) = p_i$.

An intuitive interpretation of a set of probability values is that the *larger* the probability value of a particular outcome, then the *more likely* it is to happen. If two outcomes have identical probability values assigned to them, then they can be thought of as being equally likely to occur. On the other hand, if one outcome has a larger probability value assigned to it than another outcome, then it can be thought of as being more likely to occur than the other outcome.

If a particular outcome has a probability value of one, then the interpretation is that it is certain to occur, so that there is actually no uncertainty in the experiment. In this case all of the other outcomes must necessarily have probability values of zero.

The following examples illustrate the assignment of probability values.

EXAMPLE 1 •

Machine Breakdowns Suppose that the machine breakdowns occur with probability values of $P(\text{electrical}) = 0.2$, $P(\text{mechanical}) = 0.5$, and $P(\text{misuse}) = 0.3$. This is a valid probability assignment since the three probability values 0.2, 0.5, and 0.3 are all between zero and one and they sum to one. Figure 1.6 illustrates these probabilities in diagrammatic form by recording the respective probability value with each of the outcomes. These probability values indicate that mechanical failures are most likely, with misuse failures being more likely than electrical failures.

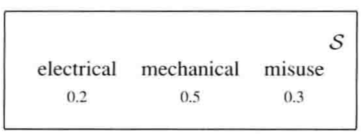


FIGURE 1.6 • Probability values for machine breakdown example

In addition, the fact that $P(\text{mechanical}) = 0.5$ indicates that about half of the failures will be attributable to mechanical causes. This does not mean that of the next two machine breakdowns, exactly one will be for mechanical reasons, nor that in the next ten machine breakdowns, exactly five will be for mechanical reasons. However, it means that in the **long run**, the manager can reasonably expect that roughly half of the breakdowns will be for mechanical reasons. Similarly, in the long run, the manager will expect that about 20% of the breakdowns will be for electrical reasons, and that about 30% of the breakdowns will be attributable to operator misuse.

EXAMPLE 3

Software Errors Suppose that the numbers of errors in a software product have probabilities

$$P(0 \text{ errors}) = 0.05, \quad P(1 \text{ error}) = 0.08, \quad P(2 \text{ errors}) = 0.35,$$
$$P(3 \text{ errors}) = 0.20, \quad P(4 \text{ errors}) = 0.20, \quad P(5 \text{ errors}) = 0.12,$$
$$P(i \text{ errors}) = 0, \quad \text{for } i \geq 6.$$

These probabilities show that there are at most 5 errors since the probability values are zero for 6 or more errors. In addition, it can be seen that the most likely number of errors is 2, and that there are, for instance, equally likely to be 3 or 4 errors in the software product.

It is reasonable to ask how anybody would ever know the probability assignments in the above two examples. In other words, how would the engineer know that there is a probability of 0.2 that a breakdown will be due to an electrical fault, or how would a computer programmer know that the probability of an error-free product is 0.05? In practice these probabilities would have to be estimated from a collection of data and prior experiences. Later in this book it will be shown how statistical analysis techniques can be employed to help the engineer and programmer conduct studies to **estimate** probabilities of these kinds.

In some situations, notably games of chance, the experiments are conducted in such a way that all of the possible outcomes can be considered to be equally likely, so that they must each be assigned identical probability values. If there are n outcomes in the sample space that are equally likely, then the condition that the probabilities sum to one requires that each probability value be $1/n$.

GAMES OF CHANCE

For a coin toss, the probabilities will in general be given by

$$P(\text{head}) = p, \qquad P(\text{tail}) = 1 - p,$$

for some value of p with $0 \leq p \leq 1$. A fair coin will have $p = 0.5$ so that

$$P(\text{head}) = P(\text{tail}) = 0.5,$$

with the two outcomes being equally likely. A biased coin will have $p \neq 0.5$. For example, if $p = 0.6$ then

$$P(\text{head}) = 0.6, \qquad P(\text{tail}) = 0.4,$$

as shown in Figure 1.7, and the coin toss is more likely to record a head.

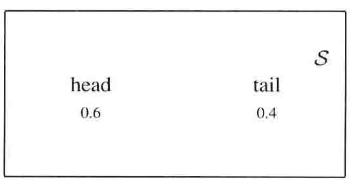


FIGURE 1.7 Probability values for a biased coin

A fair die will have each of the six outcomes equally likely with each being assigned the same probability. Since the six probabilities must sum to one, this implies that each of the six outcomes must have a probability of $1/6$, so that

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}.$$